Comparaison de l'usage d'un corpus et de WordNet pour l'extension de normes lexicales

Nadja Vincze[1], Yves Bestgen[2]

[1]UCL – CENTAL – B-1348 Louvain-la-Neuve – Belgique
[2]Chercheur qualifié du F.R.S – UCL – CECL – B-1348 Louvain-la-Neuve – Belgique

Abstract

Both in the field of psychology and in natural language processing, norms related to semantic properties, such as concreteness, polarity or emotionality, are important resources. These norms have consistently been obtained by asking judges to rate the words on scales, for example from very concrete to very abstract. It's very expensive, hence automatic construction methods have been developed. They can be divided into two types: those based on language resources and those that are based on collections of texts.
Our goal is to compare the use of a corpus and a lexical resource (WordNet) for a same method designed to increase norms on the basis of word similarities. We show that the similarities calculated from information on co-occurrences of words in texts are more effective. We also show that the choice of the corpus has little influence on the results, at least for general corpora.

Résumé

En psychologie tout comme en traitement automatique des langues, les normes qui portent sur des propriétés sémantiques des mots, comme le degré d'abstraction, l'imagerie ou la polarité, sont importantes. Ces normes ont systématiquement été obtenues en demandant à des juges d'évaluer les mots sur des échelles, allant par exemple de très concret à très abstrait. Ce mode de récolte étant lent et coûteux, des méthodes de construction automatique ont vu le jour. Elles peuvent être divisées en deux types : celles qui se basent sur des ressources linguistiques et celles qui se basent sur des corpus.
Notre objectif est de comparer, pour une même méthode d'accroissement de normes lexicales basée sur les similarités entre les mots, l'utilisation d'un corpus et d'une ressource lexicale (WordNet) pour estimer ces similarités. Nous montrons que les similarités calculées à partir d'informations sur les cooccurrences des mots dans les textes sont plus efficaces, et ce pour 4 des 5 normes étendues. Nous montrons également que le choix du corpus influence peu les résultats, du moins pour des corpus généraux.

Mots-clés : normes lexicales, similarités, analyse sémantique latente, WordNet, corpus.

1. Introduction

Parmi les nombreuses méthodes d'analyse automatique de textes, une des plus anciennes vise à déterminer si certaines catégories de mots (mots exprimant une opinion, un fait) ou certaines catégories grammaticales (pronoms personnels) sont plus fréquentes dans certains types de textes (Biber, 1988 ; Popping, 2000 ; Stone, 1997). En analyse du contenu assistée par ordinateur, cette méthode est fréquemment employée pour inférer des caractéristiques de

l'auteur d'un texte sur la base des thèmes qu'il aborde dans son discours, mais aussi de la façon dont il s'exprime (Bestgen, 1994 ; Cohen *et al.*, 2009 ; Pennebaker et al., 2003 ; Popping, 2000).

La première étape de ce genre d'analyses consiste à construire des lexiques contenant des listes de mots qui relèvent des différentes catégories à étudier. Ces catégories peuvent correspondre à des classes grammaticales, mais aussi à des regroupements thématiques ou sémantiques de mots. Ensuite, ces lexiques sont comparés aux mots présents dans chaque texte à analyser afin de déterminer la fréquence de chaque catégorie dans chacun de ceux-ci. Ces fréquences sont utilisées pour construire une matrice dont les lignes correspondent aux segments de textes et les colonnes aux différentes catégories. Finalement, cette matrice est analysée pour déterminer si certaines catégories sont plus fréquentes dans certains textes.

La première étape est évidemment la plus coûteuse en ressources, surtout lorsque les propriétés lexicales pertinentes sont sémantiques comme le degré d'abstraction, la familiarité, l'âge d'acquisition, l'imagerie ou encore la polarité (caractère plus ou moins agréable de ce qui est exprimé). Pour construire de tels lexiques, on n'a classiquement d'autre recours que de demander à des juges, souvent plusieurs dizaines, d'évaluer les mots sur des échelles en 7 ou 9 points, allant par exemple de «ce mot vous semble très concret» à «ce mot vous semble très abstrait». En psycholinguistique, de nombreuses listes de mots, appelées *normes*, ont été recueillies (pour des index de ces normes voir par exemple Bradshaw, 1984, Lavaur et Font, 2000 et Desrochers et Saint−Aubin, 2008). Leur principale limitation est qu'elles portent sur peu de mots, les plus grandes ne dépassant pas deux ou trois mille mots. Si une telle taille est suffisante pour les fonctions pour lesquelles elles sont recueillies en psychologie, comme identifier les propriétés des mots qui affectent l'efficacité des processus cognitifs à l'œuvre lors de la lecture, elle pose problème en analyse de contenu et en linguistique computationnelle, étant donné que seule une petite partie du vocabulaire – celui commun à la norme et au texte – peut être pris en compte (Bestgen, 1994). Il serait donc intéressant de pouvoir étendre de telles normes sans avoir à faire appel à des juges supplémentaires.

