proxem
from text to data

alpage

Antelope

une plate-forme de TAL permettant d'extraire les sens du texte

Théorie et applications de l'interface syntaxe-sémantique

François-Régis Chaumartin – 25/09/2012

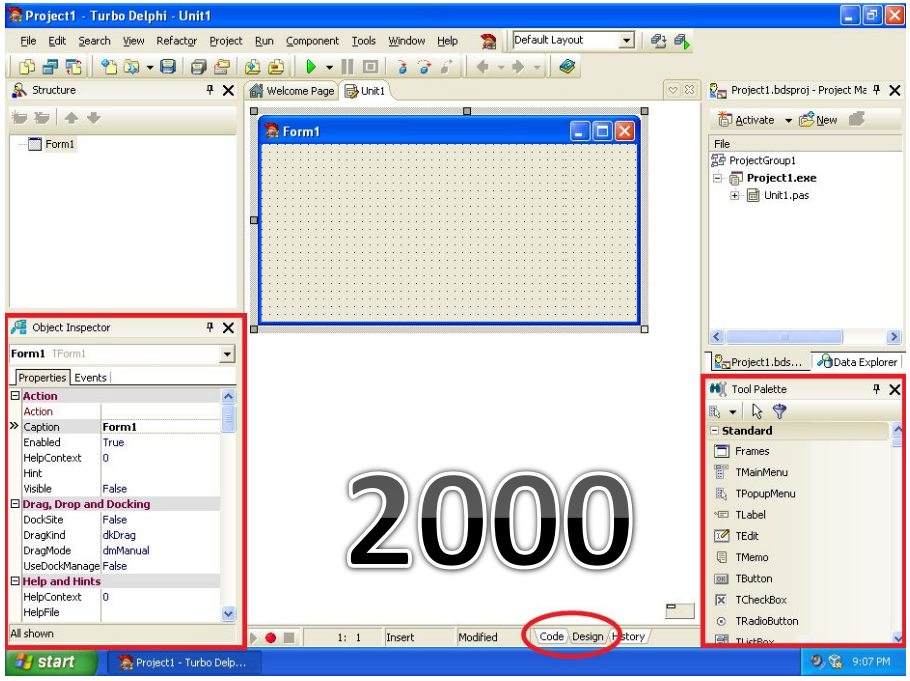
Soutenance de doctorat – Université Paris Diderot

Introduction

- Découverte du TAL en autodidacte vers 2003/2004
  - Si le TAL a régulièrement fait des progrès…
  - … l'assemblage de plusieurs composants de TAL pour réaliser une application complète reste complexe
- Parallèle avec les progrès en développement informatique

```
LRESULT CALLBACK WndProc (HWND hwnd, UINT message, WPARAM wParam, LPARAM lParam)
{
     HDC         hdc ;
     PAINTSTRUCT ps ;
     RECT        rect ;

     switch (message)
     {
     case WM_CREATE:
          PlaySound (TEXT ("hellowin.wav"), NULL, SND_FILENAME | SND_ASYNC) ;
          return 0 ;

     case WM_PAINT:
          hdc = BeginPaint (hwnd, &ps) ;

          GetClientRect (hwnd, &rect) ;

          DrawText (hdc, TEXT ("Hello, Windows 98!"), -1, &rect,
                    DT_SINGLELINE | DT_CENTER | DT_VCENTER) ;

          EndPaint (hwnd, &ps) ;
          return 0 ;

     case WM_DESTROY:
          PostQuitMessage (0) ;
          return 0 ;
     }
     return DefWindowProc (hwnd, message, wParam, lParam) ;
}
```

1990



Introduction

- Objectifs de mon activité de recherche
  - Créer une plate-forme de TAL pour analyser du texte tout-venant, utilisable même par des non-linguistes
  - Développer rapidement et simplement des applications par assemblage et paramétrage de briques existantes
  - Offrir des ressources sémantiques à large couverture
- Activité de R&D ancrée dans le développement applicatif
  - Thèse menée de front avec le développement de Proxem
  - Analyse sémantique abordée en largeur plutôt qu'en profondeur
  - Navigation entre l'approche linguistique et celle venant de l'IA
  - Hésitation entre la démarche de l'ingénieur et celle du chercheur

