Utilisation des documents en provenance d'Internet pour l'apprentissage de modèles de langage

Dominique Vaufreydaz

Utilisation des documents en provenance d'Internet pour l'apprentissage de modèles de langage

Dominique Vaufreydaz
Laboratoire CLIPS-IMAG, équipe GEOD
Université Joseph Fourier
Campus Scientifique, BP 53 38041 Grenoble Cedex 9 France
tel: .: 33 + (0)4.76.51.45.26 - Fax: 33+ (0)4.76.63.55.52
email: Dominique.Vaufreydaz@imag.fr

1. INTRODUCTION

L'un des composants principaux d'un système de reconnaissance de la parole (SRAP) est le modèle de langage (ML). Son rôle est de réduire l'espace de recherche pour accélérer le processus de reconnaissance. Dans un SRAP basé sur une reconnaissance phonémique, le ML est aussi un pont entre les représentations textuelle et phonétiques des mots. Les ML sont calculés sur des textes de grande taille tout en contrôlant le vocabulaire, la longueur employée (2, 3 ou plus de mots) pour les *contextes* et les méthodes de *backoff* pour l'intégration des mots hors vocabulaire. L'adéquation d'un ML à une tâche est généralement décrite par la mesure de la *perplexité* [1]. La taille du corpus est déterminante pour le calcul du ML. En effet, en l'augmentant, il est possible d'accroître le nombre de contextes rencontrés pendant l'apprentissage, et d'avoir une meilleure estimation des probabilités qui leur sont associées. Ainsi, il est évident que les textes écrits ne peuvent pas directement servir au calcul de ML appropriés à la reconnaissance de la parole spontanée.

Pour cette raison, nous portons notre attention sur d'autres sources de textes pour l'apprentissage : les documents en provenance d'Internet. En effet, de nos jours, beaucoup de personnes accèdent à Internet dans le cadre de leur travail, de leur école, ou à partir de chez eux. Non seulement ils utilisent les informations du réseau mais ils publient aussi leurs propres documents. Ceux-ci sont donc de différentes natures (professionnelle, personnelle, et.) et représentent diverses manières de s'exprimer. Il est donc possible d'y trouver un vocabulaire et des expressions de la vie courante qui ne sont pas présentes dans les textes écrits.

Dans la section 2, nous décrivons comment, à l'aide de robots et de filtres appropriés, nous avons collecté beaucoup de données utilisables pour le calcul de ML. Dans la section 3, nous étudions les résultats, en termes de *nombre de contextes* et de *perplexité*, exprimés en fonction de la taille du corpus.

2. ANALYSE DES DONNÉES

2.1 Collecte et traitements des données

Nous avons la possibilité d'utiliser plusieurs sources Internet de documents. Pour simplifier notre tâche, nous n'avons conduit nos expérimentations qu'en utilisant des pages *Web* et des articles postés sur les *news*. Nous avons pour cela employé deux robots développés par nos soins : **CLIPS-Index** et **CLIPS-News**. Nous avons obtenu deux corpus nommés respectivement **WebFr** (1550000 documents) et **NewsFr** (400000). Ces deux corpus représentent 16 Go de données.

2.2 Génération de corpus textuel

Toutes ces données, acquises à partir d'Internet, ne sont pas sous une forme utilisable directement pour le calcul de ML: nombre de tags et autres diacritiques sont superflus et doivent être enlevés. De plus, il est à noter que, même si nous avions ciblé nos collectes, les documents ne sont pas tous en français ou sont multilingues. Le premier pas est d'extraire les textes français de ces documents. Nous avons choisi BDLex [2], un dictionnaire de 245000 entrées, comme lexique. Il a été enrichi par le dictionnaire de l'ABU [3] (Association des Bibliophiles Universels) pour atteindre 400000 formes lexicales.

