Modèles neuronaux pour la modélisation statistique de la langue

Introduction

Les modèles de langage ont pour but de caractériser et d'évaluer la qualité des énoncés en langue naturelle. Leur rôle est fondamentale dans de nombreux cadres d'application comme la reconnaissance automatique de la parole, la traduction automatique, l'extraction et la recherche d'information. La modélisation actuellement état de l'art est la modélisation "historique" dite n-gramme associée à des techniques de lissage. Ce type de modèle prédit un mot uniquement en fonction des $n-1$ mots précédents. Pourtant, cette approche est loin d'être satisfaisante puisque chaque mot est traité comme un symbole discret qui n'a pas de relation avec les autres. Ainsi les spécificités du langage ne sont pas prises en compte explicitement et les propriétés morphologiques, sémantiques et syntaxiques des mots sont ignorées. De plus, à cause du caractère éparse des langues naturelles, l'ordre est limité à $n = 4$ ou 5. Sa construction repose sur le dénombrement de successions de mots, effectué sur des données d'entrainement. Ce sont donc uniquement les textes d'apprentissage qui conditionnent la pertinence de la modélisation n-gramme, par leur quantité (plusieurs milliards de mots sont utilisés) et leur représentativité du contenu en fonction de thématique, époque ou de genre.

L'usage des modèles neuronaux ont récemment ouvert de nombreuses perspectives. Le principe de projection des mots dans un espace de représentation continu permet d'exploiter la notion de similarité entre les mots: les mots du contexte sont projetés dans un espace continu et l'estimation de la probabilité du mot suivant exploite alors la similarité entre ces vecteurs. Cette représentation continue confère aux modèles neuronaux une meilleure capacité de généralisation et leur utilisation a donné lieu à des améliorations significatives en reconnaissance automatique de la parole et en traduction automatique.

Pourtant, l'apprentissage et l'inférence des modèles de langue neuronaux à grand vocabulaire restent très couteux. Ainsi par le passé, les modèles neuronaux ont été utilisés soit pour des tâches avec peu de données d'apprentissage,

soit avec un vocabulaire de mots à prédire limités en taille. La première contribution de cette thèse est donc de proposer une solution qui s'appuie sur la structuration de la couche de sortie sous forme d'un arbre de classification pour résoudre ce problème de complexité. Le modèle se nomme *Structure OUtput Layer* (SOUL) et allie une architecture neuronale avec les modèles de classes. Dans le cadre de la reconnaissance automatique de la parole et de la traduction automatique, ce nouveau type de modèle a permis d'obtenir des améliorations significatives des performances pour des systèmes à grande échelle et à état l'art. La deuxième contribution de cette thèse est d'analyser les représentations continues induites et de comparer ces modèles avec d'autres architectures comme les modèles récurrents. Enfin, la troisième contribution est d'explorer la capacité de la structure SOUL à modéliser le processus de traduction. Les résultats obtenus montrent que les modèles continus comme SOUL ouvrent des perspectives importantes de recherche en traduction automatique.

La structure de cette thèse est organisée de la manière suivante. D'abord, dans le Chapitre 1, nous présentons les principes de la modélisation du langage ainsi que ses approches de l'état de l'art. Puis ses avantages et ses inconvénients sont introduits afin d'analyser l'impact et la contribution de l'approche basée sur les modèles neuronaux et les espaces de représentations continus . Dans le Chapitre 2, une nouvelle architecture pour les modèles neuronaux qui s'appelle Structured OUtput Layer (SOUL) est proposée. Dans ce chapitre, nous présentons une vue d'ensemble de cette approche, tandis que le Chapitre 3 donne une description détaillée du modèle et spécifie les propriétés de la structure SOUL. La représentation continues des mots du modèle est analysée dans le Chapitre 4, tandis que le Chapitre 5 explore la capacité de ces modèles à prendre en compte un large contexte. Enfin son extension pour le modèle de traduction est finalement examinée et évaluée dans le Chapitre 6.

