TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Structure des trigrammes inconnus et lissage par analogie

Julien Gosme[1] Yves Lepage[2]
(1) GREYC, université de Caen Basse-Normandie, France
Julien.Gosme@unicaen.fr
(2) IPS, université Waseda, Japon
Yves.Lepage@aoni.waseda.jp

Résumé. Nous montrons dans une série d'expériences sur quatre langues, sur des échantillons du corpus Europarl, que, dans leur grande majorité, les trigrammes inconnus d'un jeu de test peuvent être reconstruits par analogie avec des trigrammes hapax du corpus d'entraînement. De ce résultat, nous dérivons une méthode de lissage simple pour les modèles de langue par trigrammes et obtenons de meilleurs résultats que les lissages de Witten-Bell, Good-Turing et Kneser-Ney dans des expériences menées en onze langues sur la partie commune d'Europarl, sauf pour le finnois et, dans une moindre mesure, le français.

Abstract. In a series of experiments in four languages on subparts of the Europarl corpus, we show that a large number of unseen trigrams can be reconstructed by proportional analogy using only hapax trigrams. We derive a simple smoothing scheme from this empirical result and show that it outperforms Witten-Bell, Good-Turing and Kneser-Ney smoothing schemes on trigram models built on the common part of the Europarl corpus, in all 11 languages except Finnish and French.

Mots-clés : analogie, trigrammes inconnus, trigrammes hapax, modèle de langue trigrammes, Europarl.

Keywords: proportional analogy, unseen trigrams, hapax trigrams, trigram language models, Europarl.

1 Introduction

Les techniques de lissage de modèles de langue reposent habituellement sur des hypothèses purement statistiques pour estimer la probabilité des évènements inconnus. Il y a dix ans, (Rosenfeld, 2000) constatait que :

> *Ironically, the most successful SLM techniques use very little knowledge of what language really is. The most popular language models (n-grams) take no advantage of the fact that what is being modeled is language.*

Nous présentons ici une technique de lissage pour les modèles de langue trigrammes qui repose sur la structure des évènements inconnus, c'est-à-dire la manière dont les trigrammes inconnus peuvent être construits à partir des trigrammes connus en utilisant une opération structurelle linguistiquement justifiée, l'analogie.

Le but du lissage des modèles de langue est d'attribuer des probabilités non-nulles aux évènements inconnus. Habituellement, les probabilités attribuées dépendent d'une caractérisation théorique des événements inconnus. L'hypothèse à l'origine de ce travail est que les trigrammes inconnus peuvent être caractérisés, dans une large mesure, par la similitude de leurs structures avec des trigrammes rares. Plus précisément nous montrons ci-dessous que, dans une large mesure, les trigrammes inconnus sont analogues aux trigrammes hapax.

En guise d'illustration, dans une de nos expériences préliminaires, le trigramme de mots *opportunité de servir* était un trigramme de notre jeu de test absent du corpus d'entraînement. Il se trouvait que ce trigramme pouvait être reconstruit par analogie à l'aide de trois trigrammes du corpus d'entraînement de la manière suivante :

opportunité de servir : *opportunité de* ~~*modifier*~~ :: ~~*qui pourrait*~~ *servir* : ~~*qui pourrait modifier*~~

La ligne précédente se lit ainsi : le trigramme inconnu *opportunité de servir* est au trigramme connu *opportunité de modifier* ce qu'un autre trigramme connu, *qui pourrait servir*, est à un dernier trigramme connu, *qui pourrait*

modifier. Les différents éléments du trigramme inconnu sont obtenus par similarité avec le second et le troisième trigrammes (*opportunité de* et *servir*) et peuvent être assemblés par différence avec le quatrième trigramme (mots barrés). En plus de permettre la reconstruction, les trois trigrammes ci-dessus étaient tous hapax dans le corpus d'entraînement.

La relation, telle celle donnée ci-dessus entre trigrammes de mots, qui énonce qu'A est à B ce que C est à D est appelée analogie. Un certain nombre de travaux en traitement automatique des langues exploitent l'analogie. Nous n'en citons que quelques-uns ici. Par exemple, sur des tâches de segmentation morphologique, (Lavallée & Langlais, 2010) ont récemment obtenu d'excellents résultats dans la découpe des mots par analogie. (Claveau & L'Homme, 2005), entre autres auteurs, avaient auparavant étudié, en faisant usage de l'analogie, dans quelle mesure la similarité liait la forme et le sens des mots : *connector* : *to connect* :: *editor* : *to edit*. En plus des analogies entre mots eux-mêmes, (Hathout, 2009) a récemment exploité les analogies entre définitions extraites du TLFi pour construire automatiquement des familles de mots liés par la forme et le sens. Dans le même ordre d'idée, (Langlais *et al.*, 2008) avaient proposé d'utiliser l'analogie pour forger de nouvelles équivalences terminologiques dans le domaine médical à cheval sur deux langues. Sur le seul plan sémantique, (Turney, 2008) a quant à lui présenté une approche générale au problème de l'association entre mots utilisant l'analogie entre vecteurs contextuels : *mason* : *stone* :: *carpenter* : *wood*, approche qu'il prétend généralisable aux relations de synonymie et d'antonymie. (Lepage & Denoual, 2005) quant à eux ont conçu un système de traduction automatique entièrement fondé sur l'analogie. Dans le cadre de la traduction automatique aussi, (Denoual, 2007) et (Langlais & Patry, 2007) ont montré la possibilité de traduire certains mots inconnus par analogie.