Contributions

- Antelope : une plate-forme industrielle
  - Définition d'une architecture rendant possible l'interopérabilité de ressources hétérogènes
  - Proposition d'un modèle de données linguistiques unifié
  - Constitution d'un lexique sémantique à large couverture
  - Création de composants d'analyse sémantique originaux
  - Implémentation d'une interface syntaxe-sémantique
- Démarche semi-supervisée d'acquisition rapide de connaissances spécifiques à un domaine
  - Mise en œuvre sur plusieurs projets opérationnels dans des domaines applicatifs variés

Plan

- Représentation du sens souhaitée
- Architecture de la plate-forme
- Lexique sémantique à large couverture
- Composants d'analyse sémantique
- Applications de la plate-forme
- Travaux en cours : désambiguïsation & ISS
- Synthèse et perspectives

REPRÉSENTATION DU SENS SOUHAITÉE DANS LA PLATE-FORME

Représentation du sens par un moteur de recherche

- je tenais à féliciter la caissiére Céline pour son accueil chaleureux et souriant du samedi 16 février malgré la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. FELICITATIONS

Représentation du sens par un moteur de recherche

- je tenais a feliciter la caissiere celine pour son accueil chaleureux et souriant du samedi 16 fevrier malgre la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. felicitations

Représentation du sens par un moteur de recherche

- je tenais a feliciter la caissiere celine pour son accueil chaleureux et souriant du samedi 16 fevrier malgre la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. felicitations

Représentation du sens par un moteur de recherche

- je tenais a felicit er la caissiere celine pour son accueil chaleureux et souriant du samedi 16 fevrier malgre la foule incroiable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. felicitations

Représentation du sens par un moteur de recherche

ten felicit caiss celin

accueil chaleur souri samed

fevri foul incroi jour

su fair abstract gard

souri bon humeur felicit

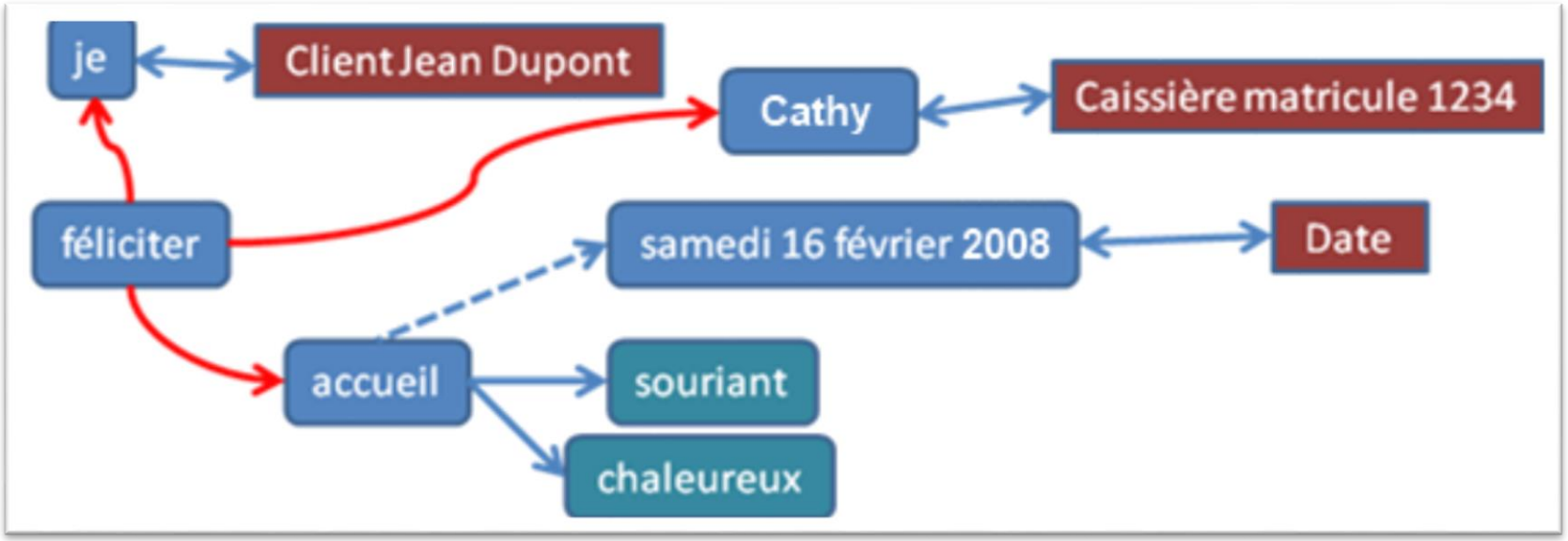
Représentation du sens par un moteur de recherche