Chapitre 1: Modélisation du langage et des approches de l'état de l'art

Dans ce chapitre, nous introduisons la modélisation du langage et nous décrivons les approches qui fondent actuellement l'état de l'art. La littérature dans ce domaine est très riches et couvrent plus de deux décennies de recherche. Ainsi, nous nous concentrons sur l'analyse de la relation entre les approches standards basées sur une représentation discrète des mots et celles plus récentes qui projettent les mots dans un espace continu.

Définition Le modèle de langage joue un rôle important dans plusieurs systèmes de traitement automatique du langage. Par exemple, pour la reconnaissance automatique de la parole, en notant X le signal de parole à reconnaitre,

et W une phrase hypothèse quelconque (une séquence de mots), l'objectif est de trouver la phrase la plus probable $\hat{W}$ donnée par l'équation suivante :

$$\hat{W} = \underset{W}{\text{argmax}}\ P(W|X) \tag{1}$$

$$= \underset{W}{\text{argmax}}\ \frac{P(X|W) \times P(W)}{P(X)} \tag{2}$$

$$= \underset{W}{\text{argmax}}\ P(X|W) \times P(W) \tag{3}$$

La recherche de $\hat{W}$ nécessite la construction d'un espace de recherche énorme et l'évaluation des hypothèses de phrases à partir des modèles statistiques. Les modèles acoustiques basés sur les modèles de Markov cachés permettent d'estimer $P(X|W)$ tandis que $P(W)$ est évaluée à partir du modèle de langage. Le rôle d'un modèle statistique du langage dans cette tâche est donc multiple, il permet :

- l'estimation de $P(W)$ pour toutes séquences de mots issues d'un vocabulaire fini;
- de guider la construction de l'espace de recherche;
- et il doit contribuer à la construction d'une phrase en langage naturel.

Le rôle du modèle de langage est équivalent en traduction automatique.
La probabilité d'une séquence de mots peut se décompose grâce à l'équation suivante:

$$P(w_1^T) = \prod_{t=1}^{T} P(w_t|w_1^{t-1}), \tag{4}$$

où w_t est le t-ème mot, est la sous-séquence $w_i^j = (w_i, \ldots, w_j)$. Le modèle de langage est souvent construit sur une hypothèse de chaine de Markov d'ordre n: le mot ne dépend que $n-1$ mots précédents. C'est à dire:

$$P(w_t|w_1^{t-1}) \approx P(w_t|w_{t-n+1}^{t-1}) \tag{5}$$

$P(w_t|w_{t-n+1}^{t-1})$ est estimée selon le maximum de vraisemblance par ratio de fréquences. Les problèmes généraux du modèle de langage sont les suivants.

- La complexité: le nombre de paramètres est très grand. Par exemple, selon Bengio dans (Bengio et al., 2003), si on veut modéliser la distribution jointe de 10 mots consécutifs avec un vocabulaire contenant 100.000 mots, il y a potentiellement $100.000^{10} - 1 = 10^{50} - 1$ paramètres libres.

- La nécessité de beaucoup de données d'apprentissage afin de pouvoir observer toutes les séquences de n mots.
- La généralisation: malgré la grande quantité de données utilisées, de nombreux n-gram ne peuvent être observés lors de l'apprentissage; des techniques de lissage existent pour y remédier mais sont insuffisantes.

Critère de l'évaluation Pour un modèle de langage, la perplexité est souvent utilisée. Cette mesure est l'exponentiel de la log-vraisemblance. Plus elle est petite, meilleure c'est. La perplexité d'un modèle de langage pour un texte **D** se calcul ainsi:

$$2^{\frac{-\sum_{w_1^T \in \mathbf{D}} \log \hat{P}(w_1^T)}{N}} \tag{6}$$

Les probabilités sont estimées à partir du corpus d'apprentissage, puis sont comparées à celles observées sur le texte **D**. Le terme exponentiel dans l'équation de la perplexité peut être considéré comme la cross-entropie des deux distributions.
On peut également évaluer un modèle de langage d'une façon indirecte mais plus pertinente en utilisant des systèmes concernées. Par exemple, on peut utiliser le taux de reconnaissance de la parole comme un critère d'évaluation. Bien que le lien entre la perplexité et ce taux d'erreur soit sujet à caution, on observe que le modèle qui réduit la perplexité améliore souvent le système de reconnaissance de la parole (Rosenfeld, 2000).

