

Short text contextualization in information retrieval : application to tweet contextualization and automatic query expansion

Liana Ermakova

▶ **To cite this version:**

Liana Ermakova. Short text contextualization in information retrieval : application to tweet contextualization and automatic query expansion. Information Retrieval [cs.IR]. Université Toulouse le Mirail - Toulouse II; Permskij gosudarstvennyj universitet (Russie), 2016. English. NNT : 2016TOU20023 . tel-01729649

HAL Id: tel-01729649
https://theses.hal.science/tel-01729649

Submitted on 12 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse - Jean Jaurès*
Cotutelle internationale *Perm State National Research University*

Présentée et soutenue le *31/03/2016* **par :**

Liana ERMAKOVA

Short Text Contextualization in Information Retrieval: Application to Tweet Contextualization and Automatic Query Expansion
Contextualisation de textes courts pour la recherche d'information : application à la contextualisation de tweets et à l'expansion automatique de requêtes

---

JURY

Olivier TESTE	Professeur d'Université	Président du Jury
Josiane MOTHE	Professeur d'Université	Membre du Jury
Elena NIKITINA	Maître de conférences d'Université	Membre du Jury
Irina OVCHINNIKOVA	Professeur d'Université	Membre du Jury
Brigitte GRAU	Professeur d'Université	Membre du Jury
Jacques SAVOY	Professeur d'Université	Membre du Jury
Eric SANJUAN	Maître de conférences d'Université	Membre du Jury

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5055)
Directeur(s) de Thèse :
Josiane MOTHE, *Elena NIKITINA* et *Irina OVCHINNIKOVA*
Rapporteurs :
Brigitte GRAU et *Jacques SAVOY*

Declaration of Authorship

I, Liana Ermakova, declare that this thesis titled, 'Short Text Contextualization in Information Retrieval: Application to Tweet Contextualization and Automatic Query Expansion' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Université Toulouse - Jean Jaurès
Perm State National Research University

Abstract

École doctorale et spécialité : MITT : Image, Information, Hypermedia
Institut de Recherche en Informatique de Toulouse (UMR 5055)

Doctor of Philosophy

Short Text Contextualization in Information Retrieval: Application to Tweet Contextualization and Automatic Query Expansion

by Liana Ermakova

The efficient communication tends to follow the principle of the least effort. According to this principle, using a given language interlocutors do not want to work any harder than necessary to reach understanding. This fact leads to the extreme compression of texts especially in electronic communication, e.g. microblogs, SMS, search queries. However, sometimes these texts are not self-contained and need to be explained since understanding them requires knowledge of terminology, named entities or related facts. The main goal of this research is to provide a context to a user or a system from a textual resource.

The first aim of this work is to help a user to better understand a short message by extracting a context from an external source like a text collection, the Web or the Wikipedia by means of text summarization. To this end we developed an approach for automatic multi-document summarization and we applied it to short message contextualization, in particular to tweet contextualization. The proposed method is based on named entity recognition, part-of-speech weighting and sentence quality measuring. In contrast to previous research, we introduced an algorithm for smoothing from the local context. Our approach exploits topic-comment structure of a text. Moreover, we developed a graph-based algorithm for sentence reordering. The method has been evaluated at INEX/CLEF tweet contextualization track. We provide the evaluation results over the 4 years of the track. The method was also adapted to snippet retrieval. The evaluation results indicate good performance of the approach.

Moreover, we extended the idea of the use of topic-comment to document re-ranking in information retrieval. While most information retrieval models make the assumption that relevant documents are about the query and that aboutness can be captured considering bags of words only, we rather consider a more sophisticated analysis of discourse to capture document relevance by distinguishing the topic of a text from what is said about the topic (comment) in the text. The topic-comment structure of texts is extracted automatically from the first retrieved documents which are then re-ranked so that the top documents are the ones that share their topics with the query. The evaluation on TREC collections showed that the method significantly improves the retrieval performance.

The second aim of our research is to provide a context to a search query, i.e. to expand a search query in order to improve information retrieval effectiveness by enhancing the query formulation. We suggested three methods of query expansion exploiting term proximity: Adaptation of Sentence Extraction Method to Query Expansion, Co-occurrence Model, and Proximity Relevance Model. These new methods estimate the importance of expansion candidate terms by the strength of their relation to the query terms. The former method is an elaboration of the method we proposed to tweet contextualization. The Co-occurrence Model combines local analysis, namely relevance feedback, and global

analysis of texts. Rather than considering feedback documents as a bag of words, it is possible to exploit term proximity information. Although there are some researches in this direction, the majority of them are empirical. The lack of theoretical works in this area motivated us to introduce a Proximity Relevance Model integrated into the language model formalism that takes advantage of the remoteness of candidate terms for expansion from query terms within feedback documents. In contrast to previous works, our approach captures the proximity directly and in terms of sentences rather than tokens. We show that the method significantly improves the retrieval performance on TREC collections especially for difficult queries. Besides, we pursue an objective to deeply analyze the results: both the initial and expanded queries and the terms they are composed of, and the cases when the expansion lowers the results and when it improves them.

Université Toulouse - Jean Jaurès
Perm State National Research University

Résumé

École doctorale et spécialité : MITT : Image, Information, Hypermedia
Institut de Recherche en Informatique de Toulouse (UMR 5055)

Doctor of Philosophy

Short Text Contextualization in Information Retrieval: Application to Tweet Contextualization and Automatic Query Expansion

by Liana Ermakova

La communication efficace a tendance à suivre le loi du moindre effort. Selon ce principe, en utilisant une langue donnée les interlocuteurs ne veulent pas travailler plus que nécessaire pour être compris. Ce fait mène à la compression extrême de textes surtout dans la communication électronique, comme dans les microblogues, SMS, ou les requêtes dans les moteurs de recherche. Cependant souvent ces textes ne sont pas auto-suffisants car pour les comprendre, il est nécessaire d'avoir des connaissances sur la terminologie, les entités nommées ou les faits liés. Ainsi, la tâche principale de la recherche présentée dans ce mémoire de thèse de doctorat est de fournir le contexte d'un texte court à l'utilisateur ou au système comme à un moteur de recherche par exemple.

Le premier objectif de notre travail est d'aider l'utilisateur à mieux comprendre un message court par l'extraction du contexte d'une source externe comme le Web ou la Wikipédia au moyen de résumés construits automatiquement. Pour cela nous proposons une approche pour le résumé automatique de documents multiples et nous l'appliquons à la contextualisation de messages, notamment à la contextualisation de tweets. La méthode que nous proposons est basée sur la reconnaissance des entités nommées, la pondération des parties du discours et la mesure de la qualité des phrases. Contrairement aux travaux précédents, nous introduisons un algorithme de lissage en fonction du contexte local. Notre approche s'appuie sur la structure thème-rhème des textes. De plus, nous avons développé un algorithme basé sur les graphes pour le ré-ordonnancement des phrases. La méthode a été évaluée à la tâche INEX/CLEF Tweet Contextualization sur une période de 4 ans. La méthode a été également adaptée pour la génération de snippets. Les résultats des évaluations attestent une bonne performance de notre approche.

Par ailleurs, nous avons étendu l'idée de l'utilisation de la structure thème-rhème au ré-ordonnancement de documents dans la recherche d'information. La structure thème-rhème est identifiée automatiquement à partir des premiers documents retrouvés qui sont ré-ordonnés selon cette structure. L'évaluation sur les collections TREC a montré que notre méthode améliore significativement les résultats de la recherche.

La deuxième tâche de notre recherche est de fournir le contexte de recherche à un moteur de recherche, i.e. étendre une requête. Nous proposons trois méthodes d'expansion automatique de requêtes basées sur la proximité des termes : l'adaptation de la méthode d'extraction des phrases que nous avons développée pour la contextualisation de tweets, un modèle de co-occurrence et un modèle appelé Proximity Relevance Model. Le modèle de co-occurrence combine l'analyse locale, i.e. le retour de pertinence, et l'analyse globale. Le manque de travaux théoriques sur l'utilisation de la proximité des termes (la plupart des travaux reste empiriques) a motivé l'introduction du modèle Proximity Relevance

intégré dans le formalisme du modèle de langage. Il estime l'importance des termes candidats selon la proximité des termes de la requête dans les premiers documents retrouvés. Contrairement aux autres travaux de la littérature, dans notre approche la proximité est calculée directement en fonction des phrases et non des mots simples. L'évaluation sur les collections TREC indique que la performance de la recherche est améliorée significativement. De plus, nous avons analysé en détail certaines requêtes et leurs extensions automatiques pour expliquer l'amélioration ou la dégradation des résultats.

Acknowledgements

I would like to acknowledge my advisers Josiane Mothe, Irina Ovhinnikova and Elena Nikitina as well as the reviewers Brigitte Grau and Jacques Savoy. I would like to thank l'Ambassade de France en Russie for funding this project (bourse de thèse en cotutelle).

Contents

Declaration of Authorship i

Abstract iii

Résumé vi

Acknowledgements ix

Contents x

List of Figures xiii

List of Tables xiv

Abbreviations xvi

Introduction 1

1 Sentence Extraction for Short Text Contextualization 4
- 1.1 Introduction . . . 4
- 1.2 Related Works . . . 7
  - 1.2.1 Query and Document Preprocessing . . . 7
  - 1.2.2 Sentence Ranking . . . 8
  - 1.2.3 Redundancy Treatment and Result Filtering . . . 13
  - 1.2.4 Sentence Re-ordering . . . 15
- 1.3 Contribution 1: Sentence Ranking Approach for Message Contextualization 16
  - 1.3.1 Sentence Ranking . . . 18
    - 1.3.1.1 Sentence Representation . . . 19
    - 1.3.1.2 Smoothing from Local Context . . . 22
  - 1.3.2 Topic-comment Relationship Integration . . . 24
  - 1.3.3 Result Filtering . . . 24
- 1.4 Contribution 2: Sentence Re-Ordering . . . 25
  - 1.4.1 Model Description . . . 26
    - 1.4.1.1 Traveling Salesman Problem for Sentence Re-Ordering . . . 26
    - 1.4.1.2 Sequential Ordering Problem . . . 27
    - 1.4.1.3 Combining Informativeness and Readability . . . 28

1.5 Evaluation Framework . . . 29
1.5.1 INEX Data . . . 29
1.5.2 Informativeness Measurement . . . 32
1.5.3 Readability Measurement . . . 34
1.5.4 System Details . . . 35
1.6 Results . . . 40
1.6.1 Informativeness . . . 40
1.6.1.1 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2011 . . . 40
1.6.1.2 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2012 . . . 41
1.6.1.3 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2013 . . . 42
1.6.1.4 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2014 . . . 42
1.6.2 Readability . . . 44
1.6.2.1 Readability Results at INEX/CLEF Tweet Contextualization Task 2011 . . . 44
1.6.2.2 Readability Results at INEX/CLEF Tweet Contextualization Task 2012 . . . 44
1.6.2.3 Readability Results at INEX/CLEF Tweet Contextualization Task 2013 . . . 45
1.6.2.4 Readability Results at INEX/CLEF Tweet Contextualization Task 2014 . . . 45
1.6.3 Result Summary . . . 46
1.7 Contribution 3: Extension to Snippet Retrieval . . . 47
1.7.1 Modifications . . . 47
1.7.2 Evaluation . . . 48
1.7.2.1 Data Description . . . 48
1.7.2.2 Measures . . . 49
1.7.2.3 Results at INEX/CLEF Snippet Retrieval Task . . . 51
1.8 Contribution 4: Topic-comment Structure for Information Retrieval . . . 56
1.8.1 Topic-comment Structure in Linguistics . . . 58
1.8.2 Discourse-level Topic vs Rhetorial Relations and Topic-comment Structure in IR . . . 59
1.8.3 Contribution 4: Document Re-ranking Algorithm Based on Topic-Comment Structure Analysis . . . 62
1.8.3.1 Automatic Topic-comment Annotation . . . 62
1.8.3.2 Topic vs Comment for Query Matching . . . 62
1.8.4 Integration of the Topic-comment Structure into Retrieval Models 63
1.8.4.1 Multi-word Expression Extraction . . . 65
1.8.5 Evaluation . . . 65
1.9 Conclusion . . . 68

2 Query Expansion **70**
2.1 Introduction . . . 70
2.2 Related Works . . . 73
2.2.1 Global Methods . . . 74

2.2.2 Query Expansion and Pseudo Relevance Feedback . . . 75
2.2.3 Proximity Based Methods . . . 76
2.2.4 Selective Query Expansion . . . 78
2.3 Models . . . 78
2.3.1 Contribution 5: LC Model . . . 78
2.3.2 Contribution 6: Co-occurrence Model . . . 80
2.3.3 Contribution 7: Proximity Relevance Model . . . 83
2.4 Evaluation Framework . . . 87
2.4.1 Data Sets . . . 87
2.4.1.1 TREC Robust . . . 87
2.4.1.2 WT10G. . . . 88
2.4.2 Evaluation Measures . . . 88
2.4.3 Systems Used For Comparison . . . 90
2.4.4 Details of the Implemented Systems . . . 92
2.4.4.1 LC Model . . . 92
2.4.4.2 Proximity Relevance Model . . . 93
2.5 Results . . . 93
2.6 Deep Analysis of the Queries . . . 99
2.6.1 Analysis of the Individual Queries from the Robust Collection . . . 101
2.6.2 Analysis of the Individual Queries from the WT10G Collection . . . 115
2.6.3 Types of initial queries . . . 125
2.7 Conclusion . . . 129

Conclusion **132**

Bibliography **136**
2.1 Publications of the Author on Short Text Contextualization . . . 136
2.2 Other Publications of the Author . . . 137
2.3 References . . . 139

List of Figures

1.1 Dependence of neighboring sentence impact on distance 23
1.2 Example of the graph representation of a text for the TSP sentence reordering method (vertices represent sentences and edges correspond to the same cosine similarity measure) 27
1.3 SOP sentence reordering algorithm 28
1.4 Documents and tweets preprocessing step for Tweet Contextulization . . . 35

2.1 Histogram of the NDCG difference with the baseline (Robust) 98
2.2 Histogram of the NDCG difference with the baseline (WT10G) 99

List of Tables

1.1 Test collections (INEX/CLEF Tweet Contextualization 2011-2014) 30
1.2 Tweet example from Tweet Contextualization Task 2014 32
1.3 Log difference to New York Times articles (Tweet Contextualization 2011) 41
1.4 Log difference with the set of relevant passages (Tweet Contextualization 2011) 41
1.5 Informativeness evaluation (Tweet Contextualization 2012) 42
1.6 Informativeness evaluation (Tweet Contextualization 2013) 43
1.7 Informativeness evaluation (Tweet Contextualization 2014) 43
1.8 Readability results with the relaxed and strict metrics (Tweet Contextualization 2011) 44
1.9 Readability results (Tweet Contextualization 2012) 44
1.10 Readability evaluation (Tweet Contextualization 2013) 45
1.11 Readability evaluation (Tweet Contextualization 2014) 46
1.12 Snippet evaluation 2013 51
1.13 Re-ranking results using topic-comment structure 68

2.1 General comparison of QE methods 94
2.2 Collections' statistics 95
2.3 # of improved and worsen queries 95
2.4 Result degradation (NDCG) 98
2.5 Results for individual queries (NDCG) 100
2.6 NDCG differences with the baseline 101
2.7 NDCG differences with Bo1 101
2.8 Query 429. BPREF differences with the baseline 103
2.9 Query 659. BPREF differences with Bo1 105
2.10 BPREF differences with the baseline 105
2.11 Query 614. BPREF differences with Bo1 106
2.12 Query 614. BPREF differences with the baseline 106
2.13 Query 415. BPREF differences with Bo1 108
2.14 Query 415. BPREF differences with the baseline 108
2.15 Query 615. BPREF differences with Bo1 109
2.16 Query 615. BPREF differences with the baseline 110
2.17 Query 350. BPREF differences with Bo1 111
2.18 Query 350. BPREF differences with the baseline 111
2.19 Query 648. BPREF differences with Bo1 113
2.20 Query 648. BPREF differences with the baseline 113
2.21 Query 352. BPREF differences with the baseline 114
2.22 Query 484. BPREF differences with the baseline 116

2.23 Query 538. BPREF differences with the baseline 117
2.24 Query 531. BPREF differences with the baseline 119
2.25 Query 531. BPREF differences with Bo1 119
2.26 Query 504. BPREF differences with the baseline 121
2.27 Query 486. BPREF differences with Bo1 122
2.28 Query 529. BPREF differences with Bo1 124
2.29 Query 548. BPREF differences with Bo1 125
2.30 Query 548. BPREF differences with the baseline 125
2.31 Query 317. BPREF differences with the baseline 127

Abbreviations

Co	**Co**-occurrence model
DFR	**D**ivergence **F**rom **R**andomness
ICF	**I**nversed **C**omment **F**requency
IDF	**I**nversed **D**ocument **F**requency
IR	**I**nformation **R**etrieval
LM	**L**anguage **M**odel
LC	**L**ocal **C**ontext
NE	**N**amed **E**ntity
POS	**P**art **O**f **S**peech
PRF	**P**seudo **R**elevance **F**eedback
PRM	**P**roximity **R**elevance **M**odel
RF	**R**elevance **F**eedback
SOP	**S**equential **O**rdering **P**roblem
TF	**T**erm **F**requency
TSP	**T**raveling **S**alesman **P**roblem
QE	**Q**uery **E**xpansion

Introduction

The efficient communication tends to follow the principle of the least effort. According to this principle, using a given language interlocutors do not want to work any harder than necessary to reach understanding. This fact leads to the extreme compression of texts especially in electronic communication, e.g. microblogs, SMS, search queries. However, sometimes these texts are not self-contained and need to be explained since understanding of them requires knowledge of terminology, named entities or related facts. The main goal of this research is to provide a context to a user or a system.

The first aim of this work is to help a user to better understand a short message by extracting a context from an external source like a text collection, the Web or the Wikipedia by means of text summarization. To this end we developed an approach for automatic multi-document summarization and we applied it to short message contextualization, in particular to tweet contextualization. The proposed method is based on named entity recognition, part-of-speech weighting and sentence quality measuring. In contrast to previous research, we introduced an algorithm for smoothing from the local context. Our approach exploits topic-comment structure of a text. In linguistics, the **topic** is what the clause is about, while the **comment** is what is being said about the topic [Büring, 2011] (the detailed description is given in Section 1.8). Moreover, we developed a graph-based algorithm for sentence reordering. The method we proposed for short text contextualization has been evaluated at INEX/CLEF tweet contextualization track. We provide the evaluation results over the 4 years of the track. According to informative evaluation, in 2011 and 2013 we obtained the best results among all automatic systems that participated. In 2013 it was the best in terms of readability among all participants according to all metrics except redundancy. The method was also adapted to snippet retrieval. In 2013 our system showed the best results in the INEX Snippet Retrieval Track. Thus, the evaluation results indicate good performance of the approach. The proposed method

is described in the Chapter 1. The approach was presented at several national and international conferences [Ermakova and Faessel, 2013, Ermakova and Mothe, 2012a,b, 2013, 2014]. It obtained the best paper award at the international conference CLEF-2015 [Ermakova, 2015].

Moreover, we extended the idea of the use of topic-comment to document re-ranking in information retrieval. While most information retrieval models make the assumption that relevant documents are about the query and that aboutness can be captured considering bags of words only, we rather consider a more sophisticated analysis of discourse to capture document relevance by distinguishing the topic of a text from what is said about the topic (comment) in the text. The topic-comment structure of texts is extracted automatically from the first retrieved documents which are then re-ranked so that the top documents are the ones that share their topics with the query. The evaluation on TREC collections showed that the method significantly improves the retrieval performance.

The second aim of our research is to provide a context to a search query, i.e. to expand a search query in order to improve information retrieval effectiveness by enhancing the query formulation (see Chapter 2). Information retrieval aims at retrieving the relevant documents according to a user's need. Concretely, a search engine computes a similarity between the user's query and the indexed documents; the documents that contain the query terms are retrieved and ordered according to their decreasing similarity with the query. Retrieving relevant information to a query implies a two-step process: off line, the system indexes documents, generally using a bag of words representation; online, the system computes the similarity between the user's query and the document representations (indexing terms) to retrieve the most similar documents. Matching is difficult because the terms used by the authors of documents and the search engine users to represent a concept may be different. It is also difficult because users express their needs using just a few words, making the query difficult to "understand" by the system. Various approaches have been developed to face these challenges. One of them is to diversify the results. On the other hand, query expansion techniques also aim at improving system results [Carpineto and Romano, 2012b]. The principle of query expansion is to add new query terms to the initial query in order to enhance the users' need formulation. Automatic methods for QE were firstly proposed by Maron and Kuhns in 1960. QE based on RF makes the hypothesis that relevant documents are key components to decide which terms are important to formulate an enhanced query regardless to an information need.

While improving the effectiveness of search, the method however implies that document relevance is collected. Buckley et al. [Buckley, 1995] went a step further by assuming the top-retrieved documents are relevant. The so-called blind or pseudo relevance feedback is now commonly used in IR evaluation campaigns.

We suggested three methods of query expansion based on pseudo relevance feedback: Adaptation of Sentence Extraction Method to Query Expansion, Co-occurrence Model, and Proximity Relevance Model. These new methods estimate the importance of expansion candidate terms by the strength of their relation to the query terms. The former method is an elaboration of the method we proposed to tweet contextualization. The Co-occurrence Model combines local analysis, namely relevance feedback, and global analysis of texts. This approach was presented at the international conference Dialog-2013 [Ermakova et al., 2014]. Rather than considering feedback documents as a bag of words, it is possible to exploit term proximity information. Although there are some researches in this direction, the majority of them are empirical. The lack of theoretical works in this area has motivated us to introduce a Proximity Relevance Model integrated into the language model formalism that takes advantage of the remoteness of candidate terms for expansion from query terms within feedback documents. In contrast to previous works, our approach captures the proximity directly and in terms of sentences rather than tokens. We show that the method significantly improves the retrieval performance on TREC collections especially for difficult queries. The proposed model was described in the paper accepted at the international conference SAC-2016 [Ermakova et al., 2016]. Besides, we pursue an objective to deeply analyze the results: both the initial and expanded queries and the terms they are composed of, and the cases when the expansion lowers the results and when it improves them.

The rest of the thesis is organized as follows. Chapter 1 is devoted to short text contextualization from a textual resource. It also provides the extensions of the proposed method to snippet retrieval and document re-ranking. Chapter 2 presents the second axis of the research, namely query expansion. The last part concludes the thesis. The list of the author's publications is given after the conclusion.

Chapter 1

Sentence Extraction for Short Text Contextualization

1.1 Introduction

The communication in a natural language tends to follow the principle of the least effort, i.e. interlocutors try to minimize their efforts to the limits allowing to reach understanding. That leads to the extreme compression of messages, especially in micro-blogs, SMS, search queries. One of the examples is a micro-blogging service Twitter allowing users to share short posts called "Tweets". Twitter is a widely used web service. In 2013 it had about 200 million users sending over 400 million tweets daily. In December 2015 around 6,000 tweets are tweeted per second, which corresponds to over 500 million tweets per day [Twitter Usage Statistics - Internet Live Stats, accessed date: 14/12/2015]. Media organizations are among the most-followed users on Twitter [Wu et al., 2011]. Tweets are more and more used in relation with various types of events such as conferences, political conventions, and even in case of emergency (community evacuation, wildfire, hurricanes, terrorist attacks, road closures) [Hughes and Palen, 2009]. However, consisting of 140 characters, a tweet is also hard to understand, specifically for those users who do not know the event it relates to or more generally the tweet context. Since understanding of a message presupposes indemnification of ellipses in speech, reconstruction of the sense that the speaker intended to pass and requires background knowledge (knowledge of terminology, named entities or related facts), contextualization seems to be a good mean to

help users to understand short messages that are not self-contained. In this work contextualization of a short message is viewed as its explanation, providing details of related entities and events. This research aims at developing an approach to contextualize short messages.

The idea to contextualize short texts like micro-blogs or tweets is quite recent. Several systems automatically discover the wide-range of vocabulary used in tweets, including topic tags, and they use linguistic processing to collect and summarize the thousands of ways people have of saying the same thing (e.g., Linguamatics) [1]. Other researches are more targeted at providing a context to a tweet, e.g. Meij et al. mapped a tweet into a set of Wikipedia articles but in their work, no summary is provided to the user, rather a set of related links [Meij et al., 2012]. San Juan et al. went a step further and introduced Tweet Contextualization as an INEX task which became the CLEF lab in 2012 [Bellot et al., 2013, SanJuan et al., 2012].

Following the task suggested at INEX Tweet Contextualization track, the main motivation of our research on this topic is to help a user to better understand a short message by extracting a context from an external source like the Web or the Wikipedia by means of text summarization. Thus, we consider a short text as a query while the context is represented by a summary biased to this query. A summary is defined as a "condensed version of a source document having a recognizable genre and a very specific purpose: to give the reader an exact and concise idea of the contents of the source" [Saggion and Lapalme, 2002]. A summary is either an "extract", if it consists in the most important passages extracted from the original text, or an "abstract", if these sentences are rewritten, generating a new text. Abstract generation is usually based on extraction which implies searching for relevant sentences [Erkan and Radev, 2004]. Actually, often even human beings firstly extract relevant information before writing a summary. Moreover, most of real-world summarization systems are extractive since abstraction requires strong natural language generation tools. The development of these tools is a very difficult task. Besides, human-like approach needs internal semantic representation. While some work has been done in abstractive summarization [Hahn and Mani, 2000, Radev and McKeown, 1998], extractive methods remains more in focus of current research [Bellot et al., 2013, Giannakopoulos et al., 2011].

[1] http://www.linguamatics.com/

Summarization implies two tasks: searching for relevant information and organizing it into a summary either by paraphrasing (in case of abstracts) or reordering (in case of extracts). In our approach searching for relevant information implies sentence retrieval based on TF-IDF measure enriched by named entity recognition, part of speech weighting, smoothing from local context and sentence quality measuring. Moreover, our algorithm takes advantage of topic-comment structure of sentences. The topic-comment structure have already got the attention of linguists in the 19-th century, however, it is hardly applied in information retrieval tasks. To our knowledge, the topic-comment analysis was never exploited in the summarization task.

The proposed approach demonstrated better performance than other systems like Cortex, Enertex, REG, etc. Cortex combines such metrics as word frequency, overlap with query terms, entropy of the words, shape of text etc. [Torres-Moreno et al., 2012b]. In Enertex sentence score is calculated from text energy matrix [Torres-Moreno et al., 2012b]. REG is an enhancement of Cortex which uses query expansion [Vivaldi and da Cunha, 2012].

As Barzilay et al. showed, sentence order is crucial for readability [Barzilay et al., 2002]. Moreover, sentence reordering is the only way to improve the readability of a text produced by an extraction system. Barzilay et al. proposed to order the sentences by searching for the Hamiltonian path of maximal length in a directed graph where vertices are themes and edges corresponds to the number of times a theme precedes the other one. This approach requires a training corpus. In contrast to this, we hypothesized that in a coherent text neighboring sentences should be somehow similar to each other and the total distance between them should be minimal. Therefore, we propose an approach to increase global coherence of text on the basis of its graph model, where the vertices correspond to the extracted passages and the edges represent the similarity measure between them. Under these assumptions, sentence ordering implies searching for the minimal path that visits each vertex exactly once. This task is known as the traveling salesman problem. However, this method does not consider chronological constraints therefore we introduce another method based on the sequential ordering problem. In contrast to [Barzilay et al., 2002], our approach is not limited to the news articles on the same topic and it takes advantages of the similarity between sentences.

The rest of the chapter is organized as follows. Section 1.2 describes related works. Section 1.3 presents our method of sentence ranking. Section 1.4 describes two approaches

we propose to sentence re-ordering. Section 1.5 provides the details of the evaluation framework. Section 1.6 contains the results and their analysis. Section 1.7 suggests the application of the proposed sentence retrieval method to snippet generation. Section 1.8 proposes the extension of the idea of the use of the topic-comment structure in information retrieval. Section 1.9 concludes this chapter.

1.2 Related Works

The general steps of the extractive summarization are the following:

1. query and document pre-processing;
2. sentence ranking;
3. result filtering;
4. sentence re-ordering.

Let us discuss them in details.

1.2.1 Query and Document Preprocessing

In the case of a subject related summary, like tweet contextualization, the subject may be considered as a query and the summary is made of the sentences relevant to this query which can be expanded e.g. by synonyms from the WordNet [Soriano-Morales et al., 2011]. A query may be also expanded by terms from headers and the most frequent words [Amini et al., 2007].

Amini and Usunier proposed to expand title words with the respective cluster terms extracted by EM algorithm based on co-occurrence measure [Amini and Usunier, 2007]. Candidate sentences were filtered by Marcu's alignment technique [Marcu, 2000]. Marcu's algorithm implies at each iteration the removal of a sentence that maximizes the similarity between the query and the rest of the sentences in that set.

Schiffman presented an approach that incorporates corpus-driven semantic information and query expansion by log likelihood ratio [Schiffman, 2007]. He used a window of 3

sentences as a unit of summarization. This approach showed very low results on DUC-2007.

The common methods of query expansion are described in details in the chapter 2.

1.2.2 Sentence Ranking

Apparently, the first article on automated summarization was published by Luhn in 1958 [Luhn, 1958]. H.P. Luhn proposed to order sentences by the number of the most frequent meaningful words. This approach was extended by taking into account inverse document/sentence frequency, sentence position in the text, key word and key phrase occurrence [Erkan and Radev, 2004, Radev and McKeown, 1998, Seki, 2005]. The further extension was made in [Sun et al., 2005] by computing the frequency of a word both in a document and in the set of the query terms collected from the click-through data.

Stokes et al. combined the following metrics: term similarity, named entity similarity, centroid similarity, similarity to the query expanded by WordNet synonyms and the most frequent words form the pseudo-relevance feedback, density of numeric references, noun phrase similarity, sentence position [Stokes et al., 2007].

Gusev et al. proposed to use scan statistics in order to test whether word distribution fits the uniform one [Gusev et al., 2005]. If sentences form a cluster according to a specific word (cluster forming lexical units), the cluster is interpreted as a supra-phrasal entity reflecting the semantics of the fragment. Sentences are weighted according to the number of clusters forming lexical units.

The best result at Document Understanding Conference DUC-2007[2], that aims at evaluation of text summarization method, was obtained by the approach proposed by Pingali et al. [Prasad Pingali and Varma, 2007] that combines query-dependent and query-independent sentence score. This approach implies terms clustering. Terms are clustered together if they have similar probability distribution in an elite set of documents and a random document pool. Each sentence is expanded by the words co-occurring with its terms within an elite set.

Gotti et al. took into account not only word-based similarity, but also the depth of the node within a syntactic tree [Gotti et al., 2007].

[2] http://duc.nist.gov/duc2007/tasks.html

Blake et al. considered lexical diversity [Blake et al., 2007].

Madnani et al. introduced a Multiple Alternative Sentence Compressions algorithm which produces several variants of compressed sentences to be added into the pool of candidates [Madnani et al., 2007].

Frequency-based methods are a subset of **statistical methods**. Besides term frequency analysis [Erkan and Radev, 2004, Radev and McKeown, 1998, Seki, 2005], statistical methods can be referred to **machine learning** [Lin and Hovy, 1997], **graphical models** [Erkan and Radev, 2004, Shen et al., 2007] or using lexical chains [Morris and Hirst, 1991, Silber and Mccoy, 2002].

Probabilistic graphical models are widely used for summarization. One of the most common models is conditional random fields [Shen et al., 2007]. Within the model, the summarization task is considered as a sequence labeling problem. A document is represented by a sequence of sentences and to each sentence is assigned a value 0 or 1 depending on the assignment of labels of others. A forward-backward algorithm can solve this problem. However, additional parameters like in Dual Wing Factor Graph (e.g. combining social content with documents) may cause loops in the model and in this case loopy-sum-product or max-sum algorithm may be applied [Yang et al., 2011].

Another very efficient model is Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. LDA is a graphical topic model where a document is viewed as a mixture of topics and a topic is considered as a mixture of words [Arora and Ravindran, 2008]. Sentences are scored according to their probability to represent the topics. In the LexRank algorithm, a document is viewed as a graph where vertices correspond to the sentence and the edges represent the similarity measure between them [Erkan and Radev, 2004]. Sentences are scored by expected probability of a random walker visiting each sentence [Paul et al., 2010]. In [Paul et al., 2010] edges correspond to the probability of two sentences to represent the same point of view. As LDA, weighted feature subset non-negative matrix factorization allows to obtain the most representative terms among the topics [Wang et al., 2010].

Lin et al. introduced a timestamped graph model [Lin et al., 2007]. In their approach a time stamp is viewed as a position of a sentence within the source document. Sentences are ranked by the page rank score and the similarity with a query.

Witte et al. introduced a fuzzy co-reference graph for multi-document summarization [Witte et al., 2007].

Zhang et al. proposed a summarization method based on graph representation of subtopics [Zhang et al., 2007]. The idea is to rank subtopics and to find sentences to support them. Sentence score is estimated by its length, position, chronological order, linguistic patterns, and word-based features.

Applying **machine learning** for summarization requires a corpus consisting of original texts and corresponding summaries. Text corpora provide much useful information on features which should be kept in a summary, how long a text should be, etc. [Lin and Hovy, 1997, Radev and McKeown, 1998]. A classifier should divide a set of all sentences into two parts, namely relevant (appearing in a summary) and not relevant. The aim is to minimize the mathematical expectation of loss function, i.e. the number of misclassified sentences [Amini et al., 2007].

$$L_C(h) = E([[yh(s) < 0]]) \tag{1.1}$$

where $L_C(h)$ is a loss function; $h(s)$ is a classification function for a sentence s that is equal to 1, if the sentence is considered relevant and -1 otherwise; y is the true class of a sentence s; and $[[predicate]]$ is equal to 1, if the *predicate* holds and 0 otherwise. A set of sentences can be partially ordered in the following way: $s > s' \leftrightarrow h(s) > h(s')$. Thus, the learning goal is to minimize the loss of the ranking function L_R:

$$L_R(h, D) = \frac{1}{|D|} \times \sum_{d \in D} \frac{1}{|S_d^1||S_d^{(-1)}|} \sum_{s \in S_d^1} \sum_{s \in S_d^{(-1)}} [[h(s) > h(s')]] \tag{1.2}$$

where d is a document from a collection D [Amini et al., 2007].

Hickle et al. applied machine reading framework for multi-document summarization. Their approach called GISTEXTER presupposes knowledge acquisition from a text collection and knowledge base by recognition of textual entailment relationships between discourse commitments [Hickl et al., 2007]. Textual entailment recognition is used to filter candidate sentences that entail or contradict hypotheses in the current knowledge base. This approach obtained very competitive results at DUC 2007.

The PYTHY Summarization System presented by Microsoft Research is a supervised learning algorithm that uses sentence features (position, length, etc.) and term frequency features in the original and reduced sentences [Toutanova et al., 2007]. Fisher and Roark also applied machine learning, namely, perceptron classifier with query-neutral and query-focused features [Fisher and Roark, 2006]. Li et al. used support vector machine based on word frequencies, named entities, WordNet semantics, centroid features and sentence position [Li et al., 2007b].

Supervised machine learning (e.g. decision trees, Bayes classifier etc.) could help to extract key words. Usually features are represented by the frequencies of unigrams, bigrams and trigrams [Ercan and Cicekli, 2007, Turney, 2000], named entities, relative sentence length [Turney, 2000], position within a text (including the first and the last occurrences), document structure and lexical chains [Angheluta et al., 2002, Ercan and Cicekli, 2007].

Lexical chains could be used to analyze lexical coherence of texts [Morris and Hirst, 1991].

Lexical chains are computed by Silber and Mccoy proposed a linear algorithm to compute lexical chains [Silber and Mccoy, 2002].

Morris and Hirst introduced the idea of lexical chain implementation based on Roget dictionary [Morris and Hirst, 1991]. Hirst and St-Onge proposed to use WordNet to the same purpose [Hirst and St-Onge, 1998]. Barzilay and Elhadad were the first who applied lexical chains for single document summarization [Barzilay and Elhadad, 1997]. Li et al. enhanced this strategy by applying lexical chains with WordNet similarity for multi-document summarization [Li et al., 2007a]. They used nouns, noun compounds and named entities to build lexical chains and to select sentences. Lexical chains are build with regard to word frequencies and synsets' similarity. Sentences are ranked by the total of lexical chain score of their words, similarity with a query and named entity similarity with it. Chali and Joty's approach includes lexical chains and basic element extraction [Chali and Joty, 2007].

Another direction of sentence ranking is using linguistic knowledge.

Linguistic methods fall into several categories:

- rule-based approaches, which may be combined with statistics [Lin, 1998, Lin and Hovy, 1997];
- methods based on genre features, text structure etc. [Barzilay et al., 1999, Lin and Hovy, 1997, Seki, 2005, Teufel and Moens, 2002]; methods based on syntax analysis [Barzilay et al., 1999, Teufel and Moens, 2002].

Let us look at some examples.

One of the first summarizers was the domain-specific rule-based system SUMMONS which had an extraction component and a linguistic module for checking the syntax [Radev and McKeown, 1998]. The multi-document summarization system SUMMARIST combined statistical approach with rules for key-phrases, e.g. the most important, to conclude, to summarize etc. [Lin, 1998, Lin and Hovy, 1997]. The main idea was to identify the subject by frequency of words, their position within a text, key-phrases. Related terms were generalized, e.g. the notions waiter, meal and menu refer to the same concept restaurant.