La définition de l'analogie que nous utilisons dans ce travail est détaillée et justifiée dans (Lepage, 2004). Nous l'appliquons aux trigrammes de mots. Un quadruplet de trigrammes de mots A, B, C et D est une analogie lorsque les contraintes suivantes sont vérifiées :

$$\begin{cases} d(A,B) &= d(C,D) \\ d(A,C) &= d(B,D) \\ |A|_m - |B|_m &= |C|_m - |D|_m, \forall m \end{cases}$$

Ici, d est la distance d'édition qui compte le nombre minimum d'insertions et de suppressions de mots nécessaires à la transformation d'un trigramme en un autre. [1] $|A|_m$ est le nombre d'occurrences du mot m dans le trigramme A. En reprenant l'exemple précédent :

$$\begin{array}{rcl} A &=& \textit{opportunité de servir} \\ B &=& \textit{opportunité de modifier} \\ C &=& \textit{qui pourrait servir} \\ D &=& \textit{qui pourrait modifier} \end{array}$$

on peut vérifier que $d(A,B) = d(C,D) = 2$ et $d(A,C) = d(B,D) = 4$. La relation entre nombres d'occurrences est vérifiée pour chaque mot :

mot m	$\lvert A\rvert_m - \lvert B\rvert_m = \lvert C\rvert_m - \lvert D\rvert_m$
opportunité	$1 - 1 = 0 - 0$
de	$1 - 1 = 0 - 0$
servir	$1 - 0 = 1 - 0$
modifier	$0 - 1 = 0 - 1$
qui	$0 - 0 = 1 - 1$
pourrait	$0 - 0 = 1 - 1$

Le bon sens veut que les trigrammes inconnus apparaissant dans un jeu de test qui peuvent être reconstruits par analogie avec des trigrammes d'un corpus d'entraînement soient considérés plus sûrs que ceux qui ne peuvent pas l'être. Une technique de lissage basée sur de simples décomptes devrait donc donner une plus forte ré-estimation aux trigrammes pouvant être reconstruits par analogie et une plus faible ré-estimation aux autres. Si, en plus, les trigrammes reconstruits peuvent l'être à l'aide de trigrammes hapax, la ré-estimation de leur effectif devrait être proche de 1 puisqu'ils sont alors proches structurellement des trigrammes hapax.

1. La distance d'édition de Levenshtein (Levenshtein, 1966) prend en compte la substitution comme opération d'édition supplémentaire.

La suite de l'article est divisée en deux parties : la section 2 est consacrée à vérifier l'hypothèse que les trigrammes inconnus sont structurellement analogues aux trigrammes hapax. Des expériences successives menées sur quatre langues européennes confirment, les unes après les autres, cette hypothèse. La section 3 présente alors une technique de lissage reposant sur cette propriété, et directement inspirée des techniques élémentaires de Lidstone et de Laplace. Des comparaisons effectuées avec quatre autres techniques de lissage classiques sur onze langues européennes montrent son efficacité, voire sa supériorité.

2 La structure des trigrammes inconnus d'un jeu de test

Nous menons des expériences sur les quatre langues suivantes : l'anglais, le français, l'allemand et le finnois. Ces langues ont été choisies pour leurs différentes richesses morphologiques. Sur une échelle croissante, on peut en effet placer successivement l'anglais, le français, l'allemand puis le finnois qui a la morphologie la plus riche.

Le corpus Europarl (Koehn, 2005) offre des textes alignés dans ces quatre langues,[2] ce qui permet de mener des expériences véritablement comparables. Pour les expériences de cette section, de l'ensemble de toutes les phrases correspondantes dans toutes les langues, nous avons extrait aléatoirement 100 000 phrases. Parmi elles, 90 000 phrases ont été sélectionnées aléatoirement, les mêmes dans toutes les langues, pour servir de corpus d'entraînement. Le jeu de test est constitué des 10 000 phrases restantes.

2.1 Proportion de trigrammes inconnus reconstruits

Pour vérifier dans quelle mesure les trigrammes inconnus d'un jeu de test peuvent être reconstruits par analogie, nous effectuons une première série d'expériences par validation croisée. Nous comptons simplement le nombre total de trigrammes inconnus du jeu de test reconstructibles par analogie à l'aide de trois trigrammes du corpus d'entraînement. Dès lors que la reconstruction est possible, le processus est interrompu pour ce trigramme.

Les résultats obtenus, reportés dans la table 1, montrent que la proportion de trigrammes inconnus dans le jeu de test reconstructibles par analogie avec trois autres trigrammes du corpus d'entraînement est supérieure à 80 % en anglais et en français et supérieure à 70 % en allemand. Cette proportion, appelée μ ici, est donc importante. Elle est calculée sur le nombre total de trigrammes inconnus différents (sans répétition) dont la proportion relativement au nombre total de trigrammes du jeu de test est appelée λ. Des expériences non rapportées ici montrent qu'en augmentant la taille des données, les valeurs de λ baissent tandis que les valeurs de μ augmentent. Les paramètres λ et μ seront exploités dans la section 3.1.

TABLE 1 – Nombre de trigrammes inconnus différents dans le jeu de test et proportion de trigrammes inconnus différents reconstructibles par analogie à l'aide de trois trigrammes du corpus d'entraînement.

	Trigrammes inconnus	
	du jeu de test (λ)	reconstruits (μ)
anglais	114,566 (60,04 %)	83,67 %
français	116,922 (57,81 %)	81,87 %
allemand	140,226 (68,97 %)	72,14 %
finnois	132,931 (83,33 %)	44,93 %

En finnois, la proportion de trigrammes inconnus différents reconstruits est faible avec seulement 45 %. Cette faible valeur est certainement explicable par la richesse morphologique de cette langue et donc l'absence relative de mots-fonctions permettant plus de commutations de mots dans les trigrammes. Nous pouvons dès à présent nous attendre à des résultats différents en finnois dans toute la suite de notre étude.

2. http://www.statmt.org/europarl

2.2 Patrons d'analogie les plus fréquents

Un trigramme de mots donné peut être obtenu par analogie à l'aide d'autres trigrammes de plusieurs façons. Par exemple, pour le trigramme *opportunité de servir*, on a, entre autres, les deux analogies suivantes qui utilisent des trigrammes différents du corpus d'entraînement et respectent bien la définition de l'analogie vue en introduction :

opportunité de servir : *opportunité de modifier* :: *qui pourrait servir* : *qui pourrait modifier*

opportunité de servir : *opportunité pour dire* :: *de servir le* : *pour dire le*

Ces deux analogies exemplifient deux patrons d'analogie différents donnés dans la table 2 et numérotés 1 et 2. Le patron 1 correspond au remplacement du bigramme correspondant à la partie gauche du premier trigramme par un autre bigramme et le remplacement de l'unigramme restant à droite par un autre unigramme : *opportunité de* est remplacé par *qui pourrait*, et *servir* est remplacé par *modifier*. Le patron 2 revient à trouver deux bigrammes dans les mêmes contextes droit et gauche : *de servir* et *pour dire* existent dans les mêmes contextes *opportunité* $\sim$ et $\sim$ *le*.