- je tenais à féliciter la caissiére Céline pour son accueil chaleureux et souriant du samedi 16 février malgré la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. FELICITATIONS

[abstract, accueil, bon, caiss, celin, chaleur, fair, felicit*2, fevri, foul, gard, humeur, incroi, jour, samed, souri*2, su, ten]

Représentation du sens souhaitée dans la plate-forme

- Représentation inspirée par la Théorie Sens-Texte
  - Relations prédicat-argument
  - Désambiguïsation des unités lexicales
- Repérage des entités nommées

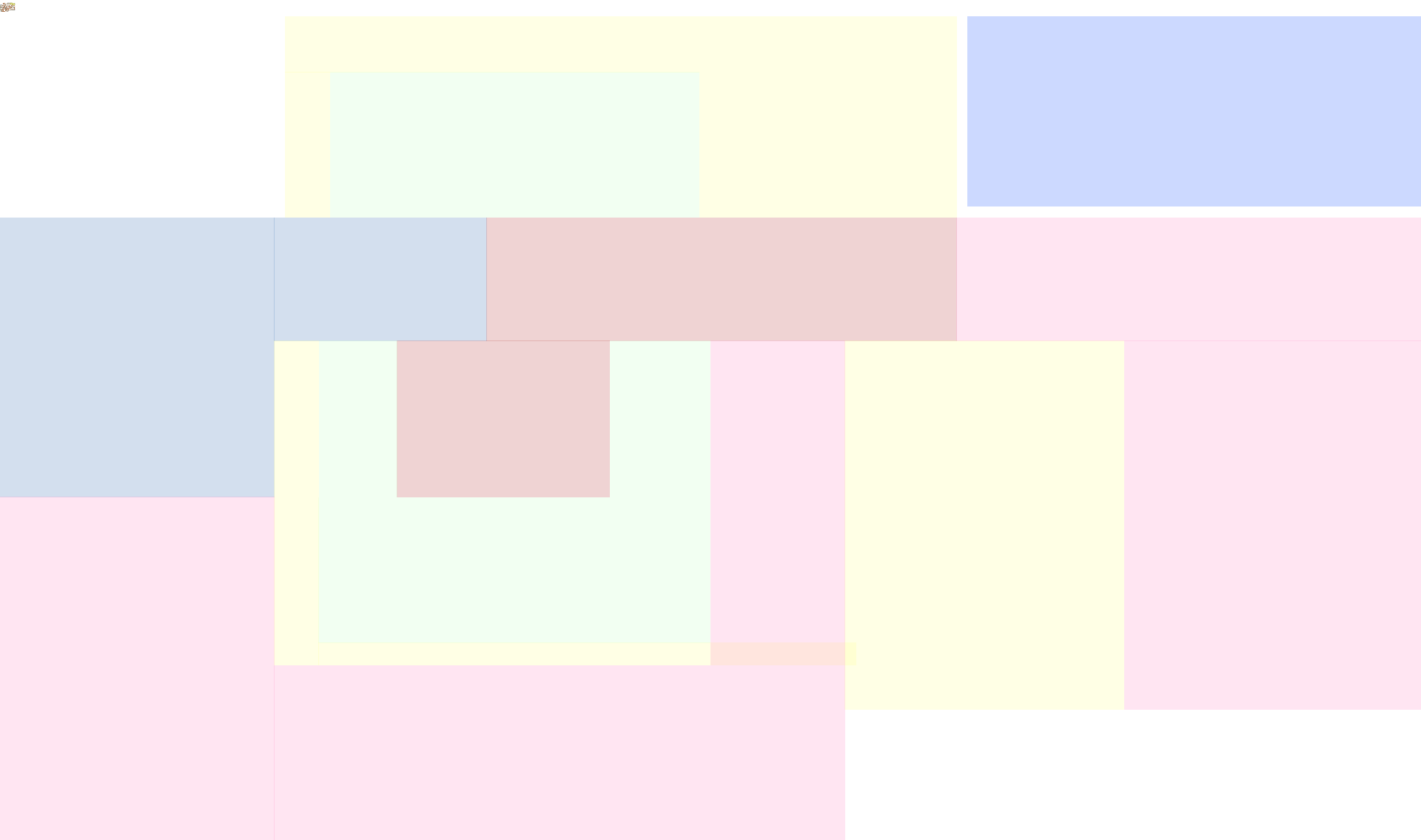


ARCHITECTURE DE LA PLATE-FORME

Plate-forme Antelope

- CHAUMARTIN (TAL 49:2, 2008) « ANTELOPE, une plate-forme industrielle de traitement linguistique »
- Effort d'intégration informatique
  - Conception logicielle pour articuler des composants hétérogènes
  - Application de « bonnes pratiques » du génie logiciel
  - Architecture technique autorisant les traitements parallèles
  - Passerelle avec l'architecture UIMA : CHAUMARTIN ET AL. (RMLL 2009) « Bridging .NET and Java in the SCRIBO Project: UIMA.NET »
- Effort d'intégration linguistique
  - Prise en charge du multilinguisme
  - Constitution d'un lexique sémantique à large couverture
  - Modèle de données linguistiques unifié
- Capacité à « extraire les sens des textes »
  - Texte→RMorphS→RMorphP→RSyntS→RSyntP→RSém→RConcept

Modèle de données linguistiques



LEXIQUE SÉMANTIQUE À LARGE COUVERTURE

Conception du lexique sémantique

- Exploiter au maximum les ressources existantes
  - WordNet et son écosystème
  - Ontologies SUMO et MILO
  - Grande hétérogénéité des formats et des informations
- Proposer des mécanismes d'enrichissement
- Contributions personnelles
  - Extension de WN à partir d'articles encyclopédiques
  - Détection de cadres de sous-catégorisation en relation de paraphrase
  - Acquisition sur WN de règles de polysémie régulière

Etendre WN à partir de la Wikipédia

- CHAUMARTIN (RECITAL 2007) « Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement »
- Rapprochement entre synsets et articles encyclopédiques
  - Analyse syntaxique des définitions + un ensemble d'heuristiques
  - Quand un article correspond à une entrée du lexique, correspondance entre les deux (précision = 92 %)