Genre related features such as text structure are useful for summarization purposes [Seki, 2005]. For example, in news the most important information is written at the beginning of an article, while scientific papers have an abstract [Lin and Hovy, 1997]. Moreover, in scientific texts a sentence has a specific rhetorical status: research goals, methods, results, contribution, scientific argumentation or attitude toward other people's work. A rhetorical status may be assigned according to matching to a linguistic pattern, position in the text, use of key words, grammatical features (verb tenses, modal verbs etc.) [Teufel and Moens, 1998, 2002]. As for news articles, multiple descriptions of the same events are rather typical for them [Barzilay et al., 1999, Teufel and Moens, 2002]. So in the news articles the most important information tends to be mentioned several times [Barzilay et al., 1999]. However, the same idea may be expressed in different ways and therefore in the system MultiGen sentences are clustered by comparison of predicate-argument structure [Barzilay et al., 1999]. Besides that, different genres should be compressed at different rate, e.g. a news article may retain 25-30% of the original size while for a scientific paper this coefficient is about 3% [Teufel and Moens, 2002].

Semi-structured documents, e.g. documents in XML format, provide a lot of **metadata** as well as structural information that could be used in summarization. Structural features

include depth of the element in which the sentence is contained, sibling number of the element, number of sibling elements, position within the element etc. [Amini et al., 2007]. The Wikipedia is one of the largest sources of semi-structured documents. Tags referring to headers, categories, info-boxes, entities etc. can be used for summarization needs [Janod and Mistral, 2011].

As it is showed in [Delort et al., 2003], the context obtained through hyperlinks and integrated into the original document may also improve the quality of a summary. Moreover, various social networks provide a lot of user generated content associated with regular documents which can be useful for summarization task, e.g. users' comments and URLs posted on Twitter show the parts of a document the most interesting for users [Yang et al., 2011]. Comments may be linked by topic, quotation (one quotes another) and mention relations (replies) [Hu et al., 2008]. The importance of a comment can be estimated by the PageRank algorithm or tensor decomposition. The words appearing in many important comments are considered to make big contribution. Another option is to integrate the valuable comments into a document. Comments may be considered themselves as a set of documents to be summarized [Lu et al., 2009a]. In [Lu et al., 2009a] a summary is a set of tuples: topic aspect, its rank and a representative phrase. One of the way to define topic aspects is Topic-Aspect Model which is an extension of Latent Dirichlet Allocation [Paul and Girju, 2010].

1.2.3 Redundancy Treatment and Result Filtering

Redundancy treatment may be performed by applying Manifold-Ranking algorithm, which implies iterative selection of candidate sentences and is based on the assumption that sentences should provide different information and therefore they should not be similar [Wan et al., 2007]. For abstracts, graph-based approaches can be applied [Ganesan et al., 2010].

In the context sensitive approach SumBasic introduced in [Nenkova and Vanderwende, 2005-01] the term probability is reduced when the term occurs in a selected sentence so the terms with lower probability are more likely to be found within newly selected sentences bearing non-redundant information. A similar reasoning underlies Maximal Marginal Relevance (MMR) that presupposes the minimal similarity of a candidate sentence with the sentences already included into the summary [Carbonell and Goldstein,

1998b]. Fillipova et al. slightly modified the MMR method by eliminating redundant sentences that are very similar to a query [Filippova et al., 2007].

Reeve and Han compared the distribution of terms within a source text with the one of a summary [Reeve and Han, 2007].

Conroy et al. introduced an algorithm for redundancy treatment based on Traveling Salesman Problem [Conroy et al., 2007]. Amini and Usunier applied this algorithm at DUC 2007 by filtering sentences that have more than 8 terms in common [Amini and Usunier, 2007].

Hickle et al. proposed to cluster sentences and for each cluster keep only one sentence that contains the most information [Hickl et al., 2007]. Clustering was also used in [Ying et al., 2007]. The difference is that from each cluster they kept a sentence that maximizes the difference with other summary sentences.

Toutanova et al. defined a dynamic sentence score, which is the score of a sentence as a continuation of a given partial summary where the values of some features are discounted to avoid redundancy [Toutanova et al., 2007].

Verma et al. grouped candidate sentences by pair-wise distances threshold and selected highest-ranked one from each group [Verma et al., 2007].

For news summarization, Chali and Joty proposed to discount sentences with the same dates since the respective documents often describe the same event [Chali and Joty, 2007]. In addition, they filtered out the sentences with the basic element overlap greater than the predefined threshold.

Madnani et al. compared the word frequencies within a summary with those in the general language [Madnani et al., 2007].

Stokes et al. filtered out sentences if their cosine similarity to the sentences already included in a summary is greater than a predefined threshold [Stokes et al., 2007].

1.2.4 Sentence Re-ordering

In single-document summarization systems it is possible to use original sentence order. The idea was adopted by Majority Ordering algorithm for multi-document summarization. Subjects (sentences expressing the same meaning) T_i are organized into a directed graph were edges present the number of documents where T_i is followed by T_j and the best order corresponds to the Hamiltonian path of maximal length [Barzilay et al., 2002]. Another approach is to assign time stamp to each event and to order them chronologically. The use of chronological ordering is restricted to the news articles on the same topic [Barzilay et al., 2002]. Diversity topics in the news demand another way to arrange sentences extracted for multi-document summarization. Application of a text corpus provides the ground for improving readability. In this case the optimal order is found by the greedy algorithm maximizing the total probability [Lapata, 2003]. In a narrative text verbs and adjectives play an important role in the semantic relations between sentences [Asher and Lascarides, 2003]. Specific ordering is applied to verb tenses [Lapata, 2003]. We took advantage of the graph representation and chronological ordering in our algorithm.

In [Zhang et al., 2007] sentence re-ordering is based on document time-stamps and sentence position within a document. No further details are provided. Ying et al. re-ordered sentences according to its timestamps in the original document [Ying et al., 2007]. In Filippova's approach sentences from the same document are bunched up and ordered as in the original document [Filippova et al., 2007]. To ensure local coherence of a summary, Hickle et al. used a hierarchical clustering algorithm to re-order sentences that contain similar information [Hickl et al., 2007]. Mihalcea used directed backward graph where the edges are oriented from a sentence to previous sentences in the text [Mihalcea, 2004].

Although automatic text summarization task has been studied for about 60 years, the majority of existing approaches try to deal only with a passage ignoring its context and quality. For a short summarization unit like a sentence context is crucial to understanding since often a sentence without a context is meaningless or ambiguous. Therefore, we introduce an algorithm for sentence ranking that considers not only a candidate sentence but also its left and right neighbors. Besides, the approach that we propose takes into account sentence appropriateness for summarization. To our knowledge there is no works that exploits the topic-comment structure of a sentence for automatic summarization. We

incorporate the knowledge of the topic-comment structure to our algorithm. Moreover, there are few works focusing on sentence re-ordering despite it seems to be extremely important to readability. Thus, we propose two novel graph-based methods for sentence re-ordering.

1.3 Contribution 1: Sentence Ranking Approach for Message Contextualization

Sentence retrieval in our method is based on the similarity to the query, i.e. a short text (tweet) to be contextualized, which is one the most wide spread approaches for summarization [Amini and Usunier, 2007, Amini et al., 2007, Shen and Li, 2011-12, Soriano-Morales et al., 2011, Torres-Moreno et al., 2012b]. The most widespread retrieval models are Vector Space Model (VSM), namely TF-IDF, and Language Modelling (LM) [Lu, 2013]. Although some researchers showed the superiority of TF-IDF [Abdulmutalib and Fuhr, 2010], the others proved that LM and TF-IDF are strong correlated and achieves almost the same effect [Robertson, 2004] since LM weighting is similar to TF-IDF weighting scheme [Lu, 2013]. The difference is that the LM uses the collection frequency, while TF-IDF approach is based on document frequency. Moreover, LM does not directly allow weighting query terms. Linguistic features are easier to integrate in TF-IDF model. Thus, we preferred TF-IDF weighting scheme over LM.

The application of linguistics, especially named entity (NE) recognition, may improve information retrieval performance, including tweet study [de Oliveira et al., 2013, Mohammed and Omar, 2012, Nadeau and Sekine, 2007]. Moreover, we hypothesize that grammar analysis, namely part-of-speech tagging, may also ameliorate results. We assume that part-of-speech (POS) tagging can ameliorate results since in general some POS provide more information than others (e.g. nouns are more informative that adverbs or functional words). As in [Lioma and Blanco, 2009], we integrated POS weights into the TF-IDF measure.

Not all sentences are suitable for summarization purpose (e.g. headers, labels etc.). To avoid trash passages we enriched our method by sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio.

Usually, a sentence is viewed as a unit in summarization task. However, often a single sentence is not sufficient to catch its meaning and even human beings need a context. In contrast to [Yang et al., 2011], we believe that a context does not provide redundant information, but allows to precise and extend sentence meaning. Therefore, we introduce an algorithm to smooth a candidate sentence by its local context, i.e. the neighboring sentences from the source document. Neighboring sentences influence the sentence of interest, but this influence decreases as the remoteness of the context increases, which differs from the previous approaches where the dependence is considered to be binary (i.e. a neighboring sentence influences the sentence of interest or not) [Murdock, 2006]. The binary understanding of the influence of the context assumes that the influence is the same for all sentences. Hence, in this research *context* may refer to either neighboring sentences or the resulting summary explaining a tweet.

To contextualize tweets we consider both the tweet and a textual resource from which a summary is built. The summary is constructed after extraction of presumably the most relevant sentences from the textual resource. Usually a limited number of documents is sufficient for a summary [Filippova et al., 2007] since the most important information tends to be repeated in several documents [Barzilay, 2003]. Since sentences are much smaller than documents, general information retrieval systems provide worse results to sentence retrieval [Murdock, 2006]. Moreover, document retrieval systems are based on the assumption that the relevant document is about the query. However this is not enough for sentence retrieval, e.g. in question-answering systems the sentence containing the answer is much more relevant that the sentence which is about the subject. Without a context, a sentence often is hard to understand or ambiguous. Therefore, we believe that a sentence ranking algorithm should considers not only a candidate sentence but also its left and right context.

Our approach is based on three main procedures: POS tagging, named entity recognition and sentence scoring [Ermakova and Mothe, 2012a]. Unfortunately, the whole process is time consuming. For this reason, we do not apply this process on the entire textual resource but rather on a sub-set of documents extracted from it. Thus, we first filter the documents, focusing on presumably the most related to the targeted tweet. Then, we apply the process mentioned above to select and order the main sentences from these filtered documents, using linguistic features, and to arrange the order of the extracted sentences.

The main steps of the algorithm are the following:

- preprocessing;
- sentence ranking;
- readability improvement.

The method we use to contextualize tweets is described in details in the following subsections.

The preprocessing step depends on the collection therefore we will present it later in evaluation section 1.5.

1.3.1 Sentence Ranking

The objective of this step is to evaluate the sentences from the retrieved documents according to the degree of their importance for generating summary. We apply the descending order of ranking, thus the top sentences will be used to compose the summary.

Sentence retrieval is based on the similarity between a sentence and a targeted tweet. However, rather than just considering sentences as bag of words, we prefer to enrich sentences in order to get more information from their content. We enrich both the tweet to be contextualized and the retrieved documents by parsing them and annotating them using POS tagger and named entity recognizer. We expand a tweet by the terms for top ranked documents with the highest TF-IDF score.

The various parts of text are weighted differently according to their supposed importance. In addition, since sentence meaning depends on the context, we used an algorithm for smoothing from the local context. Thus, we use the term context in two senses: the context of a tweet and the context of a passage. The latter could be defined as "the parts of something written or spoken that immediately precede and follow a word or passage and clarify its meaning" [Pearsall, 2002].

Thus, the total sentence score (let call it informativeness) $Informativeness(S, Q)$ is a function of the query-independent sentence quality measure $Qual(S)$ and computed

query-dependent sentence score $score(S, Q)$:

$$Informativeness(S, Q) = f(Qual(S), score(S, Q)) \tag{1.3}$$

Let us discuss this formula in details.

1.3.1.1 Sentence Representation

Sentence quality measure $Qual(S)$ is used to avoid trash passages from real web collections.

We define it as the function of the lexical diversity $LexDiv(S)$, meaningful word ratio $Meaning(S)$ and punctuation score $PunctScore(S)$:

$$Qual(S) = \phi(LexDiv(S), Meaning(S), PunctScore(S)) \tag{1.4}$$

Lexical diversity allows avoiding sentences that do not contain terms except those from a query. Lexical diversity in our approach is defined as the number of different lemmas used within a sentence divided by the total number of tokens in this sentence.

Meaningful word ratio is also aimed to penalized sentences that either have no sense at all or are not comprehensible without large context. Meaningful word ratio is the number of non-stop words within a sentence over the total number of tokens in this sentence.

Besides unreadable passages, many symbols usually used as punctuation marks can be found in emoticons. Emoticons represents humans' attitude towards something. However, they are not relevant for informative, navigational nor transactional queries. Hence, $PunctScore(S)$ penalizes sentences containing many punctuation marks.

Punctuation score is estimated by the formula:

$$PunctScore(S) = 1 - \frac{PunctuationMarkCount(S)}{TokenCount(S)} \tag{1.5}$$

where $PunctuationMarkCount(S)$ is a total number of punctuation marks in the sentence, and $TokenCount(S)$ – is a total number of tokens in S. Thus, $PunctScore(S)$ shows the ratio of tokens which are not punctuation marks.

Thus, we believe that a good sentence should have high ratio of different meaningful words and reasonable ratio of punctuation.

Sentence quality is query-independent, while sentence score $score(S, Q)$ shows how well a sentence matches a query.

We assume that relevant sentences come from relevant documents. However, in real world search engines we do not know which documents are actually relevant. Therefore in our method sentence score depends on pseudo-relevance $DocRel(d, Q)$ of the corresponding document d assigned by a search engine (i.e. document rank, score or their combination), computed smoothed sentence relevance $R(S, Q)$, and a topic-comment score $TC(S, Q)$:

$$score(S, Q) = \omega(DocRel(d, Q), R(S, Q), TC(S, Q)) \tag{1.6}$$

We model a sentence as a set of vectors. The first vector represents the tokens occurred within the sentence (unigram representation). Tokens are associated with lemmas. A lemma has the following features: POS, frequency and IDF. The second vector corresponds to bigrams. In both vector representation stop-words are retrieved. However, functional words, such as conjunctions, prepositions and determiners, are not taken into account in the unigram representation. NE comparison is hypothesized to be very efficient for contextualizing tweets. Therefore, the third vector refers to found named entities. Thereby, the same token may appear in several vectors.

For each vector, we store only these components, whereas a token, a bigram or a named entity tends to appear no more than once within a sentence. Thus, it is no need to store the frequency of a component within the sentence.

We also exploit a sparse representation i.e. store only occurring components. The only operation is component-wise comparison. In order to perform it, components are sorted. This vector set representation allows combining the similarity measures obtained for different information types.

Thus, a prior sentence score is calculated as the function of unigram sim_{uni}, bigram sim_{bi} and NE sim_{NE} similarities:

$$sim_{total}(S, Q) = g(sim_{uni}(S, Q), sim_{bi}(S, Q), sim_{NE}(S, Q)) \tag{1.7}$$

As a restriction to prevent irrelevant results we can apply grammar filters such as POS distribution and syntactic constructions. For example, in the system for Automatical Genre Classification (AGC) the subset of POS are used in order to maintain performance across changes in the topical distribution [Petrenz and Webber]. According to our approach, several methods can be used to assign scores to words. The first method identifies stop-words by frequency threshold. The second method assigns different weights to different parts of speech (POS rank). Researchers asserts that nouns provide the most valuable information since they have the maximal generalizing capacity [Silber and Mccoy, 2002]. Though [Sokolov, 1968] argues that verbs represent the relationship between things and thus they may be key words as well. One can specify whether unigram vector components should be multiplied by this POS rank. POS ranking makes it possible to penalize unresolved anaphora and other readability shortcomings.

It is also possible to consider or not IDF.

Thus, sim_{uni} is a similarity measure between unigram vectors $uni(w)$ that takes into account term, POS weights and IDF of the term w:

$$uni(w) = IDF(w) \times POS(w) \tag{1.8}$$

where $IDF(w)$ is IDF of the term w and $POS(w)$ is the corresponding POS weight.

NE vectors are treated in the following way. For each NE in a query we searched for corresponding NEs in the candidate sentences. If a query does not contain NEs, all candidate sentences are considered to match the query with regard to this information type. The NE similarity measure is computed by the formula:

$$sim_{NE}(S, Q) = \frac{NE_{common} + NE_{weight}}{NE_{query} + 1} \tag{1.9}$$

where NE_{weight} is a positive floating point parameter given by a user (by default it is equal to 1.0), NE_{common} is the number of NE appearing in both query and sentence, NE_{query} is the number of NE appearing in the query.

The sentence may not contain a NE from the query and it can be still relevant. To avoid sim_{NE} to be equal to 0, NE_{weight} is used. We also add 1 to the denominator to avoid devision by zero. However, if smoothing is not performed the coefficient will be zero. We considered only the exact matches of NE.

1.3.1.2 Smoothing from Local Context

A text should be integral and coherent. In discourse, integrity and coherence are implemented by contextual predictability, i.e. a text unit depends on its left neighbors [Yagunova, 2008]. This assumption is often used in speech recognition and text generation, namely as Markov chains [Rabiner, 1990]. In these models context importance is viewed as a step function equal to 1 when the distance is smaller than k (symbols/-words/sentences) and 0 otherwise. Moreover, the majority of the models consider only left context. In contrast, we assume that the importance of the context linearly reduces as the distance increases. We contend that the right context should also be taken into account. This statement may be supported by the following facts:

- In language, besides anaphora, there exists an opposite phenomenon, i.e. cataphora which is used to insert an expression or word that co-refers with a later expression [Cutting, 2002].
- POS are interrelated and if it is impossible to see a verb after a preposition, it is also impossible to see a preposition before a verb.
- The same interaction is observed in lexics.

General approach to document IR is underlined by TF-IDF measure. In contrast, usually the number of each query term in a sentence is no more than one [Murdock, 2006]. Traditionally, sentences are smoothed by the entire collection, but the method proposed in [Murdock, 2006] the same weight to all sentences from the context within a window. In contrast, we assume that the importance of the context reduces as the distance increases.

The proposed method of smoothed sentence relevance estimation is based on the first-stage local context analysis. So, our main hypothesis is that the nearest sentences should produce more effect on the target sentence meaning than others. For sentences with the distance greater than k this coefficient is zero. The total of all weights should be equal to one.

The system allows taking into account k neighboring sentences with the weights depending on their remoteness from the target sentence. In this case the total target sentence score $R(S, Q)$ is a weighted sum of scores of neighboring sentences r_i and the target

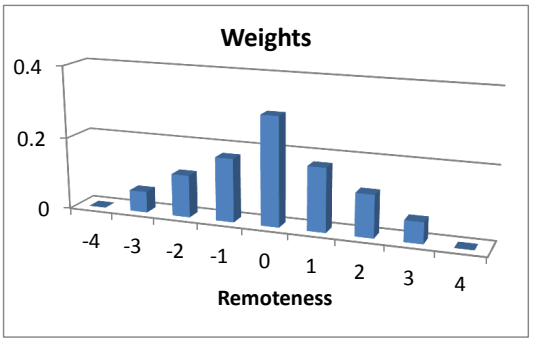


FIGURE 1.1: Dependence of neighboring sentence impact on distance

sentence r_0 itself:

$$R(S,Q) = \sum_{i=-k}^{k} w_i \times r_i \tag{1.10}$$

$$w_i = \begin{cases} \frac{1-w(S)}{k+1} \times \frac{k-|i|}{k} & 0 < |i| \leq k \\ w(S), & i = 0 \\ 0, & |i| > k \end{cases} \tag{1.11}$$

$$\sum_{i=-k}^{k} w_i = 1 \tag{1.12}$$

where $w(S)$ is the weight of the sentence S that is a tuning parameter, w_i and r_i are respectively the weights and the prior scores of the sentences from the context of S of k length. The weights become smaller as the remoteness increases (see figure 1.1). If the sentence number in left or right context is less than k, their weights are added to the target sentence weight $w(S)$. This allows keeping the sum equal to one. That is important since otherwise a sentence with a small number of neighbors (e.g. the first or last sentences) would be penalized (even if first sentences of a document are often considered to be very informative).

Besides smoothing from local context, we use smoothing from section beginning since we believe that first sentences from a section provides the most valuable and concise information about entire section content.

1.3.2 Topic-comment Relationship Integration

Linguistics establishes the difference between the clause-level topic and the discourse-level topic. However, within the bound of this research we are interested in clause-level topic only. The *topic* (or *theme*) is the phrase in a clause that the rest of the clause is understood to be about, and the *comment* (also called *rheme* or *focus*) is what is being said about the topic. In simple English clause the topic usually coincides with the subject, however it is not a case of the passive voice. In most languages the common means to mark topic-comment relation are word order and intonation. Moreover, there exist special constructions to introduce the comment. However, the tendency is to use so-called topic fronting, i.e. to place topic at the beginning of a clause.

We hypothesize that topic-comment relationship identification is useful for contextualization from the perspective of the related entities. Quick query analysis provides evidence that an entity is considered as a topic, while tweet content refers rather to comment, i.e. what is said about the entity. Moreover, we assume that providing the context to an entity implies that this context should be about the entity, i.e. the entity is the topic, while the retrieved context presents the comment. We used these assumptions for candidate sentence scoring. We double the weight of sentences in which the topic contains the entities E_i under consideration. Thus, the topic-comment score $TC(S,Q)$ is estimated as follows:

$$TC(S,Q) = \begin{cases} 2, & if E_i \in Topic(S) \\ 1, & otherwise \end{cases} \tag{1.13}$$

where $Topic(S)$ is the topic part of the sentence S. Topic identification is performed under assumption of topic fronting. We simplify this hypothesis by assuming that topic should be placed at the sentence beginning. Sentence beginning is viewed as the first half of the sentence.

1.3.3 Result Filtering

For a case of contextualization from the perspective of a related entity, we propose to apply entity filtering at the stage of document retrieval. We propose to keep documents that are relevant to the entities of interest only.

In order to deal with redundancy, we adopted the idea of H. G. Silber and K. F. Mccoy that nouns provide the most valuable information [Silber and Mccoy, 2002]. In our approach a sentence was mapped into a noun set. These sets were compared pairwise and if the normalized intersection was greater than a predefined threshold, the sentence with lower score is rejected.

1.4 Contribution 2: Sentence Re-Ordering

The retrieved sentences should be organized into a coherent text. If an extraction system deals with entire passages (which is our case), locally they may have higher readability than generated phrases since they are written by humans. Nevertheless, it is important to keep in mind the global readability of extracted passages. The only way to improve the readability of a text produced by an extraction system is to reorder the extracted passages. As Barzilay et al. showed, sentence ordering is crucial for readability [Barzilay et al., 2002]. That is why our next goal is to define an algorithm for sentence reordering, although sentence ordering was not evaluated at INEX.

Barzilay et al. proposed to order the sentences by searching for the Hamiltonian path of maximal length in a directed graph where vertices are themes and edges corresponds to the number of times when a theme precedes the other one. This approach requires a training corpus. In contrast to this, we hypothesized that in a coherent text neighboring sentences should be somehow similar to each other and the total distance between them should be minimal. Therefore, we propose an approach to increase global coherence of text on the basis of its graph model, where the vertices correspond to the extracted passages (i.e. isolated sentences or bunches of sequential sentences) and the edges represent the similarity measure between them. Under these assumptions the sentence ordering task implies searching for the minimal path that visits each vertex exactly once. This task is known as the traveling salesman problem. However, this method does not consider chronological constraints therefore we introduce another method based on the sequential ordering problem. In contrast to [Barzilay et al., 2002], our approach is not restricted by the news articles on the same topic and it takes advantages of the similarity between sentences.

In our approach we adopted an idea similar to [Soriano-Morales et al., 2011] and [Amini et al., 2007] since a summary was made of the sentences relevant to a query. The authors apply the query expansion technique to multi-document summarization. However, the technique neglects sentences from the entire texts, which include contextual synonyms to the query words. That is why we consider expanding, instead of the query, a candidate sentence by contextual synonyms to its words. Co-references may be viewed as contextual synonyms. Thus the list of the contextual synonyms was obtained by anaphora resolution performed by Stanford Core NLP. With regard to sentence ordering, we propose to combine graphical approaches with chronological constraints. Unlike Barzilay et al.'s method [Barzilay et al., 2002], we do not search for the Hamiltonian path of maximal length, but for the minimal one.

1.4.1 Model Description

As Barzilay et al. showed in 2002 sentence ordering is crucial for readability [Barzilay et al., 2002]. In single document summarization the sentence order may be the same as the initial relative order in the original text. However, this technique is not applicable to multi-document summarization. Therefore, we propose an approach to increase global coherence of text on the basis of its graph model, where vertices represent sentences and edges correspond to the same cosine similarity measure as in searching for relevant sentences. If two relevant sentences are neighbors in the original text, they are considered as a single vertex. The hypothesis is that neighboring sentences should be somehow similar to each other and the total distance between them should be minimal. Firstly, we computed the similarity between sentences and reduced sentence ordering task to traveling salesman problem [Morozenko, 2008].

1.4.1.1 Traveling Salesman Problem for Sentence Re-Ordering

The traveling salesman problem (TSP) is an NP-hard problem in combinatorial optimization. Given a list of cities and their pairwise distances, the task is to find the shortest possible route that visits each city exactly once and returns to the origin city. In the symmetric case, TSP may be formulated as searching for the minimal Hamiltonian cycle in an undirected graph (see 1.2). Asymmetric TSP implies a directed graph. The obvious solution is to use brute force search, i.e. find the best solution among all possible

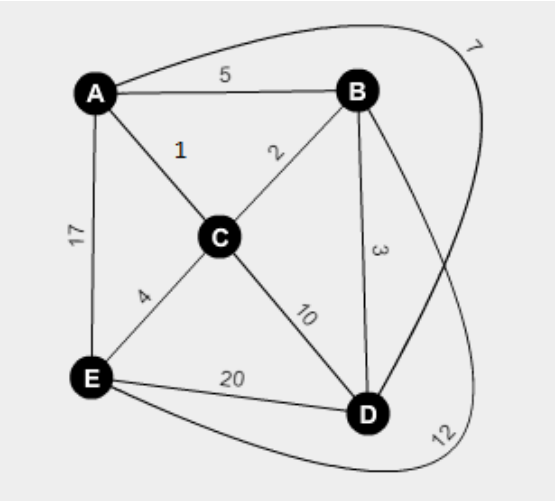


FIGURE 1.2: Example of the graph representation of a text for the TSP sentence re-ordering method (vertices represent sentences and edges correspond to the same cosine similarity measure)

permutations. The complexity of this approach is $O(n!)$ while other exact algorithms are exponential. Therefore, we chose the greedy nearest neighbor algorithm with minor changes. Since sentence ordering does not request to return to the starting vertex and the starting vertex is arbitrary but the choice of the starting vertex is crucial for the greedy algorithm, we tried every vertex as the starting one and chose the best result, i.e. the starting vertex giving the path of the minimal length. However, this method does not consider chronological constraints. So, we modified the task and it gave us the sequential ordering problem (SOP).

1.4.1.2 Sequential Ordering Problem

SOP "is a version of the asymmetric traveling salesman problem (ATSP) where precedence constraints on the vertices must also be observed" [Hernádvölgyi, 2003]. SOP is stated as follows. Given a directed graph, find a Hamiltonian path of the minimal length from the starting vertex to the terminal vertex observing precedence constraints.

Usually SOP is solved by the means of integer programming. Integer programming is NP-hard and these methods achieved only limited success [Hernádvölgyi, 2003]. Therefore, we solve the problem as follows.

Let $S = \{s_i\}_{i=1,n}^{-}$ be a set of sentences, where n is the total number of sentences. As in case of our TSP approach for sentence re-ordering, vertices and edges represents sentences

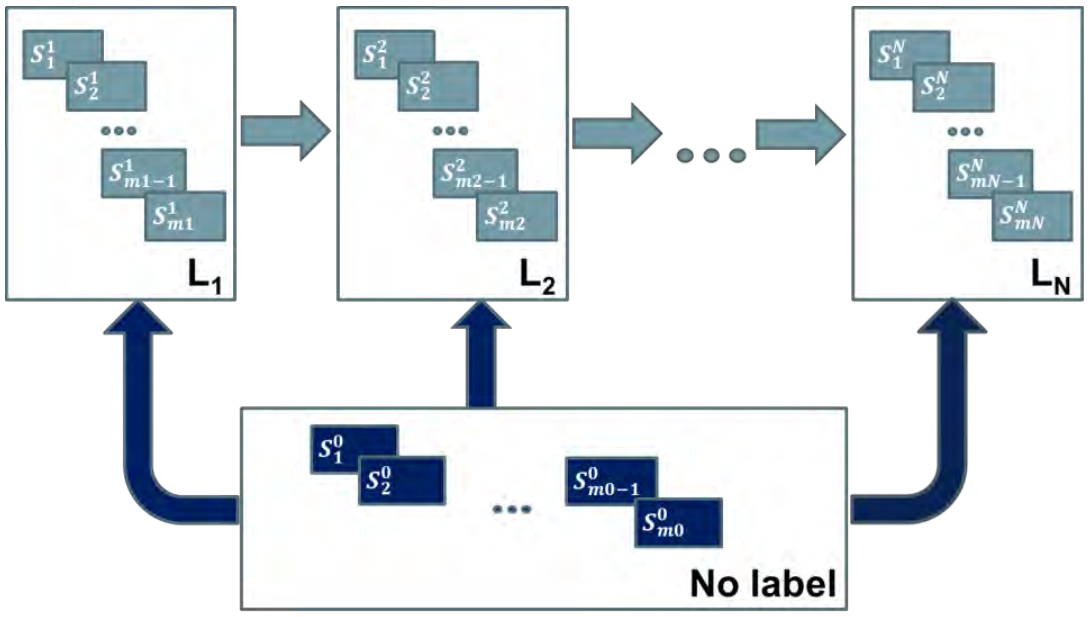


FIGURE 1.3: SOP sentence reordering algorithm

and the pair-wise similarity measures between them. Firstly, we group together sentences with identical time stamps assigned by a parser (Stanford CoreNLP in this case). Thus, $L = \{L_j\}_{j=1,\overline{N}}$, where N is the number of different time stamps, $L_j = s_1^j - s_2^j - \ldots - s_{mj}^j$, mj is the number of sentences with a given time stamp. Sentences without time stamp were added to the set $L_0 = s_1^0 - s_2^0 - \ldots - s_{m0}^0$, where $m0$ is the number of sentences without time stamp. We order groups L_j. Within each group L_j, we order sentences as in TSP approach. Then in each pair $s_k^j - s_{k+1}^j$ we try to insert each sentence x_0 from L_0 by turns. If the path $s_k^j - x_0 - s_{k+1}^j$ is smaller than the path $s_k^j - s_{k+1}^j$, we insert x_0, we remove it from L_0 and we try to insert the next sentence x_1 such as the path $s_k^j - x_0 - x_1 - s_{k+1}^j$ is smaller than the path $s_k^j - x_0 - s_{k+1}^j$. The process stops when L_0 is empty or the insertion does not diminish the path. If the set L_0 is not empty when all groups are ordered, we search for the shortest path passing from the last sentence in L_N through all vertices in L_0 to the first sentence in L_1 and the edge with the maximal weight is removed. The description of the algorithm is given in the figure 1.3.

1.4.1.3 Combining Informativeness and Readability

We defined F-measure to combine informativeness and readability:

$$F = \frac{Informativeness \times Readability}{\alpha \times Informativeness + (1-\alpha) \times Readability} \tag{1.14}$$

where $Readability = 1 - Length(Path)$, $Length(Path)$ is the length of the best path, $Informativeness$ is the total informativeness, and α is a parameter. To get the best score by this F-measure it is possible to apply rucksack problem where integral measure of relevance and readability corresponds to value and word number refers to weight. As candidate sentences we took the top relevant sentences that have in total twice more words than the maximal length n of the resulting summary. After selecting the most relevant sentence of $2n$ words, we obtained a rucksack problem. The knapsack problem or rucksack problem is stated as follow: given a set of items, each with a weight and a value, find the subset of this set to pack the rucksack so that the total weight is less than or equal to a given capacity and the total value is as large as possible [Kellerer, Hans et al., 2004]. As weight, we considered the number of words in a sentence, and the F-measure of relevance and readability represented value. We applied the branch and bound method, but it is possible to find more efficient way to solve the problem bearing in mind that triangle inequality is not hold.

1.5 Evaluation Framework

The method has been evaluated at INEX/CLEF tweet contextualization track. We report our evaluation results over the 4 years of the track. In this section we provide an evolution framework that we used. Firstly, we shall describe the data collection. Then we shall present the evaluation measures. The final subsection will provide system details.

1.5.1 INEX Data

We use INEX (Initiative for the Evaluation of XML Retrieval)[3] data for evaluation. INEX is an evaluation forum for XML IR that provides large structured test collections and scoring methods for IR system evaluation. INEX campaign aims at the evaluation of focused retrieval including passage retrieval from a long document, element retrieval from an XML document, page retrieval from books and question answering. It became a CLEF (Conference and Labs of the Evaluation Forum)[4] lab in 2012.

[3] http://inex.mmci.uni-saarland.de/
[4] http://clef2012.clef-initiative.eu/

TABLE 1.1: Test collections (INEX/CLEF Tweet Contextualization 2011-2014)

	INEX 2011	INEX 2012	INEX 2013	INEX 2014
Corpus	XML dump of English Wikipedia			
	April 2011	November 2011	November 2012	November 2012
Queries	132 tweets (tweet = title + 1-st snt of a NYT article)	1000 tweets from informative accounts	598 tweets from informative accounts	240 topics from RepLab 2013 (tweet + entity + category)
Evaluation (informativeness/ readability)	50 tweets / 53 tweets	50 tweets / 18 tweets	50 tweets / 10 tweets with the largest text references	50 tweets/12 summaries per run
Gold standards	New York Times articles Pool of relevant passages	Pool of relevant passages	Prior set of relevant pages Pool selection of submitted passages All relevant texts+10 random tweets	Pool of relevant sentences Pool of noun phrases

In 2011, the Question Answering Track aimed at evaluating tweet contextualization in terms of relevance of the retrieved information to tweets and readability of the presented results [SanJuan et al., 2012]. In 2012, this track was renamed to Tweet Contextualization.

Test collections are described in the table 1.1.

In 2011, the query data set included 132 tweets. A tweet consisted of the id (*id*), the title (*title*) and the first sentence (*txt*) of a New York Times (NYT) article released in July 2011.

Example 1.1. *Topics from Tweet Contextualization Task 2011*

```
<xml>
 <topic id="2011001">
                <title>At Comic-Con, a Testing Ground for Toymakers</title>
                <txt> THIS summer's hottest toys won't be coming to a toy aisle
   near you. The  only place to get them will be at Comic-Con International in
   San Diego.
                </txt>
 </topic>
 <topic id="2011003">
                <title>Obama to Back Repeal of Law Restricting Marriage</title>
                <txt> WASHINGTON - President Obama will endorse a bill to repeal
   the law that limits the legal definition of marriage to a union between a man
    and a woman, the White House said Tuesday, taking another step in support of
    gay rights.
                </txt>
 </topic>
</xml>
```

For each tweet, participants had to provide a summary up to 500 words in the TREC format as an answer that contextualized the tweet, i.e. answer the question "what is this

tweet about?". The summary should contain as much relevant information as possible, but not include irrelevant or redundant passages.

Example 1.2. *Output format for INEX/CLEF Tweet Contextualization Task*

```
<tid> Q0 <file> <rank> <rsv> <run_id> <text of passage 1>
<tid> Q0 <file> <rank> <rsv> <run_id> <text of passage 2>
<tid> Q0 <file> <rank> <rsv> <run_id> <text of passage 3>
```

The summary should be made solely of extracts from the XML dump of English Wikipedia (April 2011), totally 3,217,015 non-empty pages. All notes, history and bibliographic references were removed. Thus, a page was composed of a title (*title*), an abstract (a) and sections (s). A section had a header ((h)). Abstract and sections contained paragraphs (p) and entities (t) referring to other pages. The documents had the following DTD scheme:

Example 1.3. *Document DTD scheme for INEX/CLEF Tweet Contextualization Task*

```
<!ELEMENT xml (page)+>
<!ELEMENT page (ID, title, a, s*)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT a (p+)>
<!ELEMENT s (h, p+)>
<!ATTLIST s o CDATA #REQUIRED>
<!ELEMENT h (#PCDATA)>
<!ELEMENT p (#PCDATA | t)*>
<!ATTLIST p o CDATA #REQUIRED>
<!ELEMENT t (#PCDATA)>
<!ATTLIST t e CDATA #IMPLIED>
```

The summaries submitted by participants were compared to each other, to the baseline summary made of sentences (BaselineSum) and to the key terms (BaselineMWT). The baseline system was based on Indri index without stop word list and stemming (language model). Part of speech tagging was performed by TreeTagger. Summarization algorithm was TermWatch [SanJuan et al., 2012].

In 2012, the text corpus was presented by an updated Wikipedia dump from November 2011. The query set was dramatically changed. It consisted of approximately 1000 real tweets written in English collected from informative accounts such as @CNN, @TennisTweets, @PeopleMag, @science etc. However, the task remained the same: to provide a summary up to 500 words in the TREC format.

TABLE 1.2: Tweet example from Tweet Contextualization Task 2014

tweet_id	category	entity	topic	content
213051315880869888	automotive	Fiat	sales	Seeing a lot of #Fiat cars downtown these days. #Traffic

Example 1.4. *Topics from Tweet Contextualization Task 2012*

```
169125414692851713      "The European Commission approved our proposed
    acquisition of Motorola Mobility, moving us closer to closing the deal http
    ://t.co/1XJKvMFR"
169123516791263232      "For Valentine's Day, how @googlemaps can connect you to
    the people & places you love, even @ the country's biggest mall http://t.co/
    H39WJcwT"
```

In 2013 there were 598 tweets in English to be contextualized from the Wikipedia dump of November 2012.

In 2014 there were 240 tweets in English collected by the organizers of CLEF RepLab 2013. In 2014 participants should provide a context to tweets from the perspective of the related entities. Tweets were at least 80 characters long and do not contain URLs. A tweet had the following annotation types: the category (4 distinct), an entity name from the Wikipedia (64 distinct) and a manual topic label (235 distinct) (see an example Table 1.2). The context had to explain the relationship between a tweet and an entity. As in previous years it should be a summary extracted from a Wikipedia dump.