Dans le double but d'énumérer les patrons existant réellement en corpus et d'en déterminer les fréquences d'apparition respectives, nous énumérons simplement toutes les analogies existantes entre trigrammes à partir d'un échantillon aléatoire de 10 000 phrases dans chaque langue. Pour cela, nous avons contraint la méthode d'énumération de toutes les analogies d'un texte proposée dans (Gosme & Lepage, 2009) pour n'énumérer que les analogies entre trigrammes de mots. Ensuite, nous regroupons les instances d'analogies obtenues par patron et les comptons.

La table 2 donne les patrons d'analogie listés par ordre décroissant de fréquences pour l'anglais. Un résultat remarquable est que les cinq patrons d'analogie les plus fréquents dans les quatre langues apparaissent dans le même ordre avec des proportions semblables.

TABLE 2 – Patrons d'analogie entre trigrammes de mots dans un échantillon anglais de 10 000 phrases du corpus Europarl triés par proportions relatives sur l'ensemble des analogies entre trigrammes. Les symboles utilisés dans l'écriture des patrons d'analogie sont distincts deux à deux. Ces patrons respectent la définition de l'analogie donnée en introduction.

N°	A : B :: C : D	Proportion
1	$a\,b\,c$: $a\,b\,d$ :: $e\,f\,c$: $e\,f\,d$	12,6 %
2	$a\,b\,c$: $a\,d\,e$ :: $b\,c\,f$: $d\,e\,f$	9,1 %
3	$a\,b\,c$: $d\,b\,c$ :: $e\,f\,a$: $e\,f\,d$	3,1 %
4	$a\,b\,c$: $a\,e\,c$ :: $b\,c\,d$: $e\,c\,d$	2,7 %
5	$a\,b\,c$: $a\,b\,d$ :: $b\,c\,e$: $b\,d\,e$	2,6 %
6	$a\,b\,c$: $a\,d\,e$ :: $f\,b\,c$: $f\,d\,e$	2,4 %
7	$a\,b\,c$: $a\,d\,c$ :: $b\,e\,f$: $d\,e\,f$	1,3 %
8	$a\,b\,c$: $a\,b\,d$ :: $a\,e\,c$: $a\,e\,d$	0,9 %
⋮	⋮	⋮

2.3 Patrons d'analogie les plus rentables

Jusqu'à présent, nous avons montré, d'une part, qu'une grande majorité des trigrammes inconnus peuvent être reconstruits par analogie à l'aide de trigrammes issus du corpus d'entraînement (section 2.1) ; et nous avons identifié, d'autre part, les patrons d'analogie de trigrammes de mots les plus fréquents dans un même corpus (Section 2.2). L'étape suivante est d'identifier les patrons d'analogie qui permettent de reconstruire le plus de trigrammes inconnus d'un jeu de test à l'aide de trigrammes du corpus d'entraînement, autrement dit, les patrons les plus rentables. À cette fin, nous conduisons une nouvelle série d'expériences. En raison de la lourdeur en temps de calcul, nous limitons notre expérience aux cinq patrons d'analogie les plus fréquents listés dans la table 2, et nous procédons de la sorte : pour chaque trigramme inconnu du jeu de test, chaque patron d'analogie est essayé successivement dans l'ordre de la table 2. Dès lors qu'un patron d'analogie permet de reconstruire le trigramme inconnu en question, nous notons son rang et passons au trigramme inconnu suivant.

Les résultats sont présentés dans les tables 3((a))–(d)). Ils montrent les contributions cumulées des patrons d'analogie à la reconstruction des trigrammes inconnus. Le patron 1 contribue seul à la majorité de la reconstruction des trigrammes inconnus : plus de 70 % en anglais, français et allemand, mais seulement 61,5 % pour le finnois. Les patrons 1 et 2 suffisent à reconstruire environ 95 % des trigrammes inconnus en anglais, français et allemand, et presque 90 % en finnois.

TABLE 3 – Cumul des contributions des cinq patrons d'analogie les plus fréquents à la reconstruction des trigrammes inconnus dans les quatre langues étudiées. Les pourcentages présentés sont relatifs au nombre total de trigrammes inconnus.

(a) Anglais

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	72 426 (63,22 %)	75,55 %
2	19 952 (17,42 %)	96,37 %
3	3 411 (2,98 %)	99,93 %
4	46 (0,04 %)	99,97 %
5	25 (0,02 %)	100,00 %
Total	95 860 (83,67 %)	100,00 %

(b) Français

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	71,466 (61,98 %)	74,66 %
2	20,475 (17,51 %)	96,05 %
3	3 655 (3,13%)	99,87 %
4	92 (0,08 %)	99,97 %
5	35 (0,03 %)	100,00 %
Total	95 723 (81,87%)	100,00 %

(c) Allemand

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	71 150 (50,74 %)	70,34 %
2	23 810 (16,98 %)	93,87 %
3	6 003 (4,28 %)	99,81 %
4	156 (0,11 %)	99,96 %
5	37 (0,03 %)	100,00 %
Total	101 156 (72,14 %)	100,00 %

(d) Finnois

N° de patron	Trigrammes reconstruits (μ)	Proportion cumulée
1	36 717 (27,62 %)	61,48 %
2	16 064 (12,08 %)	88,37 %
3	6 227 (4,68 %)	98,80 %
4	548 (0,41 %)	99,72 %
5	169 (0,13 %)	100,00 %
Total	59 725 (44,93 %)	100,00 %

Pour les quatre langues, les cinq patrons suffisent à reconstruire l'intégralité des trigrammes ; notre restriction se justifie donc a posteriori. Quelques exemples de reconstructions de trigrammes sont donnés dans les figures 1 et 2.

en justice et : *en est ,* :: *justice et de* : *est , de*
débat en tant : *débat de ce* :: *en tant qu'* : *de ce qu'*
coûts de la : *coûts et les* :: *de la plus* : *et les plus*

FIGURE 1 – Exemples de trigrammes de mots du corpus français d'Europarl respectant le patron 2, c'est-à-dire $\underline{a}\ b\ c : \underline{a}\ d\ e :: b\ c\ \underline{f} : d\ e\ \underline{f}$.

debate and we : *debate and far* :: *but as we* : *but as far*
Union have set : *Union have that* :: *a committee set* : *a committee that*
but they do : *but they must* :: *so we do* : *so we must*

FIGURE 2 – Exemples de trigrammes de mots du corpus anglais d'Europarl respectant le patron 1, c'est-à-dire $a\ b\ \underline{c} : a\ b\ \ \underline{d} :: e\ f\ \ \underline{c} : e\ f\ \ \underline{d}$.