  - Sinon, création d'une nouvelle entrée en la rattachant au meilleur hyperonyme existant (précision = 85 %)
- Détection de paraphrases et de cadres de sous-catégorisation en relation de paraphrase
  - Exploite la présence d'entités nommées E1 et E2 dans des configurations similaires à l'intérieur d'articles comparables

Détection de paraphrases

- Corpus monolingue d'articles comparables

The Alabama River, in the U.S. state of Alabama, is formed by the Tallapoosa and Coosa rivers, which unite six miles above Montgomery. The Alabama River flows west as far as Selma, then southwest until, about 45 miles from Mobile. The Alabama River unites with the Tombigbee to form the Mobile and Tensas rivers, which discharge into Mobile Bay.	The Alabama River is formed by the Coosa and Tallapoosa rivers northeast of Montgomery. The Alabama River winds westward to Selma and then flows south for a length of 318 mi. The Alabama River is joined above Mobile by the Tombigbee to form the Tensaw and Mobile rivers, which flow into the Gulf of Mexico.	The Alabama River is a river, 315 mi long, formed in central Alaska by the confluence of the Coosa and Tallapoosa rivers north of Montgomery. Flowing southwest to Mobile, Alaska, the Alabama River joins the Tombigbee to form the Mobile River.

(RIVIÈRE#1 riv1, COULER#2, VILLE#1 v1) ~ (RIVIÈRE#1 riv1, SERPENTER#1, VILLE#1 v1)
(RIVIÈRE#1 riv1, UNIR#4, RIVIÈRE#1 riv2) ~ (RIVIÈRE#1 riv2, REJOINDRE#5, RIVIÈRE#1 riv1)

Acquisition de règles de polysémie régulière (métonymie et métaphore)

- BARQUE, CHAUMARTIN (TALN 2008) « La polysémie régulière dans WordNet »
- A partir de définitions de WordNet
  - {cerise#1, cherry#4} *[couleur]* = *the red color of cherries*
- Observation de patrons réguliers de polysémie
  - Spécialisation, métaphore ou métonymie
  - Extraction de 2000 relations obéissant à 60 patrons
    - (Couleur → Fruit), (Vin → Région), (Quantité → Conteneur)…
- Ces patrons s'appliquent dans des situations où une coercition de type est imposée par un prédicat à l'un de ses arguments (création dynamique d'acceptions)
  - *Friends don't let friends drink Bordeaux[vin]*
  - **Friends don't let friends drink Bourgogne[région]*