Modèle n-gram Chaque probabilité du modèle n-gram est estimée grâce à la fréquence relative calculée sur un corpus d'apprentissage en utilisant plusieurs fonctions d'interpolation. Le modèle n-gram a des inconvénients:

1. Il est inefficace avec n grand ($n > 5$) à cause de l'éparsité des données, i.e., il est impossible d'observer toutes les séquences de n mots et la plupart des séquence apparaissent rarement.
2. La structure inhérente au langage est ignorée, en particulier, la notion de similarité entre les mots. Par exemple, si on observe les phases suivantes dans les données d'apprentissage:
   - L'étudiant commence son stage recherche.
   - Le thésard démarre sa thèse.

   Que dire de la phrase?
   - Le thésard commence sa thèse.

Parce qu'il n'y a aucun lien entre deux mots (commence, démarre), on ne peut pas prédire efficacement sa probabilité. C'est un défaut de généralisation.

Réseau neuronal classique L'article (Bengio et al., 2003) introduit l'usage des réseaux neuronaux pour la modélisation du langage. L'idée principale est d'associer (ou de projeter) chaque mot à un vecteur dans un espace continu, puis d'utiliser le réseau neuronal pour apprendre la distribution comme une fonction des vecteurs. Ses propriétés sont:

1. La représentation continue des mot est apprise par le modèle.
2. La distribution à estimer est une fonction continue de la similarité entre les représentations continues des mots:
   (a) Si deux mot à prédire ont des représentations similaires dans un même contexte, leurs probabilités seront similaires.
   (b) Ainsi, l'observation d'un n-gram permet d'apprendre des informations pour des exemples similaires.
3. L'algorithme d'entrainement est la descente de gradient stochastique.

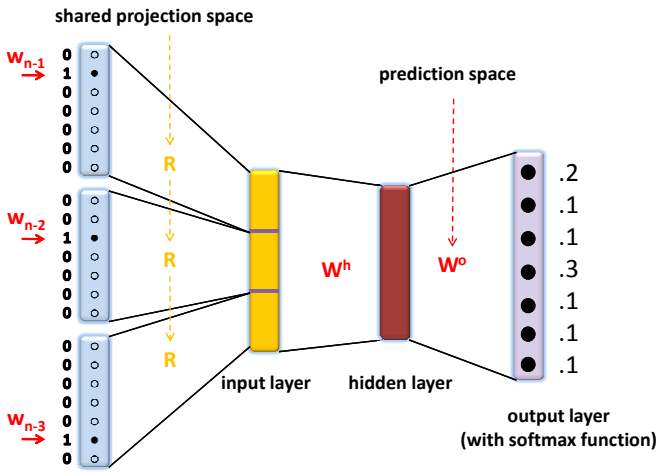

Figure 1: *Modèle du réseau neuronal classique pour la prédiction de* $P(w_n|w_1^{n-1})$

La Figure 1 représente une architecture neuronale permettant la mise en œuvre d'un tel modèle. Le réseau comporte trois couches. D'abord, les mots du contexte (w_1^{n-1}) sont projetés dans un espace continu afin d'obtenir leurs représentations vectorielles. Puis, les vecteurs du contexte sont concaténés pour créer l'entrée de la couche cachée. La fonction d'activation de la couche

cachée est non linéaire (tangent hyperbolique ou sigmoid). Sur la couche de sortie, chaque neurone correspond à la probabilité d'un mot (calculée par la fonction softmax afin garantir que la somme de toutes les probabilités soit égale à 1). La taille de cette dernière couche est donc égale à la taille du vocabulaire.

L'obstacle prépondérant est le temps d'exécution, surtout dans la phase d'entrainement. Selon (Schwenk, 2007), le calcul se concentre à la couche de sortie et le temps de calcul est donc linéaire par rapport à la taille du vocabulaire. Comme dans l'article (Schwenk, 2007), pour l'abaisser, il est possible d'utiliser une *shortlist*, i.e, seuls les mots les plus fréquents sont prédits.