1.5.2 Informativeness Measurement

For all test collections, 50 tweets were selected to evaluate the informativeness of the summaries [SanJuan et al., 2012]. For each of those topics, all submitted passages were merged into a pool. Passages were sorted in alphabetic order and therefore each passage was judged whether it was relevant independently from others. Submitted summaries were compared with the corresponded pools of relevant passages. In 2011 summaries were also evaluated according to the overlap with the original New York Times articles. In 2013 the informativeness was estimated as the overlap of a summary with 3 pools of relevant passages [Bellot et al., 2013]:

- prior set (PRIOR) of relevant pages selected by organizers (40 tweets, 380 passages);
- pool selection (POOL) of the most relevant passages (1,760) from participant submissions for 45 selected tweets;
- all relevant texts (ALL) merged together with extra passages from a random pool of 10 tweets (70 tweets, 2,378 relevant passages).

In 2014, 2 gold standards (1/5 of the topics/tweets) were used:

- pool of relevant sentences per topic/tweet (SENT);
- pool of noun phrases (NOUN) extracted from these sentences together with the corresponding Wikipedia entry.

In 2011 the informativeness was estimated as the log difference:

$$Div(S,T) = \sum_{t \in T} \left| \log\left(\frac{f_{T(t)}}{f_T} + 1\right) - \log\left(\frac{f_{S(t)}}{500} + 1\right) \right| \tag{1.15}$$

where T is the set of terms in the pool of relevant passages, $f_{T(t)}$ is the frequency of a term t in the pool, f_T is the total number of terms in the pool, $f_{S(t)}$ is the frequency of a term t in a summary, f_S is the total number of terms in a summary. A term may refer to a unigram, a bigram (two consecutive lemmas in the same sentence) or a bigram allowing a gap up to two lemmas between its component (with 2-gap). The lower values of $Div(S,T)$ corresponds to higher matching of tokens in a pool and a summary.

Since 2012 the informativeness was evaluated by the following formula:

$$Dis(S,T) = \sum_{t \in T} \frac{f_{T(t)}}{f_T} \times \left(1 - \frac{\min(\log P, \log Q)}{\max(\log P, \log Q)}\right) \tag{1.16}$$

where P and Q are computed as:

$$P = \frac{f_{T(t)}}{f_T} + 1 \tag{1.17}$$

$$Q = \frac{f_{S(t)}}{f_S} + 1 \tag{1.18}$$

Since $\frac{f_{T(t)}}{f_T} \in (0,1]$ and $\frac{f_{S(t)}}{f_S} \in (0,1]$, $P > 1$ and $Q > 1$. Therefore, $\max(\log P, \log Q) > 1$.The complement of this dissimilarity measure $1 - Dis(S,T)$ has similar properties than usual IR Interpolate Precision measures. The logarithm allows dealing with highly frequent words. The evaluation toolkit was based on Porter stemmer. The lower values of $Dis(S,T)$ correspond to the higher informativeness.

1.5.3 Readability Measurement

The same topics/tweets were used to evaluate readability of the summaries [SanJuan et al., 2012]. The readability evaluation was performed manually. For each passage in each summary, assessors should indicate if the passage contained one of the following drawbacks:

- The passage has syntactical problems (e.g. bad segmentation).
- The passage contains an unresolved anaphora.
- The passage has redundant information (that is to say, information which is already mentioned).
- The passage is meaningless in the given context (i.e. it is marked as trash).

Assessors were not asked to evaluate the relevance of the summaries. There were two metrics:

- Relaxed metric: a passage was considered valid if it was not marked as trash.
- Strict metric: a passage was considered valid if it did not have any problems mentioned above.

In 2011, the readability of summaries was estimated as the number of words (up to 500) in valid passages [SanJuan et al., 2012]. Since 2012, the score of a summary was the average normalized number of words in valid passages [Bellot et al., 2013]. Sentence ordering was not judged by conference organizers, however it is quite important for text understanding [Barzilay et al., 2002].

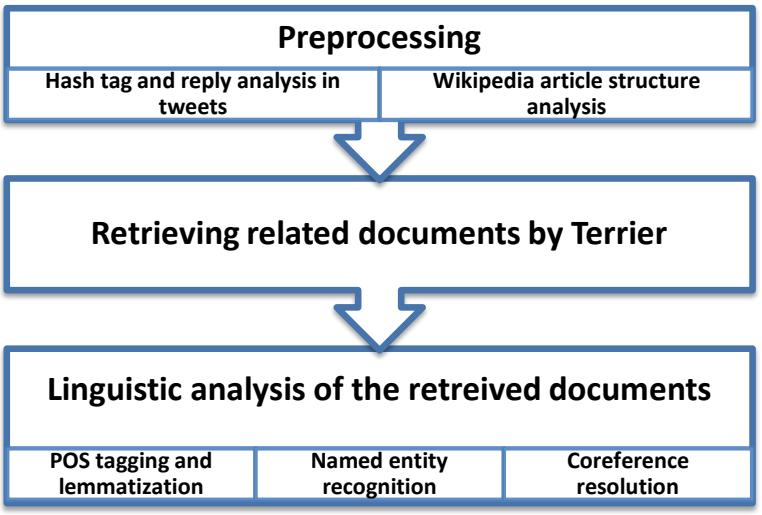


FIGURE 1.4: Documents and tweets preprocessing step for Tweet Contextulization

1.5.4 System Details

The first step of the algorithm is the preprocessing phase presented in the figure 1.4.

Query (tweet) preprocessing involves **hashtag and reply treatment** as well as combining different query parts.

While perceiving a text, a human uses various pivots such as spaces, punctuation marks, repeating words or phrases etc. [Залевская, 2001]. In our system Twitter hash tags and replies as well as article structure are viewed as pivots.

The hashtag symbol # "is used to mark keywords or topics in a Tweet. It was created organically by Twitter users as a way to categorize messages" and facilitate a search [Twitter Help Center | What Are Hashtags ("#" Symbols)?, accessed date: 02/08/2012]. Hashtags are inserted before relevant keywords or phrases anywhere in tweets ("At least 18 people injured in Kansas City, #Missouri, gas explosion", "RT @BBCNewsUS: Do you know Tony Mendez? As we head towards the #Oscars, get to know the man behind the true story of #Argo", "Patient with mysterious #SARS-like virus has died in British hospital: http://t.co/ICExnRbE via @AP", "PM #Rajoy confirms at #DEN2013 that Spanish #deficit for 2012 will be under 7% thanx to Government's plan to save more than 21.000 million"). Popular hashtags often represents trending topics. Bearing it in mind, we put higher weight to words occurring in hashtags. Usually key phrases are marked as a single hashtag.

Important information may be found in @replies, e.g. when a user reply to the post of a politician or other famous person. "An @reply is any update posted by clicking the "Reply" button on a Tweet" [Twitter Help Center | What are @Replies and Mentions?, accessed date: 02/08/2012]. Since people and organizations may use their names as Twitter accounts we treat them analogically to hashtags, i.e. they are split by capitalized letters (e.g. "New Course from @waikato, New Zealand - Data Mining with Weka. Starts Sep 9" refers to The University of Waikato, in "3 Divas at @VanityFair Oscars party- @BarbraStreisand @OfficialAdele and @DameShirley Bassey" it is mentioned the USA magazine Vanity Fair, singers Barbra Streisand, Adele and Shirley Bassey).

We split hashtags and replies by capitalized letters. An initial tweet is expanded by the words obtained from tweet hashtags and replies as stated above. Tweet preprocessing involves hashtag and reply treatment as well as combining different tweet parts. Thus, a tweet *RT StateDept: #SecKerry: Europe is strong, and stronger together. Europe and the US together have an opportunity to create jobs, build a stronger future* is expanded by *State, Dept, Sec, Kerry.*

The next step is **filtering the documents** that are supposed to contain relevant information to the tweet components from the textual resource. For this, we simply use a search engine. We use the tweet as a query and the textual resource as the document collection the query is evaluate on.

In order to obtain preliminary ranking we used Terrier[5] [Ounis et al., 2006b], an open-source search engine developed by the School of Computing Science, University of Glasgow. This platform considers documents as bags of words. It implements various weighting and retrieval models and allows stemming and blind relevance feedback.

As a retrieval model we applied $InL2c1.0$. It is a default retrieval model in Terrier. InL2c1.0 is a DFR (divergence from randomness) model based on $TF - IDF$ measure with $L2$ term frequency normalization [Amati and Van Rijsbergen, 2002b, He and Ounis, 2005]. This model is based on the assumption that informative words are relatively more frequent in relevant documents than in others. $InL2$ demonstrates better performance at many recall levels and in average precision than traditional retrieval models such as $BM25$ [Amati, 2003]. $L2$ normalization is less sensitive to document length. In the In

[5] http://terrier.org/

model the weight $weight(t, d)$ of the term t in the document d is estimated as follows:

$$weight(t, d) = \frac{1}{tf + 1} \times Inf_1(tf) \tag{1.19}$$

$$Inf_1(tf) = tf \times \log_2 \frac{N + 1}{n + 0.5} \tag{1.20}$$

where tf is the initial frequency of the term t in the document d, N is the total number of documents in the collection, and n is the number of documents containing the term t. $L2$ normalization of the term frequency tf_n is computed as:

$$tf_n = tf \times \log_2 \left(1 + c\frac{avg_l}{l}\right), \; c > 0 \tag{1.21}$$

where l is the length of the document d, avg_l is the average document length, and c is the normalization parameter. Thus, $weight(t, d)$ determined by $InL2$ is:

$$weight(t, d) = \frac{tf_n}{tf_n + 1} \times \log_2 \frac{N + 1}{n + 0.5} \tag{1.22}$$

We used the default value of $c = 1.0$ ($InL2c1.0$). Stemming was performed by Porter's algorithm [Porter, 1997a].

The third step is linguistic analysis. Retrieved **texts and tweets are parsed by Stanford CoreNLP**[6] which integrates such tools as POS tagger [Toutanova et al., 2003], named entity recognizer [Finkel et al., 2005], parser and the co-reference resolution system. It uses the Penn Treebank tag set [Marcus et al., 1993].

The last step of the preprocessing is **merging of the annotation obtained by the parser and Wikipedia tags**.

The sentence informativeness $Informativeness(S, Q)$ is computed as the product of sentence quality measure $Qual(S)$ and computed sentence score $score(S, Q)$:

$$Informativeness(S, Q) = Qual(S) \times score(S, Q) \tag{1.23}$$

[6] http://stanfordnlp.github.io/CoreNLP/

Sentence quality measure is estimated as the product of the lexical diversity $LexDiv(S)$, meaningful word ratio $Meaning(S)$ and punctuation score $PunctScore(S)$:

$$Qual(S) = LexDiv(S) \times Meaning(S) \times PunctScore(S) \tag{1.24}$$

The sentence score $score(S, Q)$ is calculated as the product of $DocRel(d, Q)$, $R(S, Q)$, and $TC(S, Q)$:

$$score(S, Q) = DocRel(d, Q) \times R(S, Q) \times TC(S, Q) \tag{1.25}$$

The similarity between a sentence and a query was computed as follows:

$$sim_{total}(S, Q) = sim_{uni} \times sim_{NE} \tag{1.26}$$

For unigram and bigram vectors, we computed three similarity measures, namely cosine, Jaccard and dice coefficients, between a sentence and a target tweet (sim_{uni} and sim_{bi} respectively).

Determiners have zero weights, proper names have the highest weights, and nouns have greater weights than verbs, adjectives and adverbs.

Thus, a sentence score similarity is estimated as the weighted sum or the product of sim_{uni}, sim_{bi} and sim_{NE}:

$$sim_{total}(S, Q) = w_{uni} \times sim_{uni} + w_{bi} \times sim_{bi} + +w_{NE} \times sim_{NE} \tag{1.27}$$

where w_{uni}, w_{bi}, w_{NE} are coefficients showing the impact of each component into the total.

$$sim_{total}(S, Q) = sim_{uni} \times sim_{bi} \times sim_{NE} \tag{1.28}$$

Example 1.5. *Example of the summary produced for Tweet Contextualization Task 2014*

```
264424350810263552      Q0      4265    1       26.3521 irit_etc_entity Total
    Chrysler vehicle production was about 1.58 million that year.
264424350810263552      Q0      4265    2       24.4135 irit_etc_entity Chrysler
    plans for Lancia to codevelop products, with some vehicles being shared.
264424350810263552      Q0      4265    3       23.8969 irit_etc_entity Chrysler
    is the smallest of the Big Three U.S. automakers(Chrysler Group LLC, Ford
    Motor Company, and General Motors).
```

```
264424350810263552      Q0      4265    4       21.4059 irit_etc_entity The sale
    of substantially all of Chrysler's assets to New Chrysler, organized as
    Chrysler Group LLC was completed on June 10, 2009.
264424350810263552      Q0      4265    5       21.3911 irit_etc_entity Chrysler
    continues to develop the Ram hybrid.
264424350810263552      Q0      4265    6       18.6103 irit_etc_entity Its core
    brands are: Chrysler, Jeep, Dodge, Ram, SRT, Fiat, and Mopar vehicles and
    products.
264424350810263552      Q0      4851104 7       18.0581 irit_etc_entity The
    vehicles were the electric-only Dodge EV sports car, the range-extended
    Chrysler EV minivan and the range-extended Jeep EV.
264424350810263552      Q0      4265    8       16.6391 irit_etc_entity Chrysler
    is the world's 13th largest vehicle manufacturer as ranked by OICA in 2010.
264424350810263552      Q0      4851104 9       15.8236 irit_etc_entity Chrysler'
    s new owner Fiat SpA disbanded the division in November 2009, The first
    electric vehicle planned from Fiat-Chrysler is an electrified Fiat Doblo
    light commercial van.
264424350810263552      Q0      4265    10      15.5848 irit_etc_entity Following
     the introduction of the Chrysler, the Maxwell brand was dropped after the
    1925 model year.
264424350810263552      Q0      4265    11      15.1500 irit_etc_entity Chrysler
    has also been experimenting with a Hybrid Diesel truck for military
    applications.
264424350810263552      Q0      4265    12      15.0538 irit_etc_entity Chrysler
    acquired the Jeep brand as part of the purchase of American Motors(AMC) on
    August 5, 1987, for somewhere between US$ 1.7 billion and $ 2 billion,
    depending on how costs were counted.
264424350810263552      Q0      4265    13      14.8875 irit_etc_entity Earlier
    in October, 2012, inaccurate reports had suggested that Chrysler's Jeep brand
     is considering moving all production to China.
264424350810263552      Q0      4265    14      13.5487 irit_etc_entity Chrysler
    is in the Advisory Council of the PHEV Research Center.
264424350810263552      Q0      4265    15      13.4989 irit_etc_entity In March
    2011, Chrysler Group LLC filed a lawsuit against Moda Group LLC(owner of Pure
     Detroit clothing retailer) for copying and selling merchandise with the
    Imported from Detroit slogan.
264424350810263552      Q0      4265    16      13.2261 irit_etc_entity Yanase Co
    ., Ltd. is currently the exclusive retailer of all imported Chrysler products
    (Chrysler, Jeep, Dodge) to Japanese consumers.
264424350810263552      Q0      4265    17      12.7996 irit_etc_entity Under
    DaimlerChrysler, the company was named DaimlerChrysler Motors Company LLC,
    with its U.S. operations generally called the Chrysler Group.
264424350810263552      Q0      1700208 18      12.7397 irit_etc_entity This
    plant was owned and operated by Chrysler before the acquisition of Jeep by
    Chrysler.
264424350810263552      Q0      4265    19      12.1359 irit_etc_entity After
    Chrysler's restructuring, the warranty program was replaced by five-year/100
    ,000 mile transferrable warranty for 2010 or later vehicles.
```

```
264424350810263552      Q0      4851104 20      12.0538 irit_etc_entity Most
    references to ENVI were removed from Chrysler web sites in November 2009.
264424350810263552      Q0      4265    21      11.6204 irit_etc_entity During
    World War II, essentially all of Chrysler's facilities were devoted to
    building military vehicles(the Jeep brand came later, after Chrysler acquired
     American Motors Corporation).
264424350810263552      Q0      4851104 22      11.6200 irit_etc_entity In
    September 2008, ENVI revealed three " production intent" electric vehicles to
     the public and announced that Chrysler Group LLC will start bringing a
    portfolio of electric vehicles to showrooms in 2010.
```

1.6 Results

1.6.1 Informativeness

1.6.1.1 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2011

For the first run we used default settings (DEFAULT), namely: NE were considered with a coefficient 1.0; abstract had weight equal to 1.0, sections had score 0.8; headers, labels, ... were not taken into account; we removed stop-words; cosine similarity was applied; POS were ranked; each term frequency was multiplied by IDF. In the second run we changed the similarity measure to Dice similarity (DICE). The section weight was reduced to 0.7. The context was extended to two sentences in each direction and the target sentence weight was equal to 0.7. For NE we kept the weight equal to 1.0. In the third run we applied Jaccard similarity measure (JAC) and we set the weight to sections equal to 0.5.

Table 1.3 presents the comparison of the baseline systems and the submitted runs with regards to New York Times articles. All three runs are ranked higher than the baseline systems. The best result is given by JAC.

Table 1.4 provides comparison referring to the pool of relevant sentences. According to these evaluations, all runs we submitted are more relevant than the baselines. However, the best results were provided by the run with the default settings. We think that the opposite evaluation results obtained for New York Times articles and the pool of relevant passages from the Wikipedia may be explained by the different language models of these

TABLE 1.3: Log difference to New York Times articles (Tweet Contextualization 2011)

Rank	Run	Unigram	Bigram	With 2-gap	Average
1	JAC	**0.0447**	**0.076644**	**0.104925**	**0.076629**
2	DICE	0.044728	0.076659	0.104933	0.076646
3	DEFAULT	0.044739	0.076668	0.104937	0.076653
6	BaselineSum	0.046049	0.078101	0.10646	0.078084
15	BaselineMWT	0.047508	0.079385	0.10766	0.079387

TABLE 1.4: Log difference with the set of relevant passages (Tweet Contextualization 2011)

Rank	Run	Unigram	Bigram	With 2-gap	Average
1	DEFAULT	**0.048639**	**0.07867**	**0.105506**	**0.078697**
2	DICE	0.048781	0.078857	0.105747	0.07889
3	JAC	0.049083	0.079249	0.106195	0.079277
10	BaselineSum	0.053691	0.085915	0.114346	0.085881
19	BaselineMWT	0.055786	0.088604	0.117854	0.088701

collections. The pool of the relevant sentences from the Wikipedia contained 103 889 tokens, which gave a vocabulary of 19 037 words, and the original news articles with a vocabulary of 26 481 words contained 154 355 tokens [SanJuan et al., 2012]. So, the average word frequency differs for 9%. Moreover, these two corpora have different genres and consequently different structure. In our approach NE matching was extremely important and therefore we preferred to select sentences with proper nouns, but not pronouns and other type of references (e.g. American President instead of Barack Obama). In a news article authors try not to repeat themselves and they substitute NE by other words. Since relevant passages were selected without context, the majority of them tended to contain NE. Thus, there exist two main explanations of the opposite ranks: different language models of the collections and the pool peculiarities.

1.6.1.2 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2012

We submitted three runs to INEX/CLEF 2012. The first run A considered the unigram cosine between a query and a sentence only. The second run B took into account the linear combination of the unigram and bigram similarity measures but did not imply anaphora resolution. The third one C differed from B by resolved anaphora. All runs had the same hashtag processing and sentence reordering (SOP). Informativeness results

TABLE 1.5: Informativeness evaluation (Tweet Contextualization 2012)

Rank	Run	Unigrams	Bigrams	Skip bigrams	Average
4	Baseline	0.7864	0.8868	0.8887	0.854
15	C	0.8484	0.9294	0.9324	0.9034
16	B	0.8513	0.9305	0.9332	0.9050
17	A	0.8502	0.9316	0.9345	0.9054

for the submitted runs are presented in Table 1.5 (the ranking is given for automatic runs). Column Run corresponds to the run id, Unigrams, Bigrams and Skip bigrams represents the proportion of shared unigrams, bigrams and bigrams with gaps of two tokens respectively. According to informativeness evaluation, the impact of the linear combination of the unigram and bigram similarity measures is smaller than the impact of anaphora resolution.

1.6.1.3 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2013

In our run SMOOTH each sentence is smoothed by its local context and first sentences from Wikipedia article which it is taken from. The run NOSMOOTH has the same parameters except it does not have any smoothing. In our best run NONUM punctuation score is not taken into account, it has slightly different formula for NE comparison and no penalization for numbers. Among automatic runs our best run NONUM was ranked first (PRIOR and POOL) and second (ALL) over 24 runs submitted by all participants. Table 1.6 provides results of the best automatic systems presented by the participants. Our results are marked by *. The best results are set off in bold. According to bigrams and skip bigrams, our best run is NONUM, while according to unigrams the best run is SMOOTH. So, we can conclude that smoothing improves Informativeness. Another conclusion is that ranking is sensitive to the pool selection as well as to the choice of divergence.

1.6.1.4 Informativeness Results at INEX/CLEF Tweet Contextualization Task 2014

The first run (ETC) was performed by the system developed in 2013. Three fields (entity, topic and content) were treated as a query. An entity was treated as a single phrase. The

TABLE 1.6: Informativeness evaluation (Tweet Contextualization 2013)

Rank	Run	All.skip	All.big	All.uni	Pool.skip	Pool.big	Pool.uni	Prior.skip	Prior.big	Prior.uni
1	258	**0.894**	**0.891**	**0.794**	0.880	0.877	0.792	0.929	0.923	0.799
2	*NONUM**	0.897	0.892	0.806	**0.879**	**0.875**	0.794	**0.917**	**0.911**	0.790
3	*SMOOTH**	0.897	0.892	0.800	0.880	0.875	**0.792**	0.924	0.916	**0.786**
4	*NOSMOOTH**	0.897	0.892	0.801	0.881	0.875	0.793	0.923	0.915	0.787

TABLE 1.7: Informativeness evaluation (Tweet Contextualization 2014)

Rank	Run	SENT.uni	SENT.big	SENT.skip	NOUN.uni	NOUN.big	NOUN.skip
3	361	**0.7632**	**0.8689**	**0.8702**	**0.7903**	**0.9273**	**0.9461**
4	360	0.782	0.8925	0.8934	0.8104	0.9406	0.9553
5	*ETC**	0.8112	0.9066	0.9082	0.8088	0.9322	0.9486
6	*ENT**	0.814	0.9098	0.9114	0.809	0.9326	0.9489
8	*RESTR**	0.8152	0.9137	0.9154	0.8131	0.936	0.9513

second run (ENT) differed from ETC by double weight for sentences where the entity represented the topic. The third run (RESTR) was based on document set retrieved for the tweet and filtered by the results obtained for the entity. Thus, the document retrieved by using the field content as a query were rejected if they did not coincide with top-ranked documents retrieved by using the field entity. According to the evaluation performed on the pool of sentences, our runs ETC, ENT and RESTR were ranked 3-rd, 4-nd and 6-th; while according to the evaluation based on noun phrases, they got slightly better ranks, namely 2, 3 and 5 respectively. Thus, the best results among our runs were obtained by the system that merges fields entity, topic and content into a single query. The run #360 is better than our runs according to sentence evaluation; nevertheless, it showed worse results according to noun phrase evaluation. Our system is targeted at nouns and especially NEs. This could provoke the differences in ranking with respect to sentences and noun phrases. The run based on entity restriction showed worst results. This could be explained by the fact that filtering out the documents that are considered irrelevant to the entity may cause a big loss of relevant documents if they are not top-ranked according to entities. The results of ETC and ENT are very close. However, topic-subject identification slightly decreased the performance of the system. Yet we believe that finer topic-comment identification procedure may ameliorate the results.

TABLE 1.8: Readability results with the relaxed and strict metrics (Tweet Contextualization 2011)

Relaxed metric			Strict metric		
Rank	Run id	Score	Rank	Run id	Score
1	BaselineSum	447.3019	1	BaselineSum	409.9434
4	DEFAULT	417.3462	6	JAC	344.1154
8	JAC	409.4038	7	DEFAULT	339.9231
9	DICE	406.3962	8	DICE	338.7547
25	BaselineMWT	137.8000	24	BaselineMWT	148.2222

TABLE 1.9: Readability results (Tweet Contextualization 2012)

Rank	Run	Relevance	Syntax	Structure	Average
4	Baseline	0.6975	0.6342	0.5703	0.634
15	C	0.4964	0.4705	0.4204	0.4624
20	B	0.449	0.4203	0.3441	0.4045
21	A	0.4911	0.3813	0.3134	0.3953

1.6.2 Readability

1.6.2.1 Readability Results at INEX/CLEF Tweet Contextualization Task 2011

Table 1.8 reports readability results according to the relaxed and strict metrics that we obtained at INEX 2011. Though the system showed the best results according the relevance judgment, it was worse than the baseline in terms of readability. The major drawback was unresolved anaphora. Trash passages refer not only to readability, but also to relevance. Therefore relevance improvement and sentence reordering may solve this problem.

1.6.2.2 Readability Results at INEX/CLEF Tweet Contextualization Task 2012

Readability evaluation results of 2012 are presented in the Table 1.9. As informativeness score, readability evaluation also provides evidence that anaphora resolution has a stronger influence on average score than the use of bigram cosine: there are four other runs between the run B and the run C, which differ only by resolved anaphora. It increases dramatically the structure score.

TABLE 1.10: Readability evaluation (Tweet Contextualization 2013)

Rank	Run	MA	T	R	A	S
1	NONUM	**72.44**%	**76.64**%	67.30%	**74.52**%	**75.50**%
2	NOSMOOTH	71.71%	74.66%	**68.84**%	71.78%	74.50%
3	SMOOTH	71.35%	75.52%	67.88%	71.20%	74.96%

1.6.2.3 Readability Results at INEX/CLEF Tweet Contextualization Task 2013

In 2013 according to all metrics except redundancy our approach was the best among all participants (see Table 1.10). Runs were officially ranked according to mean average scores. Readability evaluation also showed that the run NONUM is the best by relevance, soundness and syntax. However, the run NOSMOOTH is much better in terms of avoiding redundant information. The runs SMOOTH and NOSMOOTH are close according readability assessment as well.

1.6.2.4 Readability Results at INEX/CLEF Tweet Contextualization Task 2014

In 2014 we received very low score for diversity and structure. This may be related to the fact that we decide not to treat this problem since in previous years their impact was small. Despite we retrieved the entire sentences from the Wikipedia, unexpectedly we received quite low score for syntactical correctness.

ENT demonstrated slightly higher results according to all readability measures except diversity. The differences of readability scores between RESTR and ETC are very small since these runs are very similar. The only difference is the documents used as sources of the retrieved sentences. However, all readability scores of RESTR are lower. This can be caused by lower quality of the documents or the influence of the informativeness on the assessor perception of readability.

TABLE 1.11: Readability evaluation (Tweet Contextualization 2014)

Rank	Run	Readability	Syntax	Diversity	Structure	Average
6	ref2013	91.74%	69.82%	60.52%	85.80%	76.97%
7	ref2012	91.39%	69.58%	60.67%	85.56%	76.80%
12	ETC	90.88%	68.89%	56.59%	80.88%	74.31%
13	ENT	91.23%	69.47%	54.93%	81.56%	74.30%
14	RESTR	90.10%	68.30%	53.83%	80.70%	73.23%

1.6.3 Result Summary

In 2011 our system showed the best results according the relevance judgment. In 2012 we modified our method by adding bigram similarity, anaphora resolution, hashtag processing, redundancy treatment and sentence reordering. However, we obtained lower results than in the previous year. Therefore, in 2013 we decided to not consider bigram similarity, anaphora resolution, nor redundancy treatment. We also used generalized POS (e.g. we merge regular adverbs, superlative and comparative into a single adverb group). To avoid trash passages we enriched our method by sentence quality measure based on Flesch reading ease test, lexical diversity, meaningful word ratio and punctuation ratio. Lexical diversity allows avoiding sentences that do not contain terms except those from the query. We define it as the number of different lemmas used within a sentence divided by the total number of tokens in this sentence. Meaningful word ratio over the total number of tokens in the sentence is aimed at penalizing sentences that either have no sense at all or are not comprehensible without large context. The punctuation score penalizes sentences containing many punctuation marks. Thus, we believe that a good sentence should have high ratio of different meaningful words and reasonable ratio of punctuation. In 2014 we integrated the analysis of the topic-comment structure. However, the best results among our runs was obtained by the system 2013. The worst results corresponds to the method that uses filtering. Nevertheless, we believe that further study of the topic-comment structure could improve results.

1.7 Contribution 3: Extension to Snippet Retrieval

1.7.1 Modifications

Our approach is generic enough to be applied for various tasks. Here, we consider one of them, namely snippet retrieval. Another extension (query expansion) is given in Section 2.3. A search engine returns a larger number of results that a user cannot examine all. Therefore, a search engine provides a user with snippets (small text passages appearing under a search result extracted from the document) to help in evaluating web page relevance before browsing it. Ideally, a snippet provides the information a user is searching for. Good snippets should contain the basic information units (e.g. sentence or XML entities), they should bounded in size and distinguish the given document from other search results.

We slightly modified the method applied for tweet contextualization for the INEX Snippet Retrieval Track 2012-2013:

- nominal sentences were not penalized;
- sentences were not re-ordered;
- we did not treat redundancy since in the single-document summarization the probability of redundant information is much lower, and snippets are short and should be generated fast.

In addition to these modifications to sentence scoring, we used two algorithms for the candidate passage selection. The first one is modeling sentence selection as a **knapsack problem** which we solved by the dynamic programming approach. The second one is to apply the **moving window** algorithm.

A snippet is limited up to 1-2 sentences (~150-300 symbols) but it should provide as much information about the underlying document as possible. Therefore, snippet retrieval can be viewed as a task of selecting passages of the maximal total importance under the restriction of the total weight. This task is known as a knapsack problem.

Definition 1.1. The knapsack problem or rucksack problem is stated as follow: given a set of items (sentences), each with a weight (the number od words/symbols, i.e. its

length) and a value (score), find the subset of this set to pack the rucksack so that the total weight is less than or equal to a given capacity and the total value is as large as possible.

As a weight, we consider the number of symbols, and the score represents a value. We are dealing with $0-1$ knapsack problem, which restricts the number of each kind of item to zero or one, since otherwise a snippet would have redundant information. The knapsack problem is also applicable to multi-document summarization including tweet contextualization.

We solve this problem by the basic dynamic programming algorithm $DP-1$ with an overall running time $O(nc)$ where n is the number of items (the number of candidate sentences in our case) and c is the knapsack capacity [Kellerer, Hans et al., 2004].

However, if each sentence within a document was greater than a predefined threshold (i.e. all sentences have more words/symbols than the maximal allowed number of words/symbols), the snippet would be an empty string. Therefore, we used the moving window algorithm to find the best scored passage (that may contain just a part of a sentence). At each step the first token is removed from a candidate passage and the tokens following the candidate passage are added while its total weight is no greater than a predefined threshold. The passage with the maximal score is selected as a snippet. Despite the most relevant information may occur in the too long sentences, snippets beginning in the middle of a sentence have lower readability. That is why, we penalize them. As opposed to the knapsack algorithm, the moving window is not suitable to tweet contextualization, as it is efficient only for very small extractive summaries. Summaries built by MW are exclusively made of consecutive sentences.

1.7.2 Evaluation

1.7.2.1 Data Description

For the Snippet Retrieval Track 2012, the data collection consists of the dump of the Wikipedia of October 2008 annotated with YAGO [Schenkel et al., 2007b] and 35 topics. Participants should provide 20 snippets per topic limited to 180 characters [Trappett et al., 2012b]. In 2013 the Snippet Retrieval track was using the same document collection

as the Tweet Contextualisation track, based on a dump of the English Wikipedia from November 2012. The set of topics is the same as in 2012. The DTD for the submission format is as follows.

Example 1.6. *The DTD for the submission format for Snippet Retrieval Task*

```
<!ELEMENT inex-snippet-submission (description,topic+)>
<!ATTLIST inex-snippet-submission
  participant-id CDATA #REQUIRED
  run-id CDATA #REQUIRED
>
<!ELEMENT description (#PCDATA)>
<!ELEMENT topic (snippet+)>
<!ATTLIST topic
  topic-id CDATA #REQUIRED
>
<!ELEMENT snippet (#PCDATA)>
<!ATTLIST snippet
  doc-id CDATA #REQUIRED
  rsv CDATA #REQUIRED
>
```

1.7.2.2 Measures

Evaluation was performed manually by the organizers of INEX Snippet Retrieval Track 2012-2013 [Bellot et al., 2013]. In order to determine the effectiveness of a snippet to provide sufficient information about the corresponding document, the relevance of the documents was judged apart from the relevance of the snippets. Thus, assessors should evaluate results in two ways:

- relevance evaluation of documents;
- relevance evaluation of snippets.

The topic title, description, and narrative (intent) provide the idea of the user information need (see 1.7).

Example 1.7. *Topic 2013001 from Snippet Retrieval Task*

```
<topic id="2013001" ct_no="1">
```

```
<title>Death of John Lennon</title>
<phrasetitle>Death of "John Lennon"</phrasetitle>
<description>Information about John Lennon's death</description>
<narrative>I want to know how where and when (including time of
day) when John Lennon died. Now I know he was shot, but what was the name of
the guy who shot him?</narrative>
</topic>
```

Assessors should go through the snippets, and decide whether the underlying document seems relevant to the topic reading only the snippet. They put 1, if it seems to be relevant, and 0 otherwise. After that, they should read the entire documents and judge their relevance. Then snippet-based relevance judgments were compared with the document-based relevance judgments (ground truth), i.e. a good snippet should be judged the same as the corresponding document. Then these judgments were integrated by the following measures:

- Mean prediction accuracy (MPA) — the average percentage of results the assessor correctly assessed:

$$MPA = \frac{TP + TN}{TP + FN + TN + FP} \tag{1.29}$$

  where TP refers to true positive, TN means true negative, FN and FP corresponds to false negative and false positive respectively.

- Mean normalized prediction accuracy (MNPA) is the average of the relevant results correctly assessed and the irrelevant results correctly assessed:

$$MNPA = 0.5 \times \frac{TP}{TP + FN} + 0.5 \times \frac{TN}{TN + FP} \tag{1.30}$$

- Recall is the average percentage of relevant documents correctly assessed:

$$R = \frac{TP}{TP + FN} \tag{1.31}$$

- Negative recall (NR) is the average percentage of irrelevant documents correctly assessed:

$$NR = \frac{TN}{TN + FP} \tag{1.32}$$

TABLE 1.12: Snippet evaluation 2013

Rank	Run	MPA	MNPA	Recall	NR	PA	NA	GM
1	*knapsack**	**0.8300**	**0.6834**	**0.4190**	0.9477	**0.4921**	0.8673	**0.5352**
2	Focused	0.8171	0.6603	0.3507	0.9700	0.4210	0.8675	0.4774
3	Focused_Split	0.8214	0.6549	0.3684	0.9413	0.4358	0.8624	0.4732
4	*MW**	0.8300	0.6459	0.3852	0.9067	0.4283	0.8572	0.4605
5	Baseline	0.8171	0.6414	0.2864	**0.9964**	0.3622	**0.8711**	0.4025

- Positive agreement (PA) is the conditional probability of agreement between snippet assessor and document assessor, given that one of the two judged relevant:

$$PA = 2 \times \frac{TP}{2 * TP + FP + FN} \tag{1.33}$$

- Negative agreement (NA) is the conditional probability of agreement between snippet assessor and document assessor, given that one of the two judged irrelevant:

$$NA = 2 \times \frac{TN}{2 * TN + FP + FN} \tag{1.34}$$

- Geometric mean (GM) of recall and negative recall:

$$GM = \sqrt{R \times NR} \tag{1.35}$$

The official ranking was based on GM.

1.7.2.3 Results at INEX/CLEF Snippet Retrieval Task

The results are given in the Table 1.12 (our results are marked by *, the best values are set off in bold). Our approach demonstrated the highest performance. As we hypothesized, the knapsack algorithm provided better results since it searches for the most valuable information regardless its position.

Here are the examples of the retrieved snippets.