2.4 Effectif suffisant pour la reconstruction d'un trigramme inconnu

Puisque l'hypothèse de la reconstruction massive des trigrammes inconnus par analogie est confirmée par les expériences précédentes, nous passons maintenant à l'étude des effectifs des trigrammes impliqués dans les reconstructions. Nous cherchons à savoir quels effectifs ont les trigrammes qui permettent la reconstruction des

trigrammes inconnus. Une supposition naturelle serait que les trigrammes d'effectifs semblables aient tendance à apparaître dans les mêmes analogies. Suivant cette supposition, on peut faire l'hypothèse que les trigrammes inconnus, c'est-à-dire apparaissant zéro fois dans le corpus d'entraînement, pourraient être reconstruits à l'aide de trigrammes apparaissant une fois dans le corpus, c'est-à-dire les trigrammes hapax. Nous confirmons ici cette hypothèse.

Nous effectuons une nouvelle série d'expériences afin d'obtenir les effectifs des trigrammes en relation d'analogie avec les trigrammes inconnus. En raison de la lourdeur des calculs, nous nous limitons à l'analyse du patron 1 : *a b c* : *a b d* :: *e f c* : *e f d*. Pour chaque instance de ce patron, nous définissons son *effectif maximum* comme l'effectif du trigramme le plus fréquent parmi les quatre trigrammes de l'analogie (comme le premier trigramme est inconnu, son effectif dans le corpus d'entraînement est évidemment zéro). Pour chaque trigramme inconnu reconstruit, nous mémorisons le minimum sur les effectifs maximums de toutes les analogies permettant de le reconstruire (effectif min-max). De cette mémorisation, et par inversion, pour chaque effectif min-max, nous pouvons compter le nombre de trigrammes inconnus reconstruits. Chaque effectif min-max est donc l'effectif suffisant à considérer pour trouver à coup sûr des trigrammes permettant la reconstruction de tant de trigrammes inconnus.

Ces décomptes sont donnés dans la table 4. Pour chaque *effectif min-max*, la table présente la quantité de trigrammes reconstruits et un pourcentage cumulé. Selon ces résultats, les instances d'analogie du patron 1 impliquant trois trigrammes hapax (effectif min-max = 1) permettent la reconstruction de plus de 95 % des trigrammes inconnus en anglais, 94 % en français ou en allemand et 91 % en finnois.

TABLE 4 – Pourcentages cumulés des trigrammes reconstruits, classés par effectif suffisant des trigrammes formant analogie pour leur reconstruction (colonne *effectif min-max*).

(a) Anglais

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	54 227 (96,24 %)	96,24 %
2	1 288 (2,29 %)	98,53 %
3	345 (0,61 %)	99,14 %
4	127 (0,23 %)	99,36 %
5	99 (0,18 %)	99,54 %
⋮	⋮	⋮
523	1 (0,00 %)	100,00 %
TOTAL	56 345 (100,00 %)	—

(b) Français

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	59 050 (94,07 %)	94,07 %
2	2 167 (3,45 %)	97,52 %
3	608 (0,97 %)	98,49 %
4	302 (0,48 %)	98,97 %
5	167 (0,26 %)	99,24 %
⋮	⋮	⋮
576	1 (0,00 %)	100,00 %
TOTAL	62 771 (100,00 %)	—

(c) Allemand

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	41 272 (94,01 %)	94,01 %
2	1 475 (3,36 %)	97,36 %
3	465 (1,06 %)	98,42 %
4	219 (0,50 %)	98,92 %
5	124 (0,28 %)	99,21 %
⋮	⋮	⋮
412	1 (0,00 %)	100,00 %
TOTAL	43 904 (100,00 %)	—

(d) Finnois

Effectif min-max	Trigrammes reconstruits	Pourcentage cumulé
1	13 382 (91,02 %)	91,02 %
2	760 (5,217 %)	96,18 %
3	238 (1,62 %)	97,80 %
4	101 (0,68 %)	98,49 %
5	56 (0,38 %)	98,87 %
⋮	⋮	⋮
458	1 (0,01 %)	100,00 %
TOTAL	56 542 (100,00 %)	—

L'ensemble des résultats expérimentaux précédents conduit à la conclusion que non seulement les analogies entre trigrammes structurent les trigrammes inconnus, mais qu'en plus, la reconstruction des trigrammes inconnus est massivement possible avec des trigrammes d'effectif semblable, c'est-à-dire d'effectif 1.

Pour résumer l'étude empirique présentée ci-dessus, on peut donc dire que : *dans leur grande majorité les trigrammes inconnus sont analogues aux trigrammes hapax ; leurs structures et leurs effectifs sont semblables.*

3 Lissage de modèles trigrammes par analogie

Dans cette deuxième partie, nous allons exploiter les résultats de l'étude empirique précédente pour proposer une technique de lissage de modèles de langue. Notre proposition est volontairement simple et s'inspire de méthodes de lissage élémentaires : les lissages de Lidstone et de Laplace.

Habituellement, lorsqu'on utilise directement des outils tels que *SRILM* (Stolcke, 2002), on a l'habitude d'utiliser les techniques de lissage classiques connues pour donner des résultats acceptables. Des techniques de lissage plus élaborées ont été proposées afin de réduire la taille des modèles de langue, nous pensons en particulier au *clustering* (Brown *et al.*, 1992). Cependant, de telles techniques requièrent une phase de pré-traitement complexe, ce qui accroît le coût de calcul (Matsuzaki *et al.*, 2003). En comparaison, la méthode que nous proposons dans cet article n'extrait pas de connaissances supplémentaires des données d'entraînement. La structure des trigrammes inconnus est vérifiée au fil du calcul. Les principaux avantages de cette méthode sont sa simplicité et sa facilité d'utilisation.