L'extraction en elle-même commence par un filtrage des documents et produit un texte formaté. Certains tags sont utilisés pour avoir une meilleure correspondance entre les données originales et la sortie obtenue. Ainsi, les débuts et fins de phrases ne sont pas toujours clairement indiqués. Il est possible d'en extrapoler un certain nombre en utilisant les marqueurs de tableau, de titre ou de paragraphe . Dans la suite du processus, pour augmenter la qualité du texte produit, des corrections sont apportées. Par exemple 'Ecole' devient 'école'. Dans cet exemple, la première lettre avait perdu l'accentuation du fait de la mise en majuscule. Plus simplement, le programme recherche le mot de remplacement le plus proche. A cette étape, le texte est aussi mis entièrement en minuscule. La dernière étape est celle du traitement, en fonction du vocabulaire, de ce que nous avons appelé *les blocs minimaux*. Un bloc minimal d'ordre *n* est une séquence consécutive de mots du vocabulaire. Il est aussi possible de ne demander que des phrases complètes. Pour étudier l'impact de ces options, nous avons conduit des expériences avec un lexique limité à 3000 mots. Le tableau ci-dessous présente ces résultats.

Options choisies	Taille du corpus obtenu
bloc minimal = 3 mots	145 Millions de mots
bloc minimal = 5 mots	47 Millions de mots
phrases complètes de longueur minimale 5	46500 mots

Tableau 1: Impact des options de filtrage sur la taille du corpus avec un vocabulaire de 3000 mots.

Il est évident que plus les limitations sont fortes et plus la taille du corpus filtré diminue. Malgré cela, avec un vocabulaire suffisant, il est très facile d'obtenir de très grands corpus de plusieurs centaines de Mo.

3. EXPÉRIENCES

3.1 Contribution de WebFr et NewsFr

Nous avons comparé l'utilisation des pronoms personnels français dans trois sources différentes : le corpus de l'action Grace [4] (20 Mo de texte extrait du journal *Le Monde*), WebFr et NewsFr. La figure ci-dessous présente les résultats obtenus.

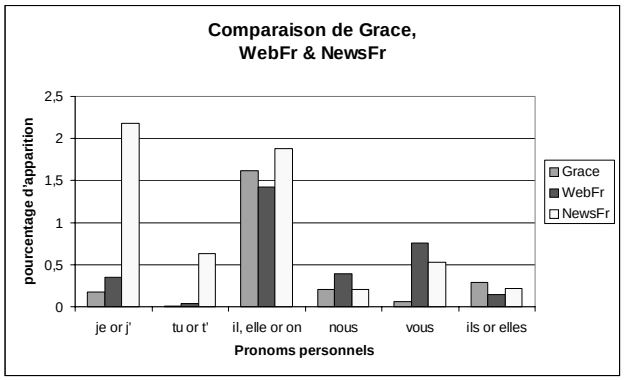

Figure 1: Utilisation des pronoms personnels français dans Grace, WebFr et NewsFr.

Il est aisé de voir que la richesse apportée par les données provenant d'Internet concerne les formes personnelles utiles dans les dialogues. Pourtant les autres formes sont au moins aussi représentées que dans les corpus écrits. Ces sources sont donc intéressantes en elles-mêmes ou conjuguées à des sources plus conventionnelles.

3.2 Evaluation de la taille de corpus

Pour évaluer la richesse des corpus WebFr et NewsFr, nous avons donc réalisé plusieurs mesures afin de voir l'évolution des caractéristiques des ML en fonction de la taille du corpus d'apprentissage. Celui-ci, décrit dans la section 2.2, comprend plus de 47 millions de mots. Nous l'avons découpé en paquets de 2 Millions de mots (avec une coupure au bloc minimal supérieur). Nous calculons à chaque fois le ML sur les *n* premiers paquets et réestimons le résultat avec ce nouveau ML. Ainsi il est possible de voir la progression du phénomène à étudier en fonction de la taille du corpus pour le calcul du ML.

3.2.1 Perplexité et nombre de trigrames

La mesure que nous détaillons dans cette partie, concerne la perplexité et le nombre de trigrames connus par le ML. Ainsi, comme cela a déjà été évoqué, l'augmentation de la taille du corpus n'a pour but que d'améliorer le ML en termes de diversité des contextes connus et d'adéquation du ML à une tâche. La figure 2 présente l'évolution de ces deux phénomènes en fonction de la taille du corpus d'apprentissage.