Example 1.8. *Snippets made by the knapsack algorithm*

```
<description>KnapSack</description>
        <topic topic-id="2013001">
                <snippet doc-id="7286939" rsv="2306.47">
```

```
                John Lennon was murdered in December 1980 and George
Harrison succumbed to lung cancer in 2001. George Harrison had written " All
those Years ago " before the death of Lennon.
            </snippet>
            <snippet doc-id="1760504" rsv="2246.13">
                Lennon was portrayed by actor Mark Lindsay Chapman.
Reportedly, Moran asked, "Are you John Lennon ? " Lennon was honored with a
Grammy Lifetime Achievement Award in 1991.
            </snippet>
            <snippet doc-id="2412327" rsv="2207.53">
                Mimi sternly criticised Cynthia for divorcing Lennon ???
Lennon gave Mimi an allowance of ?? as were Paul McCartney, George Harrison
and Ringo Starr ??? and he has got John 's sense of humor.
            </snippet>
            <snippet doc-id="5351246" rsv="2151.31">
                Lennon 's Aunt Harriet and Uncle Norman Birch were made
legal guardians of the girls ??? as it was still in Lennon 's name ??? Lennon
 was not told about his death for months afterwards.
            </snippet>
            <snippet doc-id="1121203" rsv="2130.24">
                Goldman implies that Mark David Chapman 's murder of John
 Lennon may have been part of a conspiracy by fundamentalist Christians.
            </snippet>
            <snippet doc-id="2691820" rsv="2102.34">
                Lennon was not told about Victoria 's birth ??? as it was
 still in Lennon 's name ??? After Lennon 's death and Harriet died, Lennon '
s wife, Yoko Ono, wanted to sell the house ???
            </snippet>
            <snippet doc-id="10284" rsv="2089.22">
                Lennon 's most intense feelings were reserved for
McCartney. The story is told in the documentary " The U.S. vs. John Lennon".
            </snippet>
            <snippet doc-id="2412361" rsv="2084.08">
                Lennon would later meet Paul McCartney for the first time
 at St. Peter 's Church, where Smith was buried. s death the McCartney family
 moved to 20 Forthlin Road, which is only ??
            </snippet>
            <snippet doc-id="5800903" rsv="2079.72">
                But Lennon 's musicianship went far beyond guitar and
piano. Later, the piano was on charity tour. In 2000, this piano was bought
by George Michael at an auction for ??
            </snippet>
            <snippet doc-id="996715" rsv="2071.78">
                The front and back covers for " The John Lennon
Collection " were taken by famed photographer Annie Leibovitz on 8 December
1980, the day Lennon was murdered.
            </snippet>
            <snippet doc-id="2595203" rsv="2069.51">
```

```
                    The DVD was released on February 13, 2007 in the United
States. The U.K. release was on December 8, 2006, 26 years to the day after
the death of John Lennon.
            </snippet>
            <snippet doc-id="1120848" rsv="2011.26">
                    He is best known for his bestselling book on Lenny Bruce
and his controversial biographies of Elvis Presley and John Lennon.
            </snippet>
            <snippet doc-id="1177386" rsv="2010.37">
                    Lennon 's father was second cousins with singer John
Lennon. Lennon has also written for comic books.
            </snippet>
            <snippet doc-id="2210027" rsv="2001.62">
                    Lennon started the "Tux Announcer " style that night.
Lennon was inducted into the World Boxing Hall of Fame. He was conversing
with St. John 's clergy prior to his death.
            </snippet>
            <snippet doc-id="4456430" rsv="1992.17">
                    John has since performed the song several times at
Madison Square Garden.
            </snippet>
            <snippet doc-id="553183" rsv="1984.06">
                    As Lennon had previously had cats in Liverpool ??? Lennon
 called Bob Gruen ??? It was later updated and renamed, " John Lennon : The
Lost Weekend ".
            </snippet>
            <snippet doc-id="688252" rsv="1982.41">
                    She said at the time : "Jim has never felt he 's living
in John Lennon 's shadow. Lennon then spent twice the original ?? Lennon and
the other Beatles publicly renounced drugs ???
            </snippet>
            <snippet doc-id="380400" rsv="1978.92">
                    Musicians listed in booklet for John Lennon Anthology for
 I 'm Losing You Following the birth of his son Sean in 1975, Lennon had put
his career on hold to raise him.
            </snippet>
            <snippet doc-id="9303640" rsv="1967.2">
                    Eventually this world 's John Lennon found it out but
could not tell anybody on threat of imprisonment, so he starts to put clues
in the Beatles ' songs albums and etc..
            </snippet>
            <snippet doc-id="1323797" rsv="1963.38">
                    Lennon had the closest personal relationship with Epstein
 and was the most affected by his death. Lennon and McCartney 's artistic
venues for the Beatles became more disparate.
            </snippet>
    </topic>
```

Example 1.9. *Snippets made by the moving window algorithm*

```
<description>MW</description>
    <topic topic-id="2013001">
        <snippet doc-id="7286939" rsv="2306.47">
                        John Lennon was murdered in December 1980 and George
    Harrison succumbed to lung cancer in 2001. There have been numerous tributes
    to both of them. Lennon was murdered in New York City
                </snippet>
        <snippet doc-id="1760504" rsv="2246.13">
                        Death of John Lennon. Lennon was pronounced dead on
    arrival at St. Luke 's - Roosevelt Hospital Center, where it was stated that
    nobody could have lived for more than a few minutes
                </snippet>
        <snippet doc-id="2412327" rsv="2207.53">
                        After Lennon 's death, Ono and Sean Lennon visited Mimi
    in Liverpool, where she was staying at her sister Anne 's house because of a
    heart condition. She said, "Sean is like John
                </snippet>
        <snippet doc-id="5351246" rsv="2151.31">
                        's death she wrote "John Lennon, My Brother" (with
    Geoffrey Giuliano)and gave up working in 2004 to write "Imagine This -
    Growing up with my brother John Lennon"
                </snippet>
        <snippet doc-id="1121203" rsv="2130.24">
                        The Lives of John Lennon. When first published, "The
    Lives of John Lennon" was controversial because of its portrayal of Lennon in
     a highly critical light. Lennon was presented
                </snippet>
        <snippet doc-id="2691820" rsv="2102.34">
                        Lennon was named after his paternal grandfather, and
    Winston Churchill. Alf was not present at Lennon 's birth, as he was at sea.
    The infant Lennon started at his first school in
                </snippet>
        <snippet doc-id="10284" rsv="2089.22">
                        John Lennon. Born and raised in Liverpool, Lennon became
    involved as a teenager in the skiffle craze ; his first band, the Quarrymen,
    evolved into the Beatles in 1960. As the group
                </snippet>
        <snippet doc-id="2412361" rsv="2084.08">
                        During 1942 ??? 1943, Mimi 's sister Julia lived with
    Lennon at "The Dairy Cottage"; 120a Allerton Road, Woolton, which was owned
    by the Smith family. John Lennon Lennon lived with
                </snippet>
        <snippet doc-id="5800903" rsv="2079.72">
                        John Lennon 's musical instruments. John Lennon played
    various guitars with The Beatles and during his solo career, including the
    Rickenbacker  (four variants thereof),
```

```
            </snippet>
    <snippet doc-id="996715" rsv="2071.78">
                    The John Lennon Collection. The album was released on
vinyl in 1982 by Parlophone through EMI, and by Geffen Records in the United
States, later being remastered and released on
            </snippet>
    <snippet doc-id="2595203" rsv="2069.51">
                    after the death of John Lennon. The DVD was released on
February 13, 2007 in the United States. The film made its cable television
debut in the U.S. on August 18, 2007 on VH1 Classic
            </snippet>
    <snippet doc-id="1120848" rsv="2011.26">
                    and John Lennon. Albert Goldman was born in Dormont,
Pennsylvania and raised in Mount Lebanon, Pennsylvania. Albert Goldman
briefly studied theater at the Carnegie Institute of Technology
            </snippet>
    <snippet doc-id="1177386" rsv="2010.37">
                    Thomas Lennon  (actor). Lennon is a native of Oak Park,
Illinois, and the son of Kathleen and Timothy Lennon. He is a 1988 graduate
of Oak Park River Forest High School
            </snippet>
    <snippet doc-id="2210027" rsv="2001.62">
                    His boss liked Lennon 's performance so well, he hired
Lennon as the regular fight announcer, tuxedo and all. Lennon started the "
Tux Announcer" style that night. Lennon appeared
            </snippet>
    <snippet doc-id="4456430" rsv="1992.17">
                    John rarely performs the song live, as he has said it
brings back many painful memories of Lennon 's death, though he does add it
to set lists from time to time, often when playing
            </snippet>
    <snippet doc-id="553183" rsv="1984.06">
                    It was later updated and renamed, "John Lennon : The Lost
 Weekend". The original 500-page "Loving John" book focused more on Pang 's
role on Lennon 's albums and sessions.
            </snippet>
    <snippet doc-id="688252" rsv="1982.41">
                    John Lennon song)and, in April 1989, a restaurant named
Lennon 's ??? at 13\/14 Upper St. Martin 's Lane, Covent Garden ??? which had
 menu items such as "Rubber Sole" (a play
            </snippet>
    <snippet doc-id="380400" rsv="1978.92">
                    on the "John Lennon Anthology" collection released in
1998.)Unimpressed with its cosy domesticity, critical reaction to the album
was largely scathing ??? "a self-obsessed
            </snippet>
    <snippet doc-id="9303640" rsv="1967.2">
```

```
                    Eventually this world 's John Lennon found it out but
could not tell anybody on threat of imprisonment, so he starts to put clues
in the Beatles ' songs albums and etc.. The record
            </snippet>
        <snippet doc-id="1323797" rsv="1963.38">
                    After John Lennon 's death in 1980, McCartney, Harrison,
and Starr reconvened for Harrison 's "All Those Years Ago". The trio reunited
 as the Beatles for the "Anthology"
            </snippet>
</topic>
```

1.8 Contribution 4: Topic-comment Structure for Information Retrieval

Information retrieval (IR) is usually grounded on the hypothesis that relevant documents are *about* the query; the query being supposed to reflect properly the user's information need [Wong et al., 2001].

Aboutness is not as simple to define as it seems and IR suggested various definitions. For example, Cummins [Cummins] mentions that the term-occurrence frequency is "a measure of the degree to which a document is about a specific term". Concretely, most of IR models make the hypothesis that aboutness can be caught by matching the query terms and the document terms, both considered as bags of words or considering other term/language modeling means [Nie et al.][Wong et al., 2001]. Aboutness is thus seen at a general level, considering the discourse topic, that is to say what the entire text or paragraph (in case of focused or XML passage retrieval) is about.

In linguistics, the notion of aboutness is more complex and is related to the **topic** (or **theme**), which is what the text (typically a sentence) is about, while the **comment** (or **rheme** or **focus**) is what is being said about the topic [Büring, 2011].

Definition 1.2. A clause-level topic is the phrase in a clause that the rest of the clause is understood to be about, and the comment is what is being said about the topic.

According to W. Mathesius [Mathesius and Vachek, 1975], the topic does not provide new information but it connects the sentence to the context. Thus, the *topic* and the

comment are opposed in terms of the given/new information. This contraposition is called **information structure** (i.e. the *topic-comment structure*).

Let's consider two examples:

Example 1.10. *Anna married Sam 3 years ago.*

Example 1.11. *Sam married Anna 3 years ago.*

The sentence in ex. 1.10 is about Anna, while the sentence in ex. 1.11 is about Sam. Thus, the topic of ex. 1.10 is Anna, while the topic ex. 1.11 is Sam. The comment is the answer on the question *What's about the topic?*

As a matter of fact, when seeking for information using a search engine, the user is generally interested by the comment not by the topic. A comment may be viewed as a context for the corresponding topic. Although, the topic is mandatory to make the link between the user's information need and the text aboutness. Current IR models do not distinguish these two aspects in texts.

In this research, our goal is to improve the ranking of retrieved document by taking advantage of the information structure, i.e. the topic-comment structure of texts. More precisely, in our approach the notion of aboutness is first considered at the discourse-level using current IR model and then at the clause level in order to re-order the retrieved documents so that the top ones are more likely to bring useful comments on the query topic. According to our model, rather than matching the query terms with the document terms wherever they occur in the information structure, we promote an approach in which the query terms should match differently the topic and the comment parts of the sentences.

In most languages the common means to mark topic-comment relations are word order and intonation. However, since we are considering only textual documents in this study, we do not look at intonation annotation. In texts, the prominent construction for topic-comment is the so-called *topic fronting*. Topic fronting refers to placing the topic at the beginning of a clause regardless whether it is marked or not [Büring, 2011][M.A.K.Halliday, 1994]. Thus, even if complex linguistic-based methods could be used to extract topic-comment structure from sentences, the topic fronting feature can

be used as a simpler way to extract the information structure. Moreover too sophisticated linguistic methods would not be applicable at a large scale to analyze document sentences for IR purposes.

In this research, we focus on automatic annotation based on the topic fronting assumption. The method we proposed requires only shallow parsing, namely sentence chunking and part-of-speech (POS) tagging to automatically extract the information structure. Even if this is a light NLP function, POS tagging can be a challenging issue if applied to an entire document collection. For that reason, we rather use the knowledge on information structure as a mean to re-rank documents that have been retrieved considering more traditional matching, although our algorithm is not limited by re-ranking.

We evaluate our method on two collections: TREC Robust and WT10G. We compare our method considering several commonly used measures (MAP, $NDCG$ and $BPREF$) both to a strong baseline consisting of an initial retrieval performed by Divergence from Randomness model $InL2$ and the $Bo2$ pseudo-relevance feedback method implemented in Terrier platform which provides state-of-the-art effective retrieval mechanisms [Macdonald et al., 2012].

1.8.1 Topic-comment Structure in Linguistics

Apparently, Henri Weil could be the one who introduced the topic-comment opposition in 1844 [Weil, 1844]. He established the connection between topic-comment structure and word order. At that time the topic was called a psychological subject, while the comment was defined as psychological predicate.

Topic-comment influence has been studied on speech technology. Research work investigates intonational focus assignment or the relation between discourse structure and posture and gesture in order to design embodied conversational agents.

Information structure in texts presupposes the dichotomy of information units, namely topic and comment [Hartmann and Winkler, 2013]. These information units are triggers for syntactic and semantic processes, namely word order (dislocation), prosody ((de) accentuation), and interpretation. Dislocation and accentuation mainly appear within sentence bounds, while discourse linking put a sentence into a discourse context and thus influence the interpretation.

The collaborative research cluster (SFB) 632 proposed guidelines for the annotation of information structure [Got, 2007] as follows:

Definition 1.3. A Noun Phrase (NP) X is the Aboutness Topic of a sentence S containing X if

1. S would be a natural continuation to the announcement **Let me tell you something about** X
2. S would be a good answer to the question **What about** X**?**
3. S could be naturally transformed into the sentence **Concerning** X**,** S^* where S^* differs from S only insofar as X has been replaced by a suitable pronoun.

Cook and Bildhauer [Cook and Bildhauer, 2011] shows that despite using the same guideline, annotator agreement on topic-comment is sometimes difficult to obtain.

Actually, manual annotation of information structure in texts challenges the identification of the focus of a sentence or the discourse topic [Versley and Gastel, 2012]. Versley and Gastel proposed to chunk texts into topic segments since the discourse relations are usually bounded by topic segments [Versley and Gastel, 2012]. Relations (subordinating or coordinating) fall into the following categories: contingency, expansion, temporal, comparison, and reporting.

Some work has been carried out for automatic topic segmentation in broadcast news and has been applied for example in the Topic Detection and Tracking (TDT) program mainly based on word usage [Allan et al., 1998] or using prosodic clues [Purver, 2011].

1.8.2 Discourse-level Topic vs Rhetorial Relations and Topic-comment Structure in IR

Matching the discourse-level topic referring to the notion of aboutness of a document has been well studied in IR literature [Hjørland, 2001][Wong et al., 2001][Nie et al.]. However, modern search engines are essentially key word oriented and, thus, do not consider the relationships between terms [Nie et al.] nor between topics [Suwandaratna and Perera, 2010a]. On the other hand, linguistic analysis is crucial for text interpretation; as an

example rhetorical relationships indicated how the parts of a coherent text are linked to each other.

Various parsers have also been developed in order to parse discourses such as HILDA [Hernault et al., 2010] which implements topic changes or SPADE [Soricut and Marcu, 2003]. Both parsers were trained at the RST-DT corpus annotated according to Rhetorical Structure Theory [Carlson and Marcu, 2001]. Although the original set of discourse relations were limited to 24, the RST-DT corpus contains about one hundred relations. This set is usually reduced by the integration of relations into classes. Thus, in SPADE discourse parser, 18 rhetorical relations are taken into account: attribution, background, cause-result, comparison, condition, consequence, contrast, elaboration, enablement, evaluation, explanation, manner-means, summary, temporal and topic-comment. However, the topic-comment relation in the RST-DT corpus (and therefore in SPADE and HILDA parsers) is defined in a different way. Indeed, we can find the following definition: topic-comment is "a general statement or topic of discussion is introduced, after which a specific remark is made on the statement or topic ⟨...⟩ When the spans occur in the reverse order, with the comment preceding the topic, the relation comment-topic is selected. While comment-topic is not a frequently used mean in English, it is seen in news reporting, for example, when someone makes a statement, after which a reference is given to help the reader interpret the context of the statement ⟨...⟩ Ex. [As far as the pound goes,] [some traders say a slide toward support at 1.5500 may be a favorable development for the dollar this week.]" [Carlson and Marcu, 2001]. These parsers are based on deep analysis of linguistic features and are hardly usable when large quantities of texts are involved. Importantly enough, in texts, there exist special constructions to introduce the comment: topic fronting, placing the topic at the beginning of the clause is prominent. In this research, rather than using discourse parser which is too time consuming for large amount of texts, we develop a simpler way of extracting topic-comment structure for IR.

Lioma et al. use rhetorical relations from SPADE parser to re-rank documents [Lioma et al., 2012]. The authors introduced a query likelihood retrieval model based on the probability of generating the query terms from (1) a mixture of the probabilities of generating q from a document and its rhetorical relations and (2) the probability of generating rhetorical relations from a document. One of the limitations of this approach is that not all types of texts can be parsed this way (e.g. legal texts or item lists

have a few rhetorical relations). In addition, the rule-based parsers even if they take into account some statistics, are not extensible to other languages. An even more problematic drawback is related to the shortcomings of the discourse parser since such parsers are very time consuming and cannot be applied on large volumes of data. Lioma et al. state that topic-comment relations as defined by SPADE are extremely sparse in the benchmark IR collections [Lioma et al., 2012], while in our approach topic-comment structure is common for all types of texts as well as for all genres.

In the Subsection 1.3.2 we proposed to exploit topic-comment structure for text summarization. There, the assumption of topic fronting was simplified by viewing a topic as the first half of a sentence. However, the topic-comment analysis did not improve results. In contrast to that here, we propose to apply information structure for document re-ranking. Moreover, we introduce another algorithm for topic-comment chunking, namely we assume that a topic should be placed before a personal verb while the rest of the sentence is considered as a comment.

To the best of our knowledge, the closest related work is [Bouchachia and Mittermeir, 2003]. The authors propose to apply topic-comment structure for document classification while our approach aims at document re-ranking (but can be easily applied for document retrieval). They hypothesize that the important information belongs to the theme and that relevant documents to a query should share themes. The approach is underlain by the notions of topicality power and explanatory power that allows estimating document topicality by the cascade of neural networks. In contrast to this approach we propose to integrate the topic-comment structure into the classical retrieval models such as $BM25F$ which is a variant of $BM25$ that takes into account document structure and multiple weighted fields. We choose $BM25F$ as a simplest and elegant way to assign different weights to different document parts. In contrast to $BM25F$ we do not use fields (structural components) but the set of the oppositions between topic and comment. Bouchachia and Mittermeir do consider only features within a document while we believe that it is important to take into account collection features. That is why we introduced the notion of Inversed Comment Frequency which is analogous of the concept of Inversed Document Frequency. The topic-comment annotation process in their approach requires syntax parsing, although other details are not provided in their paper.

1.8.3 Contribution 4: Document Re-ranking Algorithm Based on Topic-Comment Structure Analysis

1.8.3.1 Automatic Topic-comment Annotation

The topic-comment structure is opposed to formal structure with grammatical elements as the constituents. The difference between "topic" and grammatical subject is that topic refers to the information or pragmatic structure of a clause and how it is related to other clauses, while the subject is a merely grammatical category.

In simple English clause the topic usually coincides with the subject, even if it is not always the case as for expletives (e.g. *it is snowing*) that do not have topics at all [Got, 2007]. Moreover, the unmarked word order in English is Subject - Verb - Object (SVO). Thus, it is possible to make an assumption that, as a rule, the topic is placed before the verb. We make an additional assumption, that if a subordinate clause provides details on an object, it is rather related to the comment. Thus, the main idea of the proposed method is to split a sentence into two parts by a personal verb.

Example 1.12. *{The Bengal Standard}$_{topic}$ {is a description of the ideal Bengal and therefore is used to define the quality of each cat}$_{comment}$.*

1.8.3.2 Topic vs Comment for Query Matching

State-of-the-art models in IR consider the document ranking function as a matching function between the terms in the documents and the query without considering term relationships. In our model, we hypothesize that the topic-comment structure could be useful in the matching process. Moreover, we argue that topic matching would be more effective than term matching; thus giving more importance to words that correspond to topic during matching.

First of all, we consider that a user expresses the information need by topic only, that is to say that there is no comment in a user's query. For this reason, any query term is considered as a topic in our approach. On the contrary document sentences contain both topic and comment parts. Since users are supposed to be interested by comments about their topic of interest, we argue that the matching model should consider differently topic/query and comment/query matching.

Furthermore, we can assume that matching topics induce comments be considered relevant information. Thus, the importance of each topic in a document depends not only on its frequency, but also on the number of related comments, i.e. how well the topic is explained in a document. We propose to take the logarithm of this number in order to smooth the influence. On the other hand, some topics may be too specific and thereby linked to few comments. Therefore we introduced the measure of specificity of the topic t Inversed Comment Frequency $ICF(t)$:

$$ICF(t) = \log \frac{CommentCount(t)}{\sum_{t_j \in T} CommentCount(t_j)} \tag{1.36}$$

where $CommentCount(t)$ is the number of comments related to the topic t in the collection, $T = \{t_j\}_{j=1,|T|}$ refers to all topics in the collection, $|T|$ is the total number of topics.

The integration of this proposition in most of IR models is quite simple: a specific document term is considered differently whether it occurs in the topic or the comment part of the sentence. We give the example of the integration into the BM25F retrieval model in the next section.

1.8.4 Integration of the Topic-comment Structure into Retrieval Models

We integrated topic-comment structure into BM25F retrieval model. Originally BM25F is an extension of Okapi's BM25 to multiple weighted fields in contrast to linear combination of scores for structured documents [Robertson et al., 2004]. BM25 is calculated as follows:

$$bm25(d) = \sum_{i=1}^{n} \frac{IDF(q_i) \times TF_d(q_i) \times (k_1 + 1)}{TF_d(q_i) + k_1 \times (1 - b + b \times \frac{|d|}{avgDL})} \tag{1.37}$$

where q_i are the terms of the query Q, n is the number of query terms, $IDF(q_i)$ is an inverse document frequency of the term q_i, $TF_d(q_i)$ is a term frequency in the document d, $|d|$ is the length of the document d in terms, $avgDL$ is the average document length in the collection, k_1 and b are free parameters.

BM25 model is based on the assumption that term frequencies follow 2-Poisson distribution and for each term the collection is split into two categories: elite and non-elite. As

Robertson et al. assert, this relation may be considered from the opposite point of view, namely, the terms of a given document are labeled as elite or non-elite [Robertson et al., 2004]. A term is elite in a document if the document is about the concept denoted by the term. The elite terms refer to the topics of the document. Bag-of-words based approaches presuppose the independence from the position of a term but the boosted probabilities of elite terms. Robertson et al. assumed that for some parts of structured documents the probabilities of the elite terms are boosted even more. Thus, they proposed to assign different weights to the term coming from different document parts.

However, document structure is not uniform and therefore is hard to analyze. In contrast to document fields, topic-comment structure is common for all texts and genres. Thus, we compute document score as follows:

$$score(d) = \sum_{i=1}^{n} \frac{ICF(q_i) \times TC \times (k_1 + 1)}{TC + k_1 \times (1 - b + b \times \frac{len_{topic}(d)}{avgDL_{topic}})} \tag{1.38}$$

$$TC = tw \times explRate(q_i) f(q_i, T_d) + (1 - tw) \times f(q_i, C_d)$$

$$explRate(q_i) = \log(CommentCount_d(t) + 1)$$

where tw is the topic weight which is the analogue to the field weight in the classical BM25F model, $f(q_i, T_d)$ is q_i's term frequency in the topic set of the document d, $f(q_i, C_d)$ is q_i's term frequency in the comment set of the document d, $len_{topic}(d)$ is the length of the document d in topics, and $avgDL_{topic}$ is the average document length in the collection in topics, k_1 and b are free parameters, and $CommentCount_d(t)$ refers to the number of comments related to the topic t in the document d. tw is a parameter in the model. It could be assigned or learnt. We introduced the notion of the explanation rate $explRate(q_i)$ showing how well the topic is explained in the document. This notion is similar to the topicality power of a term proposed in [Bouchachia and Mittermeir, 2003] which is considered within a document and shows how strong it is explained (i.e. the number of comments it has). The first difference is that we propose to use the logarithm instead a raw sum. Explanatory power in [Bouchachia and Mittermeir, 2003] is viewed as the number of times a term is occurring at a comment regardless the topic within a single document while we are looking for comments to a specific topic. Moreover, in contrast to [Bouchachia and Mittermeir, 2003], we consider the collection features by introducing the notion of Inverted Comment Frequency.

1.8.4.1 Multi-word Expression Extraction

In order to match query terms with topics from documents, after having extracted topic-comment structure, we incrementally extract multi-word expressions based on normalized point-wise mutual information $npmi(x, y)$ [Bouma, 2009]:

$$npmi(x, y) = \frac{pmi(x, y)}{-\log[p(x, y)]} \tag{1.39}$$

$$pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \tag{1.40}$$

where $pmi(x, y)$ is the point-wise mutual information of the terms x and y, $p(x, y)$ is the joint probability of x and y, $p(x)$ and $p(y)$ are the probabilities of the terms x and y respectively.

Candidates made of exclusively functional words are rejected as well as candidates containing punctuation marks. We hypothesized that multi-word expression matching should be more important than a single word. Therefore, we integrated the length in terms of tokens of the expression $length(q_i)$ into the final score:

$$score(d) = \sum_{i=1}^{n} \frac{length(q_i) \times ICF(q_i) \times TC \times (k_1 + 1)}{TC + k_1 \times (1 - b + b \times \frac{len_{topic}(d)}{avgDL_{topic}})} \tag{1.41}$$

1.8.5 Evaluation

The evaluation was performed on two TREC datasets:

- Robust TREC;
- WT10G.

Robust TREC set consists of about 528,000 news articles and 1,904 MB of text of TREC Disk4&5 (except Congressional Record data) and 249 topics with relevance judgments. Robust TREC set is "pure" collections since the documents have almost the same format and there is no spam. WT10G is 10GB subset of the web snapshot and of Internet Archive. WT10G contains more than 1.6 million of documents. There are 98 topics with relevance judgments. In contrast to Robust, WT10G is a snapshot of the web with real documents in HTML format, some of which are spam.

The system performance was evaluated using several measures implemented in trec_eval[7] software provided by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results. We considered the following evaluation measures:

- MAP (Mean Average Precision) over all queries which is the arithmetic mean of average precision values for individual queries and has been shown to have very good discrimination and stability.
- $NDCG$ (Normalised Discounted Cumulated Gain). Since the gain of each document is discounted at lower ranks, this measures is suitable for re-ranking evaluation.
- $BPREF$ (Binary Preference) computes a preference of whether judged relevant documents have higher rank than judged non-relevant documents. Thus, $BPREF$ does not treat non-assessed documents as non-relevant while MAP does. This is important for large collections where the probability of retrieving non-assessed documents is higher.

The further description of the collections and evaluation measures is given in the section 2.4.

We compared our system with a baseline implemented in the Terrier platform [Ounis et al., 2006a], namely $InL2$ weighting model with $Bo2$ query expansion algorithm. $InL2$ is a DFR (divergence from randomness) model based on TF-IDF measure with L2 term frequency normalization [Amati and Van Rijsbergen, 2002a]. This model is based on the assumption that informative words are relatively more frequent in relevant documents than in others. $InL2$ demonstrates better performance at many recall levels and in average precision than traditional retrieval models such as $BM25$ [Amati, 2003]. $L2$ normalization is less sensitive to document length. According our preliminary study, with the default Terrier's parameters, on the used collections $InL2$ showed better results than Okapi's BM25 and Hiemstra's implementation of the language model. This was the reason why we did not compare our results with those of BM25. $Bo2$ is a pseudo-relevance feedback algorithm for query expansion based on Bose-Einstein statistics and

[7] http://trec.nist.gov/trec_eval/

DFR model. On the chosen collections, this method outperformed RM3 model implemented in Indri, a search engine from the Lemur project mainly built on the language modeling information retrieval[8]. RM3 is an Indri's adaptation of Lavrenko and Croft's relevance models [Lavrenko and Croft, 2001b]. The stemming was performed by Porter algorithm. We parsed the document retrieved by the baseline system by the Stanford POS tagger which also allows sentence chunking [Manning et al., 2014]. The detailed description of the DFR models is presented in the subsection 2.4.3.

For our model, we used top 20 documents for re-ranking. The re-ranking was performed within blocks of 5 documents. Our first hypothesis was that the topics should have more weight than the comments. However, the preliminary study indicated the opposite. This could be explained by the fact that the comments are usually much longer than the topics. Thus, the prior probability to find a term within comments is higher than in topics. Higher values of topic weight decrease comment weight. This leads to the lost of documents that just mention relevant information but are not entirely about the subject. That is why the topic weight was set to $tw = 0.2$. The coefficients $k_1 = 10$ and $b = 0.2$. We considered only unigrams and bigrams. We also excluded the lower order expressions from the query term list if they are parts from a higher order expression. For example, a query $q = safety\ plastic\ surgery$ is presented as $q = \{q_1, q_2\}$, where $q_1 = safety$ and $q_2 = plastic\ surgery$ and the unigrams *plastic* and *surgery* are ignored.

Table 1.13 provides evaluation results. The differences with the baseline marked by * are significant according to the Student T-test at the level $p = 0.05$. According to all evaluation measures for both data sets our method (TC) outperformed the baseline.

On Robust collection our method excelled the baseline on 107 queries and it was bellow it on 101 queries. The lower performance was observed for queries with higher values of NDCG in average (0.64) while the better results were demonstrated for more difficult queries ($NDCG_{avg} = 0.56$).

On the WT10G our method showed better results for 40 queries ($NDCG_{avg} = 0.56$) and it was less efficient for 22 queries ($NDCG_{avg} = 0.628$). Thus, we can conclude that our approach is more suitable for difficult queries.

[8] http://www.lemurproject.org/

TABLE 1.13: Re-ranking results using topic-comment structure

Collection	**Measure**	**Baseline**	**TC**
Robust	MAP	0.2801	**0.2884***
	BPREF	0.2782	**0.2863***
	NDCG	0.5549	**0.5597***
WT10G	MAP	0.2152	**0.219***
	BPREF	0.2056	**0.2138***
	NDCG	0.4861	**0.4917***

1.9 Conclusion

In this chapter we presented an approach for short message contextualization from an external source based on query-biased summarization by sentence retrieval. Sentence retrieval is based on NE recognition, POS weighting and sentence quality measuring. We introduced an algorithm of smoothing from the local context. We also integrated the knowledge of topic-comment structure into the sentence retrieval model. Moreover, we developed a graph-based algorithm for sentence re-ordering. The method has been evaluated at INEX/CLEF Tweet Contextualization track. We obtained the best results in 2011 according to informative evaluation. In 2013 according to informative evaluation our system was ranked first (PRIOR and POOL) and second (ALL) over all automatic systems that participated. At the same time in terms of readability it was the best among all participants according to all metrics except redundancy. Run comparison showed that smoothing improves informativeness. Another conclusion is that ranking is sensitive to the pool selection as well as to the choice of divergence. Despite the topic-comment analysis did not improve results, we believe that small changes in implementation may produce positive effect on the system performance. In 2014 the worst results among our runs were shown by the run based on entity restriction that could be explained by the loss of the recall. The results were published in the INEX/CLEF working notes 2011-2014 [Ermakova and Mothe, 2012a,b, 2013, 2014].

The sentence retrieval method was also adapted to snippet retrieval and QE. In 2013 our

system showed the best results in the INEX Snippet Retrieval Track. The approach was published in the INEX/CLEF-2012 working notes [Ermakova and Mothe, 2012b] and the workshop EGC-2013 [Ermakova and Faessel, 2013].

In this chapter we also proposed a novel approach to document re-ranking in information retrieval based on topic-comment structure of texts. Although it can be easily generalized to document retrieval. To the best of our knowledge, this information structure was never applied to the ad hoc information retrieval nor re-ranking.

We introduced an automatic topic-comment annotation method based on the topic fronting assumption that requires only shallow parsing, namely sentence chunking and POS tagging. The main idea of the proposed method is to split a sentence into two parts by a personal verb.

We integrated topic-comment structure into BM25F retrieval model. Firstly, we hypothesized that the topics should have more weight than the comments. However, the preliminary studies demonstrated that high values of this coefficient decreased the results in average. The possible explanation is that the comments are usually much longer than the topics and therefore the prior probability of a query term to occur within comments is higher. Higher values of topic weight could lead to the lost of documents that just mention relevant information but are not entirely about the subject.

We evaluated our approach on two TREC data sets. According to all used evaluation measures for both test collections, our method significantly outperformed the strong baseline provided by the Terrier platform. Experiment results allow drawing a conclusion that the approach proposed in this chapter is more suitable for difficult queries.

Since our method makes the difference between sentences where the topic and the comment are inversed (as in 1.10 and 1.11), we believe that our approach makes sense for question answering and focused IR. In future work we are going to investigate these tracks.

Chapter 2

Query Expansion

2.1 Introduction

Information Retrieval (IR) systems aim at retrieving information that answers a user's need s/he expresses through a query. Because real queries are short and because natural language is ambiguous such matches can be wrong or incomplete. The average query length remains between 2.4 and 2.7 words [Gabrilovich et al., 2009, Lau and Horvitz, 1999]. To face these challenges, IR systems consider several strategies. On the one hand, a user may prefer to get documents treating of various aspects of her information need rather than possibly redundant aspects within documents [Carbonell and Goldstein, 1998a, Santos et al., 2013a]; on the other hand by providing document related to the various senses of query terms, the system optimizes the chance of providing relevant information [Clarke et al., 2008, Vargas et al., 2013]. Semantic indexing and search aim at tackling the problem of term ambiguity. Some solutions rely on knowledge resources such as ontologies to use concepts rather than terms or stems, both during indexing and matching. Term ambiguity has also been treated with positive results as a classification or clustering problem, in which documents that share the same sense with the query terms are retrieved whereas documents that use the query terms but in a different meaning are filtered out [Schütze, 1998]. On the other hand Query Expansion (QE) has driven many works in IR (see Carpineto's survey on QE [Carpineto and Romano, 2012b]). QE aims at adding new terms to the initial query that will improve retrieval based on some knowledge, either extracted from the term collection distribution, user's profile (e.g.

topics of interest), or relevance feedback (RF) [Carpineto and Romano, 2012b]. Thus, QE in a search engine may be also viewed as contextualization of the initial query.

The initial query can be expanded using term co-occurrences in the documents [Amati, 2003, Amati and Van Rijsbergen, 2002b, Xu and Croft, 1996] or based on WordNet definition [Voorhees, 1994]. Candidate terms for expansion are either extracted from external resources such as WordNet [Mandala et al., 1998] or from the documents themselves; based on their links with the initial query terms. The former type of approaches is collection independent whereas the latter has the advantage of taking into account the document collection and thus the capability of the collection to contain the relevant information. In the latter types of methods, the most popular one is the pseudo-relevance feedback [Buckley, 1995]. The initial method was to add terms from relevant documents [Rocchio, 1971]; since this information is not easily available, Buckley suggested to consider the first retrieved document as relevant and to select candidate terms from these documents. Pseudo-relevance feedback is now common practice and used in many expansion methods [Carpineto and Romano, 2012b].

This chapter pursues two objectives; first of all, we suggest three new automatic methods of query expansion:

- Adaptation of our Sentence Extraction Method described in the Section 1.3 to Query Expansion (LC);
- Co-occurrence Model (Co);
- Proximity Relevance Model (PRM).

The Co-occurrence Model retrieves candidates from the relevance feedback and scores the by applying the global analysis of texts. Per contra LC and PRM exploits term proximilty within PRF. LC is an extension of the method we developed for tweet contextualization. Selecting the most appropriate terms from the relevant -or considered as such- documents is a challenge. While weighting the term candidates considering their frequency or their weight calculated during the indexing phase is an intuitive and widely used approach, we suggest that a deeper analysis of document content can be useful. Our first hypothesis is that terms that occur closely to query terms within the documents should be good candidates for QE; the closer the better candidate. The second hypothesis is that natural

language considerations should help to decide the best candidate terms, that is to say that some types of terms should be better candidates (e.g. noun being better than adverbs). To study these hypotheses, we propose a method that considers a term windows surrounding query terms from feedback documents. In addition, our method considers Part of Speech (POS) information to weight differently the QE term candidates.

In its turn, PRM is a formal model for QE based on PRF.

There is relatively little studies of formal models using positional heuristics for QE. One of the formal approach for QE is the positional relevance model [Lv and Zhai, 2010] which is an extension of the relevance language model (LM) [Lavrenko and Croft, 2001b]. In the positional relevance model query likelihood is estimated as the product of the probabilities of the query terms in the position within pseudo-relevant documents. However, in this approach the term proximity is captured indirectly by weighting the positions within PRF.

As in the approaches based on the term proximity, we hypothesize that the closer a term to a query term, the better the QE term candidate is. Nevertheless, we believe that it is more appropriate to estimate the distance not in terms of tokens, but rather in terms of sentences. This is motivated by the following facts:

- In linguistics a sentence is viewed as a minimal set of words that in principle tells a complete thought;
- Within a sentence, words could be reordered without meaning shift (e.g. paraphrasing);
- Synonyms and associations are usually considered as good expansion candidates. However, synonyms usually do not co-occur within a sentence unlike other semantically related words.

One of the main contribution of this work is that it provides a novel formal LM for QE that directly captures the term proximity rather than by weighting term positions, and the distance is computed at sentence level.

The remainder of the chapter is organized as follows. Section 2.2 presents related works. Section 2.3 describes the first QE method we promote, namely LC. Section 2.3.2 details the co-occurence model which is the second contribution in QE we made. Section

introduces the PRM model. Section 2.4 presents the experimental framework as well as the collections and performance measures we used. Section 2.5 reports the results and discusses them. Finally, section 2.7 concludes this paper and draws up some future works.

2.2 Related Works

QE techniques are divided into five main groups:

- dictionary-based or ontology-based methods [Bhogal et al., 2007];
- methods using other textual sources besides the original collection (e.g. in QA terms from FAQ texts are often used for QE data [Agichtein et al., 2004, Harabagiu and Lacatusu, 2004]);
- cross-lingual methods [Cao et al., 2008c];
- global analysis (corpus analysis for the purpose of word relationships detection) [Carpineto and Romano, 2012b];
- local analysis or local feedback (analysis of documents retrieved by the initial query) [Rocchio, 1971, Xu and Croft, 1996].