COMPOSANTS D’ANALYSE SÉMANTIQUE ORIGINAUX

Composants de traitement

- Implémentation de plusieurs composants
  - Extraction d'informations (entités nommées et relations)
  - Analyse de sentiments
  - Regroupement de documents (*clustering*)
  - Résolution d'anaphores et de coréférences
- Optimisation des performances et travail sur la robustesse
- Limites de l'exercice
  - Travail réalisé en largeur plutôt qu'en profondeur
  - Temps de développement limité pour chaque composant
  - Pas toujours d'évaluation avec comparaison à l'état de l'art
- Découverte des méthodes d'apprentissage automatique
  - Glissement de paradigme pendant la durée de la thèse

Reconnaissance d'entités nommées

- Paramétrage en fonction de l'application à réaliser
  - Skype : numéros de téléphone
  - Veille : personnes, lieux, organisations
  - RH : compétences, métiers, expériences, diplômes
  - 100 classes dans la hiérarchie proposée dans (Sekine, 2002)
- Amorçage à partir d'une liste de termes établie par classe
  - Entrées associées à des contextes de désambiguïsation
- Apprentissage utilisant les CRF
- Résultats obtenus dans des applications spécifiques
  - Allant de corrects (personnes, lieux)
  - A très bons (97% sur les entités des avis de consommateurs)

Reconnaissance d'entités nommées

- On a appelé la cliente pour venir chercher une commande : Lait Après Soleil de Garnier .
- On ne trouve pas de gel gommant Roger Gallet autre que parfum gingembre .
- Acheter merguez le 27/08/2010 horrible pas beau , pas appétissant sur le barbecue de marque PRIM' GRILL
- CDE larroche mazet Cabernet 2 cartons
- j' ai reçu le sommier et matelas de marque ivana
- bonjour , je voulais commander un automatisme pour portail a mon magasin de bethune

MARQUES – PRODUITS – CONCEPTS – **SANS CRF**

Reconnaissance d'entités nommées

- On a appelé la cliente pour venir chercher une commande : Lait Après Soleil de Garnier .
- On ne trouve pas de gel gommant Roger Gallet autre que parfum gingembre .
- Acheter merguez le 27/08/2010 horrible pas beau , pas appétissant sur le barbecue de marque PRIM' GRILL
- CDE larroche mazet Cabernet 2 cartons
- j' ai reçu le sommier et matelas de marque ivana
- bonjour , je voulais commander un automatisme pour portail a mon magasin de bethune

MARQUES – PRODUITS – CONCEPTS – **AVEC CRF**

Etiquetage de rôles

- Distinction entre rôles thématiques et sémantiques
  - Arg0/Arg1 vs. Agent/Patient vs. Acheteur/Achetée
- Etiquetage de rôles thématiques
  - CHAUMARTIN (RECITAL 2006) « Construction automatique d'une interface syntaxe-sémantique utilisant des ressources à large couverture en langue anglaise »
- Extraction de relations entre entités nommées
  - Recherche de patrons dans le graphe des dépendances
  - S'applique sur la RSyntS ou (mieux) sur la RSyntP (rappel)
  - Définition par l'exemple de patrons morphosyntaxiques

Autres composants de traitement

- Détection d'opinions
  - CHAUMARTIN (SemEval-2007 - ACL) « A knowledge-based system for headline sentiment tagging »
  - Extensions de WordNet-Affect et de SentiWordNet
  - Règles heuristiques et analyse syntaxique
- Regroupement automatique de documents
  - Regroupement spectral
  - Regroupement par cliques
- Résolution d'anaphores
  - Implémentation d'algorithmes classiques
  - Testé dans le cadre particulier d'articles encyclopédiques

APPLICATIONS DE LA PLATE-FORME

Applications

- « Preuves » de l'intérêt de la plate-forme
  - Applications développées par l'équipe Proxem…
  - … ou par des tiers académiques ou industriels
- Démarche semi-supervisée de constitution de ressources spécifiques au domaine
  - Approche *top-down* : exploitation et enrichissement sélectif du lexique sémantique
  - Approche *bottom-up* : analyse du corpus disponible (extraction terminologique…)