Pour cette raison, différentes méthodes d'extension automatique ont été proposées en traitement automatique des langues, plus particulièrement en fouille d'opinion (Pang et Lee, 2008). Ce champ, né il y a une dizaine d'années, vise à déterminer automatiquement le caractère subjectif de phrases ou de textes, les émotions qui y sont exprimées ou encore la polarité. De nombreuses méthodes proposées pour réaliser ces tâches calculent la polarité d'un document selon les orientations de mots ou de groupes de mots qui le composent et ont donc besoin de lexiques où à chaque entrée sont associés une polarité ou un degré de polarité. Bien entendu, plus la couverture du lexique est large, plus le nombre de mots identifiés dans les textes est élevé, plus la méthode est efficace, ce qui justifie le développement de méthodes automatiques pour étendre un lexique de polarité.

De telles méthodes estiment la polarité d'un mot au moyen de la similarité sémantique entre ce mot et d'autres mots dont la polarité est connue. Parmi celles-ci, nous pouvons distinguer deux approches : celles basées sur des ressources linguistiques, comme des dictionnaires ou des thésaurus, et celles basées sur des corpus. Les approches qui s'appuient sur des ressources linguistiques calculent généralement la similarité entre les mots à partir de leur relation de synonymie (Kamps et Marx, 2002 ; Esuli et Sebastiani, 2006 ; Hu et Liu, 2004 ; Kim et Hovy, 2004). Elles procèdent en partant de quelques mots dont la polarité est connue, qui servent de

points de repère, et lancent un algorithme d'amorçage (*bootstrapping*) qui parcourt les liens synonymiques et antonymiques de la base en attribuant la même polarité aux mots synonymes et vice-versa. Les approches qui s'appuient sur des corpus de textes calculent les similarités entre les mots sur la base de statistiques textuelles. Turney et Littman (2002, 2003) et Bestgen (2002, 2006) ont ainsi proposé d'employer l'analyse sémantique latente (ASL, *Latent Semantic Analysis*, Deerwester *et al.*, 1990) pour estimer les similarités entre des mots et des points de repère dont la polarité est connue. Tout récemment, nous avons développé une méthode d'extension de normes qui identifie automatiquement les points de repère (germes) optimaux dans un lexique de polarité (ASG, pour *Apprentissage Supervisé de Mots Germes*, Vincze et Bestgen, 2011). Pour ce faire, un modèle de régression multiple est construit sur la base de vecteurs de similarité, obtenus par une ASL, entre les mots présents simultanément dans le lexique et dans le corpus. L'avantage majeur de cette méthode est qu'elle peut être aisément employée avec d'autres dimensions que la polarité puisqu'il n'est pas nécessaire de définir a priori les mots qui serviront de points de repère. Employer cette méthode pour estimer d'autres normes que la polarité est notre premier objectif. Pour ce faire, nous travaillons sur cinq dimensions sémantiques : la polarité, l'activité et la dominance à partir des normes ANEW (*Emotional Norms for English Words*, Bradley et Lang, 1999), et le caractère concret et imagé à l'aide des normes de Gilhooly et Logie (1980).

L'avantage principal des approches basées sur des corpus, et donc d'ASG, est qu'elles n'ont pas besoin pour fonctionner de WordNet (Miller *et al.*, 1990), une ressource spécifiquement développée par des linguistes pour une exploitation automatique des liens sémantiques entre les mots. Il reste néanmoins à démontrer que les corpus permettent d'extraire les similarités sémantiques nécessaires aussi efficacement que WordNet. Une telle comparaison est aisée à mener avec la méthode ASG parce que celle-ci peut être adaptée afin d'employer des similarités issues de WordNet en lieu et place de celles extraites d'un corpus. Réaliser cette comparaison est notre deuxième objectif. Ceci nous permettra aussi d'évaluer l'impact du corpus employé et de sa taille sur l'efficacité de la méthode. Pour ce faire, nous comparons trois corpus qui peuvent être considérés comme des corpus de référence en analyse sémantique latente, en linguistique de corpus et en traitement automatique des langues.

Dans la suite de ce rapport, nous présentons la méthode que nous avons reprise pour nos expérimentations, ainsi que son adaptation avec les similarités issues de WordNet. Nous abordons ensuite le matériel nécessaire (normes et corpus) pour terminer avec les résultats obtenus.