Thus, QE techniques are either based on the analysis of a document collection [Carpineto and Romano, 2012b] or they imply dictionary-based or ontology-based methods [Bhogal et al., 2007]. Verma et al. used WordNet and Unified Medical Language System for query expansion [Verma et al., 2007]. S. Tratz and E. Hovy proposed to use Basic Elements (BEs) as paraphrases [Hovy and Tratz, 2008]. A BE is a syntactic unit up to three words with associated tags such as NER (Named Entity Recognition) and POS (Part-of-Speech). BEs can take into account lemmas, synonyms, hyponyms and hyperonyms, identical prepositional phrases, spelling variants, nominalization and denominalization (derivation in WordNet), transformations like prenominal noun - prepositional phrase, noun swapping for IS-A type rules, pronoun transformations, and pertainym[1] adjective transformation. Chali and Joty kept only nouns for a query [Chali and Joty, 2007]. Besides WordNet synonyms, they proposed to apply topic signature based on likelihood

[1] a pertainym is an adjective, which can be defined as "relating to" or "pertaining to" another word

ratio for binomial distribution tests for related terms as well as the terms from the strongest lexical chains. The results of the methods that need external resources to be used (dictionary-based or ontology-based methods, methods that uses other sources besides the original collection such as FAQ texts in question-answering systems) can highly depend on these resources.

Some researchers also payed a lot of attention to cross-lingual methods [Cao et al., 2008c].

On the contrary, the local and global analyses are centered on the document collection.

2.2.1 Global Methods

The analysis of a document collection may be either (1) global (corpus analysis for the purpose of word relationships detection) [Carpineto and Romano, 2012b] or (2) local feedback (analysis of documents retrieved by the initial query) [Rocchio, 1971, Xu and Croft, 1996]. Global methods work alike but in that case candidate terms come from the entire document collection rather than just (pseudo-) relevant documents.

Rather considering a document as a bag-of-word, one can consider term co-occurrence by applying latent semantic analysis (LSA). LSA implies that related words co-occur in similar context [Landauer et al., 1998]. Term co-occurrence may be discovered by a cluster algorithm, e.g. the Naive-Bayes maximizing the classification maximum likelihood criterion, where each word is presented as a vector with the components corresponding to the number of occurrences of the word in each document [Amini et al., 2007].

Schiffman presented an approach that incorporates corpus-driven semantic information and query expansion by log likelihood ratio [Schiffman, 2007].

Similar approach was proposed by Gabrilovich and Markovitch: the strength of the relation between two terms is computed as $TF \times IDF$ value within a Wikipedia page [Gabrilovich and Markovitch, 2007].

Milne and Witten also proposed to use the Wikipedia to estimate the strength of the relation $w(s \to t)$ between the terms by counting the number of outgoing links in the

corresponding articles (s and t are source and target articles respectively):

$$w(s \to t) = \begin{cases} log\frac{|W|}{|T|}, & if s \in T \\ 0, & otherwise \end{cases} \tag{2.1}$$

where T is the set of articles with links to t, and W is the set of all Wikipedia pages [Milne and Witten, 2008]. Wikipedia redirects are useful source of synonyms [Niemann and Gurevych, 2011]. Wikipedia page structure may be also applied to get related terms from headers, categories and the first passage [Niemann and Gurevych, 2011].

However, current methods use blind methodologies and uses learning methods as black boxes. On the contrary, we think that a deep analysis of queries and of query expansion terms could help understanding when QE would be useful and if there are some sort of typology of QE usefulness.

2.2.2 Query Expansion and Pseudo Relevance Feedback

Local analysis or local feedback methods rely on the hypothesis that relevant documents contain terms that could be useful to reformulate an enhanced query regardless to an information need.

The use of relevance information for QE was suggested first by Rocchio [Rocchio, 1971] who defines the Relevance Feedback (RF) principle. Users are supposed to judge some of the retrieved documents and this feedback information is used in turn either to re-weight query terms or to expand the query with the most important terms from relevant documents. Using the vector space model, Rocchio defined a method to re-weight query terms and thus to add new terms to the initial query - terms that were initially associated with a null value. The term weights are re-computed so that the terms that occur in relevant documents contribute positively to the new query whereas the weight of the terms that occur in non-relevant documents are lowered. A balance between the initial query and feedback information is involved in the weighting.

Rocchio's method implies to know document relevance. To avoid users' judgment that can be difficult to collect and to make the process fully automatic, Buckley et al. [Buckley, 1995] suggested to consider the first initially retrieved documents as relevant, i.e. pseudo-relevance feedback (PRF). Pseudo-relevance feedback has then been implemented

in the various IR models such as the probabilistic model or the language model [Lavrenko and Croft, 2001b]. Many studies have shown that this method is efficient in average; however, it can lower results for some queries [Amati et al., 2004c, Carpineto and Romano, 2012b, Chen et al., 2012b]. For example, it is most probable that for poor performing queries query expansion is helpless since it will be based on the first retrieved documents that are probably non-relevant documents. It is thus important to know in advance if QE will be helpful or on the contrary if it will degrade the results. Selective query expansion aims at making this decision [Cronen-Townsend and Croft, 2002b].

Singh and Sharan combined co-occurrence and semantic similarity of terms [Singh and Sharan, 2015].

Rather than considering the top-retrieved documents all as relevant, some works aim at distinguishing relevant from non-relevant documents before using them in PRF. Xu et al. [Xu et al., 2009] suggested that top documents should not be considered in a blind way but non-relevant documents should cluster as relevant documents do. In addition, they consider that query terms should occur in the relevant document cluster and that some documents from the non-relevant cluster do not contain any of the query terms. Lee *et al.* propose a *resampling method* using top-retrieved document clustering [Lee et al., 2008]. Another range of works focuses on selecting the best feedback information. Rather than focusing on how to select the best documents to used in PRF, some approaches focus on how to select the best terms to expand the initial query. Selecting the most appropriate terms from the relevant -or considered as such- documents is indeed a challenge [Cao et al., 2008c, Lv and Zhai, 2010].

2.2.3 Proximity Based Methods

Local analysis or local feedback methods rely on the hypothesis that relevant documents contain terms that could be useful to reformulate an enhanced query. In the majority of previous works local context is viewed as an entire document presented as a bag of words and the proximity of terms is not captured.

Xu and Croft use a feature selection based on co-occurrence of terms, considering that the best terms are the ones that co-occur with as many query terms as possible within the top-ranked documents or document passages [Xu and Croft, 1996]. In addition, they

consider nouns and noun phrases as the expansion terms. Xu and Croft's co-occurrence measure is not a probability in the strict sense, while mutual information shows the joint probability of terms to co-occur within a text. Distinctly from Xu and Croft's approach that considers the distance between the candidates and the query terms as binary (i.e. terms either co-occur within a text passage or not), we hypothesize that the dependence of the probability to find good candidates for QE on the distance is more sophisticated and that it should be considered at the sentence level.

Other empirical studies have shown that the term proximity is effective for selecting expansion terms. Cao *et al.* suggested a term classification method based on SVM to predict the usefulness of expansion term candidates [Cao et al., 2008c] based on the term distribution, co-occurrence with query terms, and the proximity from them. Wan *et al.* suggested to combine ontology-based methods with the proximity heuristics [Wan et al., 2012]. Miao *et al.* proposed an extension of the Rocchio's approach by introducing a concept of proximity-based term frequency that focuses on the proximity of terms rather than positional information unlike the positional relevance model [Miao et al., 2012]. They provide 3 approaches to estimate the proximity-based term frequency, namely (1) moving window; (2) kernel-based and (3) Hyperspace Analogue to Language (HAL) methods. The approach of Miao *et al.* is rather empirical and is an elaboration of the TF-IDF model. Unlike [Cao et al., 2008c, Miao et al., 2012], we propose a theoretical reasoning of our approach.

Some works take into account only ordered or unordered n-grams within the window of N-terms [Metzler and Croft, 2007, Song and Croft, 1999] capturing the proximity in binary sense. Tao and Zhai [Tao and Zhai, 2007] explore only the proximity of query terms resting upon the hypothesis that in relevant documents query terms should be closer to each other. In contrast to the cumulative proximity expansions retrieval model [Vuurens and de Vries, 2014] that does not require any co-occurrence statistics, we combined proximity and co-occurrence statistics within the language model (LM) formalism.

There is relatively little studies of formal models using positional heuristics for QE. The only formal approach for QE we are aware of is the positional relevance model [Lv and Zhai, 2010] which is an extension of the relevance LM [Lavrenko and Croft, 2001b]. In the positional relevance model query likelihood is estimated as the product of the probabilities of the query terms in the position within pseudo-relevant documents.

However, in this approach the term proximity is captured indirectly by weighting the positions within PRF.

Although some researchers exploit term proximity in QE [Cao et al., 2008c, Miao et al., 2012, Tao and Zhai, 2007, Xu and Croft, 1996], their works are rather empirical. The lack of theoretical works in this area motivated us to introduce a novel method integrated into the language model formalism that takes advantage of the remoteness of candidate terms for QE from query terms within feedback documents. Thus, the main contribution of this work is that it provides a novel formal LM for QE that directly captures the term proximity rather than by weighting term positions, and the distance is computed at sentence level.

2.2.4 Selective Query Expansion

Selective QE has been introduced after some work has shown that even if in average QE improves the results, some queries can suffer from expansion specifically queries for which the system faces difficulties to retrieve relevant documents from the initial query. In selective QE, the system decides whether or not QE should be applied, [Cronen-Townsend et al., 2002b]. Current studies are based on feature analysis and learning models: queries are characterized by features and from a set of examples for which the QE decision is known (either QE should be applied or not), the system learns the binary QE model. Query features are divided into pre-retrieval and post-retrieval features; the former can be extracted before any search on the document collection whereas the latter are search dependent. Cao et al. propose a term classification method to predict the usefulness of expansion term candidates [Cao et al., 2008b]. Some methods combine the analysis of term co-occurrence and term distribution methods [Pal et al., 2013a, Pérez-Agüera and Araujo, 2008].

2.3 Models

2.3.1 Contribution 5: LC Model

The key idea of the proposed method is to search the most appropriate candidates for QE by ranking terms and sentences from the pseudo-relevance feedback, i.e. from the top

ranked documents. Both ranking procedures include local context analysis, i.e. analysis of neighboring sentences. Sentence scoring method is an elaboration of RF. We strengthen the criteria of provenance of good terms for QE used in RF. In contrast to [Wan et al., 2012] we estimate the distance in term of sentences and we evaluate the sentences that are the sources of the candidate terms.

Our approach is underlain by the following hypotheses:

1. Not always an entire document is relevant to a query, but it can contain one or several relevant passages. Term candidates should be selected from these passages.

2. Terms for QE come from appropriate sentences (in general, this hypothesis is similar to those of RF). The measure of sentence appropriateness is called sentence score and referred to $score(S, Q)$ in the rest of the section. This sentence scoring method is the adaptation of the method initially developed for query-biased multi-document summarization described in the Chapter 1. The details of $score(S, Q)$ were provided in the Section 1.3.

3. Good terms should have appropriate part of speech (POS) and high IDF. Not all POS are suitable for query expansion (e.g. functional words). Moreover, the most frequent terms are nouns. However, in some cases adjectives, verbs and numbers are indispensable. A good term should well distinguish documents from each other. POS weight and IDF may be considered as a query-independent term score.

4. The terms lying in the neighborhood of query terms are closer related to them than the remote ones.

The term score is combined with the corresponding sentence score. Thus, we used a two-step local context analysis: for sentence scoring and for estimation of term importance. In previous works local context was viewed as a single document and it was opposed to the entire collection analysis (global context) [Carpineto and Romano, 2012b, Xu and Croft, 2000]. In this research we consider local context in a stricter way, precisely we look not only to the whole document statistics, but also for terms surrounding the query terms. Thus, all candidate terms are ranked according to the following metric:

$$w_{total}(t) = f(score(S, Q), w_{pos}(t), IDF(t), importance(t, Q)) \tag{2.2}$$

where $score(S, Q)$ is score of the sentence S containing the term t, $w_{pos}(t)$ is the weight of the POS of t,

$IDF(t)$ is the inverse document frequency of the candidate term, $importance(t, Q)$ is a function of (1) the distance to the query Q terms, (2) their weights, and (3) the likelihood of the candidate term to co-occur not by chance with the query terms in the top ranked documents.

$importance(t, Q)$ allows to find terms occurring in the neighborhood of important query terms.

The next step of our method is to compute the importance of all terms in all sentences from RF:

$$importance(t, Q) = \theta(wd(t, Q), coocurrence(t, Q)) \tag{2.3}$$

$wd(t, Q)$ is a function of the distance from the candidate terms to the query Q and their weights, and $cooccurrence(t, Q)$ shows the likelihood of the candidate term to occur not by chance with the query terms in the top documents ranked according to the initial query.

The concrete functions are given in the evaluation Section 2.4.

2.3.2 Contribution 6: Co-occurrence Model

Methods based on the local feedback highly depend on the top retrieved documents. Documents that has just a small relevant part could influence badly QE. The global analysis of the collection is less sensitive to the topic shift.

The key idea of the proposed method is to estimate the importance of candidate terms by the strength of their relation to the query terms.

In contrast to DFR models [Amati, 2003] we do not compare the term frequency in RF and the entire collection. In our approach, documents from RF provide term candidates that are analyzed in two aspects: their frequency in RF and their co-occurrence with query terms in the whole collection. As it is shown in the Section 2.4.3, all DFR models are based on two metrics: term frequency in RF and the frequency of the term t in the collection. Particularly, Bo2 uses the extrapolation of term frequency in RF on the whole collection.

In our method candidate terms are selected from the RF. The strength of their relation to the query terms is proportional to the fraction of the number of the documents containing both candidate terms and query terms and the product of the number of documents containing at least one of these sets.

Thus, the underlain hypotheses are as follows:

1. The importance of the query terms depends on the number of documents where they occur.
2. The importance of query term combinations depends on their number and the importance of each term. The importance of all possible term combinations is calculated.
3. The importance of a candidate term depends on its frequency in RF.
4. The importance of a candidate term is proportional to the number of documents where it co-occurs with query terms.

The proposed algorithm implies the following steps:

1. Preprocessing.
2. The frequencies of terms from the RF are computed.
3. The importance of the query terms is calculated.
4. The importance of all possible term combinations is calculated.
5. The importance of candidate terms is estimated.
6. The best-scored candidates are selected.

A query is cleared from stop-words, punctuation; duplicate terms are removed. However if a query contains only stop-words, this could mean that a user is interested, for example, in grammar. For instance, the query "a and the" may imply that a user wants to find how to use English articles. Thus, if a query contains only stop-words, we keep all of them (it requires to keep stop-words during indexing).

Let T be a set of all possible term combinations T_j. $T_j \in T = 2^Q \setminus \emptyset$ where 2^Q is the power set of all query terms. We compute T directly, i.e. we generate all possible subsets.

The importance of term combinations W_{T_j} is estimated by the formula:

$$W_{T_j} = \sum_{t_i \in T_j} (Il(t_i) + 1) \tag{2.4}$$

$$Il(t_i) = \frac{1}{\log length(t_i)} \tag{2.5}$$

where t_i is the $i-th$ term from T_j, $length(t_i)$ is the number of documents containing the $i-th$ term. $Il(t_i)$ is similar to IDF. The difference with IDF is that we do not consider the total number of documents and the logarithm appears in the denominator. For widely-spread terms with low $Il(t_i)$ the importance of their combination is approximately equal to their number. Moreover, we hypothesize that terms occurring only in one document in the collection are not useful for query expansion. Thus, we ignore them.

The importance of candidate terms W_c is computed as follows:

$$W_c = TF(c) \times \sum_{T_j \in T} MI(T_j, c) \tag{2.6}$$

where $MI(T_j, c)$ is the analogue of non-negative point-wise mutual information calculated by the formula:

$$MI(T_j, c) = \frac{-\log_2 \left(\max \left(\frac{length(T_j,c) \times n}{length(T_j) \times length(c)}, 1 \right) \right)}{\log_2 \left(\frac{length(T_j,c)}{n} \right)} \tag{2.7}$$

where $length(c)$ is the size of the set of the documents containing the candidate term c, $length(T_j)$ is the number of the documents containing all terms from the term combination T_j, $length(T_j, c)$ is the length of the intersection between the set of the documents containing all terms from the term combination T_j and the set of the documents containing the candidate term c, and n is the total number of documents in the collection.

All weights W_c are normalized. The best-scored term candidates are selected for query expansion.

2.3.3 Contribution 7: Proximity Relevance Model

Two previous QE approaches proposed in this chapter are empirical. Here we introduced a formal QE model.

As in the approaches based on the term proximity, we hypothesize that the closer a term is to a query term, the better the QE term candidate is. However, unlike the positional relevance model [Lv and Zhai, 2010] which is a formal approach for QE extending the relevance LM [Lavrenko and Croft, 2001b], we believe that the suitability of the expansion candidates depends on rather their nature and the nature of the query terms than position within a document (e.g. synonyms usually do not co-occur within a sentence unlike other semantically related words). In contrast to [Lv and Zhai, 2010], the proximity is captured directly rather than by weighting the positions within PRF. We choose LM formalism since it is justified statistically.

We also put forward a hypothesis that it is more appropriate to estimate the distance not in terms of tokens, but rather in terms of sentences. This is motivated by the following facts:

- In linguistics a sentence is viewed as a minimal set of words that in principle tells a complete thought;
- Within a sentence, often words could be reordered without meaning shift (e.g. paraphrasing, transformation between passive and active voices);
- Synonyms and associations are usually considered as good expansion candidates. However, synonyms usually do not co-occur within a sentence unlike other semantically related words.

Thus, our approach differs from the previous works by capturing the proximity directly and in terms of sentences rather than tokens.

The proposed method aims at selecting the most appropriate expansion terms for QE from the top-retrieved documents. Our approach is grounded on the following hypotheses:

1. A candidate term can expand not only a query term, but also a combination of query terms.

2. The terms lying in the neighborhood of query terms are closer related to them than the remote ones, and are better candidates for QE; however this dependence is not binary but rather it should be described as a more complex function.

3. Since a sentence is a minimal set of words that in principle tells a complete thought (i.e. it's a minimal unit telling a complete thought), the distance should be estimated in terms of sentences rather than in terms of tokens. The probability to find semantically related words is the same sentence is usually higher. However, this probability depends on the nature of relationship (i.e. synonyms, antonyms, meronyms, associations etc.).

One of the most efficient and robust relevance model used for QE is the relevance LM that determines the probability $P(w|Q)$ of observing a word w in the documents relevant to a particular information need expressed by a query Q [Lavrenko and Croft, 2001b]:

$$P(w|Q) \propto \sum_{d \in D} P(w|d)P(d)\prod_{i=1}^{m} P(q_i|d) \tag{2.8}$$

where $Q = q_1, q_2, ..., q_m$ is a query, q_i is the $i-th$ term in Q, $P(d)$ is a prior of a document d, and D is a document set. Often, document priors $P(d)$ are assumed to be uniform and in this case they can be ignored since they do not affect ranking.

In the relevance LM the probabilities are computed over the top documents from PRF. By the definition of conditional probability and since $P(Q)$ does not depend on w:

$$P(w|Q) = \frac{P(w,Q)}{P(Q)} \propto P(w,Q) \tag{2.9}$$

In contrast to the relevance LM, we assume that considering the distance between a candidate term and query terms may improve the quality of QE. We hypothesize that good QE candidates in the neighborhood of query terms. Usually, the closer a term is to a query term, the better candidate it is. However, it is not a case of synonyms. Therefore, we introduce the random variable *dist* that expresses the probability to find a candidate term at some sentence distance from the query terms Q. Since $P(w,Q)$ may be viewed as marginal over the variable *dist*, the general proximity relevance model can be expressed as:

$$P(w|Q) \propto \sum_{dist=0}^{\infty} P(w, dist, Q) \tag{2.10}$$

where $P(w, dist, Q)$ is the joint probability of seeing w, $dist$, and Q.

We enriched the relevance model by integrating the query term combinations into it. Thus, a word w can extend a query term combination $Q_i \in \Omega = 2^Q \setminus \emptyset$ where 2^Q is the power set of all query terms meeting the condition that $(\forall i, j | i \neq j) : Q_i \cap Q_j = \emptyset$. Since the events Q_i are mutually exclusive,

$$P(w, dist, Q) = \sum_{Q_i \in \Omega} P(w, dist, Q_i) \tag{2.11}$$

Thus, formula (2.10) can be rewritten as:

$$P(w|Q) \propto \sum_{Q_i \in \Omega} \sum_{dist=0}^{\infty} P(w, dist, Q_i) \tag{2.12}$$

Applying the chain rule, $P(w, dist, Q_i)$ can be decomposed as:

$$P(w, dist, Q_i) = P(Q_i)P(dist|Q_i)P(w|dist, Q_i) \tag{2.13}$$

where $P(Q_i)$ is the probability of the query term combination Q_i, $P(dist|Q_i)$ is the probability to find any expansion term at distance $dist$ from Q_i, and $P(w|dist, Q_i)$ is the probability to find the term w at distance $dist$ from Q_i. $P(dist|Q_i)$ may be viewed as a likelihood to see an expansion term at a specified distance depending on the nature of a query term combination Q_i. $P(w|dist, Q_i)$ shows a likelihood to meet a specific term depending on the remoteness of a given Q_i i.e. it potentially captures the nature of the expansion candidate and its relationship with the query term (synonymy, meronymy, function etc.).

Substituting $P(w, dist, Q_i)$ in (2.12) by (2.13), we obtain the final formula to estimate expansion candidate scores:

$$P(w|Q) \propto \sum_{Q_i \in \Omega} \sum_{dist=0}^{\infty} P(Q_i)P(dist|Q_i)P(w|dist, Q_i) \tag{2.14}$$

The probability of a term combination $Q_i = q_1, q_2, ..., q_m$ is usually calculated as follows:

$$P(Q_i) = \prod_{j=1}^{m} P(q_j) \tag{2.15}$$

To avoid underflow, the probability is replaced by its logarithm as in [Hiemstra, 2009]:

$$P(Q_i) \propto \sum_{j=1}^{m} \log(P(q_j) + 1) \tag{2.16}$$

Assuming that the probability to find any expansion term at distance *dist* from Q_i does not depend on Q_i we can simplify the calculation of $P(dist|Q_i)$ by reducing it to $P(dist)$. The dependence of the nature of the query terms Q_i is the perspective of this research.

The distributions of many quantities follow the power law, at least in their upper tail, especially in natural languages (e.g. Zipf's law). Although it is not exactly known why the power law holds for most languages, the explanation may be statistical or related to the principle of least effort, i.e. interlocutors do not want to work any harder than necessary to reach understanding. We hypothesize that the principle of least effort holds also for topic development within a text. Thus, a topic within a text is expanded in the neighboring context and we assume that the distribution of the words used for it follows the power law. Thereby, the probability to find an expansion candidate for a topic expressed by query terms should also fit the power law.

The probability of $P(w|dist, Q_i)$ is estimated as the frequency of observing the term w at distance *dist* from Q_i:

$$P(w|dist, Q_i) \approx \frac{count(w|dist, Q_i)}{\sum_{k=1}^{|W|} count(w_k|dist, Q_i)}, \tag{2.17}$$

where W is a set of all terms, $|W|$ is the cardinality of W, i.e. the number of terms in the dictionary.

In this work the distance means the remoteness from the closest query term or their combination Q_i. Since we compute the distance in terms of sentences and the combinations of the query terms are considered only within a sentence, the remoteness does not depend on the length of the query term combination.

The set of the query term combinations $Q_i \in \Omega = 2^Q \setminus \emptyset | (\forall i, j | i \neq j) : Q_i \not\subseteq Q_j$ does not lead to the exponential complexity of the algorithm since we consider only query term combinations within a sentence and we ignore embedded combinations. Thus, the computation of the query term combinations has a linear time over the number of tokens in the PRF.

Smoothing the probability $P(w|dist, Q_i)$ by the collection probability of the candidate term $P_c(w)$ gives:

$$P_s(w|dist, Q_i) = \lambda P(w|dist, Q_i) + (1 - \lambda)P_c(w) \tag{2.18}$$

where $P_s(w|dist, Q_i)$ is a smoothed probability and λ is a smoothing parameter.

Dividing the equation by $(1-\lambda)P_c(w)$ we obtain the final ranking score of the expansion candidate terms:

$$score(w) = \frac{\lambda P(w|dist, Q_i)}{(1 - \lambda)P_c(w)} + 1 \tag{2.19}$$

2.4 Evaluation Framework

In this section the experimental framework is described. Firstly, we present the data sets we used. Then we describe the evaluation metrics and the systems used for comparison. The last subsection provides the details of the implemented system.

2.4.1 Data Sets

The evaluation was performed on two kinds of datasets: TREC (Text Retrieval Conference) Robust data sets and WT10G. TREC Robust Track data sets are a "pure" collection since the documents have almost the same format and there is no spam. In contrast, WT10G is a snapshot of the web with real documents in HTML format, some of which are spam. As it was showed in [Soboroff, 2002], WT10G "looks like" the web.

2.4.1.1 TREC Robust

For the evaluation purpose we used TREC Robust Track data sets for five years: 1997 - 2001 [Voorhees and Harman, 1998b, 2000b,c]. TREC Robust data are driven on the data on Disks 4 and 5 (except Congressional Record data) and contain 249 topics in total. There are 4 sources of documents: the news articles from

- The Financial Times, 1991-1994 (FT) - 564MB, 210,158 documents;
- Federal Register, 1994 (FR94) - 395MB, 55,630 documents;

- Foreign Broadcast Information Service (FBIS) - 470MB, 130,471 documents;
- The LA Times - 475MB, 131,896 documents.

Documents are tagged by SGML. The collected documents are not normalized and may contain spelling or other errors. Each of TREC Robust has 50 topics. A topic represents an information need and contains 4 fields:

- Topic number;
- Title (very short description of a topic – about three words);
- Description (a "normal" sentence description of a topic);
- Narrative (description of the information that should be presented at relevant documents).

The pools of relevant documents (q-rels) were merged from the top 100 documents per topic retrieved in each submitted run and assessed by humans.

2.4.1.2 WT10G.

WT10G was used at TREC Web track 2000-2001 [Hawking and Craswell, 2002b]. It is 10GB subset of the web snapshot of 1997 from Internet Archive. WT10G contains 1,692,096 documents from 11,680 servers (minimum 5 documents per server). There were 50 topics in 2000 and 2001 (total 100 topics). In total 5,953 were judged as relevant.

There are 98 topics with relevance judgments.

As in the Robust collection, documents are not normalized and tagged by SGML parser. The topics are also given in the TREC format.

2.4.2 Evaluation Measures

The performance of the systems was evaluated by several measures implemented in trec_eval software[2] provided by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results. The trec_eval

[2] http://trec.nist.gov/trec_eval/

software enables to evaluate ranked retrieval results. In this study, we report the following measures:

- Mean Average Precision (MAP) over all queries;
- Normalized discounted cumulative gain (NDCG);
- Binary preference (BPREF).

Precision (P) is the fraction of retrieved documents that are relevant [Manning et al., 2008]:

$$P = \frac{\#RelevantRetrievedItems}{\#RetrievedItems} \tag{2.20}$$

Precision at k ($P@k$) the fraction of the top k retrieved documents that are relevant. Interpolated average precision is the ratio of relevant retrieved documents over the number of documents that gives a certain percentage of recall. Recall (R) shows the fraction of relevant retrieved documents over all relevant documents:

$$R = \frac{\#RelevantRetrievedItems}{\#RelevantItems} \tag{2.21}$$

Mean average precision is calculated as follows:

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P@k \times rel(d_k) \tag{2.22}$$

where $|Q|$ is the number of queries, m_j is the number of relevant documents for the j-th query, $P@k$ is the precision at k, and $rel(d_k)$ is the relevance of the document d_k. MAP may be viewed as one of the main measures since it has very good discrimination and stability [Manning et al., 2008].

Discounted cumulative gain DCG is a measure of effectiveness of information retrieval that penalizes highly relevant documents appearing lower in a search result [Manning et al., 2008]. The graded relevance value is discounted logarithmically proportional to the position of the result:

$$DCG_k(Q_j) = \sum_{i=1}^{k} \frac{2^{rel_i^{(j)}} - 1}{log_2(i+1)} \tag{2.23}$$

Normalized discounted cumulative gain $NDCG$ is normalized over Ideal DCG $IDCG$, i.e. the maximum possible DCG till the position k:

$$NDCG_k(Q_j) = \frac{DCG_k(Q_j)}{IDCG_k(Q_j)} \tag{2.24}$$

$NDCG$ can be averaged over all queries and all positions:

$$NDCG(k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} NDCG_k(Q_j) \tag{2.25}$$

$$NDCG = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} NDCG_k(Q_j) \tag{2.26}$$

Since the gain of each document is discounted at lower ranks, NDCG is suitable for non-binary judgments.

Binary preference computes a preference of whether judged relevant documents have higher rank than judged non-relevant documents. Thus, BPREF does not treat non-assessed documents as non-relevant while MAP does. This is important for large collections where the probability of retrieving non-assessed documents is higher.

2.4.3 Systems Used For Comparison

For comparison purpose we used several PRF methods, namely Divergence from Randomness (DFR) models implemented in an open-source search engine Terrier [Amati, 2003, Ounis et al., 2006b].

During QE the best-scored terms from the top-ranked documents are extracted. Terms are ranked using one of the DFR weighting model. We compare our systems with the following DFR models:

- Baseline presented by InL2c1.0 model without any query expansion which is the default model in Terrier and based on $TF-IDF$ measure with $L2$ term frequency normalization (InL2);
- Kullback-Leibler divergence model (KL);
- Chi-square divergence model (CS);

- Bose-Einstein 1 model (Bo1);
- Bose-Einstein 2 model (Bo2).

According to our preliminary study, with the default Terrier's parameters, on the used collections $InL2$ showed better results than Okapi's BM25 and Hiemstra's implementation of the language model. For the detailed description of the $InL2$ model see Subsection 1.5.4.

In the DFR models QE is performed by ordering the candidate terms by their information content given the query Q [Amati, 2003]:

$$Inf(t \mid Q) = Inf_{D_Q}(t) = -\log P(t \mid Q) \tag{2.27}$$

where t is the candidate term.

In the Kullback-Leibler model $P(t \mid Q)$ is viewed as binomial distribution:

$$\begin{aligned} P(t \mid Q) &= B(C_{D'(t)}, C_{D'}, \frac{C_D(t)}{C_D}) \\ &= \binom{C_{D'}}{C_{D'}(t)} (\frac{C_D(t)}{C_D})^{C_{D'}(t)} (1 - \frac{C_D(t)}{C_D})^{C_{D'} - C_{D'}(t)} \end{aligned} \tag{2.28}$$

where D' is a subset of the original collection D, $C_{X(t)}$ is the number of time the term t occurs in X, C_X – the total number of terms in X; it can be approximated via the divergence function. In this case the information content of the term t is proportional to:

$$Inf(t \mid Q) \sim TF_{D'}(t) \times \log \frac{TF_{D'}(t)}{TF_D(t)} \tag{2.29}$$

Chi-square divergence implies that the information content of the term t is estimated as [Amati, 2003]:

$$\begin{aligned} Inf(t \mid Q) \sim TF_{D'}(t) \times TF_{D'} \times (\log \frac{TF_{D'}(t)}{TF_D(t)} + \log \frac{1 - TF_{D'}(t)}{1 - TF_D(t)}) \\ + 0.5 \times (2\pi \times TF_{D'} \times (1 - \frac{TF_{D'}(t)}{TF_D(t)})) \end{aligned} \tag{2.30}$$

Bose-Einstein 1 (Bo1) and 2 (Bo2) models are the best DFR models implemented in Terrier [Amati, 2003]. By default they are parameter-free, but Rocchio's query expansion

mechanism can be also applied.

$$Bo1 = TF_{D'}(t) \times \log \frac{1+f1}{f1} + \log(1+f1) \tag{2.31}$$

$$f1 = \frac{TF_D(t)}{|D|} \tag{2.32}$$

$$Bo2 = TF_{D'}(t) \times \log \frac{1+f2}{f2} + \log(1+f2) \tag{2.33}$$

$$f2 = \frac{TF_{D'}(t) \times TF_{D'}}{TF_D} \tag{2.34}$$

In Bo1 $f1$ presents the average frequency of the term t in a document from the collection, as well as $f2$ in Bo2. The difference is that $f1$ is actually calculated as the average frequency of the term t, while in Bo2 the frequency of the term t in RF is extrapolated to the entire collection.

Moreover, we compared our method with RM3 model implemented in Indri, a search engine from the Lemur project mainly built on the language modeling information retrieval[3]. RM3 is an Indri's adaptation of Lavrenko and Croft's relevance models [Lavrenko and Croft, 2001b]. RM3 is a well-known relatively strong baseline.

2.4.4 Details of the Implemented Systems

All systems used InL2c1.0 model for relevance feedback, 5 documents from which 10 best scored terms were extracted.

Our approach requires PRF. In order to obtain preliminary ranking we used Terrier with the following parameters: words are stemmed using Porter's algorithm, as a retrieval model we applied $InL2c1.0$. The sentence chunking was performed by Stanford CoreNLP[4].

2.4.4.1 LC Model

Candidate terms are ranked according to the following metric:

$$w_{total}(t) = score(S) \times w_{pos}(t) \times IDF(t) \times importance(t, Q) \tag{2.35}$$

[3] http://www.lemurproject.org/
[4] nlp.stanford.edu/software/corenlp.shtml

$$importance(t, Q) = wd(t, Q) \times cooccurrence(t, Q) \tag{2.36}$$

where $score(S)$ is score of the sentence S containing t computed by (1.25), $w_{pos}(t)$ is the weight of the POS of t, $IDF(t)$ is the inverse document frequency of the candidate term, $wd(t, Q)$ is a function of the distance in terms of tokens from the candidate terms to the query Q and their weights, and $coocurence(t, Q)$ shows the likelihood of the candidate term to occur not by chance with the query terms in the top documents ranked according to the initial query.

2.4.4.2 Proximity Relevance Model

In our experiments we assumed that the probability $P(dist)$ of a candidate to occur at the given distance follows the power law. We set the limit of distance $MaxDist = 9$ sentences and thus we calculated the $P(dist)$ as:

$$P(dist) = \begin{cases} \frac{1}{(MaxDist+2)^{0.5}} & \text{if } dist > MaxDist \\ \frac{1}{(dist+1)^{0.5}} & \text{if } dist \leq MaxDist \end{cases} \tag{2.37}$$

The smoothing parameter λ was set to 0.3. This parameter should be learnt and optimized in future work.

In order to test the hypothesis that it is preferable to estimate the distance in terms of sentences rather than tokens, we compared our approach with the same method in which the distance was calculated at word level (PRM_W). The $MaxDist$ parameter was also set to be 9 sentences, the estimation of the probability $P(dist)$ was slightly different:

$$P(dist, wdist) = \begin{cases} \frac{1}{((MaxDist+1)\times avgSntLen+1)^{0.5}} & \text{if } dist > MaxDist \\ \frac{1}{(wdist+1)^{0.5}} & \text{if } dist \leq MaxDist \end{cases} \tag{2.38}$$

where $wdist$ is a word distance, $avgSntLen$ is an average sentence length.

2.5 Results

Table 2.1 provides information about the results obtained for the Robust and WT10G data sets applying 4 methods we proposed:

TABLE 2.1: General comparison of QE methods

		InL2	**KL**	**CS**	**Bo1**	**Bo2**	**RM3**	**LC**	**Co**	**PRM_W**	**PRM_SNT**
Robust	MAP	0.2407	0.2829	0.263	0.2822	0.2801	0.2602	$0.2852^{*\#}$	$0.2859^{*\#}$	$0.2795^{*\#}$	$\mathbf{0.2884}^{*+\#}$
	BPREF	0.2506	0.2807	0.2635	0.28	0.2782	0.2585	$0.2832^{*\#}$	$0.2845^{*\#}$	$0.28^{*\#}$	$\mathbf{0.2863}^{*+\#}$
	NDCG	0.5124	0.5566	0.5329	0.5561	0.5549	0.5268	$0.5598^{*\#}$	$0.5615^{*\#}$	$0.5549^{*\#}$	$\mathbf{0.5614}^{*\#}$
WT10G	MAP	0.1894	0.2121	0.2013	0.2179	0.2174	0.2142	$0.2239^{*\#}$	$0.2276^{*+\#}$	$0.2247^{*\#}$	$\mathbf{0.2345}^{*+\#}$
	BPREF	0.1895	0.2016	0.2009	0.2107	0.2071	0.2084	$0.2156^{*\#}$	$0.2234^{*+\#}$	$0.2153^{*\#}$	$\mathbf{0.2275}^{*+\#}$
	NDCG	0.4624	0.4888	0.4497	0.4927	0.4885	0.4699	$0.4958^{*\#}$	$0.4955^{*\#}$	$0.5005^{*+\#}$	$\mathbf{0.5076}^{*+\#}$

- Adaptation of Sentence Extraction Method to Query Expansion (LC);
- Co-occurrence Model (Co);
- Proximity Relevance Model based on word-level distance (PRM_W).
- Proximity Relevance Model based on sentence-level distance (PRM_SNT).

We performed the Student's t-test to verify the statistical significance of the difference of the results obtained by our method and the baseline (this test is applicable since the performance results follow a normal distribution according to Chi-square test). We also compared our results with those of the best approach implemented in Terrier, namely Bo1 (although KL is slightly better on the Robust data set, it has much lower results on WT10G), and the RM3 implementation of Lemur's LM. The differences with the baseline, Bo1 and RM3 marked by *, + and # respectively are significant at the level $p = 0.05$.