Applications

- Application de veille économique
  - Détection d'entités nommées, extraction de relations
  - SP3 SCRIBO: personnes, lieux, organisations, montants
- Analyse d'avis de consommateurs
  - Détection des entités pertinentes et des opinions
  - Présentation de tableaux de bord synthétiques
- RH : rapprochement entre CV et offres d'emploi
- Démos industrielles à TALN 2011 et 2012

Applications

- Travaux académiques utilisant la plate-forme
  - Plus de 2500 téléchargements
  - Antelope utilisée pour développer différentes applications socio-économiques

AMMARI, DIMITROVA, DESPOTAKIS (UMAP 2011). Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models.

DESPOTAKIS (UMAP 2011). Multi-perspective Context Modelling to Augment Adaptation in Simulated Learning Environments.

DOUMIT, MINAI (ICCS 2011). Online News Media Bias Analysis using an LDA-NLP Approach.

FERREIRA, DA SILVA (Wikis4SE 2008). Wiki Supported Collaborative Requirements Engineering.

FITRIANIE, YANG, DATCU, CHITU, ROTHKRANTZ (2010). Context-Aware Multimodal Human-Computer Interaction. *Interactive Collaborative Information Systems.*

NGUYEN, PHUNG, ADAMS, TRAN, VENKATESH (2010). Classification and Pattern Discovery of Mood in Weblogs. *Advances in Knowledge Discovery and Data Mining*.

ROUILLARD, TARBY (2011). How to communicate smartly with your house? *Journal Ad Hoc and Ubiquitous Computing.*

SOTO, FLORES HERNÁNDEZ, DE LOS ÁNGELES, DIEZ (MICAI 2009). Using Ontologies to generate Learning Objects automatically.

VAN WILLEGEN, ROTHKRANTZ, WIGGERS (ICTSD 2009). Lexical Affinity Measure between Words.

VARGA, FURLAN, JAKUS, MILUTINOVIC (2010). Document Filter Based on Extracted Concepts. *Transactions on Internet Research.*

TRAVAUX EN COURS
DÉSAMBIGUÏSATION & ISS

Désambiguïsation

- Gestion des ambiguïtés avec résolution tardive
  - Certaines ambiguïtés sont levées au plus tôt
  - La résolution d'autres peut être retardée
- Implémentation de plusieurs algorithmes
  - Désambiguïsation lexicale : Lesk étendu
  - Résolution d'anaphores : Lappin & Leass et Mitkov
  - Désambiguïsation de rattachement prépositionnel à l'aide de n-grammes
- Expériences préliminaires

Interface syntaxique-sémantique

- Construction automatique un module d'ISS indépendant de l'analyseur syntaxique dont on exploite les résultats
  - Automatiser l'adaptation à un nouvel analyseur syntaxique
  - CHAUMARTIN, KAHANE (TALN 2010) « Une approche paresseuse de l'analyse sémantique (construire une ISS à partir d'exemples) »
- Méthode pour induire des règles de calcul de structure sémantique à partir d'exemples annotés
  - Annotation syntaxique produite par l'analyseur choisi
  - Extraction de portion de structures qui sont des règles de l'ISS
  - Mécanisme de « soustraction de règles »

SYNTHÈSE ET PERSPECTIVES

Synthèse

- Plate-forme opérationnelle
  - Utilisée par l'équipe Proxem (millions de documents traités)
  - Mais aussi par d'autres équipes académiques ou industrielles
- Environnement intégré offrant une bonne productivité
  - Composant Antelope d'extraction terminologique : 400 lignes
  - Acabit (Daille, 1994) : 4 000 lignes de code par langue
- Composante d'ingénierie marquée
- Contribution méthodologique à l'ingénierie linguistique
  - Intégration de ressources de large couverture hétérogènes
  - Démarche semi-supervisée d'acquisition de connaissances
- Une dizaine de publications (conférences et revues)

Perspectives

- Travaux futurs
  - Poursuivre le développement de l'ISS
    - Trouver les failles des modèles théoriques et les combler
    - Améliorer la désambiguïsation
  - Extension à d'autres langues
- Enseignements tirés
  - La création et la maintenance d'une plate-forme nécessitent un effort significatif sur la durée
  - Nécessaire partage des ressources