3.1 Ré-estimation des effectifs

Redisons une vérité élémentaire : tout évènement inconnu apparaissant dans le jeu de test a un effectif nul dans le corpus d'entraînement. Immédiatement au-dessus de la classe des évènements d'effectif nul, vient la classe des évènements observés une seule fois dans le corpus d'entraînement : ce sont les hapax. Or, il est classique pour une technique de lissage d'essayer d'estimer la probabilité lissée des évènements inconnus en se basant sur les propriétés des événements classés selon leurs fréquences d'apparition : c'est la base du lissage de Good-Turing (Gale, 1994). Nous exploitons la même idée mais dans une mise en application plus simple.

Dans le lissage de Laplace, tout événement voit son effectif augmenté de 1. Dans notre technique de lissage, nous gardons cet incrément de 1 pour les événements connus. L'essence de notre technique de lissage tient dans la distinction faite entre évènements inconnus selon qu'ils peuvent être reconstruits par analogie ou non.

Nous donnons un fort avantage aux évènements inconnus qui peuvent être reconstruits par analogie au détriment de ceux qui ne peuvent pas l'être. Les résultats des expériences présentées en section 2.4 conduisent à proposer un effectif très proche de 1 pour les trigrammes reconstructibles puisqu'ils sont analogues aux trigrammes hapax Nous fixons leur effectif à $1-\alpha$ avec α proche de 0. Ils deviennent donc de nouveaux quasi-hapax, alors que les anciens hapax sont ré-estimés avec un effectif de 1 + 1 = 2.

En désespoir de cause, nous affectons comme estimation des effectifs des trigrammes inconnus qui ne peuvent pas être reconstruits une valeur très proche de 0. Pour simplifier, nous utilisons la valeur α. On peut dire que cette partie du lissage est en fait un lissage de Lidstone.

Au total donc, la probabilité lissée d'un trigramme $h_i.m_i$ (h_i représente les deux mots précédant m_i) est ré-estimée selon chacun des trois cas suivants, avec N la longueur du texte, $|V|$ la taille du vocabulaire et δ restant à déterminer :

– trigrammes connus : $\dfrac{C(h_i.m_i)+1}{N+\delta\times|V|}$

– trigrammes inconnus pouvant être reconstruits par analogie : $\dfrac{1-\alpha}{N+\delta\times|V|}$

– trigrammes inconnus ne pouvant être reconstruits par analogie [3] : $\dfrac{\alpha}{N+\delta\times|V|}$

En reprenant les notations de la section 2.1 et de la table 1, nous notons λ la proportion de trigrammes inconnus différents et μ la proportion relative de trigrammes inconnus différents reconstruits par analogie. Les valeurs de λ et μ sont comprises entre 0 et 1. Avec ces notations :

– $(1-\lambda)$ est la proportion de trigrammes connus dans le jeu de test, λ étant la proportion de trigrammes inconnus dans le jeu de test ;
– $\mu\lambda$ est la proportion, sur l'ensemble du jeu de test, de trigrammes inconnus qui peuvent être reconstruits, μ étant la proportion de trigrammes inconnus reconstructibles ;

3. C'est en particulier le cas de tout trigramme contenant un mot inconnu. Un tel trigramme ne peut en effet être reconstruit par analogie de par la définition donnée en introduction (test sur le nombre d'occurrences des mots).

– et $(1-\mu)\lambda$ est le reste sur l'ensemble des trigrammes du jeu de test, c'est-à-dire la proportion de trigrammes inconnus ne pouvant être reconstruits par analogie.

La somme des probabilités de tous les trigrammes devant faire 1, la valeur de δ peut être déterminée :

$$\begin{aligned}\delta &= (1-\lambda)\times 1 \;+\; \mu\lambda\times(1-\alpha) \;+\; (1-\mu)\lambda\times\alpha \\ &= 1-(2\alpha\mu-\alpha-\mu+1)\lambda\end{aligned}$$

3.2 Estimation des paramètres

Dans la pratique, les paramètres λ et μ sont estimés dans une phase de pré-traitement. Le corpus d'entraînement est divisé en deux parties, l'une comprenant les neuf dixièmes du corpus, l'autre comprenant le dixième restant. La proportion de trigrammes inconnus dans la plus petite partie ainsi que la part de trigrammes inconnus reconstruits par analogie sont estimées par échantillonnage. Ces estimations deviennent les valeurs des paramètres λ et μ.

Concernant le paramètre α, des résultats d'expériences non présentés dans cet article nous ont conduits à le fixer à 10^{-6} pour toutes les langues.

3.3 Temps d'exécution

Afin de déterminer si un trigramme inconnu peut être reconstruit ou non par analogie, le corpus d'entraînement est mémorisé sous forme de deux tableaux de suffixes (sens de lecture normal et miroir). Lorsque la reconstruction d'un trigramme doit être testée, pour chaque patron d'analogie, une recherche appropriée à ce patron est effectuée dans ces tableaux de suffixes. Par exemple, pour le patron 1, le trigramme candidat *a b c* est décomposé en une partie gauche *a b* et une partie droite *c*. La recherche de ces séquences dans les tableaux de suffixes réduit aux trigrammes hapax est très rapide. Il suffit alors de prendre l'intersection en termes de positions de l'ensemble des unigrammes *d* qui suivent *a b* (sens de lecture normal) et de l'ensemble des bigrammes *e f* qui précèdent *c* (miroir). Dès qu'une position est trouvée dans l'intersection, nous en déduisons qu'il existe au moins un trigramme *e f d* dans le corpus d'entraînement et nous pouvons conclure en l'existence d'une analogie *a b c* : *a b d* :: *e f c* : *e f d*. Cela signifie que le trigramme *a b c* peut être reconstruit par analogie à l'aide de trigrammes du corpus d'entraînement. Une procédure similaire a été implantée pour le patron 2.