Chapitre 2: Structured Ouput Layer(SOUL)

L'apprentissage et l'inférence des modèles de langage neuronaux à grand vocabulaire restent très couteux. Pour résoudre ce problème, Une nouvelle architecture est introduite dans (Le et al., 2011), qui s'appelle *Structured OUtput Layer* (SOUL). Elle se base sur une représentation hiérarchique du vocabulaire de sortie. Ce type de modèle combine deux approches qui font leurs preuves: les modèle neuronaux et les modèles de classes de mots. En utilisant l'information des classes des mots, la couche de sortie est structurée pour que l'estimation de la distribution soit faisable avec les vocabulaires de n'importe quelle taille.

La structure hiérarchique générale Pour factoriser la probabilité conditionnelle, le vocabulaire de sortie est structuré par un arbre où chaque mot est associé à un chemin unique de la racine à sa feuille. Si U est le profondeur de l'arbre, La séquence $x_0^U(w_n) = x_0, \ldots, x_U$ peut être utilisée pour encoder le chemin du mot w_n . Dans la séquence $x_0^U(w_n)$, x_0 est la racine, les nœuds x_u pour $u = 1, \ldots, U-1$ correspondent aux classes et sous-classes de w_n et x_U est sa feuille associée. La probabilité de w_n sachant son histoire w_1^{n-1} est calculée comme suit:

$$\begin{aligned} P(w_n|w_1^{n-1}) &= P(x_0^U(w_n)|w_1^{n-1}) \\ &= P(x_0|w_1^{n-1}) \times \prod_{u=1}^{U} P(x_u|w_1^{n-1}, x_0^{u-1}) \end{aligned} \tag{7}$$

La structure SOUL Dans la structure SOUL, l'arbre utilisé et peu profond afin de ne pas dégrader les performances comme cela a été observé avec des arbres binaires. De plus, comme le modèle de langage neuronal utilisant une shortlist a montré des améliorations significatives pour des systèmes de l'état de l'art, nous pensons qu'avec cette structure hiérarchique les performances ne pourront qu'être améliorées. Dans la structure SOUL, les mots de la shortlist

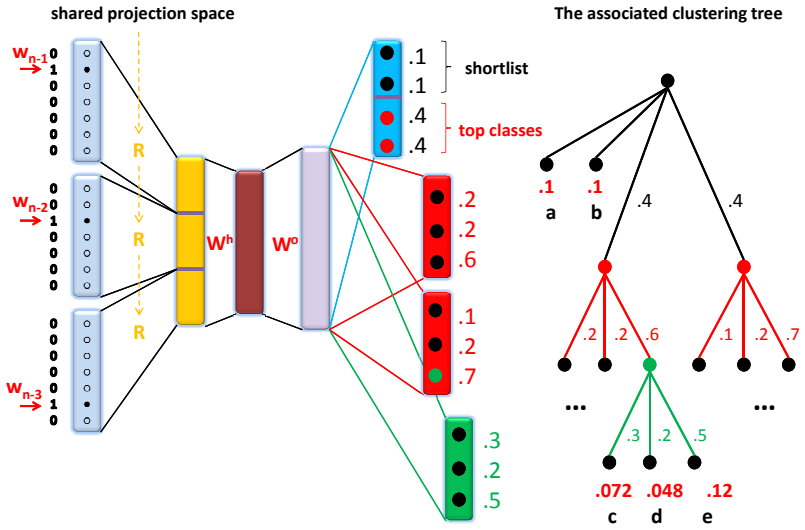

Figure 2: *L'architecture du modèle SOUL.*

(les mots plus fréquents) sont gardés au premier niveau de l'arbre, c'est à dire qu'il forme leurs propres classes.

Un exemple pour l'architecture SOUL est présenté à la Figure 2. Les différences par rapport à une architecture standard se situent au niveau de la sortie qui se compose de plusieurs couches utilisant la fonction softmax comme activation. Cet ensemble de couches softmax se divisent en deux parties:

1. *La couche principale* qui estime $P(x_1|w_1^{n-1})$, la probabilité des classes et des mots de la shortlist;

2. *Les autres couches softmax* estiment $P(x_u|w_1^{n-1}, x_0^{u-1}), u = 2 \dots (U)$, la probabilité des sous-classes et des mots hors de la shortlist.