According to all evaluation measures on both test collections all our systems significantly outperformed the baseline and showed better results in average than all the QE models implemented in Terrier as well as RM3. The difference between all our systems and RM3 is significant in all cases. On Robust data set RM3 performed worse than the DFR QE models. In case of WT10G, RM3 was comparable with DFR QE approaches but remained significantly lower than the methods proposed in this research.

TABLE 2.2: Collections' statistics

	Robust	**WT10G**
Total # of queries	249	98
# of very difficult queries ($MAP(InL2) \leq 0.1$)	79	38
# of difficult queries ($MAP(InL2) \leq 0.25$)	145	69
# of easy queries ($MAP(InL2) \geq 0.5$)	30	9

TABLE 2.3: # of improved and worsen queries

		All/Very hard/Hard/Easy				
		Bo1	**LC**	**Co**	**PRM_W**	**PRM_SNT**
Robust	> **InL2**	182/53/100/21	183/57/103/21	172/53/98/18	188/58/107/20	177/52/99/21
	< **InL2**	65/25/44/8	65/22/42/8	75/26/47/11	60/21/38/9	71/27/46/8
	> **Bo1**	-	147/49/80/14	136/45/78/14	126/41/76/13	124/42/75/12
	< **Bo1**	-	101/30/65/15	112/34/67/15	122/38/69/16	123/37/70/16
WT10G	> **InL2**	58/21/40/5	64/23/45/4	57/18/39/4	63/22/45/5	62/19/43/5
	< **InL2**	36/15/27/3	31/13/22/4	38/18/28/4	32/14/22/3	33/17/24/3
	> **Bo1**	-	50/15/37/4	42/12/31/3	50/16/35/5	54/15/40/4
	< **Bo1**	-	44/21/30/3	52/24/36/4	44/20/32/2	41/21/27/4

Table 2.3 reports the detailed statistics of the amelioration/degradation of results for all, very difficult ($MAP(baseline) \leq 0.1$), difficult ($MAP(baseline) \leq 0.25$) and simple ($MAP(baseline) \geq 0.5$) queries.

On the Robust collection our LC method ameliorated the highest number of queries (all, difficult, very difficult and easy) comparing to Bo1 while it had the minimal rate of the result degradation. On WT10G it showed the best improvement for all, difficult and very difficult queries with regard to the baseline while keeping the lowest degradation rate for this type queries.

Co demonstrated the highest degradation of the results with regard to the baseline for both test collections (Robust - 75, WT10G - 38). It had the biggest number of the performance lower than Bo1 on WT10G. However, the average results are better than LC.

On the Robust collection the word-based Proximity Relevance Model (PRM_W) improved the most of all queries (188), very difficult (58) and difficult queries (107) relatively to the baseline. It has the lowest rate of the degradation of results regarding the baseline among all queries (60), very difficult (21) and difficult queries (38). However, the amelioration of results towards Bo1 is the lowest among all our methods for very difficult queries (41). PRM_W showed worse results than PRM_SNT and KL according to *BPREF* but outperformed other DFR models. PRM_W is much better than the baseline and it is comparable with the DFR QE models according to other metrics.

On WT10G PRM_W was worse than Bo1 only for 20 of very difficult queries and for 2 easy queries which is the lowest rate of the degradation. It improved the highest number of easy queries (5) and very difficult queries (16) with regard to Bo1.

Our Proximity Relevance Model based on sentence-level distance (PRM_SNT) demonstrated the best results according to all metrics for both data collections.

Considering the Robust collection, in comparison with Bo1 our method PRM_SNT showed better results for 124 queries and lower performance for 123 queries. Our method outperformed Bo1 for 75 (60% of all improved results) difficult queries and for 42 (34%) very difficult queries. Among ameliorated results 12 (10%) of queries were simple. Thus, we can conclude that PRM_SNT is better than the state-of-the-art QE model even in case of difficult queries. The degradation of results in comparison with the baseline was

observed in 71 cases. Among the latter for 37% of queries the degradation of the results relative to the baseline without QE was observed for all QE methods; this feature leds us to conclude that either these queries should not be expanded or the methods based on co-occurrence are not suitable. Although PRM_W raised much more queries (all, difficult, and very difficult) than PRM_SNT and showed lower degradation of results with regard to the baseline, its average improvement is lower. Both our systems significantly exceeded RM3 by all metrics.

For the WT10G data set our PRM_SNT system was better than Bo1 for 54 queries. PRM_SNT was worse than the baseline for 33 queries and among them for 16 queries (49%) any applied QE method worsened the results. It lowers performance compared to Bo1 for 41 queries. This allows to draw a conclusion that our method may be significantly improved by selective QE since it has lower results than the DFR models mainly for queries that should not be expanded at all. Word-based PRM surpassed all DFR models but it was inferior to PRM_SNT.

For 51% (Robust) and 59% (WT10G) of queries improved by Bo1, our system outperformed the DFR models.

On the Robust collection for 26 queries (10%) all systems showed the degradation of performance regarding the baseline. On WT10G the same effect was observed for 16 queries (16%).

Table 2.4 reports the statistics of result degradation. *NotExp* refers to the queries all systems showed the degradation of performance relative to the baseline. *DegradQ* corresponds to the degraded queries. For the Robust collection approximately 40% of the queries our systems demonstrated worse results than the baseline was decreased by all QE methods under consideration. For WT10G this percentage is almost 50%. The average degradation of results of our systems is much lower for the rest of queries on the Robust data set. We can observe the same trend on WT10G while for Bo1 it is an opposite tendency. Thus, we can conclude that our methods fail when all other statistical approaches also fail. Since the hypotheses underlain the methods are different (the divergence of word frequencies in the elite set and the rest of collection for DFR models, the proximity to the query terms for LC and PRM models, the strength of their relation to the query terms for Co), we can also draw a conclusion that statistical methods are not suitable for these queries.

TABLE 2.4: Result degradation (NDCG)

		Bo1	LC	Co	PRM_W	PRM_SNT
Robust	# of DegradQ	65	65	75	60	71
	% of NotExp over DegradQ	40	40	35	43.3	36.6
	Avg degradation for NotExp	-0.0477	-0.0564	-0.0607	-0.0499	-0.0732
	Avg degradation for DegradQ\NotExp	-0.033	-0.0323	-0.0356	-0.0313	-0.0416
WT10G	# of DegradQ	36	31	38	32	33
	% of NotExp over DegradQ	44	52	42	50	49
	Avg degradation for NotExp	-0.0478	-0.0533	-0.0744	-0.0532	-0.0786
	Avg degradation for DegradQ\NotExp	-0.0785	-0.0561	-0.0662	-0.0562	-0.0646

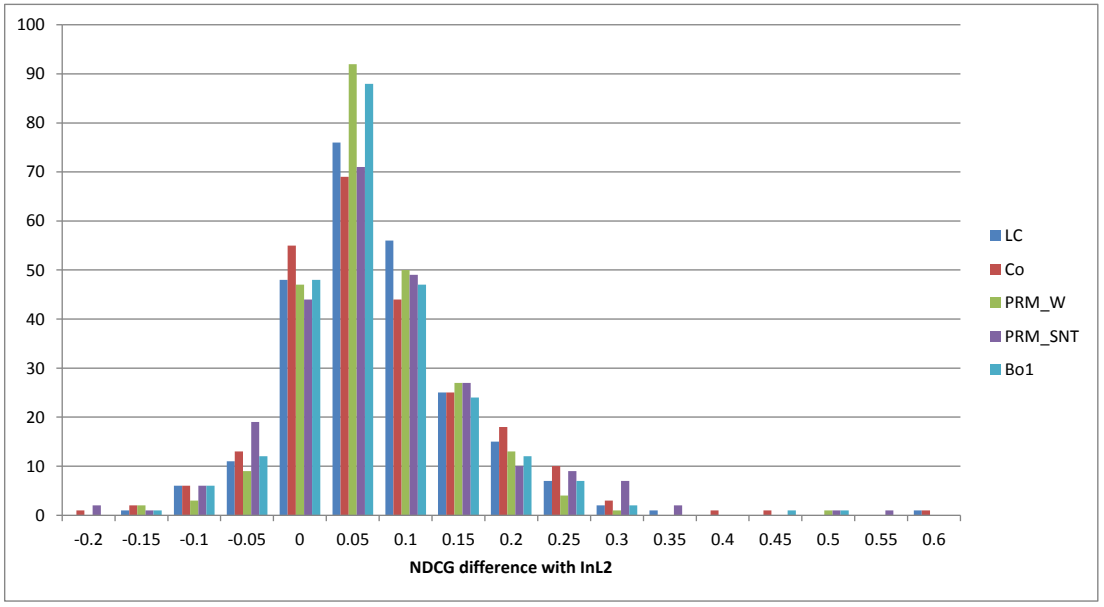


FIGURE 2.1: Histogram of the NDCG difference with the baseline (Robust)

Figures 2.1 and 2.2 present the histogram of the NDCG difference with the baseline on Robust and WT10G collections respectively. As it is evident from the charts, the difference follows the normal distribution. Bo1, LC and PRM_W tend to have small amelioration of results while PRM_SNT and Co are further from 0.

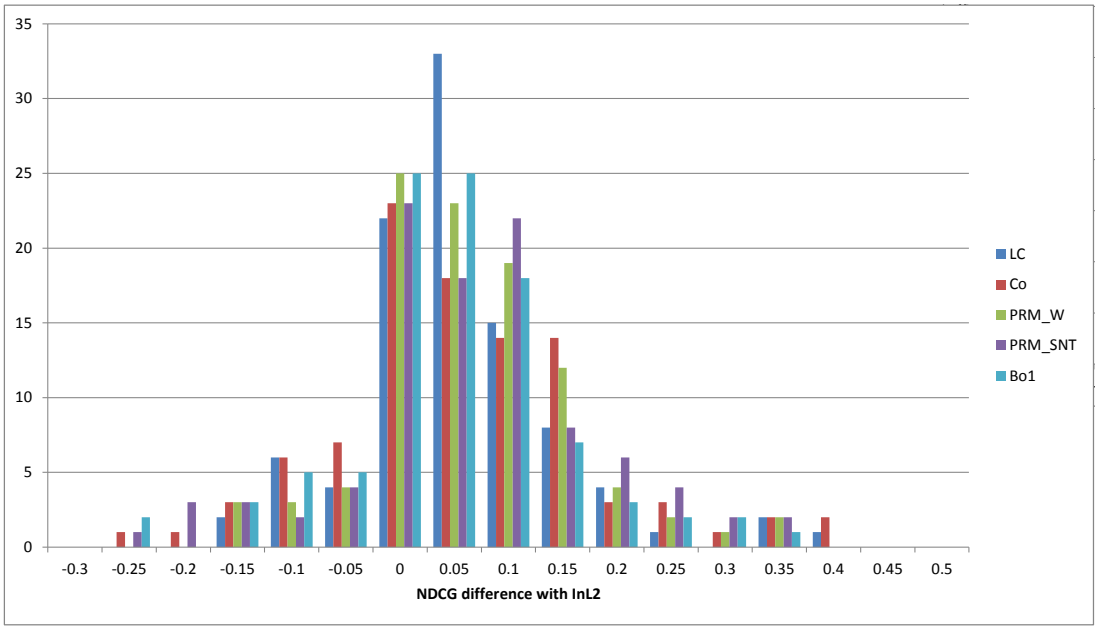


FIGURE 2.2: Histogram of the NDCG difference with the baseline (WT10G)

2.6 Deep Analysis of the Queries

A few studies have reported analysis of results. The deeper analysis has been conducted in the RIA Workshop that took place in 2004. One of the objectives of the workshop was to analyze the variability in systems: some systems answering well on some queries and badly on others; some other systems behaving oppositely. One of the conclusion of the workshop was that the comprehension of variability is complex because of various parameters: query formulation, the relation between the query and the documents as well as the characteristics of the system [Harman and Buckley, 2009a]. Moreover, they conducted failure analysis for 45 of the TREC topics. After using various systems on "hard" topics, the workshop participants analysed why the system failed. For 39 topics out of 45 the systems failed for the same reason. Moreover, even if they did not retrieve the same documents, they were missing the same aspect in the top documents. During the same workshop, the fact that systems reached an optimal in results using a different number of pseudo-relevant documents has been studied as well as a different number of terms. It has been shown that when choosing the optimal number of terms in the expanded query, the results can be improved up to 30% compared to using the same fixed number of terms for all queries [Harman and Buckley, 2004].

Some studies focus on the when QE is useful. Indeed it has been shown that if RF

TABLE 2.5: Results for individual queries (NDCG)

Collection	Query	# of rel docs	InL2	Bo1	LC	Co	PRM_W	PRM_SNT
Robust	429	11	0.4498	0.1818	0.1736	0.0661	0.2397	0.2397
	659	16	0.4065	0.3242	0.3398	0.4414	0.4961	0.4688
	614	30	0.1791	0.5189	0.2278	0.4756	0.4711	0.5978
	415	136	0.2351	0.5723	0.3646	0.3683	0.3721	0.3979
	615	12	0.078	0.0486	0.2292	0.2361	0.1111	0.0625
	350	68	0.0742	0.34	0.2571	0.4412	0.1256	0.2978
	648	57	0.2574	0.5078	0.5543	0.5106	0.5349	0.6962
	352	246	0.0347	0.3831	0.4269	0.4462	0.4126	0.4137
WT10G	484	13	0.1943	0	0.1775	0.1538	0.1716	0.1479
	538	2	0.3929	0.25	0	0	0	0.25
	531	22	0.1098	0	0.5661	0.4421	0.3017	0.657
	504	18	0.4183	0.358	0.3488	0.25	0.4444	0.2377
	486	4	0.85	0.8125	0.5	0.5	0.5625	0.6875
	529	39	0.2847	0.4602	0.3649	0.3636	0.1696	0.3241
	548	2	0.5	1	1	1	1	0.5

successfully improves the system performance in average [Voorhees and Harman, 1998b], in some cases, QE worsens the quality of the retrieval [Amati et al., 2004c].

In the previous subsection, we reported the results when averaged over the set of topics. In this subsection, we aim at analyzing the results deeper.

Tables 2.6 and 2.7 provides the maximal and minimal values of the NDCG differences between the systems and the baseline and Bo1 respectively. The maximal enhancement in comparison with the baseline was observed for PRM_SNT while it also demonstrated the minimal degradation of the performance. At the same time Bo1 showed lower improvement over the baseline and it had higher lost of performance. For both

TABLE 2.6: NDCG differences with the baseline

	Bo1	**LC**	**Co**	**PRM_W**	**PRM_SNT**
Robust max	0.4816	0.5618	0.5759	0.4682	0.5084
Robust min	-0.1633	-0.1872	-0.2023	-0.1328	-0.2481
WT10G max	0.3491	0.3969	0.3837	0.3491	0.3314
WT10G min	-0.2766	-0.1655	-0.2654	-0.1812	-0.2705

TABLE 2.7: NDCG differences with Bo1

	LC	**Co**	**PRM_W**	**PRM_SNT**
Robust max	0.1155	0.2011	0.2782	0.2629
Robust min	-0.1969	-0.2073	-0.2366	-0.1749
WT10G max	0.6538	0.6134	0.5243	0.5751
WT10G min	-0.2296	-0.1862	-0.1988	-0.2904

data sets PRM_SNT indicated the highest amelioration of results regarding Bo1 and it kept the minimal lost on Robust collection. The minimal lost on WT10G was observed for PRM_W.

Let's provide the examples of the reformulation for those queries.

2.6.1 Analysis of the Individual Queries from the Robust Collection

Query 429

Example 2.1. *Query 429: Initial query*

```
<top>

<num> Number: 429
<title> Legionnaires' disease

<desc> Description:
Identify outbreaks of Legionnaires' disease.

<narr> Narrative:
```

```
To be relevant, a document must discuss a specific outbreak of Legionnaires'
    disease.  Documents that address prevention of or cures for the disease
    without citing a specific case are not relevant.

</top>
```

Example 2.2. *Query 429: Bo1 reformulation*

```
legionnair^1.804569662 diseas^1.482883114 legionella^0.636239403 pneumophila
    ^0.558617881 infect^0.227070259 outbreak^0.210052526 amoeba^0.165899622
    chlorin^0.143418258 pneumonia^0.112580615 water^0.111270052
```

Example 2.3. *Query 429: LC reformulation*

```
legionnair^2.0146209689198065 diseas^1.8911322950158538 legionella
    ^0.6830360594367585 legionella^0.6830360594367585 pneumophila
    ^0.5067319102381689 water^0.32210003427033346 infect^0.30570263056332364
    outbreak^0.27820978995931267 case^0.2227261848005783 chlorin
    ^0.21961963787580077 hospit^0.21003013807269566 patient^0.1994115622483936
```

Example 2.4. *Query 429: Co reformulation*

```
legionnair^2.0 legionella^2.0 diseas^1.5134640718113959 infect^0.3940050560402672
     water^0.3874685818828991 outbreak^0.3648333691693237 pneumophila
    ^0.3313340667472908 chlorin^0.2683662387112157 patient^0.24802786837612692
    hospit^0.24172943280864295 tower^0.22450585573370108
```

Example 2.5. *Query 429: PRM_ W reformulation*

```
legionnair^3.0 diseas^2.9183354796666667 nosocomi^0.5286469750203374 center
    ^0.5053074898420538 medicin^0.5000574364908077 definit^0.4917742460981773
    control^0.47531432410657065 chlorin^0.47000430090050777 health
    ^0.4661053502779508 case^0.45965540070895783 surveil^0.456962705478709 caus
    ^0.4555249360833738
```

Example 2.6. *Query 429: PRM_ SNT reformulation*

```
diseas^3.0 legionnair^2.9098803602971515 chlorin^0.7849228166812063 health
    ^0.7629129558785768 infect^0.7620206959420539 water^0.7560286059809243
    disinfect^0.7516944009519004 center^0.7416048549060725 case
    ^0.7371036835753165 outbreak^0.736943034776419 caus^0.7364951954447273 amoeba
    ^0.7287685252650867
```

The maximal degradation of results of all our systems and Bo1 in comparison with the baseline was observed for the query 429. The BPREF differences with the baseline are given in the table 2.8. Both PRM models have lower loss of performance than Bo1.

TABLE 2.8: Query 429. BPREF differences with the baseline

Bo1	**LC**	**Co**	**PRM_W**	**PRM_SNT**
-0.268	-0.2762	-0.3837	-0.2101	-0.2101

LC showed the results slightly lower than Bo1. The major comedown was detected for Co. Apparently, it is related to the high weight of the term *legionella*. Bo1, LC and Co expanded the query by a term *outbreak* which should be highly relevant according to the narrative of the topic. However, all three models added primary the terms related to the typology of the disease and its causes (*legionella, pneumophila, infect, amoeba, pnemonia etc.*). Although these terms are strongly related to the query terms, they are very rare in the collection and therefore they are considered to be very important misleading the retrieval. Probably, the lower degradation of results of PRM_W and PRM_SNT could be explained by the fact that they extracted less specific terms. The information need is not clearly expressed by the query. We believe that this is the main cause of the fail of all systems for this query.

Query 659

Example 2.7. *Query 659: Initial query*

```
<top>

<num> Number: 659

<title>
cruise health safety

<desc>
What standards do cruise ships use for health and safety maintenance?

<narr>
Relevant documents refer to health and safety practices and standards for
    recreational cruise ships.  Not relevant are standards for small pleasure
    craft or commercial freight ships, tankers, etc.  Documents referring to a
    specific ship's problems are not relevant.

</top>
```

Example 2.8. *Query 659: Bo1 reformulation*

```
cruis^1.303755462 health^1.000000000 safeti^1.000000000 ship^0.321389416 sail
    ^0.116974298 inspect^0.116003620 passeng^0.094697144 sanit^0.073082594 vessel
    ^0.068586475 port^0.066333262 line^0.055063305 sea^0.053340215
```

Example 2.9. *Query 659: Query 659: LC reformulation*

```
cruis^1.7519672213759834 safeti^1.2124865710441546 health^1.1518448055648127
    icebreak^1.086526643960179 3 ship^0.9180418021960389 power^0.7021132852871275
    kuchiyev^0.6750950166732226 struzhentsov^0.45659893731737294 fleet
    ^0.30015958934520043 inspect^0.30013352889268685 sail^0.29656750779355456
    passeng^0.26206595659335336 sea^0.2156437658102652
```

Example 2.10. *Query 659: Co reformulation*

```
cruis^2.0 safeti^1.7897828025361338 health^1.7559235744588189 icebreak^1.0 ship
    ^0.698087832370346 power^0.43239587464154294 fleet^0.24372184120429 32 sail
    ^0.24095777772725843 inspect^0.21711288978425802 passeng^0.19223197616073157
    academician^0.1836395599384964 sea^0.1474030054729126 port
    ^0.13598087895249514
```

Example 2.11. *Query 659: PRM_ W reformulation*

```
cruis^3.0 safeti^2.8544067608864556 health^2.7910606371666704 ship
    ^0.5818492620795515 inspect^0.5705751336486626 royal^0.5266487743361852
    transport^0.5044367524646532 passeng^0.4995080846520235 emerg
    ^0.4979492006843004 earlier^0.4861624133116 lo^0.4848770053975582 vessel
    ^0.47324907555052326 line^0.4682171312506147
```

Example 2.12. *Query 659: PRM_ SNT reformulation*

```
cruis^2.776903479765094 safeti^2.639150492185416 health^2.599447409110432 ship
    ^1.0 inspect^0.838043539697097 passeng^0.7313219495278382 pass
    ^0.6810221896919779 line^0.6340575134327232 sail^0.6206002188985422 vessel
    ^0.615465457306640 5 earlier^0.6128238969712001 room^0.5857926796872691 water
    ^0.580221365439473
```

PRM_W demonstrated the maximal superiority over Bo1 for the query 659. BPREF differences between our systems and Bo1 are reported in the table 2.9. Bo1 and LC showed small degradation relative to the baseline while Co, PRM_W and PRM_SNT improved results. The terms extracted by Bo1 are related mainly to ships in general while LC had very specific but wrong terms (*kuchiyev, struzhentsov*). PRM_W expanded the query by highly semantically related terms (*ship, inspect, transport, passeng, emerg, vessel*). The term *lo* could be mapped into *'line of sight'* or *'loss of signal'* that could also occur in the relevant documents.

TABLE 2.9: Query 659. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
0.0156	0.1172	0.1719	0.1446

TABLE 2.10: BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
-0.0823	-0.0667	0.0349	0.0896	0.0623

Query 614

Example 2.13. *Query 614: Initial query*

```
<top>

<num> Number: 614
<title> Flavr Savr tomato

<desc> Description:
Find information about the first genetically modified food product to go on the
    market, Flavr Savr (also Flavor Saver) Tomato developed by Calgene.

<narr> Narrative:
Documents about genetically engineered food in general are not relevant; relevant
     documents must include specifics regarding the Flavr Savr tomato.

</top>
```

Example 2.14. *Query 614: Bo1 reformulation*

```
flavr^1.500545745 savr^1.500545745 tomato^1.540276242 calgen^0.490575331 tm
    ^0.321191024 pg^0.175662406 food^0.104973154 gene^0.093340475 antisens
    ^0.076225038 genet^0.072076565
```

Example 2.15. *Query 614: LC reformulation*

```
tomato^1.8545771050547137 flavr^1.3672831848530111 savr^1.3672831848530111 tm
    ^0.27476206299682876 food^0.16482417595186108 pg^0.1595979781137451 varieti
    ^0.130127687244464 gene^0.10097707123341029 ripe^0.09906326257718012 plant
    ^0.08991836634393088 request^0.08894338227587754 regul^0.0882567318120294
    agenc^0.08632530767088847
```

Example 2.16. *Query 614: Co reformulation*

TABLE 2.11: Query 614. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
-0.2911	-0.0433	-0.0478	0.0789

TABLE 2.12: Query 614. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
0.3398	0.0487	0.2965	0.292	0.4187

```
flavr^2.0 savr^2.0 tomato^1.1923111694540367 tm^0.6809654083793093 pg
    ^0.3568850349170204 food^0.24460826307502045 varieti^0.20920057512305787 gene
    ^0.18581417852467508 ripe^0.1847189478644446 oils^0.15313450198130424 rape
    ^0.14680108950833834 request^0.14557447483211147 slice^0.14463453477508037
```

Example 2.17. *Query 614: PRM_ W reformulation*

```
tomato^3.0 savr^2.8752917447303084 flavr^2.759452695496538 calgen
    ^0.4587843454740932 us^0.40964639475673825 longer^0.39931330985104924 regul
    ^0.3624193132186865 engin^0.36235505305692123 genet^0.35903460868248505 slice
    ^0.3511052936471268 4 research^0.3438449745499505 uk^0.341107231204365
```

Example 2.18. *Query 614: PRM_ SNT reformulation*

```
tomato^3.0 flavr^2.8637586255035807 savr^2.8637586255035807 calgen
    ^0.8788467304734948 longer^0.6788497015687273 genet^0.6676712265774216 us
    ^0.6669452358951007 varieti^0.6662385804341597 patent^0.6564389374138181
    plant^0.6460880399023902 produc^0.642128625592035 properti^0.6406298078458327
```

The maximal loss of performance of LC in comparison with Bo1 was observed for the query 614. BPREF differences between our systems and Bo1 are presented in the table 2.11. The degradation of results for PRM_W and Co is small while PRM_SNT was slightly better than Bo1. Nevertheless all QE systems outperformed the baseline.

Query 415

Example 2.19. *Query 415: Initial query*

```
<top>

<num> Number: 415
```

```
<title> drugs, Golden Triangle

<desc> Description:
What is known about drug trafficking in the "Golden Triangle", the area where
    Burma, Thailand and Laos meet?

<narr> Narrative:
A relevant document will discuss drug trafficking in the Golden Triangle,
    including organizations that produce or distribute the drugs; international
    efforts to combat the traffic; or the quantities of drugs produced in the
    area.

</top>
```

Example 2.20. *Query 415: Bo1 reformulation*

```
drug^1.457383467 golden^1.365921652 triangl^1.518868975 burma^0.384646574
    thailand^0.324059968 suppress^0.291647683 narcot^0.277702502 lao^0.248346170
    traffick^0.154929325 control^0.124460639
```

Example 2.21. *Query 415: LC reformulation*

```
drug^1.6257241624928187 triangl^1.5779946180764446 golden^1.4853453258726765
    heroin^0.7285890710409573 narcot^0.35263447800121817 burma^0.3476912190324811
     polic^0.29831132336772664 suppress^0.2224464022494208 traffick
    ^0.19851967133627246 lao^0.1890405051825568 opium^0.16593849558960 31 control
    ^0.16542947516109593 kilogram^0.15931734497610617
```

Example 2.22. *Query 415: Co reformulation*

```
triangl^2.0 golden^1.8053091304652558 drug^1.7111123851673247 heroin^1.0 narcot
    ^0.4746024936438949 burma^0.44373537947238806 polic^0.34564562180768593
    suppress^0.2948573124963363 traffick^0.26959189262955674 lao
    ^0.24850134144540723 opium^0.2311055940089396 kilogram^0.2170198137362629 kg
    ^0.1965435981566 5648
```

Example 2.23. *Query 415: PRM_ W reformulation*

```
triangl^3.0 golden^2.8676387871636857 drug^2.8499017592992555 suppress
    ^0.5646747765107035 burma^0.5513050676547326 thailand^0.5339456563240104
    narcot^0.4817358138571512 text^0.47202601592703536 abl^0.4594065149205492 lao
    ^0.45299740327309723 heroin^0.452926970162884 bureau^0.4508098380962942
    cooper^0.4483558490327689
```

Example 2.24. *Query 415: PRM_ SNT reformulation*

```
drug^3.0 triangl^2.9123108845623644 golden^2.9123108845623644 thailand
    ^0.873865826398 8418 burma^0.822045108791076 traffick^0.7996386914969046 lao
    ^0.7983312090759819 suppress^0.7578428435008487 narcot^0.7488810409819456
    cooper^0.7399435023100265 control^0.7130002162979792 china^0.698823121459116
    reach^0.6978977872025081
```

TABLE 2.13: Query 415. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
-0.2077	-0.204	-0.2002	-0.1744

TABLE 2.14: Query 415. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
0.3372	0.1295	0.1332	0.137	0.1628

For the query 415 PRM_SNT and Co showed the worse results with regard to Bo1 however all systems outperformed the baseline.

Query 615

Example 2.25. *Query 615: Initial query*

```
<top>

<title> timber exports Asia

<desc> Description:
What is the extent of U.S. raw timber exports to Asia, and what effect do these
    exports have on the U.S. lumber industry?

<narr> Narrative:
Documents containing information about economic or environmental concerns related
     to the export of timber to Asia are relevant. Documents must specifically
    address exports to Asia, rather than the timber industry in general, to be
    relevant.

</top>
```

Example 2.26. *Query 615: Bo1 reformulation*

```
timber^1.462342379 export^1.000000000 asia^1.090534478 tropic^0.382211087 log
    ^0.281165765 malaysia^0.272611242 lim^0.168755300 pacif^0.138850136 sarawak
    ^0.109591161 yaik^0.087078809 criticis^0.083236502
```

Example 2.27. *Query 615: LC reformulation*

TABLE 2.15: Query 615. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
0.1806	0.1875	0.0625	0.0139

```
timber^1.7558611390146648 export^1.4981250384646643 asia^1.3142567401870537 log
    ^0.7230574633380481 trade^0.2499438098959911 mr^0.2189963609276134 countri
    ^0.21011901301192829 ban^0.19812677392953637 lim^0.1903104871259828 region
    ^0.1763481465797501 land^0.15957269344460828 forest^0.1557339444260987 cent
    ^0.15457372014814452
```

Example 2.28. *Query 615: Co reformulation*

```
timber^2.0 asia^1.816043072566756 export^1.753407171009809 log^1.0 lim
    ^0.28666087342498003 ban^0.22054030460504476 mr^0.2057435141570501 forest
    ^0.18393409431240532 trade^0.18355342973956834 land^0.1534356929747892 region
    ^0.14092121069098026 countri^0.13916404494932494 cent^0.1387508190904011
```

Example 2.29. *Query 615: PRM_ W reformulation*

```
export^3.0 timber^2.7488280676839154 asia^2.628676826044251 log
    ^0.5745147821081881 pacif^0.5544982139230267 ban^0.5497152616113912 tropic
    ^0.5014409075902035 us^0.4746935340367266 zealand^0.47352171814541544 produc
    ^0.4612733510316243 attack^0.45513601184937114 amount^0.45238325353711983
    opportun^0.4501032957639282
```

Example 2.30. *Query 615: PRM_ SNT reformulation*

```
timber^3.0 export^2.9511454560409374 asia^2.8483329451786243 log
    ^0.8989217074383455 tropic^0.8439192408064655 ban^0.7427315148815021 malaysia
    ^0.7009685667080805 us^0.688243678552848 pacif^0.6857148765794292 opportun
    ^0.6783319333546328 criticis^0.6656153955421972 forest^0.66401536656771
    zealand^0.6549351753262446
```

The biggest improvement over Bo1 for Co and LC was observed for the query 615. At the same time both PRM_W and PRM_SNT showed very low amelioration in comparison with Bo1. Bo1 and PRM_SNT demonstrated the degradation relatively the baseline caused by the occurrence of the unrelated named entities (*yaik, zealand, us*).

Query 350

Example 2.31. *Query 350: Initial query*

TABLE 2.16: Query 615. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
-0.0294	0.1512	0.1581	0.0331	-0.0155

```
<top>

<title> Health and Computer Terminals

<desc> Description:
Is it hazardous to the health of individuals to work with computer terminals on a
      daily basis?

<narr> Narrative:
Relevant documents would contain any information that expands on any physical
    disorder/problems that may be associated with the daily working with computer
     terminals. Such things as carpel tunnel, cataracts, and fatigue have  been
    said to be associated, but how widespread are these or
other problems and what is being done to alleviate any health problems.

</top>
```

Example 2.32. *Query 350: Bo1 reformulation*

```
health^1.182267662 comput^1.345704741 termin^1.235950424 vdt^0.849697664 occup
    ^0.205628845 wrist^0.178391443 problem^0.173695037 safeti^0.144124097 adjust
    ^0.127159051 injuri^0.126431477
```

Example 2.33. *Query 350: LC reformulation*

```
comput^1.5396806832532275 health^1.3731939486518805 termin^1.317706395805149
    computer^0.5396806832532276 vdt^0.4931358068520746 occup^0.26405449953261O1
    problem^0.2587744428172878 workstat^0.25257608808317944 injuri
    ^0.23925664890718396 worker^0.20873624359202142 studi^0.2053239876455065
    report^0.19307649487377446 system^0.18447737884340468
```

Example 2.34. *Query 350: Co reformulation*

```
termin^2.0 comput^1.9355127629782112 health^1.8964656710865921 vdt
    ^0.9677605621176558 workstat^0.755780739766757 occup^0.6354251465692401
    injuri^0.5971864378823498 wrist^0.502865164489338 worker^0.44361709524378Ol
    problem^0.4267608698060099 repetit^0.42143292079025313 studi
    ^0.4003210740452786 safeti^0.39022711329011794
```

Example 2.35. *Query 350: PRM_ W reformulation*

TABLE 2.17: Query 350. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
-0.0829	0.1012	-0.2144	-0.0422

TABLE 2.18: Query 350. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
-0.268	-0.2762	-0.3837	-0.2101	-0.2101

```
termin^3.0 comput^2.981399679560267 health^2.973983458634137 board
    ^0.5754612209518796 computer^0.575450023595181 contract^0.5606920378540494
    studi^0.5602512408211342 safeti^0.558373205329485 labor^0.5579233723949151
    begin^0.553400436929554 occup^0.546630142932005 univers^0.546630142932005
    system^0.5446868854698105
```

Example 2.36. *Query 350: PRM_ SNT reformulation*

```
termin^3.0 health^2.952437100061932 comput^2.939825042851446 computer
    ^0.9988279787526647 vdt^0.8609816715536798 occup^0.8460368040706719 safeti
    ^0.8454210928375558 system^0.823037451540038 studi^0.8220783811176574 injuri
    ^0.8214632955290793 worker^0.82066976000047 problem^0.8145880473484853
    editori^0.8140892680021105
```

The greatest loss of performance of PRM_W with regard to Bo1 was shown for the query 350. PRM_W extended the query with low related terms (*begin, univers, labor, contract*). However, Co outperformed Bo1. Co has better weighting than Bo1. Besides highly related terms extracted by Bo1 (*vdt, injury, wrist*), it also added the term (*workstat*). All QE methods outperformed the baseline.

Query 648

Example 2.37. *Query 648: Initial query*

```
<top>
<num> Number: 648
<title> family leave law

<desc> Description:
Identify documents that discuss details of a family leave law, such as
how long, compensation, if any, for what reason allowed, etc.
```

```
<narr> Narrative:
A relevant document must contain some detail about a family leave law
to be relevant.  The mere mention of the existence of such a law
is not relevant.
</top>
```

Example 2.38. *Query 648: Bo1 reformulation*

```
famili^1.147023442 leav^1.397604348 law^1.000000000 sick^0.492591012 employe
    ^0.203405559 opm^0.185752078 630^0.154542791 bereav^0.108872876 care
    ^0.083682448 member^0.061241753 feder^0.052912594
```

Example 2.39. *Query 648: LC reformulation*

```
leav^1.6055537706261886 famili^1.187922430861951 law^1.047706012536678 sick
    ^0.5711239672301337 employe^0.30739375614198844 opm^0.20496191800893915
    recredit^0.14541813134280202 care^0.11810908974812119 regul
    ^0.10108765797263124 member^0.09984135237602529 agenc^0.09111629410766252
    hour^0.0816923523406791 purpos^0.07796466455168101
```

Example 2.40. *Query 648: Co reformulation*

```
famili^2.0 leav^1.992453565057158 law^1.9643928973154874 sick^1.0 opm
    ^0.4726368345315722 employ^0.44821346347040325 recredit^0.207931243982234
    care^0.14638383478402778 bereav^0.14570097583889724 regul^0.13798078952696444
     agenc^0.11624231183407094 member^0.09779034906465875 purpo
    ^0.09599080766179181
```

Example 2.41. *Query 648: PRM_ W reformulation*

```
leav^3.0 famili^2.840387454671628 law^2.775488869209177 employe
    ^0.5755486697144645 friendli^0.5111940623795045 opm^0.4681663656924842 act
    ^0.458640780617361 care^0.451385420280l2113 sick^0.4444307982541633 unpaid
    ^0.4442546035381996 septemb^0.43200309414317295 sign^0.42867510747062526
```

Example 2.42. *Query 648: PRM_ SNT reformulation*

```
leav^3.0 famili^2.6461797160454488 law^2.0 employe^0.7023635894754549 sick
    ^0.6586175668313367 unpaid^0.4952379730775164 care^0.47702495502051967 opm
    ^0.47180421286613256 decemb^0.47170757388744644 republican
    ^0.45775548907829694 purpos^0.45566287912728254 permit^0.4548907652270009
    congress^0.4458222882429474
```

For the query 648 PRM_SNT showed the best performance regarding the baseline and Bo1. PRM_SNT had better scoring of initial terms. It did not extract terms that could biased the retrieval. PRM_W managed to find the terms *act, unpaid* while unrelated

TABLE 2.19: Query 648. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
0.0465	0.0028	0.0271	0.1884

TABLE 2.20: Query 648. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
0.2504	0.2969	0.2532	0.2775	0.4388

terms were quite frequent in the collection (*friendli, septemb, sign*). All systems including Bo1 were much better than the baseline. All our methods outperformed Bo1 that retrieved the terms *630, bereav*.