L'implantation des lissages classiques de *SRILM* permet de lisser environ 1 000 phrases par seconde en français quelle que soit la méthode de lissage et quelle que soit la taille du corpus d'entraînement sur une machine équipée d'un processeur 16 bits cadencé à 2 GHz et ayant 4 Go de mémoire.

Notre méthode de lissage effectue des recherches dans des tableaux de suffixes et nous nous attendons à ce que la vitesse de lissage dépende de la taille du corpus d'entraînement. Sur le même type de machine, nous mesurons la vitesse de notre méthode de lissage en fonction de la taille du corpus d'entraînement pour deux variantes : patron 1 seul et patrons 1 et 2. Nous utilisons des échantillons de la partie française d'Europarl avec des tailles variant de 900 à 320 000 phrases. Dans la seconde variante, le patron 2 est utilisé en deuxième instance dans la cas où le patron 1 n'a pas permis de reconstruire le trigramme.

Les courbes de la figure 3 donnent le nombre de phrases du jeu de test traitées par seconde en fonction de la taille du corpus d'entraînement. La vitesse de lissage de notre méthode dépend nettement de la taille du corpus d'entraînement. Pour de petits corpus, notre implantation traite 300 phrases par seconde. Cette vitesse chute à 100 phrases par seconde pour les corpus de plus grande taille et n'évolue plus vraiment à partir de 180 000 phrases. Les deux variantes sont similaires, ce qui signifie que la variante patrons 1 et 2 n'engendre qu'un faible surcoût de temps de traitement.

L'implantation actuelle de notre méthode de lissage par analogie, en Python, est donc dix fois plus lente que les implantations de *SRILM* en C++ des méthodes classiques de lissage. On peut raisonnablement espérer des temps comparables avec une implantation en C++ si l'on se fie aux règles très grossières donnant des accélérations par dix lors de réécritures de Python en C++. [4]

4. http://shootout.alioth.debian.org/u32q/benchmark.php?test=all&lang=gpp&lang2=python

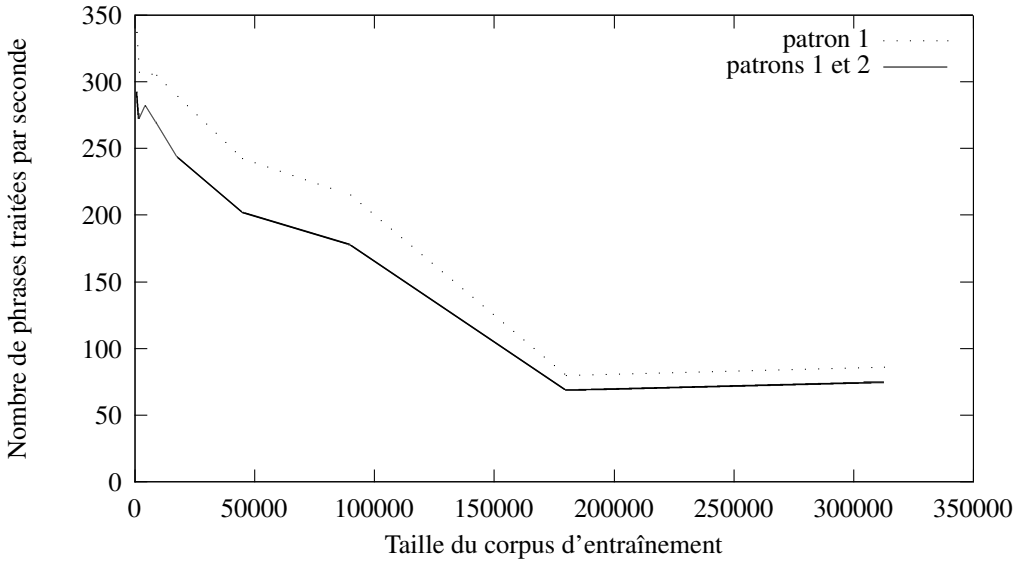

FIGURE 3 – Vitesse de la méthode de lissage par analogie pour différentes tailles du corpus d'entraînement, pour deux variantes de la méthode : patron 1 et patrons 1 et 2.

3.4 Évaluation des performances

Nous comparons notre méthode de lissage par analogie avec quatre méthodes de lissage classiques : Lidstone (Chen & Goodman, 1999), Witten & Bell (1991), Good-Turing (Gale, 1994) et Kneser & Ney (1995). Cette dernière méthode est souvent considérée comme la meilleure en pratique. [5] Pour ces quatre méthodes, nous utilisons les implantations de *SRILM* (Stolcke, 2002). Dans le cas du lissage de Lidstone, après optimisation, nous utilisons la valeur 10^{-3} pour le paramètre α pour chaque langue.

Les critères d'évaluation utilisés sont la divergence de Kullback-Leibler et la perplexité en mots. La divergence de Kullback-Leibler est définie pour chaque phrase par :

– entropie du jeu de test : $H(p) = -\sum_{i=1}^{l} p(m_i|h_i) \times \log_2 p(m_i|h_i)$;

– entropie d'un modèle de langue : $H(p,q) = -\sum_{i=1}^{l} p(m_i|h_i) \times \log_2 q(m_i|h_i)$;

– divergence de Kullback-Leibler : $D_{KL} = H(p,q) - H(p)$.

La perplexité en mots est définie comme la moyenne géométrique des inverses des probabilités réestimées. En notant n le nombre de mots du jeu de test : $PPL = 2^{\frac{-\sum_{i=1}^{n} \log_2 p(m_i|h_i)}{n}}$.

Dans ces formules, $p(m_i|h_i)$ est la probabilité conditionnelle obtenue sur le jeu de test, avec m_i le mot à la position i et h_i son histoire, c'est-à-dire les deux mots précédant m_i ; $q(m_i|h_i)$ est la probabilité conditionnelle lissée utilisant le corpus d'entraînement et l est la longueur de la phrase du jeu de test.

La comparaison est effectuée sur des données extraites d'Europarl en onze langues. Pour chaque langue, les phrases ayant une traduction en anglais sont retenues. Nous obtenons de cette manière onze corpus alignés de 383 237 phrases représentant 10 millions de mots ou plus dans chaque langue, sauf en finnois (seulement 8 millions). Chaque corpus est ensuite divisé en deux parties : 90 % du corpus pour l'entraînement, les 10 % restants servant de jeu de test. De cette manière, nos expériences sont véritablement comparables entre langues. Les statistiques concernant le corpus d'entraînement et le jeu de test pour chaque langue sont présentées dans la table 5.