Ces modèles sont évalués au sein de systèmes de reconnaissance de la parole à grande échelle qui ont obtenus des résultats "état de l'art" lors des évaluations internationales GALE pour le Mandarin et l'Arabe. Ce nouveau type de modèle a permis d'obtenir des améliorations significatives des performances en termes de perplexité et de taux de d'erreur. Ces gains ont également été observés dans le cadre de la traduction automatique pour la tâche à grande échelle définie par les évaluations internationales WMT 2011 et WMT 2012.

Chapitre 3: Analyse sur la configuration de la structure SOUL

Ce chapitre propose d'étudier la configuration du modèle neuronal SOUL afin de mieux comprendre ses caractéristiques. Nos observations sont qu'il faut d'une part traiter séparément les mots les plus fréquents et les mots plus rares, et d'autre part que la shortlist peut être de taille petite. En effet les expériences montrent que l'on peut obtenir des perplexités comparables en faisant varier la taille de la shortlist ($8k$, $12k$, $25k$ or $56k$ mots). Deuxièmement, avec une shortlist de $8k$, le nombre de classes principales pour les mots qui sont pas dans la shortlist (de 128 à $4k$) et la profondeur de l'arbre (de 1 à 3 niveaux) n'ont pas d'influence importante sur perplexité. Par contre l'usage de l'arbre est important puisqu'il permet un entrainement plus rapide et les perplexités obtenues sont meilleures que celles obtenues avec un modèle n'utilisant pas d'arbre.

Dans le cadre de reconnaissance automatique de la parole, des gains significatifs ont été constatés avec l'architecture SOUL par rapport à l'approche qui se base sur la shortlist. Le temps d'apprentissage et d'inférence est de plus similaire. Plusieurs améliorations de la structure SOUL sont également proposées. Les expériences montrent qu'avec une structure de sortie de taille réduite (à partir de $8k$ à $2k$ pour la shortlist et de $4k$ à $2k$ pour les classes principales), les temps d'inférence et d'entrainement peuvent être considérablement réduits sans impact sur les performances de reconnaissance. Ainsi, les architectures avec la nouvelle configuration obtiennent de meilleures performances et sont deux fois plus rapide que ceux basés sur la shortlist.

Les limites de profondeur des architectures de réseaux neuronaux sont ensuite examinées. Les résultats expérimentaux montrent qu' utiliser une grande couche cachée supplémentaire conduit à 5% de réduction de la perplexité. Par contre, l'utilisation de deux couches cachées n'apporte pas d'améliorations supplémentaires. Cela peut être expliqué par l'absence de pré-entrainement qui empêche modèles profonds d'échapper à des extrema locaux peu satisfaisants. Il semble donc prometteur d'explorer des techniques avancées pour l'apprentissage profond.

Chapitre 4: Analyse des espaces de représentation continus

Dans ce chapitre, nous analysons d'abord l'impact des espaces de mots induits par le modèle neuronal classique et le modèle log-bilinéaire sur les performances du système afin de montrer que, le modèle log-bilinéaire ne surpasse pas le modèle classique. Dans l'approche classique, il y a deux espaces de mots

utilisés pour représenter les deux rôles différents qu'un mot joue dans la modélisation du langage (contexte et prédiction). En outre, grâce à leur similitude, la convergence plus rapide de l'espace de prédiction peut aider la convergence de l'espace de contexte. Trois nouvelles méthodes d'entrainement sont proposées et se sont montrées efficaces. Ce travail met en évidence l'impact de l'initialisation et de la manière d'entrainer les modèles de langage neuronaux. Nos résultats expérimentaux et nos méthodes d'entraînement sont liées aux techniques de pré-entraînement pour les modèles neuronaux.

Nous étudions ensuite les propriétés des espaces de représentation continue qui sont induits par l'apprentissage des modèles SOUL pour différentes langues. Nos observations montrent que les similarités qui sont captées relèvent de différents aspects linguistiques: certaines régions de l'espace regroupent les mots selon des critères syntaxiques (adverbe par exemple) alors que pour d'autres régions les propriétés sémantiques semble être plutôt prisent en considération (les jours de la semaine par exemple). Les représentations de mots fournies par les modèles SOUL sont également intéressantes pour d'autres tâches du traitement de langage naturel. Ainsi des résultats prometteurs sont obtenus dans nos tentatives pour la tâche de mesurer la similarité sémantique entre les mots.