Query 352

Example 2.43. *Query 352: Initial query*

```
<top>

<title> British Chunnel impact

<desc>
What impact has the Chunnel had on the British economy and/or the life style of
    the British?

<narr>
Documents discussing the following issues are relevant:

 - projected and actual impact on the life styles of the British
 - Long term changes to economic policy and relations
 - major changes to other transportation systems linked with
   the Continent

Documents discussing the following issues are not relevant:

 - expense and construction schedule
 - routine marketing ploys by other channel crossers (i.e.,
   schedule changes, price drops, etc.)

</top>
```

TABLE 2.21: Query 352. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
0.3484	0.3922	0.4115	0.3779	0.379

Example 2.44. *Query 352: Bo1 reformulation*

```
british^1.000000000 chunnel^1.809930819 impact^1.000000000 tunnel^0.381230835
    channel^0.255926117 delai^0.255873041 construct^0.181659485 road^0.128543449
    traffic^0.108740623 plan^0.086204042 govern^0.079678038 east^0.075242792
```

Example 2.45. *Query 352: LC reformulation*

```
chunnel^1.4794908821798096 british^1.245760874435395 impact^1.0 tunnel
    ^0.700405514397l412 project^0.44043491220421876 channel^0.35367665204707777
    rail^0.26599475155819335 sector^0.2405511651352228 govern^0.204463308562132
    delai^0.1833023383792397 infrastructur^0.1722228761315763 risk
    ^0.16514974072791586
```

Example 2.46. *Query 352: Co reformulation*

```
chunnel^2.0 impact^1.3400805441915944 british^1.3287360443990754 tunnel^1.0 capit
    ^0.5301191313581218 channel^0.43782854353346534 project^0.4277597907984974
    minist^0.40356146728132714 rail^0.3373764292625618 todai^0.3005617252712104
    sector^0.18715017352968963 railwai^0.18069679101483843
```

Example 2.47. *Query 352: PRM_ W reformulation*

```
chunnel^3.0 british^2.769237652457146 impact^2.0 road^0.6538843570780691 tunnel
    ^0.6439131186933911 rail^0.601051459183691 channel^0.5857506428307809 delai
    ^0.5749962682856534 railwai^0.5709858742776154 infrastructur
    ^0.5649084742205032 project^0.5633620241941057 union^0.5619990173152198
```

Example 2.48. *Query 352: PRM_ SNT reformulation*

```
british^2.630198293426793 chunnel^2.6168181951505733 impact^2.0 channel^1.0
    tunnel^0.6874198817264021 delai^0.6297710310806067 rail^0.6201234636273717
    road^0.6078232874024265 project^0.6018628859739272 construct
    ^0.599440945315073 risk^0.5976025611441421 come^0.597334351840012 privat
    ^0.5965859767762698
```

The biggest improvement of the results in comparison with the baseline for all systems except PRM_SNT was observed for the query 352.

2.6.2 Analysis of the Individual Queries from the WT10G Collection

Query 484

Example 2.49. *Query 484: Initial query*

```
<top>

<num> Number: 484
<title> auto skoda

<desc> Description:
Skoda is a heavy industrial complex in Czechoslovakia.  Does it manufacture
    vehicles?

<narr> Narrative:
Relevant documents would include references to historic and contemporary
    automobile and truck production.  Non-relevant documents would pertain to
    armament production.

</top>
```

Example 2.50. *Query 484: Bo1 reformulation*

```
auto^1.134921079 skoda^1.198019711 car^0.219091604 brand^0.204727156 www
    ^0.112815536 czech^0.087212292 market^0.053922873 qualiti^0.048133936
    wholesal^0.045159863 automobil^0.038005211
```

Example 2.51. *Query 484: LC reformulation*

```
skoda^1.2062118203984598 auto^1.1747169210302568 car^0.29662123777448524 brand
    ^0.22947686907456857 market^0.10241828623906798 qualiti^0.07575941265767594
    product^0.07061530728710326 compani^0.06344849897175683 consum
    ^0.05421955047505206 wholesal^0.05141658353107324 price^0.05016011328431477
    sell^0.04872492478470938
```

Example 2.52. *Query 484: Co reformulation*

```
skoda^2.0 auto^1.574808413643687 car^0.28386201934452876 brand
    ^0.25780518286127085 market^0.10426219937405447 qualiti^0.0704778263598298
    compani^0.0659341680565217 net^0.06580612112030662 product
    ^0.0646832049706434 servic^0.06054227877898166 wholes^0.05862068608823327
    consum^0.055780072986805596
```

Example 2.53. *Query 484: PRM_ W reformulation*

```
skoda^3.0 auto^2.8454774173950357 consult^0.47953626675675604 brand
    ^0.4391937845834675 gambl^0.4355011848148875 manufactur^0.42880039795477914
    factori^0.42775420114020507 car^0.42519838856533854 compani
    ^0.42235200873822615 right^0.42172850938868434 foreign^0.4201507675295606
    number^0.42003183883027223
```

TABLE 2.22: Query 484. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
-0.1943	-0.0168	-0.0405	-0.0227	-0.0464

Example 2.54. *Query 484: PRM_ SNT reformulation*

```
skoda^2.996189516028906 auto^2.917094133665903 car^1.0 brand^0.9215989122137519
    manufactur^0.9176378894584417 factori^0.8987681819511917 bui
    ^0.8800779725859493 market^0.8597358630965145 sell^0.8540436371830348 foreign
    ^0.8537856662781781 qualiti^0.8429688142767879 price^0.8393919080308583
```

The maximal degradation of Bo1 with regard to the baseline was detected for the query 484. Terms of the marketing area prevail in the expanded query. Meanwhile according to the narrative *Relevant documents would include references to historic and contemporary automobile and truck production.* Non-relevant documents would pertain to armament production. At the same time the loss of all our systems is very small. The worse QE terms added by Bo1 are *www* and *czech*. The former term could be filtered out as a stop-word while the latter is misguiding since the information need is related only to Skoda's vehicles.

Query 538

Example 2.55. *Query 538: Initial query*

```
<top>

<title> fha

<desc> Description:
Find documents describing the Federal Housing Administration (FHA): when and why
    it was originally established and its current mission.

<narr> Narrative:
A relevant document will discuss the history and current purpose of the Federal
    Housing Administration (FHA).

</top>
```

Example 2.56. *Query 538: Bo1 reformulation*

TABLE 2.23: Query 538. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
-0.1429	-0.3929	-0.3929	-0.3929	-0.1429

```
fha^1.509078297 hud^0.583076579 mip^0.455518723 mmi^0.441275409 mortgag
    ^0.150132341 loan^0.050178803 urban^0.048426529 payment^0.047509347 opportun
    ^0.035254941 buyer^0.030610153
```

Example 2.57. *Query 538: LC reformulation*

```
fha^2.093028970996946 mip^0.508900991091095 famili^0.2737457957176758 mortgag
    ^0.23960784566167678 home^0.22858338181359678 parti^0.20971879593004053
    institut^0.1401866541323876 loan^0.13576769844749792 resid
    ^0.1109031401507361 4 payment^0.10009150306471741 requir^0.09135036919777205
```

Example 2.58. *Query 538: Co reformulation*

```
fha^2.0 mip^0.5167756408007965 famili^0.16298338755281258 mortgag
    ^0.15415360119689495 parti^0.1477208130593235 institut^0.08953514795990722
    loan^0.08003658543845979 home^0.06669714985197214 resid^0.06655176400161641
    payment^0.06069365218666244 chapter^0.05514291689265208
```

Example 2.59. *Query 538: PRM_ W reformulation*

```
fha^3.0 hud^0.588422044687347 4 parti^0.5794433765290039 reimburs
    ^0.5451166119418531 payment^0.5434156310137882 individu^0.5399731525822031
    resid^0.5353836345032943 rate^0.5331116766234237 follow^0.5265599998105929
    loan^0.5256320013055341 mortgag^0.5253138978862039
```

Example 2.60. *Query 538: PRM_ SNT reformulation*

```
fha^2.980835334492952 institut^1.0 famili^0.9658503903182041 approv
    ^0.9478100156099346 chapter^0.9403837847000324 resid^0.9220474729854227 home
    ^0.917866130630144 2 note^0.9172022963882673 limit^0.9123441564350102 parti
    ^0.9115950323971022
```

LC, Co and PRM_W showed the maximal degradation of results for the query 538. The loss of PRM_SNT is equal to the one of Bo1 and it is much lower but still significant. This query has only 2 relevant documents and therefore the performance measure is very sensitive to small changes. We believe that such queries are not suitable to the statistical QE.

Query 531

Example 2.61. *Query 531: Initial query*

```
<top>

<title> who and whom

<desc> Description:
What is the proper grammatical use of "who" versus "whom"?.

<narr> Narrative:
A relevant document will provide explicit guidance for the proper grammatical use
      of "who" and "whom".

</top>
```

Example 2.62. *Query 531: Bo1 reformulation*

```
who^1.086666041 whom^1.311822963 claus^0.280323866 parenthet^0.136861688 writer
    ^0.096387078 word^0.091244602 rel^0.087537676 passiv^0.083937488 sentenc
    ^0.071545552 correct^0.066263927
```

Example 2.63. *Query 531: LC reformulation*

```
whom^1.6117039991780309 who^1.394314925584469 claus^0.5108908894850113 rel
    ^0.3239380543201977 pronoun^0.21749056089386187 word^0.21364523509743016
    writer^0.16555170250240686 parenthet^0.14409431416459167 sentenc
    ^0.12690852666827193 passiv^0.11284540537999693 merci^0.11018076054529444
    think^0.0960023584365 4759
```

Example 2.64. *Query 531: Co reformulation*

```
whom^1.695793263717705 who^1.5710726412735085 clau^1.0 rel^0.5091877867863553
    pronoun^0.4443263548223316 word^0.3144300215463974 writer^0.26063866337150 31
    sentenc^0.21608431560192923 parenthet^0.20826949834404893 passiv
    ^0.1972532982685857 chapter^0.14505928607013988 object^0.13219145858858306
```

Example 2.65. *Query 531: PRM_ W reformulation*

```
whom^3.0 who^1.0 correct^0.8213046550888013 object^0.6249177681764913 fault
    ^0.554712965733374 9 take^0.529420530797749 meet^0.4853862796414344 rel
    ^0.4853862796414344 new^0.4853862796414344 leav^0.4853862796414344 time
    ^0.4853862796414344 refer^0.4853862796414344
```

Example 2.66. *Query 531: PRM_ SNT reformulation*

```
whom^3.0 who^1.0 correct^0.9139736716410004 object^0.8450841384384611 fault
    ^0.7919386690275101 book^0.7748820122923828 claus^0.7472274897628195 subject
    ^0.7397070237070543 take^0.7281254580862134 case^0.7211867052592538 sentenc
    ^0.7204142018361758 word^0.713956985251095
```

TABLE 2.24: Query 531. BPREF differences with the baseline

Bo1	**LC**	**Co**	**PRM_W**	**PRM_SNT**
-0.1098	0.4563	0.3323	0.1919	0.5472

TABLE 2.25: Query 531. BPREF differences with Bo1

LC	**Co**	**PRM_W**	**PRM_SNT**
0.5661	0.4421	0.3017	0.657

All our methods significantly improved the results of the baseline for the query 531 while Bo1 showed much worse results. The difference between the performance of our systems and Bo1 is maximal for this query. Moreover, PRM_SNT demonstrated the maximal enhancement in comparison with the baseline. PRM_SNT managed to retrieved the terms that answer user's information need, namely *object* and *subject* with very high weights. It did not add semantically unrelated terms. Moreover, other grammatical terms were added (*claus, sentenc*). LC and Co eliminated unrelevant documents by the term *pronoun*. Meanwhile Bo1 was misled by the terms related to other grammatical subjects.

Query 504

Example 2.67. *Query 504: Initial query*

```
<top>

<num> Number: 504
<title> information about what manatees eat

<desc> Description:
Find documents that describe the diet of the manatee.

<narr> Narrative:
Relevant documents will identify any foods providing sustenance to the manatees.

</top>
```

Example 2.68. *Query 504: Bo1 reformulation*

```
manate^1.678649351 eat^1.000000000 florida^0.082611641 speci^0.079335464 lake
    ^0.053153509 mammal^0.047307046 protect^0.046647485 endang^0.046265880 sea
    ^0.043138175 water^0.042586603 dugong^0.041717726
```

Example 2.69. *Query 504: LC reformulation*

```
manate^2.000732492437405 eat^1.0751707817851617 food^0.28225906410292184
    enviroworld^0.23182821288267977 return^0.23127228574788042 todai
    ^0.19859271861563949 state^0.19376933377825226 headlin^0.19004337167527463
    water^0.18982752143573997 anim^0.17739405893651003 speci^0.16067988972857047
    year^0.15382086750987717
```

Example 2.70. *Query 504: Co reformulation*

```
manate^2.0 eat^1.6390286208225362 enviroworld^1.0 food^0.6999874856115996 headlin
    ^0.6314754530160908 speci^0.46086537356453017 return^0.45897479532525176
    water^0.4424809990332262 anim^0.43910373653057216 state^0.4283639350804834
    cleanup^0.4278497298641111 todai^0.4123290679705141
```

Example 2.71. *Query 504: PRM_ W reformulation*

```
manate^3.0 eat^2.548077160615667 marin^0.4488988591375219 speci
    ^0.4468646513345947 4 sea^0.42993848341467417 florida^0.4195532150247278
    concentr^0.4159019687085566 food^0.414362673721884 fish^0.4137604470558857
    protect^0.412694913009162 sanctuari^0.407581139880445 popul
    ^0.4072890763742271
```

Example 2.72. *Query 504: PRM_ SNT reformulation*

```
manate^3.0 eat^2.0 summer^0.8306548752236135 protect^0.7954629265912893 save
    ^0.7835529851219353 bai^0.7824028958814484 river^0.7795401810429052 thompson
    ^0.7718377221055347 committe^0.768726714722209 todai^0.7683753750525516
    headlin^0.768311662006025 4 cover^0.7666446770181162
```

The query 504 turned out to have the maximal loss for the PRM_SNT. Meanwhile all systems, except PRM_W have the decreased performance with regard to the baseline. Although all systems retrieved semantically related terms they biased the query. Bo1 and LC assigned low weights to the QE terms and thus their loss was small. PRM_SNT retrieved the named entity *thompson*. It also added the term *headlin*. LC, Co and PRM_SNT retrieved terms such as *year, todai* that could be considered as stop words but they were not filtered out. Moreover, Co assigned the maximal score to *enviroworld* which is a rare term and therefore could affect the retrieval a lot.

TABLE 2.26: Query 504. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
-0.0603	-0.0695	-0.1683	0.0261	-0.1806

Query 486

Example 2.73. *Query 486: Initial query*

```
<top>

<num> Number: 486
<title> where is the Eldorado Casino in Reno ?

<desc> Description:
The Eldorado (El Dorado) Casino is reportedly  located in Reno.  Is this so and
    what is the address?

<narr> Narrative:
A relevant document will provide the street address of an Eldorado or El Dorado
    Casino in Reno, Nevada.

</top>
```

Example 2.74. *Query 486: Bo1 reformulation*

```
eldorado^1.369900210 casino^1.401197585 reno^1.183920315 renonv^0.367551658
    buffet^0.308709154 arcad^0.124326666 opinion^0.098698595 restaur^0.087094100
    c7^0.084868369 chef^0.078490466
```

Example 2.75. *Query 486: LC reformulation*

```
casino^1.7040931286943775 reno^1.5499818447720832 eldorado^1.336238276944807
    legaci^0.41728247693749915 silver^0.35662886016446577 hotel^0.291958921555965
     buffet^0.26278658427984963 opinion^0.14982098844457573 restaur
    ^0.13739312709788767 resort^0.12494340412262318 downtown^0.12039526191406814
    room^0.111752373344056 arcad^0.10636328954248118
```

Example 2.76. *Query 486: Co reformulation*

```
eldorado^2.0 reno^1.7653641629781993 casino^1.7422199681652952 legaci^1.0 silver
    ^0.9088459743624621 buffet^0.8110233875826147 hotel^0.6813407139174554
    restaur^0.36094388127150 78 opinion^0.2964450410998601 downtown
    ^0.28562769052380677 arcad^0.25651446445352855 room^0.23516388908933836 ski
    ^0.2121061271836777
```

Example 2.77. *Query 486: PRM_ W reformulation*

TABLE 2.27: Query 486. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
-0.3125	-0.3125	-0.25	-0.125

```
casino^3.0 eldorado^2.9260864763630132 reno^2.735470207214934 hotel
    ^0.6651033402567262 legaci^0.5820929792747783 silver^0.5365600816584596
    downtown^0.532804721011736 resort^0.5181736929302136 virginia
    ^0.5170906132281008 buffet^0.5162574857035129 restaur^0.496346647576537O7
    circu^0.4919809972768258
```

Example 2.78. *Query 486: PRM_ SNT reformulation*

```
casino^3.0 eldorado^2.8644906284088267 reno^2.820350600239473 legaci
    ^0.7062234036485537 hotel^0.694889053394174 tivoli^0.6690449679928323 silver
    ^0.6422613262523766 downtown^0.5878784929401452 renonv^0.5799162796592143
    buffet^0.5779234995704986 visit^0.5534196296493733 virginia
    ^0.5482986450656002 guid^0.5363020817856137
```

LC and Co indicated the worse performance regarding Bo1 for the query 486. However, all QE systems under consideration decreased the baseline results. There are only 4 relevant documents. Therefore small changes in ranking influence a lot the measurement. Bo1 filtered out some terms and kept only 7 additional words while our systems always have 10 QE terms. All our systems assigned very high scores to the terms *legaci, silver* and *hotel* that mislead the retrieval process.

Query 529

Example 2.79. *Query 529: Initial query*

```
<top>

<num> Number: 529
<title> history on cambodia?

<desc> Description:
Find accounts of the history of Cambodia.

<narr> Narrative:
A relevant document will provide historical information on Cambodia. Current
    events in Cambodia are not relevant.

</top>
```

Example 2.80. *Query 529: Bo1 reformulation*

```
histori^1.000000000 cambodia^1.566129497 vh^0.338465032 col^0.176080284 min
    ^0.161106293 vietnam^0.151557626 kampuchea^0.117285437 khmer^0.109662267 lao
    ^0.077478548 roug^0.053250146 war^0.047183339
```

Example 2.81. *Query 529: LC reformulation*

```
cambodia^1.8320547649784795 histori^1.1113324062363075 min^0.2398842568449682 war
    ^0.1108627439327093 refuge^0.10520048431239497 countri^0.10332231949804299
    cultur^0.09291870405461788 peopl^0.07263328268206477 polit
    ^0.06781655142472254 televis^0.06289443295187876 border^0.06134659767167929
    year^0.05657026775371622
```

Example 2.82. *Query 529: Co reformulation*

```
cambodia^2.0 histori^1.6339199630107832 cambodium^1.0 min^0.536680688958293 refug
    ^0.20625458777899092 war^0.16441831933703407 cultur^0.12746763889695664
    countri^0.12385684302316535 border^0.09954724785804847 televi
    ^0.09659409325967123 regim^0.08458863253631985 peopl^0.08242641786376183
    holocaust^0.07749687292088836
```

Example 2.83. *Query 529: PRM_ W reformulation*

```
cambodia^3.0 histori^2.6871713317618893 cambodium^1.0 lao^0.6538846190949142 min
    ^0.6327451049652792 border^0.5416575239363208 cultur^0.5324673343962415 vh
    ^0.5310665464394084 recent^0.5090035682305452 televis^0.5044644634057054
    vietnam^0.5034653555049893 asia^0.5022944293955275
```

Example 2.84. *Query 529: PRM_ SNT reformulation*

```
cambodia^3.0 histori^2.9631335189753343 cambodium^1.0 vietnam^0.9913509103791804
    war^0.9792401630685783 vh^0.9749315105314222 min^0.9738235311034161 lao
    ^0.9652445186899827 televis^0.954798126577348 border^0.9528265966709957
    cambodian^0.9451979835867599 countri^0.9343421861370426
```

The biggest loss of PRM_W was observed for the query 529. For this query all our systems were excelled by Bo1. All our systems except PRM_W slightly ameliorated the baseline results while the improvement made by Bo1 is significant. Bo1 managed to extract such terms as *kampuchea, khmer* and *roug*. Kampuchea was the name of the Khmer Rouge – controlled state that existed in present-day Cambodia. These terms are quite rare and therefore they can only appear in relevant documents since they are related to the history of Cambodia. Both PRM models have named entities such as *lao, vietnam* and *asia* that mislead the retrieval.

TABLE 2.28: Query 529. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
-0.0953	-0.0966	-0.2906	-0.1361

Query 548

Example 2.85. *Query 548: Initial query*

```
<top>

<num> Number: 548
<title> how do you use solar heat to heat a pool?

<desc> Description:
What are the methods of using solar heat to warm up the water in a swimming pool?

<narr> Narrative:
A relevant document will explain a technique or method for warming the water in a
     swimming pool using heat from the sun. General discussions of solar heating
    are not relevant; the document must describe its application to swimming
    pools

</top>
```

Example 2.86. *Query 548: Bo1 reformulation*

```
solar^1.359279048 heat^1.368377094 pool^1.140070260 water^0.226359254 collector
    ^0.214837769 temperatur^0.147179914 spa^0.120612578 system^0.099402695 energi
    ^0.096536374 hot^0.069348047
```

Example 2.87. *Query 548: LC reformulation*

```
heat^1.7919285912836798 solar^1.6837427505856992 pool^1.4750140908456943 water
    ^0.647094476301776 collector^0.4294396611699935 system^0.4134369123977827
    temperatur^0.351186617191877 spa^0.31068791400403156 energi
    ^0.2758177441616125 cost^0.2197935752504485 heater^0.189687740255769 pump
    ^0.1837192666329885 facil^0.15449132387707737
```

Example 2.88. *Query 548: Co reformulation*

```
solar^2.0 pool^1.8819275897096996 heat^1.875458486260376 water^1.0 collector
    ^0.8552177592084321 temperatur^0.6158649335946544 spa^0.5989406980223102
    system^0.5849826157319304 energi^0.42380713564467065 heater
    ^0.3683750364395994 cost^0.32657750888003284 pump^0.3183330942868988 valv
    ^0.24813017774014062
```

TABLE 2.29: Query 548. BPREF differences with Bo1

LC	Co	PRM_W	PRM_SNT
0	0	0	-0.5

TABLE 2.30: Query 548. BPREF differences with the baseline

Bo1	LC	Co	PRM_W	PRM_SNT
0.5	0.5	0.5	0.5	0

Example 2.89. *Query 548: PRM_ W reformulation*

```
pool^3.0 heat^2.9376340972506427 solar^2.9022659947261045 swim
    ^0.48178081127476674 equip^0.4568201472305853 energi^0.426243987407158 13 spa
    ^0.4238982885777929 collector^0.3919264334658048 water^0.3859818677384903
    pump^0.37010952576832057 technic^0.36890951990730675 brochur
    ^0.3608806852371273 temperatur^0.3574434880023987
```

Example 2.90. *Query 548: PRM_ SNT reformulation*

```
pool^2.6460078602410277 heat^2.5235739197487588 solar^2.4617145708953743 summer
    ^1.0 collector^0.27803447585709 73 energi^0.2768621632943583 swim
    ^0.2395773271677297 water^0.2394268529443586 pump^0.23843574609620133 filter
    ^0.2246802281108941 spa^0.21981745691676752 temperatur^0.2186067144049193 ga
    ^0.20875971388969033
```

All systems except PRM_SNT have the maximal improvement for the query 548. However, our best system PRM_SNT have not enhanced the results of the baseline. The query 548 has only 2 relevant documents. The relevant string of PRM_SNT is '1001000000' while for other systems it is '1100000000'. Apparently, the distortion term is *summer* which has the highest weight after the query terms. Although this term is strongly semantically related to the query terms, it is quite broad and having a high weight it can lead astray the retrieval.

2.6.3 Types of initial queries

Types of initial queries play an essential role in the prediction of successful information retrieval. As usual initial queries include, besides articles and other grammar words, nouns and entities, sometimes attributes and verbs. Grammatical structure of a title does

not influence on the information retrieval process, because every title while processing the query is ruined into words and simple word's chunks. So types of queries are limited by a number of words and topic. Types of the query terms are restricted to words grammatical classes, such as parts of speech, and words semantic classes, such as terminology, entities, peculiarities, etc.

Potentially a document matches the initial query thanks to one term, or one term with its attribute, or two (or more) different terms. The last possibility is the best one, since a number of documents with two (or more) unconnected terms from the initial query is less, than a number of documents with the term and its attribute (noun phrase). In other words, co-occurrence of two (or more) semantically unconnected query terms in a document guarantees more accurate matching the initial query, while occurrence of one term just presupposes matching in a topic. Thus, one term query is less informative, than two and more terms queries. Hence for a one-word initial query QE is a productive way to increase the relevance of results, however, it depends on semantics of the one-word query. Our results for QE for one-word queries are slightly better regardless of the QE methods.

As a basis for the type of the query, we consider (1) the number of words and (2) the topic (theme). The number of words has been mentioned above. The topics (themes) of the initial queries in our two collections concern more naive (Animals, Culture) and more rigid categories (Technologies) as well. The structure of the naive categories differs from the rigid one as diffusive, ambiguous, associative [Frumkina and Telia, 1984, Rosch, 1978]. Moreover, the structure of the categories reflects in texts' word diversity and distribution. Thus, the initial query in the field of a naive category provokes as a result texts with different associative connections to the topic. Associative connections are stimulated by similarity, contiguity, frequency and contrast as well, and all of them are represented in texts devoted to the naive category topics.

As a consequence of the diffusive character of the category, there are a lot of different factors which influence on document frequency of the words associated with the topic. Thus, sometimes the more texts we use for the QE in the global analysis, the more unpredictable candidates we get for the QE.

The structure of scientific categories is more compact and hierarchical. We assume that the initial query in the field of scientific categories evokes texts with less associative and

TABLE 2.31: Query 317. BPREF differences with the baseline

Bo1	**LC**	**Co**	**PRM_W**	**PRM_SNT**
0.0146	0.0299	0.1728	0.035	0.0146

more logical connections. The QE allows directing the IR process in a narrow relevant field. The title *Unsolicited Faxes* (Robust) refers to a multi-topic document, which simultaneously belongs to at least two topics in our set ("crimes" and "technology"). The results of QE performed by all systems are very good, but for our systems (except PRM_SNT) they are better (see Table 2.31).

Robust collection is more homogeneous than WT10G. The QE results for the former collection are better. From our point of view, the reason of the relevance of an initial query, as well as an expanded one, is the similarity of texts and transparent categorization, such us "culture", "technologies", "crimes", "health", etc. Meanwhile, our QE system has an advantage when applied to queries within "technology" topic in both text collections, even if "technology" co-exists with another theme.

Therefore, the topic (theme) of the initial query is a strong factor, which influence on the necessity of the QE. Within homogenous text collection, every QE system works good, producing better results, than an initial query. Within naive topics (theme) categories the simple QE system is appropriate, while our QE generates complicated associative queries. So for the IR on the topic from naive category within heterogeneous text collection our QE system is overcomplicated, and that is why it works worse.

As already have been mentioned, types of the query terms are restricted to grammatical and semantic classes of words. We are taking into consideration such semantic classes as entities, terminology, peculiarities, etc. For deep analysis we choose the extraordinary cases: the best improvement, the failure, and the problematic ones for each class from the WEB collection as more relevant to the "natural" IR.

Diversity of words and their distribution in the set of documents relevant to the query with entities and names is also more limited in comparison with words diversity and their distribution in a set of documents relevant to a query without entities and names within topics belonging to naive categories [Dalton and Dietz, 2013]. It is obvious that initial queries with entities and names work well enough, because a set of relevant documents is

restricted to significantly limited topics. The representations of this tendency depend on the entity field: culture, science, business, etc., where sometimes an entity and a name is not enough to determine the issue because of multi-topic character of documents or for another reason. To be more correct in the analysis, we choose all examples from the WEB collection, which is more relevant to the "natural" IR.

Documents are not assumed to belong to single topics, but to several topics simultaneously. In a case with entities and names multi-topical documents are processed successfully thanks to an unambiguity of the query term, its narrow or even unique reference. Thus, for the queries including entities and names, the QE often works well because of decreasing ranks of multi-topical documents.

To conclude, the entities and names restrict the topic (theme) of the documents, but it does not work as simple mechanic restriction. The limits change under influence of different factors, such as lexical and grammar polysemy and the category's structure.

Terminology is special semantic class in the query terms, which increase relevance for initial query and of results with our QE system. Usage of terminology in a specific field or types of texts decreases a set of potential associations, and hence the more specified query is applied, the more relevant result is received. Probably, thanks to terminology in technological texts from our collections, our QE system for texts on the technological topics always provide more relevant results, than other QE systems.

Thus, even queries with terminology demonstrate differences, connected with a field. As usual our system is slightly better in processing queries with terminology, but in cases within less rigid categories.

Attributes as nominations of a term (object) peculiarities, generic or specific ones, do not produce clear effect on the QE system. Probably, a specific peculiarity is strongly connected with a set of different objects, which include the peculiarity and thanks to it are cross-associated with each other. Thus, the more precise and accurate is the QE system, the more irrelevant documents with the description of the peculiarity will appear in results.

To conclude the discussion of the query types and the types of the query terms, it is important to stress, that the topic of query, the character of the category and presence

of entities, names and terminology play key role in the choice of QE system, if one is needed.

2.7 Conclusion

QE is a powerful technique in IR, though the terms that are added can bias a query and thus decrease both recall and precision.

In this research we proposed three methods for QE.

The first method (LC) is based on co-occurrence measure as well as importance estimated by analyzing local context. In contrast to previous works we treated not only entire documents, but also text passages surrounding query terms. The method was published at CLEF-2015 [Ermakova, 2015].

The second method for QE we call Co combines local analysis, namely relevance feedback, and global analysis of texts. The key idea of the proposed method is to estimate the importance of candidate terms by the strength of their relation to the query terms. In our approach, documents from RF provide term candidates that are analyzed in two aspects: their frequency in RF and their co-occurrence with query terms in the whole collection. This approach was published at the international conference on computational linguistics Dialog-2014 [Ermakova et al., 2014].

In the third approach for query expansion (PRM) we proposed incorporating term proximity information into the LM formalism. The method is based on PRF, but it differs from previous researches in several ways:

- it is formalized within LM;
- the term proximity is captured directly, and not by weighting term positions;
- the distance is computed in terms of sentences from the query terms and its combinations.

The paper devoted to the PRM method was accepted at SAC-2016 [Ermakova et al., 2016].

We evaluated our methods on two international benchmark collections: TREC Robust and WT10G. Our systems demonstrated the best results among the state-of-the-art QE method implemented in such search engines as Terrier and Lemur according to all metrics for both data collections.

LC method ameliorated the highest number of queries (all, difficult and very difficult) relative to Bo1 and it had the minimal rate of the result degradation on the Robust collection. On WT10G it showed the best improvement for difficult queries with regard to the baseline (44) and Bo1 (38) while keeping the lowest degradation rate for this type queries (23 and 28 respectively). It also has the highest improvement for very difficult queries in comparison to the baseline for both collections.

Co demonstrated the highest degradation of the results with regard to the baseline for both test collections (Robust - 87, WT10G - 38). However, the average results are better than LC. In our future work, we will work on the relationship between the types of queries and the field associated to the query in order to detect correlation with these features and the best method to treat the query. We will clarify the results with the help of clustering and evaluate them by ANOVA. From our analysis, we can conclude that QE systems need to be specified according to the peculiarities of the initial queries. We think that using more linguistic features can help in selective approaches in IR.

Experiment results showed that the proposed methods are significantly better than other PRF-based QE approaches (DFR QE models, RM3) as well as the baseline. Major improvement was observed for difficult and very difficult queries. Our method has lower results than the DFR models mainly for queries that should not be expanded.

For both test collections, our Proximity Relevance Model grounded on sentence-level distance estimation outperformed the word-based one. This fact allows concluding that distance measuring in terms of sentences is more preferable than in terms of tokens.

One of the most promising directions of further research is to differentiate the probability distribution of the distances depending on query terms.

We also showed that there is no a single method that can treat homogeneously the all set of topics and that there is no clear correlation between the topic difficulty and the method to use. However it was clear that the different methods have benefits since each of them treats the best a high number of topics.

Our methods fail when all the statistical approaches under consideration also fail. We compared several statistical methods grounded on the different hypotheses: DFR models based on the divergence of word frequencies in the elite set and the rest of collection, LC and PRM models that take into account the proximity to the query terms, Co that considers the strength of their relation to the query terms. The fail of all these methods for some queries allows drawing a conclusion that statistical approaches are not suitable for these queries. For the Robust collection approximately 50% of the queries our systems demonstrated worse results than the baseline was decreased by all QE methods under consideration. For WT10G this percentage is almost 70%. Therefore, we believe that the proposed approaches may be significantly improved by selective QE.

Previous approaches consider selective QE: the system decides whether or not QE should applied, based on some query features [Cronen-Townsend and Croft, 2002b]. The query features are either pre-retrieval or post-retrieval query features that are used to cluster queries. A training phase builds the model from queries for which the best decision is know; then the model is applied to any new query. We believe that the analysis we made is linguistically motivated and therefore it is more portable to other collections.

Finally our finer analysis shows that the type of initial query can have an influence on the success of QE. We specifically detected various cases in which QE provokes shift on topic. We assume that the initial query in the field of scientific categories evokes texts with less associative and more logical connections and thus lead to better results using QE. Initial queries with entities and names work well enough, because a set of relevant documents is restricted to significantly limited topics, even if the representations of this tendency depend on the entity field: culture, science, business, etc.

Conclusion

The principle of the least effort leads to the extreme compression of texts especially in electronic communication, e.g. microblogs, SMS, search queries making them hard to understand for a user as well as for a system, e.g. search engine. Therefore, in this research we presented methods for contextualization of short texts based on local context analysis that were applied for automatic summarization, document re-ranking and query expansion.

The first contribution we made is an approach to tweet contextualization from an external source based on query-biased summarization. Our approach implies sentence retrieval and re-ordering. Sentence retrieval is based on NE recognition, POS weighting and sentence quality measuring. We introduced an algorithm of smoothing from the local context. We also integrated the knowledge of topic-comment structure into the sentence retrieval model. Moreover, we developed a graph-based algorithm for sentence re-ordering. The method has been evaluated at INEX/CLEF TC track. We obtained the best results in 2011 according to informative evaluation. In 2013 according to informative evaluation our system was ranked first (PRIOR and POOL) and second (ALL) over all automatic systems that participated. At the same time in terms of readability it was the best among all participants according to all metrics except redundancy. Run comparison showed that smoothing improves informativeness. Another conclusion is that ranking is sensitive to the pool selection as well as to the choice of divergence. Despite the topic-comment analysis did not improve results, we believe that small changes in implementation may produce positive effect on the system performance. In 2014 the worst results among our runs were shown by the run based on entity restriction that could be explained by the loss of the recall.

Although sentence ordering was not evaluated at INEX campaign, we believe that it is crucial for readability. Thus, the our second contribution is two algorithm for sentence re-ordering based on the graph representation of text.

The sentence retrieval method was also adapted to snippet retrieval and QE which is the third contribution. In 2013 our system showed the best results in the INEX Snippet Retrieval Track.

As the fourth contribution, we introduced an automatic topic-comment annotation method based on the topic fronting assumption that requires only shallow parsing, namely sentence chunking and POS tagging. We propose to split a sentence into two parts by a personal verb and we embedded the topic-comment structure into BM25F retrieval model. According to all used evaluation measures for two TREC data sets, our method significantly outperformed the strong baseline provided by the Terrier platform.

The last three contributions are related to query expansion. Query expansion is a powerful technique in IR, though the terms that are added can bias a query and thus decrease both recall and precision.

We propose three methods for QE. The first method LC exploits the analysis of the local context. In contrast to previous works we treated not only entire documents, but also text passages surrounding query terms.

The second method Co is based on the global analysis of texts. The key idea of the proposed method is to estimate the importance of candidate terms by the strength of their relation to the query terms. In our approach, documents from RF provide term candidates that are analyzed in two aspects: their frequency in RF and their co-occurrence with query terms in the whole collection.

In the third approach for query expansion PRM we proposed incorporating term proximity information into the LM formalism.

We evaluated our methods on two international benchmark collections: TREC Robust and WT10G. Our systems demonstrated the best results among the state-of-the-art QE methods implemented in such search engines as Terrier and Lemur according to all metrics for both data collections.

Experiment results showed that the proposed methods is significantly better than other PRF-based QE approaches (DFR QE models, RM3) as well as the baseline. Major improvement was observed for difficult and very difficult queries. Our method has lower results than the DFR models mainly for queries that should not be expanded.

For both test collections, our Proximity Relevance Model grounded on sentence-level distance estimation outperformed the word-based one. This fact allows concluding that distance measuring in terms of sentences is more preferable than in terms of tokens.

One of the most promising directions of further research is to differentiate the probability distribution of the distances depending on query terms.

We also showed that there is no a single method that can treat homogeneously the all set of topics and that there is no clear correlation between the topic difficulty and the method to use. However it was clear that the different methods have benefits since each of them treats the best a high number of topics.

Our methods fail when all the statistical approaches under consideration also fail. We compared several statistical methods grounded on the different hypotheses: DFR models based on the divergence of word frequencies in the elite set and the rest of collection, LC and PRM models that take into account the proximity to the query terms, Co that considers the strength of their relation to the query terms. The fail of all these methods for some queries allows drawing a conclusion that statistical approaches are not suitable for these queries. For the Robust collection approximately 50% of the queries our systems demonstrated worse results than the baseline was decreased by all QE methods under consideration. For WT10G this percentage is almost 70%. Therefore, we believe that the proposed approaches may be significantly improved by selective QE.

Previous approaches consider selective QE: the system decides whether or not QE should applied, based on some query features [Cronen-Townsend and Croft, 2002b]. The query features are either pre-retrieval or post-retrieval query features that are used to cluster queries. A training phase builds the model from queries for which the best decision is know; then the model is applied to any new query. We believe that the analysis we made is linguistically motivated and therefore it is more portable to other collections.

Finally our finer analysis shows that the type of initial query can have an influence on the success of QE. We specifically detected various cases in which QE provokes shift on

topic. We assume that the initial query in the field of scientific categories evokes texts with less associative and more logical connections and thus lead to better results using QE. Initial queries with entities and names work well enough, because a set of relevant documents is restricted to significantly limited topics, even if the representations of this tendency depend on the entity field: culture, science, business, etc.