Les estimations des paramètres λ et μ nécessaires à notre méthode de lissage sont détaillées dans la table 6.

5. « Kneser & Ney (1995) smoothing and its variants are generally recognized as having the best perplexity of any known method for estimating N-gram language models. » (Moore & Quirk, 2009). (Chen & Goodman, 1998) ont montré qu'une première version modifiée du lissage de Kneser-Ney « consistently had the best performance » sur l'ensemble de leurs tests et qu'une seconde version modifiée « [p]erform[ed] just slightly worse ».

TABLE 5 – Statistiques des corpus d'entraînement et des jeux de tests utilisés pour la comparaison.

	Corpus d'entraînement : 347 613 phrases			Jeux de test : 38 624 phrases		
Langue	Nbr total de mots ($\times 10^6$)	Taille du vocabulaire	Mots/phrase	Nbr total de mots ($\times 10^6$)	Taille du vocabulaire	Mots/phrase
da	9,46	153 425	27,21	1,06	46 117	27,36
de	9,51	167 942	27,36	1,06	51 398	27,48
el	10,00	149 247	28,76	1,12	52 671	28,89
en	9,94	67 819	28,60	1,11	25 854	28,76
es	10,47	100 410	30,12	1,17	37 128	30,27
fi	7,18	299 116	20,65	0,80	84 964	20,74
fr	10,95	86 567	31,51	1,22	33 403	31,65
it	9,88	99 252	28,42	1,10	36 624	28,54
nl	10,01	125 565	28,80	1,12	39 728	29,00
pt	10,29	102 800	29,59	1,15	38 041	29,73
sv	8,99	157 116	25,86	1,00	48 327	25,98

TABLE 6 – Proportion de trigrammes inconnus différents (λ) et proportion de trigrammes inconnus différents reconstructibles par analogie (μ) estimées à partir d'un échantillon de 10 % du corpus d'entraînement pour chaque langue, et valeurs correspondantes de δ ($\alpha = 10^{-6}$). Lors du calcul de δ, les valeurs de λ et μ sont ramenées entre 0 et 1.

	Trigrammes inconnus différents				
	(λ)	Reconstruits (μ)		(δ)	
		Patron 1	Patrons 1 et 2	Patron 1	Patrons 1 et 2
da	55,03 %	45,84 %	70,22 %	0,702	0,836
de	61,41 %	43,24 %	69,43 %	0,651	0,812
el	59,57 %	42,98 %	69,29 %	0,660	0,817
en	51,97 %	55,72 %	79,72 %	0,770	0,895
es	48,96 %	50,34 %	73,45 %	0,757	0,870
fi	**78,24 %**	**26,68 %**	**49,25 %**	**0,426**	**0,603**
fr	49,13 %	53,40 %	79,19 %	0,771	0,898
it	58,88 %	49,82 %	75,27 %	0,705	0,854
nl	54,56 %	52,00 %	75,94 %	0,738	0,869
pt	54,72 %	47,46 %	72,84 %	0,713	0,851
sv	60,18 %	47,25 %	71,28 %	0,683	0,827

TABLE 7 – Comparaison de la technique de lissage par analogie (patron 1 et patrons 1 et 2) avec quatre techniques de lissage classiques en onze langues.

	Perplexités en mots										
	da	de	el	en	es	fi	fr	it	nl	pt	sv
Patron 1	197,5	401,5	226,9	125,6	144,5	10099,8	106,0	149,0	181,0	141,6	334,6
Lidstone	171,0	247,1	179,3	107,4	107,6	1135,9	84,5	141,0	162,1	125,6	235,3
Witten-Bell	130,1	192,0	139,5	93,2	91,9	828,3	73,7	119,9	132,3	106,2	180,0
Good-Turing	128,9	189,2	138,1	92,6	91,0	**784,6**	**73,3**	119,1	131,0	105,3	177,7
Kneser-Ney	134,7	196,3	158,3	95,6	92,0	824,3	74,6	120,1	137,3	106,9	186,4
Patron 1 et 2	**107,8**	**182,4**	**116,4**	**90,9**	**85,8**	2876,6	73,7	**81,0**	**99,2**	**79,7**	**152,6**
	Divergences de Kullback-Leibler										
Patron 1	61,2	73,1	73,5	52,2	56,7	121,9	55,8	68,8	62,9	62,3	70,3
Lidstone	54,2	66,3	63,2	44,2	47,9	95,7	45,0	56,7	54,8	53,0	60,6
Witten-Bell	47,0	60,0	56,2	40,6	43,4	89,0	41,0	52,5	49,4	48,4	54,1
Good-Turing	46,5	59,3	55,7	40,0	42,9	87,6	40,5	52,0	48,8	47,8	53,5
Kneser-Ney	46,4	58,8	55,9	40,2	43,0	**87,2**	**40,4**	51,9	48,8	47,8	53,5
Patron 1 et 2	**43,5**	**50,2**	**51,8**	**38,7**	**41,5**	105,7	41,2	**48,5**	**44,8**	**44,3**	**48,4**

Nous rappelons que λ est la proportion de trigrammes inconnus différents et que μ est la proportion relative de trigrammes inconnus différents qui peuvent être reconstruits par analogie. Nous considérons deux variantes de notre méthode : la première n'utilise que le patron 1, la seconde utilise les patrons 1 et 2. Afin de rendre la technique de lissage indépendante du jeu de test, pour chaque langue les estimations des paramètres ont été obtenues automatiquement à partir d'un échantillon aléatoire formé d'un dixième du corpus d'entraînement comme décrit dans la section 3.2. Comme le montrent les chiffres de la table 6, l'utilisation du patron 2 en plus du patron 1 augmente sensiblement la valeur du paramètre μ : plus d'un quart en valeurs absolues. À l'exception du finnois, l'utilisation conjointe des patrons 1 et 2 permet la reconstruction de 70 % à 80 % des trigrammes inconnus. Les valeurs pour le finnois, en gras dans la table, sont nettement différentes des valeurs pour les autres langues.