Chapitre 5: Étude le l'impact de la longueur du contexte

Dans ce chapitre, nous étudions plusieurs types de modèles neuronaux afin d'évaluer l'influence des dépendances dans la tâche de modélisation du langage: du modèle récurrent qui peut par récursivité traiter un contexte de longueur arbitraire au modèle n-gramme, avec n variant entre 4 et 10.

Dans un premier temps, nous proposons une méthode expérimentale afin de quantifier dans un modèle neuronal n-gram, l'influence des mots du contexte selon leurs positions dans le contexte. Les résultats montrent que d'une part, plus le mot du contexte est loin du mot à prédire, moins il a d'influence, d'autre part au-delà de 8 mots avant l'influence du contexte devient statistiquement négligeable. Par conséquent, l'hypothèse n-gramme avec $n \approx 10$ semble être justifiée. Puis, en introduisant le modèle pseudo-récurrent et en limitant le contexte du modèle récurrent à le phrase actuelle, il est possible d'utiliser des techniques d'optimisation utilisées pour les modèles n-grammes. Ainsi le temps d'entrainement d'un modèle récurrent peut être divisé par un facteur de 8 sans perte significative sur les performances. Nous avons comparé les modèles n-gramme et récurrents dans une tâche à grande échelle, avec des données monolingues contenant $\approx$ 2.5 milliards de mots. Les résultats expérimentaux montrent que l'utilisation de dépendances à longue portée

($n = 10$) avec un modèle de langage SOUL surpasse de manière significative le modèle de langage conventionnels. Dans ce contexte, l'utilisation d'une architecture neuronal récurrente ne donne pas d'améliorations sur une architecture neuronale n-gramme, à la fois en termes de perplexité et de BLEU.

Notre conclusion est que le problème principal des modèles n-gramme ne semblent pas être l'hypothèse d'indépendance conditionnelle, mais l'utilisation de valeurs trop petites pour l'historique. Tous ces résultats suggèrent également que, dans l'avenir, de nouvelles méthodes de modélisation du langage doit prendre en compte les dépendances longues.

Chapitre 6: Modèle de traduction neuronal

Le modèle de traduction à base de n-gramme Casacuberta and Vidal, 2004; Crego, Yvon, and Mariño, 2011 propose une factorisation de la probabilité jointe d'une paire de phrases (source et cible) comme suivant:

$$P(\mathbf{s}, \mathbf{t}) = \prod_{i=1}^{L} P(u_i | u_{i-1}, \ldots, u_{i-n+1}) \tag{8}$$

Un des problème majeur est que les unités considérées sont des paires bilingues (l'association d'un segment de mots source avec un segment de mots cible), le vocabulaire et le nombre de paramètres sont donc très grands, même pour une tâche de taille réduite. Le deuxième problème est que dans l'équation (8), les deux langues (source et cible) jouent des rôles symétriques, alors que le côté de source est connu et que le côté de cible doit être prédit.

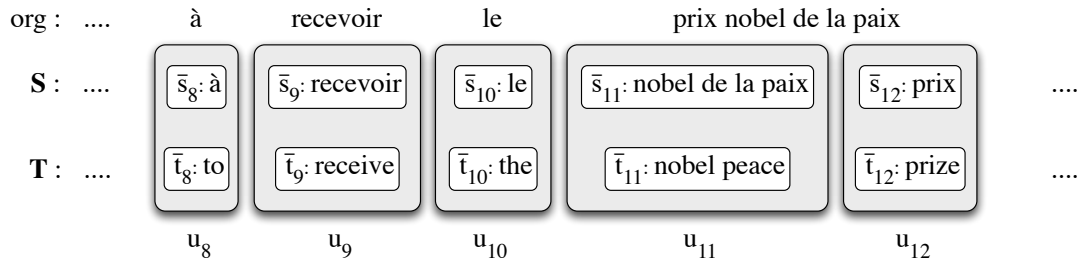

Figure 3: *Une paire de phrases segmentée en unités bilingues.*

Nous proposons trois factorisations de cette probabilité jointe et donc trois manière d'utiliser les réseaux neuronaux comme modèle de traduction. Ces trois factorisations et donc les trois modèles résultants se distinguent par le choix des unités de traduction, respectivement la paire de segment, le segment