2.1 Publications of the Author on Short Text Contextualization

Liana Ermakova. A method for short message contextualization: Experiments at CLEF/INEX (best paper award). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 352–363, 2015. doi: 10.1007/978-3-319-24027-5_38. URL http://dx.doi.org/10.1007/978-3-319-24027-5_38.

Liana Ermakova and Nicolas Faessel. Création de snippets : une application de la génération automatique de résumés (regular paper). In Patrice Bellot, Josiane Mothe, Éric SanJuan, Ludovic Tanguy, and , editors, *Atelier Contextualisation de Messages Courts (EGC), Toulouse, France, 29/01/2013*, Revue des Nouvelles Technologies de l'Information, pages 27–36, http://www.cepadues.com, janvier 2013. Cépaduès. URL http://www.irit.fr/publis/SIG/2013_EGC_EF.pdf-http://oatao.univ-toulouse.fr/12382/.

Liana Ermakova and Josiane Mothe. Irit at inex: Question answering task. In Shlomo Geva, Jaap Kamps, and Ralf Schenkel, editors, *Focused Retrieval of Content and Structure*, volume 7424 of *Lecture Notes in Computer Science*, pages 219–226. Springer Berlin Heidelberg, 2012a. ISBN 978-3-642-35733-6. doi: 10.1007/978-3-642-35734-3_19. URL http://dx.doi.org/10.1007/978-3-642-35734-3_19.

Liana Ermakova and Josiane Mothe. Irit at inex 2012: Tweet contextualization. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), Rome, Italy, 17/09/2012-20/09/2012*, page (on line). Univesity La Sapienza (Rome, Italy), 2012b. URL http%3A%2F%2Fwww.clef-initiative.eu%2Fdocuments%2F71612%2F3e9ecc64-fae6-4af3-93fd-1a6a6fabb5d6.

Liana Ermakova and Josiane Mothe. Irit at inex 2013: Tweet contextualization track. In and, editor, *INitiative for the Evaluation of XML Retrieval (INEX), Valencia, Spain, 23/09/2013-26/09/2013*, page (on line), http://www.upv.es, 2013. Polytechnic University of Valencia. URL http://www.clef-initiative.eu/documents/71612/58a64b0a-cf0c-4751-a91f-9c8aba4312e1-http://oatao.univ-toulouse.fr/12739/.

Liana Ermakova and Josiane Mothe. Irit at inex 2014: Tweet contextualization track. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), Sheffield, UK, 15/09/2014-18/09/2014*, page (on line), http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS, 2014. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-ErmakovaEt2014.pdf-http://oatao.univ-toulouse.fr/13277/.

Liana Ermakova, Josiane Mothe, and Irina Ovchinnikova. Query expansion in information retrieval: What can we learn from a deep analysis of queries? In V.P. Selegey, A.V. Baytin, I. M. Boguslavski, V.I. Belikov, and E. Hovy, editors, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue", Moscou, Russie, 04/06/2014-08/06/2014*, volume 20, pages 162–172, http://rggu.com/, juin 2014. Russian State University for Humanities. URL http://www.dialog-21.ru/digests/dialog2014/materials/pdf/ErmakovaLMMotheJ.pdf-http://oatao.univ-toulouse.fr/13097/.

Liana Ermakova, Josiane Mothe, and Elena Nikitina. Proximity relevance model for query expansion. In *ACM Symposium on Applied Computing (SAC), Pisa, Italy, 04/04/2016-08/04/2016*, http://www.acm.org/, 2016. ACM. acceptance rate 25

2.2 Other Publications of the Author

Sergei Ermakov and Liana Ermakova. Sentiment classification based on phonetic characteristics. In Pavel Serdyukov, Pavel Braslavski, Sergei Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 706–709. Springer Berlin Heidelberg, 2013. ISBN 978–3–642–36972–8. URL http://dx.doi.org/10.1007/978--3--642--36973--5_65.

L. Ermakova. Spam and phishing detection in various languages. *International Journal "Information Technologies and Knowledge"*, 4(3):216–232, 2010.

Л. Ермакова and Ю. Айдаров. Лингвистика против социальной инженерии [linguistics against social engineering]. *Открытые системы. СУБД*, (1):56–58, 2009. URL http://www.osp.ru/os/2009/01/7198515/.

В. Салимовский and Л. Ермакова. Экстремистский дискурс в массовой коммуникации Рунета [extremist discourse in runet]. *Вестник Пермского университета. Российская и зарубежная филология*, 15(3):71–80, 2011.

Л. Ермакова. О корпусном подходе к изучению ошибок при билингвизме [a corpus approach to study of bilingual errors]. *Вестник Пермского университета. Российская и зарубежная филология*, (3):34–44, 2012. URL http://www.rfp.psu.ru/archive/3.2012/ermakova.pdf.

S. Ermakov and L. Ermakova. Linguistic approach to suicide detection. *Proceedings of the Institute for System Programming*, 26(4):113–122, 2014. ISSN ISSN 2079-8156.

L. Ermakova. Spam and phishing detection in various languages. *International Journal "Information Technologies and Knowledge"*, 4(3):216–232, 2010.

2.3 References

А.А. Залевская. *Текст и его понимание:*. Тверской государственный ун–т, 2001. URL http://books.google.fr/books?id=dzocAQAAIAAJ.

Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure, 2007. ISSN 1866-4725.

Najeeb Abdulmutalib and Norbert Fuhr. Language models, smoothing, and idf weighting. In Martin Atzmüller, Dominik Benz, Andreas Hotho, and Gerd Stumme, editors, *Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivitaet*, Kassel, Germany, 2010. URL http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir2.pdf.

Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning to find answers to questions on the web. *ACM Trans. Internet Technol.*, 4(2):129–162, May 2004. ISSN 1533–5399. doi: 10.1145/990301.990303. URL http://doi.acm.org/10.1145/990301.990303.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, James Allan, Brian Archibald, Doug Beeferman, Adam Berger, Ralf Brown, Ira Carp Dragon, George Doddington, Alex Hauptmann, John Lafferty, Victor Lavrenko, Xin Liu Cmu, Steve Lowe Dragon, Paul Van Mulbregt Dragon, Ron Papka, Thomas Pierce, Jay Ponte, and Mike Scudder. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

G. Amati. *Probability Models for Information Retrieval Based on Divergence from Randomness: PhD Thesis.* University of Glasgow, 2003.

Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Query difficulty, robustness, and selective application of query expansion. *Advances in Information Retrieval*, page 127–137, 2004c. URL http://www.springerlink.com/content/2f3yht17838j3e6p.

Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*,

20(4):357–389, October 2002a. ISSN 1046-8188. doi: 10.1145/582415.582416. URL http://doi.acm.org/10.1145/582415.582416.

Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002b. ISSN 1046–8188. doi: 10.1145/582415.582416. URL http://doi.acm.org/10.1145/582415.582416.

Massih R. Amini and Nicolas Usunier. A contextual query expansion approach by term clustering for robust text summarization. In *Proceedings of DUC*. Citeseer, 2007. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.3296&rep=rep1&type=pdf.

Massih R. Amini, Anastasios Tombros, Nicolas Usunier, and Mounia Lalmas. Learning-based summarisation of XML documents. 10(3):233–255, 2007.

Roxana Angheluta, Rik De Busser, and Marie-Francine Moens. The use of topic segmentation for automatic summarization. *Proceedings of the workshop on automatic summarization*, pages 66–70, 2002.

Rachit Arora and Balaraman Ravindran. Latent dirichlet allocation based multi–document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 91–97, Singapore, 2008. ACM. ISBN 978–1–60558–196–5.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Regina Barzilay. Information fusion for multidocument summarization: Paraphrasing and generation, 2003. AAI3088294.

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information fusion in the context of multi–document summarization. *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 550–557, 1999.

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, pages 35–55, 2002. 17.

Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, Michael Preminger, Eric SanJuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, Matthew Trappett, and Qiuyue Wang. Overview of inex 2013. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 269–281. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40801-4. doi: 10.1007/978-3-642-40802-1_27. URL http://dx.doi.org/10.1007/978-3-642-40802-1_27.

J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886, July 2007. ISSN 0306–4573. doi: 10.1016/j.ipm.2006.09.003. URL http://dx.doi.org/10.1016/j.ipm.2006.09.003.

Catherine Blake, Julia Kampov, Andreas K. Orphanides, David West, and Cory Lown. UNC-CH at DUC 2007: Query expansion, lexical simplification and sentence selection strategies for multi-document summarization. In *Proceedings of Document Understanding Conference (DUC) Workshop*, 2007. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/unc-ch.blake.final.pdf.

David M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003. 3.

Abdelhamid Bouchachia and R Mittermeir. A neural cascade architecture for document retrieval. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1915–1920. IEEE, 2003.

G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen, 2009.

Daniel Büring. *Topic and Comment.* Cambridge University Press, 2011. Three entries for: Patrick Colm Hogan (ed.) The Cambridge Encyclopedia of the Language Sciences. Cambridge: Cambridge University Press.

Chris Buckley. Automatic query expansion using SMART : TREC 3. In *In Proceedings of The third Text REtrieval Conference (TREC–3). NIST Special Publication 500–226.*, page 69–80, Gaithersburg, MD, 1995. National Institute of Standards and Technology (NIST).

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 243–250, New York, NY, USA, 2008b. ACM. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390377. URL http://doi.acm.org/10.1145/1390334.1390377.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo–relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, page 243–250, New York, NY, USA, 2008c. ACM. ISBN 978–1–60558–164–4. doi: 10.1145/1390334.1390377. URL http://doi.acm.org/10.1145/1390334.1390377.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998a. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291025. URL http://doi.acm.org/10.1145/290941.291025.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336. ACM, 1998b. ISBN 1-58113-015-5. doi: 10.1145/290941.291025. URL http://doi.acm.org/10.1145/290941.291025.

Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54, 2001.

Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50, January 2012b. ISSN

03600300. doi: 10.1145/2071389.2071390. URL http://dl.acm.org/citation.cfm?id=2071389.2071390.

Yllias Chali and Shafiq R. Joty. University of lethbridge's participation in DUC-2007 main task. In *Proceedings of the Document Understanding Conference, Rochester. NIST*. Citeseer, 2007. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.770&rep=rep1&type=pdf.

Chen Chen, Hou Chunyan, and Yuan Xiaojie. Relevance feedback fusion via query expansion. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '12, pages 122–126, Washington, DC, USA, 2012b. IEEE Computer Society. ISBN 978-0-7695-4880-7. doi: 10.1109/WI-IAT.2012.48. URL http://dx.doi.org/10.1109/WI-IAT.2012.48.

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390446. URL http://doi.acm.org/10.1145/1390334.1390446.

John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary, and others. Classy 2007 at DUC 2007. In *Proceedings of the Document Understanding Conference 2007*, 2007. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/ida-umd.final.pdf.

P Cook and F Bildhauer. Annotating information structure: The case of topic. In *Proceedings of the Workshop Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena*, pages 45–56, 2011.

Steve Cronen-Townsend and W. Bruce Croft. Quantifying query ambiguity. page 104–109, March 2002b. URL http://dl.acm.org/citation.cfm?id=1289189.1289266.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, pages 299 – 306, New York,

New York, USA, August 2002b. ACM Press. ISBN 1581135610. doi: 10.1145/564376.564429. URL http://dl.acm.org/citation.cfm?id=564376.564429.

Ronan Cummins. A standard document score for information retrieval. In *ICTIR*.

J. Cutting. *Pragmatics and Discourse: A Resource Book for Students*. Routledge English language introductions series. Routledge, 2002. ISBN 9780415253574. URL http://books.google.fr/books?id=--KFELwzkFhYC.

Jeffrey Dalton and Laura Dietz. Constructing query-specific knowledge bases. In *AKBC'13*, 2013.

Diego Marinho de Oliveira, Alberto H.F. Laender, Adriano Veloso, and Altigran S. da Silva. Fs-ner: A lightweight filter-stream approach to named entity recognition on twitter data. In *Proceedings of the 22Nd International Conference onArabic named entity recognition World Wide Web Companion*, WWW '13 Companion, pages 597–604, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2038-2. URL http://dl.acm.org/citation.cfm?id=2487788.2488003.

J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi. Enhanced web document summarization using hyperlinks. In *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 208–215, Nottingham, UK, 2003. ACM. ISBN 1–58113–704–4.

Gonenc Ercan and Ilyas Cicekli. Using lexical chains for keyword extraction. *Information Processing and Management: an International Journal*, 43(6):1705–1714, 2007.

G. Erkan and D. R. Radev. LexRank: graph–based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

L. Ermakova and J. Mothe. IRIT at INEX 2012: Tweet contextualization. *CLEF 2012 / Conference and Labs of the Evaluation Forum*, 2012a. URL http://www.clef--initiative.eu/documents/71612/3e9ecc64--fae6--4af3--93fd--1a6a6fabb5d6.

Katja Filippova, Margot Mieskes, Vivi Nastase, Simone Paolo Ponzetto, and Michael Strube. Cascaded filtering for topic-driven multi-document summarization. In *Proceedings of the Document Understanding Conference*, volume 2007, 2007. URL http://duc.nist.gov/pubs/2007papers/emlr.pdf.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non–local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL http://dx.doi.org/10.3115/1219840.1219885.

Seeger Fisher and Brian Roark. Query-focused summarization by supervised sentence ranking and skewed word distributions. In *Proceedings of the Document Understanding Conference, DUC-2006, New York, USA*. Citeseer, 2006. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.333.3930&rep=rep1&type=pdf.

R.M. Frumkina and V.N. Telia. *Color, Meaning, Affinity [in Russian]*. Nauka, 1984.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia–based explicit semantic analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1606–1611, 2007.

Evgeniy Gabrilovich, Andrei Broder, Marcus Fontoura, Amruta Joshi, Vanja Josifovski, Lance Riedel, and Tong Zhang. Classifying search queries using the web as a source of knowledge. *ACM Trans. Web*, 3(2):5:1–5:28, April 2009. ISSN 1559–1131. doi: 10.1145/1513876.1513877. URL http://doi.acm.org/10.1145/1513876.1513877.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010.

George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. TAC 2011 MultiLing pilot overview. 2011. URL http://eprints.lancs.ac.uk/71274/.

Fabrizio Gotti, Guy Lapalme, Luka Nerima, Eric Wehrli, and Technologie du Langage. GOFAIsum: a symbolic summarizer for DUC. In *Proc. of DUC*, volume 7, 2007. URL http://www.researchgate.net/profile/Eric_Wehrli/publication/228992375_GOFAIsum_a_symbolic_summarizer_for_DUC/links/09e4150e851ae04b23000000.pdf.

V.D. Gusev, L.A. Miroshnechenko, and N.V. Salomatina. Thematic analysis and quasi–abstracting of text using scan statistics. *Proceedings of International Conference "Dialogue"*, pages 121–125, 2005.

Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. 33(11):29–36, 2000. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=881692.

A. Harabagiu and Finley Lacatusu. Strategies for advanced question answering. In *In HLTNAACL Workshop on Pragmatics of QA*, 2004.

Donna Harman and Chris Buckley. Sigir 2004 workshop: Ria and "where can ir go from here?". *SIGIR Forum*, 38(2):45–49, 2004. URL http://dblp.uni-trier.de/db/journals/sigir/sigir38.html#HarmanB04.

Donna Harman and Chris Buckley. Overview of the Reliable Information Access Workshop. *Information Retrieval*, 12(6):615–641, July 2009a. ISSN 1386-4564. doi: 10.1007/s10791-009-9101-4. URL http://dl.acm.org/citation.cfm?id=1644394.1644419.

Jutta M Hartmann and Susanne Winkler. Investigating the role of information structure triggers. *Lingua*, 136:1–15, 2013.

David Hawking and Nick Craswell. Overview of the TREC–2001 web track. *NIST special publication*, page 61–67, 2002b.

Ben He and Iadh Ounis. Term frequency normalisation tuning for BM25 and DFR models. In *Proceedings of the 27th European conference on Advances in Information Retrieval Research*, ECIR'05, page 200–214, Berlin, Heidelberg, 2005. Springer–Verlag. ISBN 3–540–25295–9, 978–3–540–25295–5. doi: 10.1007/978--3--540--31865--1_15. URL http://dx.doi.org/10.1007/978--3--540--31865--1_15.

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. Hilda: a discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3), 2010.

István T. Hernádvölgyi. Solving the sequential ordering problem with automatically generated lower bounds. *Proceedings of Operations Research 2003*, pages 355–362, 2003.

Andrew Hickl, Kirk Roberts, and FLCC Lacatusu. LCC's GISTexter at DUC 2007: Machine reading for update summarization. In *Proc. of DUC*, volume 7, 2007. URL http://duc.nist.gov/pubs/2007papers/lcc.final.pdf.

D. Hiemstra. Language models. In T. Özsu and L. Liu, editors, *Encyclopedia of Database Systems*, pages 1591–1594. Springer Verlag, Berlin, 2009.

Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. 305:305–332, 1998. URL http://books.google.com/books?hl=en&lr=&id=Rehu8OOzMIMC&oi=fnd&pg=PA305&dq=%22The+index+entry+of+one+points+to+a+thesaurus+category+that+contains+the%22+%22and+Hirst+showed+that+the+distribution+through+a+text+of+lexical+chains+de%EF%AC%81ned+in+this%22+&ots=IqjcKfSQj9&sig=r9_6IC6SK84_sMubOXIanNkOL78.

Birger Hjørland. Towards a theory of aboutness, subject, topicality, theme, domain, field, content …and relevance. *J. Am. Soc. Inf. Sci.*, 52(9):774–778, July 2001. ISSN 0002-8231. URL http://dl.acm.org/citation.cfm?id=380494.380507.

Eduard Hovy and S. Tratz. Summarization evaluation using transformed basic elements. *Proceedings TAC 2008*, 2008.

Meishan Hu, Aixin Sun, and Ee-Peng Lim. Comments–oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, Singapore, Singapore, 2008. ACM. ISBN 978–1–60558–164–4.

Amanda Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. In *ISCRAM Conference*, Gothenburg, Sweden, 2009. doi: citeulike--article--id:6788937. URL http://www.slideshare.net/guest8c177f/twitter--adoption--and--use--in--mass--convergence--and--emergency--events.

Killian Janod and Olivier Mistral. Overview of the 2011 QA track: Querying and summarizing with XML. *INEX 2011 Workshop Preproceedings*, pages 167–174, 2011.

Kellerer, Hans, Pferschy, Ulrich, and Pisinger, David. *Knapsack problems*. Springer–Verlag, Berlin, 2004. ISBN 3–540–40286–1.

T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, pages 259–284, 1998. 25.

Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. *Proceedings of ACL*, pages 542–552, 2003.

Tessa Lau and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. In *Proceedings of the seventh international conference on User modeling*, UM '99, page 119–128, Secaucus, NJ, USA, 1999. Springer–Verlag New York, Inc. ISBN 3–211–83151–7. URL http://dl.acm.org/citation.cfm?id=317328.317340.

Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, page 120–127, New York, NY, USA, 2001b. ACM. ISBN 1–58113–331–6. doi: 10.1145/383952.383972. URL http://doi.acm.org/10.1145/383952.383972.

Kyung Soon Lee, W Bruce Croft, and James Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 235–242. ACM, 2008.

Jing Li, Le Sun, Chunyu Kit, and Jonathan Webster. A query-focused multi-document summarizer based on lexical chains. In *Proceedings of the Document Understanding Conference, Rochester. NIST*, 2007a. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/cas-uhongkong.final.pdf.

Sujian Li, You Ouyang, Wei Wang, and Bin Sun. Multi-document summarization using support vector regression. In *Proceedings of DUC*. Citeseer, 2007b. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.6243&rep=rep1&type=pdf.

Chin-Yew Lin. Assembly of topic extraction modules in SUMMARIST. *In AAAI Spring Symposium on Intelligent Text Summarisation*, pages 53–59, 1998.

Chin-Yew Lin and Eduard Hovy. Identifying topics by position. *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290, 1997.

Ziheng Lin, Tat-Seng Chua, Min-Yen Kan, Wee Sun Lee, Long Qiu, and Shiren Ye. NUS at DUC 2007: Using evolutionary models of text. In *Proceedings of Document Understanding Conference (DUC)*, 2007. URL https://gala2014.comp.nus.edu.sg/~kanmy/papers/duc07.pdf.

Christina Lioma and Roi Blanco. Part of speech based term weighting for information retrieval. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal

Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 412–423. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00957-0. doi: 10.1007/978-3-642-00958-7_37. URL http://dx.doi.org/10.1007/978-3-642-00958-7_37.

Christina Lioma, Birger Larsen, and Wei Lu. Rhetorical relations for information retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 931–940, 2012. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348407. URL http://doi.acm.org/10.1145/2348283.2348407.

K. Lu. An insight into vector space modeling and language modeling. In *iConference 2013 Proceedings*, pages 717–721, 2013.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140, Madrid, Spain, 2009a. ACM. ISBN 978–1–60558–487–4.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, pages 159–165, 1958.

Yuanhua Lv and ChengXiang Zhai. Positional relevance model for pseudo–relevance feedback. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, page 579–586, New York, NY, USA, 2010. ACM. ISBN 978–1–4503–0153–4. doi: 10.1145/1835449.1835546. URL http://doi.acm.org/10.1145/1835449.1835546.

Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. From puppy to maturity: Experiences in developing terrier. pages 60–63, 2012. URL http://opensearchlab.otago.ac.nz/FullProceedings.pdf#page=65.

Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan, and Jimmy Lin. Multiple alternative sentence compressions for automatic text summarization. In *Proceedings of DUC*, 2007. URL http://www.researchgate.net/profile/Nitin_Madnani/publication/228631255_Multiple_alternative_sentence_compressions_for_automatic_text_summarization/links/00b7d51a61032b21f1000000.pdf.

M.A.K.Halliday. *An Introduction to Functional Grammar*. Arnold, London, 2 edition, 1994.

Rila Mandala, Tokunaga Takenobu, and Tanaka Hozumi. The use of wordnet in information retrieval. In *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 31–37, 1998.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA, 2000. ISBN 0262133725.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19, 1993. 2.

V. Mathesius and J. Vachek. *A Functional Analysis of Present Day English on a General Linguistic Basis*. Janua linguarum : Series practica / Ianua linguarum / Series practica. Mouton, 1975. ISBN 9789027930774. URL https://books.google.fr/books?id=ZdbLSkaPMJwC.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, page 563–572, New York, NY, USA, 2012. ACM. ISBN 978–1–4503–0747–5. doi: 10.1145/2124295.2124364. URL http://doi.acm.org/10.1145/2124295.2124364.

Donald Metzler and W. Bruce Croft. Latent concept expansion using markov random fields. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–318. ACM, 2007. URL http://dl.acm.org/citation.cfm?id=1277796.

Jun Miao, Jimmy Xiangji Huang, and Zheng Ye. Proximity-based Rocchio's model for pseudo relevance. In *Proceedings of the 35th international ACM SIGIR conference on*

Research and development in information retrieval, pages 535–544. ACM, 2012. URL http://dl.acm.org/citation.cfm?id=2348356.

Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 20. Association for Computational Linguistics, 2004. URL http://dl.acm.org/citation.cfm?id=1219064.

David Milne and Ian H. Witten. An effective, low–cost measure of semantic relatedness obtained from wikipedia links. *Proceedings of AAAI*, pages 25–30, 2008.

N.F. Mohammed and N. Omar. Arabic named entity recognition using artificial neural network. 8(8):1285–1293, 2012.

V.V. Morozenko. *Discrete Mathematics. Handbook*. Perm State University, Perm, 2008.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. 17(1):21–48, 1991. URL http://dl.acm.org/citation.cfm?id=971740.

Vanessa Graham Murdock. Aspects of sentence retrieval. *Dissertation*, 2006.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

A. Nenkova and L. Vanderwende. The impact of frequency on summarization, 2005-01. URL http://research.microsoft.com/apps/pubs/default.aspx?id=67448.

Jian-Yun Nie, Guihong Cao, and Jing Bai. Inferential language models for information retrieval. *Transactions on Asian Language Information Processing*.

Elisabeth Niemann and Iryna Gurevych. The people's web meets linguistic knowledge: Automatic sense alignment of wikipedia and WordNet. *International Conference on Computational Semantics*, 2011.

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006a.

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, Seattle, Washington, USA, 2006b.

Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. Query expansion using term distribution and term association. *arXiv preprint arXiv:1303.0667*, 2013a.

Michael J. Paul and Roxana Girju. A two–dimensional topic–aspect model for discovering multi–faceted topics. In Maria Fox and David Poole, editors, *AAAI*. AAAI Press, 2010. URL http://dblp.uni--trier.de/db/conf/aaai/aaai2010.html.

Michael J. Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, Cambridge, Massachusetts, 2010. Association for Computational Linguistics.

J. Pearsall. *The New Oxford Dictionary of English*. Oxford University Press, 2002. ISBN 9780195654325. URL http://books.google.fr/books?id=81Dl3gopeEgC.

José R Pérez-Agüera and Lourdes Araujo. Comparing and combining methods for automatic query expansion. *arXiv preprint arXiv:0804.2057*, 2008.

Philipp Petrenz and Bonnie Webber. Stable classification of text genres. *Comput. Linguist.*, 37(2):385–393.

M. F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, 1997a.

Rahul K. Prasad Pingali and Vasudeva Varma. Iiit hyderabad at duc 2007. 2007. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/iiit.pdf?q=iiit-hyderabad.

Matthew Purver. Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, pages 291–317, 2011.

Lawrence R. Rabiner. Readings in speech recognition. page 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1–55860–124–4. URL http://dl.acm.org/citation.cfm?id=108235.108253.

Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. 24(3):469–500, 1998.

Lawrence H. Reeve and Hyoil Han. A term frequency distribution approach for the duc-2007 update task. In *Proc. of Document Understanding Conference*. Citeseer, 2007. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.5272&rep=rep1&type=pdf.

Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. 60:503–520, 2004. URL http://research.microsoft.com/apps/pubs/default.aspx?id=67744.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 42–49. ACM, 2004. ISBN 1-58113-874-1. doi: 10.1145/1031171.1031181. URL http://doi.acm.org/10.1145/1031171.1031181.

J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System*, page 313–323. 1971. URL http://scholar.google.com/scholar?hl=en&lr=&client=firefox--a&q=relevance+feedback+in+information+retrieval&btnG=Search.

E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and categorization*, pages 27–48. Erlbaum, Hillsdale, New Jersey, 1978.

Horacio Saggion and Guy Lapalme. Generating indicative–informative summaries with SumUM. *Association for Computational Linguistics*, 28(4):497–526, 2002.

Eric SanJuan, Véronique Moriceau, Xavier Tannier, Patrice Bellot, and Josiane Mothe. Overview of the INEX 2011 question answering track (QA@INEX). In Shlomo Geva, Jaap Kamps, and Ralf Schenkel, editors, *Focused Retrieval of Content and Structure*, volume 7424 of *Lecture Notes in Computer Science*, pages 188–206. Springer Berlin Heidelberg, 2012. ISBN 978–3–642–35733–6. URL http://dx.doi.org/10.1007/978--3--642--35734--3_17.

Rodrygo L. T. Santos, Pablo Castells, Ismail Sengör Altingövde, and Fazli Can. Diversity and novelty in information retrieval. In *The 36th International ACM SIGIR conference*

on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013, page 1130, 2013a. doi: 10.1145/2484028.2484187. URL http://doi.acm.org/10.1145/2484028.2484187.

Ralf Schenkel, Fabian M. Suchanek, and Gjergji Kasneci. YAWN: A semantically annotated wikipedia XML corpus. In *BTW*, pages 277–291, 2007b.

Barry Schiffman. Summarization for q&a at columbia university for DUC 2007. In *Proceedings of the Document Understanding Conference 2007*, 2007. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/columbiau.pdf.

Heinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24 (1):97–123, 1998.

Yohei Seki. Automatic summarization focusing on document genre and text structure. *ACM SIGIR Forum*, 39(1):65–67, 2005.

Chao Shen and Tao Li. Learning to rank for query-focused multi-document summarization. pages 626–634. IEEE, 2011-12. ISBN 978-1-4577-2075-8, 978-0-7695-4408-3. doi: 10.1109/ICDM.2011.91. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6137267.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. Document summarization using conditional random fields. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2862–2867, Hyderabad, India, 2007. Morgan Kaufmann Publishers Inc.

H. Gregory Silber and Kathleen F. Mccoy. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics – Summarization*, 28(4):1–11, 2002.

Jagendra Singh and Aditi Sharan. Co-occurrence and Semantic Similarity Based Hybrid Approach for Improving Automatic Query Expansion in Information Retrieval. In *Distributed Computing and Internet Technology. 11th International Conference, ICDCIT 2015, Bhubaneswar, India, February 5-8, 2015. Proceedings*, volume 8956 of *Lecture Notes in Computer Science*, pages 415–418. Springer International Publishing, 2015.

Ian Soboroff. Does wt10g look like the web? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,

SIGIR '02, pages 423–424, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi: 10.1145/564376.564475. URL http://doi.acm.org/10.1145/564376.564475.

A.N. Sokolov. *Internal speech and thinking (in Russian)*. Prosveschenie, 1968. URL http://books.google.ru/books?id=ruChGAAACAAJ.

Fei Song and W. Bruce Croft. A general language model for information retrieval. In *Proceedings of the 1999 ACM SIGIR Conference on research and development in Information Retrieval*, pages 279–280. ACM, 1999.

Edmundo-Pavel Soriano-Morales, Alfonso Medina-Urrea, GerardoSierra Martínez, and Carlos-Francisco Méndez-Cruz. The GIL summarizers: Experiments in the track QA@INEX'10. In Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman, editors, *Comparative Evaluation of Focused Retrieval*, volume 6932 of *Lecture Notes in Computer Science*, pages 282–289. Springer Berlin Heidelberg, 2011. ISBN 978–3–642–23576–4. URL http://dx.doi.org/10.1007/978--3--642--23577--1_25.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 149–156. Association for Computational Linguistics, 2003. doi: 10.3115/1073445.1073475. URL http://dx.doi.org/10.3115/1073445.1073475.

Nicola Stokes, Jiawen Rong, and Lawrence Cavedon. NICTA's update and question-based summarisation systems at DUC 2007. In *Proceedings of the Document Understanding Conference Workshop*. Citeseer, 2007. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.1841&rep=rep1&type=pdf.

Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. Web–page summarization using clickthrough data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201, Salvador, Brazil, 2005. ACM. ISBN 1–59593–034–5.

N. Suwandaratna and U. Perera. Discourse marker based topic identification and search results refining. In *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, pages 119–125, Dec 2010a. doi: 10.1109/ICIAFS.2010.5715646.

Tao Tao and ChengXiang Zhai. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 295–302. ACM, 2007.

Simone Teufel and Marc Moens. Sentence extraction and rhetorical classification for flexible abstracts. *In Intelligent Text Summarization*, pages 16–25, 1998.

Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

Juan-Manuel Torres-Moreno, Patricia Velázquez-Morales, and Michel Gagnon. Statistical summarization at qa@inex 2011 track using cortex and enertex systems. In Shlomo Geva, Jaap Kamps, and Ralf Schenkel, editors, *Focused Retrieval of Content and Structure*, volume 7424 of *Lecture Notes in Computer Science*, pages 247–256. Springer Berlin Heidelberg, 2012b. ISBN 978-3-642-35733-6. doi: 10.1007/978-3-642-35734-3_23. URL http://dx.doi.org/10.1007/978-3-642-35734-3_23.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature–rich part–of–speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1*, NAACL '03, page 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073478. URL http://dx.doi.org/10.3115/1073445.1073478.

Kristina Toutanova, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamudi, Hisami Suzuki, and Lucy Vanderwende. The pythy summarization system: Microsoft research at duc 2007. In *Proc. of DUC*, volume 2007, 2007. URL http://research.microsoft.com:8082/pubs/69490/msrduc2007.pdf.

Matthew Trappett, Shlomo Geva, Andrew Trotman, Falk Scholer, and Mark Sanderson. Overview of the INEX 2012 snippet retrieval track. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, 2012b. URL http://ceur-ws.org/Vol-1178/CLEF2012wn-INEX-TrappettEt2012.pdf.

Peter Turney. Learning to extract keyphrases from text. *Information Retrieval*, 4(2): 303–336, 2000.

Twitter Help Center | What Are Hashtags ("#" Symbols)?, accessed date: 02/08/2012. URL https://support.twitter.com/articles/49309--what--are--hashtags--symbols.

Twitter Help Center | What are @Replies and Mentions?, accessed date: 02/08/2012. URL https://support.twitter.com/groups/31--twitter--basics/topics/109--tweets--messages/articles/14023--what--are--replies--and--mentions.

Twitter Usage Statistics - Internet Live Stats, accessed date: 14/12/2015. URL http://www.internetlivestats.com/twitter-statistics/.

S. Vargas, R. L. T. Santos, C. Macdonald, and I. Ounis. Selecting effective expansion terms for diversity. In *10th International Conference in the RIAO series (OAIR 2013)*, Lisbon, Portugal, May 2013. URL http://ir.ii.uam.es/predict/pubs/oair2013-vargas-gla.pdf.

Rakesh Verma, Ping Chen, and Wei Lu. A semantic free-text summarization system using ontology knowledge. In *Proc. of Document Understanding Conference*, 2007. URL http://www.geraldkembellec.fr/docOntology/A%20Semantic%20Free-text%20Summarization%20System%20Using%20Ontology%20Knowledge.pdf.

Yannick Versley and Anna Gastel. Linguistic tests for discourse relations in the tüba-d/z corpus of written german, 2012.

Jorge Vivaldi and Iria da Cunha. Qa@inex track 2011: Question expansion and reformulation using the reg summarization system. In Shlomo Geva, Jaap Kamps, and Ralf Schenkel, editors, *Focused Retrieval of Content and Structure*, volume 7424 of *Lecture Notes in Computer Science*, pages 257–268. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-35733-6. doi: 10.1007/978-3-642-35734-3_24. URL http://dx.doi.org/10.1007/978-3-642-35734-3_24.

E. Voorhees and D. Harman. Overview of the seventh text REtrieval conference (TREC–7). In *Text REtrieval Conference (TREC) TREC–7 Proceedings*, page 1–23. Department of Commerce, National Institute of Standards and Technology, 1998b. URL papers/overview_7.ps.gz;papers/overview_7.pdf.gz. NIST Special Publication 500–242: The Seventh Text REtrieval Conference (TREC 7).

Ellen M. Voorhees. Query expansion using lexical–semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, page 61–69, New York, NY, USA, 1994. Springer–Verlag New York, Inc. ISBN 0–387–19889–X. URL http://dl.acm.org/citation.cfm?id=188490.188508.

Ellen M. Voorhees and Donna Harman. Overview of the eighth text REtrieval conference (TREC–8). page 1–24, 2000b.

Ellen M. Voorhees and Donna Harman. *Overview of the Sixth Text REtrieval Conference (TREC–6)*. 2000c.

JeroenB.P. Vuurens and ArjenP. de Vries. Distance matters! cumulative proximity expansions for ranking documents. *Information Retrieval*, 17(4):380–406, 2014. ISSN 1386-4564. doi: 10.1007/s10791-014-9243-x. URL http://dx.doi.org/10.1007/s10791-014-9243-x.

Jing Wan, WenCong Wang, JunKai Yi, Chong Chu, and Kang Song. Query Expansion Approach Based on Ontology and Local Context Analysis. *Research Journal of Applied Sciences, Engineering and Technology*, 4(16):2839–2843, 2012.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Manifold–ranking based topic–focused multi–document summarization. *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2903–2908, 2007.

Dingding Wang, Tao Li, and Chris Ding. Weighted feature subset non–negative matrix factorization and its applications to document understanding. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 541–550. IEEE Computer Society, 2010. ISBN 978–0–7695–4256–0.

H Weil. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes: question de grammaire générale*. Joubert, 1844.

René Witte, Ralf Krestel, and Sabine Bergler. Generating update summaries for DUC 2007. In *Proceedings of the Document Understanding Conference*, volume 2007, 2007. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/ukarlsruhe.final.pdf.

Kam-Fai Wong, Dawei Song, Peter Bruza, and Chun-Hung Cheng. Application of aboutness to functional benchmarking in information retrieval. *ACM Trans. Inf. Syst.*, 19(4):337–370, October 2001. ISSN 1046-8188. doi: 10.1145/502795.502796. URL http://doi.acm.org/10.1145/502795.502796.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714, Hyderabad, India, 2011. ACM. ISBN 978–1–4503–0632–4.

Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, page 4–11, New York, NY, USA, 1996. ACM. ISBN 0–89791–792–8. doi: 10.1145/243199.243202. URL http://doi.acm.org/10.1145/243199.243202.

Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, January 2000. ISSN 1046–8188. doi: 10.1145/333135.333138. URL http://doi.acm.org/10.1145/333135.333138.

Yang Xu, Gareth J.F. Jones, and Bin Wang. Query dependent pseudo–relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, page 59–66, New York, NY, USA, 2009. ACM. ISBN 978–1–60558–483–6. doi: 10.1145/1571941.1571954. URL http://doi.acm.org/10.1145/1571941.1571954.

E.V. Yagunova. *Variation of perception strategies of verbal text (in Russian)*. Perm State University, 2008.

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 255–264, Beijing, China, 2011. ACM. ISBN 978–1–4503–0757–4.

J. C. Ying, S. J. Yen, Y. S. Lee, Y. C. Wu, and J. C. Yang. Language model passage retrieval for question-oriented multi document summarization. In *Proc. of Document Understanding Conference*, 2007. URL http://users.cis.fiu.edu/~lli003/Sum/DUC/2007/ncu-tw.pdf.

Jin Zhang, Hongbo Xu, Xiaolei Wang, Huawei Shen, and Yiling Zeng. ICT CAS at DUC 2007. In *Proceedings of the Document Understanding Conference 2007*, 2007. URL http://www-nlpir.nist.gov/projects/duc/pubs/2007papers/cas-ict.pdf.