Les résultats de l'évaluation des deux variantes de la méthode proposée sont présentés dans la table 7 :
– le patron 1 seul est insuffisant pour atteindre même le niveau du lissage de Lidstone. On obtient systématiquement les plus mauvais résultats dans les onze langues ;
– à l'exception du finnois, et dans une moindre mesure du français, l'ajout du patron 2 est suffisant pour obtenir des résultats bien meilleurs que ceux des quatre méthodes de lissage classiques.

La contre-performance sur le finnois n'est pas surprenante si l'on considère le nombre important de trigrammes inconnus et la faible proportion de ces trigrammes qui peuvent être reconstruits par analogie (voir table 6). Afin de remédier à ce problème, plutôt que d'accroître la quantité de données d'entraînement, il serait sans doute plus judicieux de segmenter les mots en morphèmes. Quant aux résultats en français, ils sont comparables à ceux des méthodes classiques.

4 Conclusion et perspectives

Dans cet article, à l'aide d'une série d'expériences sur quatre langues, nous avons montré qu'en majorité les trigrammes inconnus dans un jeu de test sont structurellement analogues aux trigrammes hapax d'un corpus d'entraînement.

De cette propriété, nous avons dérivé une méthode de lissage pour modèles de langue trigrammes. L'effectif des trigrammes connus est ré-estimé en appliquant un incrément de 1 comme dans le lissage de Laplace. Les trigrammes inconnus qui peuvent être reconstruits par analogie sont considérés comme quasi-hapax : leurs effectifs sont ré-estimés à une valeur proche de 1. Les trigrammes inconnus qui ne peuvent être reconstruits par analogie sont presque ignorés, leurs effectifs étant fixés à une valeur proche de 0 comme dans le lissage de Lidstone. En comparaison de techniques de lissage utilisant des techniques de *clustering*, cette méthode est simple ; elle ne construit que deux classes de trigrammes inconnus : ceux qui peuvent être reconstruits et les autres.

Des mesures sur onze langues ont montré que cette méthode de lissage donne de bons résultats en comparaison des techniques de lissage classiques, sauf dans le cas du finnois.

L'étude présentée ici laisse un certain nombre de points à examiner. Tout d'abord, cette étude a été consacrée aux trigrammes. Or, aujourd'hui, dans de nombreux domaines du traitement automatique des langues, comme par exemple la traduction automatique par approche statistique, on utilise des modèles de langue 5-grammes. Des expériences restent donc à effectuer avec des n-grammes d'ordres supérieurs pour savoir si de bons résultats peuvent aussi être obtenus. L'influence du nombre de patrons d'analogie sur l'entropie des modèles de langue obtenus reste elle aussi à étudier. Un autre point porte sur la taille des corpus utilisés. Les expériences rapportées ici visant à une comparaison sur plusieurs langues et les très grands corpus multilingues étant rares, la taille du corpus utilisé ici est relativement faible en regard de corpus monolingues dépassant le milliard de mots. Des expériences sur des corpus de tailles plus importantes restent donc à effectuer. Enfin, dans la perspective d'une intégration à la traduction automatique par approche statistique, les pouvoirs discriminants de la technique de lissage proposée ici restent à examiner.

5 Remerciements

Cet article décrit des résultats de recherche obtenus en partie grâce une subvention de l'université Waseda pour projets de recherche spécifiques (projet numéro : 2010A-906).

Références

BROWN P., PIETRA V., DESOUZA P., LAI J. & MERCER R. (1992). Class-based *n*-gram models of natural language. *Computational linguistics*, **18**(4), 467–479.

CHEN S. F. & GOODMAN J. (1998). *An empirical study of smoothing techniques for language modeling*. Rapport interne, Harvard university, Cambridge, Massachussetts.

CHEN S. F. & GOODMAN J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**(4), 359–394.

CLAVEAU V. & L'HOMME M.-C. (2005). Terminology by analogy-based machine learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, Copenhagen.

DENOUAL E. (2007). Analogical translation of unknown words in a statistical machine translation framework. In *Proceedings of Machine Translation Summit XI*, Copenhagen.

GALE W. (1994). Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, **2**.

GOSME J. & LEPAGE Y. (2009). A first study of the complete enumeration of all analogies contained in a text. In *4th Language and Technology Conference (LTC 2009)*, p. 401–405, Poznań, Poland.

HATHOUT N. (2009). Acquisition morphologique à partir d'un dictionnaire informatisé. In T. NAZARENKO, D. ET POIBEAU, Ed., *Actes de la 16e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2009)*, p. 10 p. : ATALA.

KNESER R. & NEY H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, p. 79–86, Phuket.

LANGLAIS P. & PATRY A. (2007). Translating unknown words by analogical learning. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 877–886.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2008). *Analogical Translation of Medical Words in Different Languages*, volume 5221/2008 of *Lecture Notes in Computer Science*, p. 284–295. Springer Berlin / Heidelberg : Springer Berlin / Heidelberg.

LAVALLÉE J. F. & LANGLAIS P. (2010). Analyse morphologique non supervisée par analogie formelle. In *TALN 2010*, p. 10 pages, Montréal, Québec, Canada.

LEPAGE Y. (2004). Analogy and formal languages. *Electronic notes in theoretical computer science*, **53**, 180–191.

LEPAGE Y. & DENOUAL E. (2005). Purest ever example-based machine translation : detailed presentation and assessment. *Machine Translation*, **19**, 251–282.

LEVENSHTEIN V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**(8), 707–710.

MATSUZAKI T., MIYAO Y. & TSUJII J. (2003). An efficient clustering algorithm for class-based language models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, p. 119–126 : Association for Computational Linguistics.

MOORE R. & QUIRK C. (2009). Improved smoothing for N-gram language models based on ordinary counts. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 349–352 : Association for Computational Linguistics.

ROSENFELD R. (2000). Two decades of statistical language modelling : where do we go from here ? *Proceedings of the IEEE*, **88**(8), 1270–1278.

STOLCKE A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3, p. 901–904.

TURNEY P. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, p. 905–912, Manchester, UK : Coling 2008 Organizing Committee.

WITTEN I. & BELL T. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions on*, **37**(4), 1085–1094.