et le mot. Ces décompositions successives permettent de réduire la taille du vocabulaire des unités à prédire tout en dissociant le rôle de chacune des langues. Ainsi, la factorisation qui semble la plus performante dissocie le côté source du côté cible et considère le mot comme unité dans chacune des langues:

$$P(\mathbf{s},\mathbf{t}) = \prod_{i=1}^{L} \Big[\prod_{k=1}^{|\bar{t}_i|} P\big(t_i^k | h^{n-1}(t_i^k), h^{n-1}(s_{i+1}^1)\big) \times \prod_{k=1}^{|\bar{s}_i|} P\big(s_i^k | h^{n-1}(t_i^1), h^{n-1}(s_i^k)\big) \Big] \tag{9}$$

Le cadre d'évaluation utilisé est dérivé de celui de l'évaluation internationale IWSLT 2011. L'objectif est de traduire des *TED talks* de l'anglais vers le français. Dans ce cadre nous avons étudié différents régimes d'entraînement de ces nouveaux modèles et nous avons explorer ses possibilités d'adaptation. Les résultats expérimentaux montrent que l'adaptation à partir du modèle de traduction générale du type SOUL est un moyen efficace et rapide. En utilisant toutes ces nouveautés, nous obtenons finalement une amélioration de 1, 6 point de BLEU.

En outre, cette approche a également été expérimenté dans les systèmes que nous avons soumis à la tâche de traduction de WMT 2012. Les résultats obtenus dans une configuration à grande échelle et pour des paires de langues différentes sont du même ordre.

Enfin, même si nos modèles proposés ont été initialement conçus pour fonctionner uniquement dans le cadre du système de traduction basé sur l'hypothèse n-gramme, les introduire dans système *phrase based* conventionnel (comme Moses) est non seulement faisable mais apporte également des gains significatifs.

Conclusion

Contributions

Depuis deux décennies, l'état de l'art en modélisation statistique du langage est dominé par les modèles n-grammes avec repli. Et si leurs limitations théoriques et pratiques sont connues, ce type de modèles sont les plus adaptés aux quantités de données toujours croissantes. Une des limitations est que les mots sont considérés comme des réalisations de variables aléatoires discrètes. Ainsi, les propriétés linguistiques ne sont pas explicitement prises en compte et la capacité de généralisation de tels modèles est limitée. Dans cette thèse nous nous intéressons à une des alternatives récentes basée sur les réseaux de neurones. Le principe est de projeter les mots dans un espace de représentation continu et

d'exploiter ainsi la notion de similarité entre les mots: les mots du contexte sont projetés dans un espace continu et l'estimation de la probabilité du mot suivant exploite alors la similarité entre ces vecteurs. Cette représentation continue confère aux modèles neuronaux une meilleure capacité de généralisation et leur utilisation a donné lieu à des améliorations significatives en reconnaissance automatique de la parole et en traduction automatique.

Dans cette thèse nous avons proposé un nouveau type de modèle de langage nommé SOUL (Structured OUtput Layer) basé sur les réseaux de neurones qui peut utiliser un vocabulaire de taille arbitraire. Pour cela, le vocabulaire des mots à prédire est structurer selon un arbre de classification et le réseau de neurones prédit un chemin dans l'arbre plutôt que directement un mot. Cette approche a été évaluée à la fois en reconnaissance automatique de la parole et en traduction automatique et les gains obtenus sont significatifs. De plus une étude approfondie du modèle SOUL montre l'impact du contexte et de sa longueur sur la qualité de la prédiction et comment le modèle SOUL se compare avec les modèles récurrents. Enfin nous avons proposé un modèle de traduction qui grâce à l'architecture SOUL peut être appliqué à des tâches de grande échelle et qui permet d'obtenir des améliorations significatives.