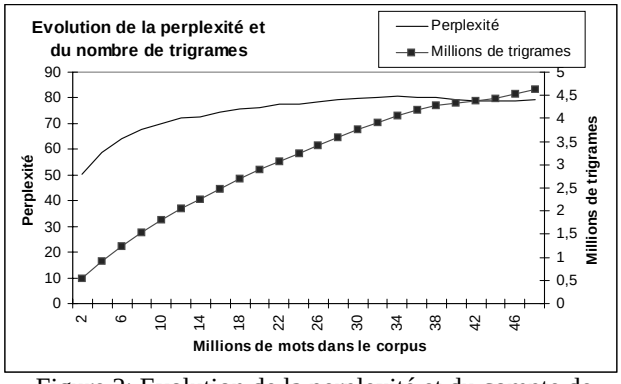

Figure 2: Evolution de la perplexité et du compte de trigrames.

Comme nous pouvons le constater sur cette figure, la perplexité évolue et se stabilise aux alentours de 80 (valeur raisonnable pour un vocabulaire de 3000 mots) après les 20 premiers millions. De plus, cette valeur décline légèrement sur la fin. Pourtant le nombre de contextes connus ne cesse de s'accroître. Nous pouvons donc dire que le ML s'améliore mais que la perplexité ne reflète pas cet état de chose. Ceci nous amène à critiquer la perplexité comme critère de qualité d'un ML. Certains auteurs, comme [5], décrivent d'autres métriques, ayant une meilleure corrélation avec la précision en termes de mots, pour évaluer un ML.

3.2.2 Impact sur la tâche de reconnaissance

Nous étudions maintenant l'influence de la taille du corpus pour le calcul d'un ML utilisé dans un SRAP. Le système choisi est le système RAPHAEL [6]. Ces tests ont porté sur des enregistrements de dialogues. L'évaluation se fait en calculant la précision en termes de mots. Nous l'avons comparé à deux corpus de référence: Grace et des textes extraits d'expériences de type Magicien d'Oz.

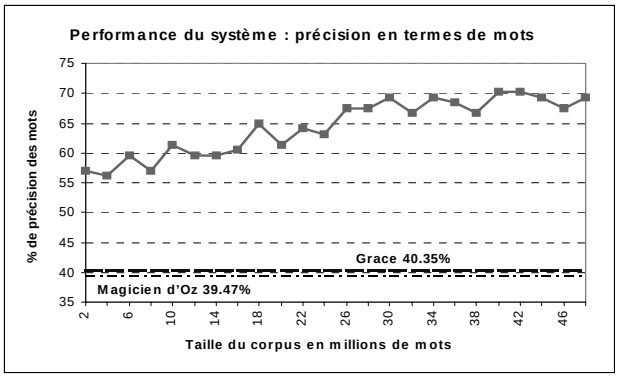

Figure 3: Gain de précision en termes de mots.

Sur la figure 3, nous pouvons constater que le gain absolu, par rapport aux corpus de référence, est important : au minimum de 15%. L'acroissement du corpus permet aussi d'augmenter ce gain.

4. CONCLUSION

Nous avons vu que notre méthode pouvait fournir de grand corpus pour l'apprentissage de ML. Ces données ont l'avantage de contenir des formes qui n'existent pas dans les textes écrits (Grace par exemple). Ces formes apportent de la richesse au ML et permettent une meilleure couverture des locutions couramment employées par les locuteurs. Notons, pour terminer, que ces données ont été utilisées avec succès pour la démonstration du projet CStar [7] (traduction automatique de dialogues parlés de renseignements touristiques) en Juillet 1999.

RÉFÉRENCES

[1] Rosenfeld R., "A maximum entropy approach to adaptive statistical language modeling", Computer, Speech and Language (1996).

[2] Pérennou G., De Calmès M., "BDLEX lexical data and knowledge base of spoken and written French", European conference on Speech Technology, pp 393-396, Edinburgh (Ecosse), Septembre 1987.

[3] voir http://cedric.cnam.fr/ABU/index.html .

[4] voir http://www.limsi.fr/TLP/grace/index.html .

[5] Ito A., Kohda M., Ostendorf M., "A new metric for stochastic language model evaluation", Eurospeech'99, pp. 1591-1594, Budapest (Hongrie), Septembre 1999.

[6] Akbar M., Caelen J., "Parole et traduction automatique: le module de reconnaissance RAPHAEL", COLLING-ACL'98, pp. 36-40, Montreal (Québec), August 1998.

[7] voir http://www.c-star.org/ .