2. Méthode

Nous avons choisi de reprendre la méthode ASG que nous avons développée récemment (Vincze et Bestgen, 2011), et ce pour deux raisons. Tout d'abord, les similarités calculées par l'ASL peuvent facilement être remplacées par d'autres similarités. Ensuite, la méthode ne nécessitant pas de points de repère définis a priori, elle peut être utilisée pour prédire n'importe quelle norme, pour peu que celle-ci puisse être estimée sur la base de similarités sémantiques, c'est-à-dire pour autant que la valeur d'un mot sur cette norme puisse être estimée à partir de la similarité sémantique entre ce mot et d'autres mots dont la valeur est connue.

Dans cette section, nous présentons la méthode telle qu'elle a été développée, c'est-à-dire avec une ASL. Ensuite, nous décrivons des mesures de similarités qui peuvent être calculées à partir de WordNet et constituer une alternative au recours à l'ASL.

2.1. ASG

ASG se base sur les proximités sémantiques entre les candidats auxquels on veut attribuer une valeur pour une dimension et des points de repère dont la valeur pour la dimension étudiée est connue. À la différence des méthodes déjà existantes, ces points de repère ne sont pas définis a priori, mais obtenus par une procédure d'apprentissage supervisé. La méthode requiert comme matériel d'apprentissage une norme lexicale pour la dimension en question et une collection de textes. Elle comporte les quatre étapes suivantes :

1. Sélectionner comme points de repère / prédicteurs potentiels les mots qui sont présents dans la norme de référence[1].
2. Effectuer ensuite une ASL sur une collection de textes afin d'obtenir un espace sémantique. Concrètement, l'ASL part d'un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chaque segment de textes, éventuellement modifié par une fonction (typiquement, la log entropie). Ce tableau fait ensuite l'objet d'une décomposition en valeurs singulières qui en extrait les dimensions orthogonales les plus importantes. Dans cet espace, le sens de chaque mot est représenté par un vecteur et l'on peut estimer la similarité sémantique entre deux mots par le cosinus entre leurs vecteurs. Dans la méthode ASG, on calcule les cosinus entre chacun des prédicteurs potentiels et tous les mots présents dans la norme. On obtient donc une matrice carrée avec en abscisse et en ordonnée les mots de la norme et en valeurs les cosinus entre leurs vecteurs dans l'espace sémantique.
3. Utiliser ensuite une procédure de régression linéaire multiple afin de construire un modèle prédictif basé sur les prédicteurs les plus efficaces pour prédire la norme. Concrètement, nous employons une procédure de régression linéaire mixte, par sélection et élimination (*stepwise*, Draper et Smith, 1981), avec un seuil de signification de 0,05.
4. Employer le modèle construit à l'étape précédente pour estimer les valeurs – pour la dimension étudiée – des mots présents dans l'espace sémantique, mais non dans la norme initiale.

Le recours à l'ASL pour calculer des distances sémantiques entre les mots à l'étape 2 n'est pas obligatoire. D'autres mesures de similarités peuvent servir, pour autant que l'on obtienne en sortie une matrice carrée comprenant des similarités entre les mots de la norme. C'est précisément ce qui a été fait lors des expérimentations, où les similarités ont été également calculées à partir de WordNet.

2.2. Similarités calculées à partir de WordNet

Les calculs de distances sémantiques à partir de WordNet ne sont pas nouveaux. WordNet est une base de données lexicales reprenant les noms, verbes, adjectifs et adverbes de la langue anglaise. Les mots y sont regroupés en concepts, appelés *synsets,* qui forment des ensembles

1 Une analyse comparative de différentes manières de sélectionner les prédicteurs potentiels pour l'extension d'une norme de polarité est donnée dans Vincze et Bestgen (2011).

de synonymes. WordNet fournit également toute une série de relations lexicales et sémantiques entre les synsets. Ces relations sont très riches pour les noms, qui sont organisés en hiérarchie, sur la base des relations d'hyperonymie et d'hyponymie. Cette structure hiérarchique, qui peut être représentée sous la forme d'un graphe, forme un système d'héritage – un mot hérite de toutes les caractéristiques de ses hyperonymes – et a permis le développement de plusieurs mesures de similarités.

La plupart de ces mesures ont été proposées à la fin des années 90 (Maki, McKinley et Thompson, 2004). On peut distinguer plusieurs approches, dont celles basées sur les arêtes (*edge-based*) et celles basées sur les nœuds (*node-based*). Les approches basées sur les arêtes reposent sur le nombre de liens entre les concepts d'un graphe, en parcourant la hiérarchie par les liens d'hyperonymie et d'hyponymie. Plus la distance entre deux nœuds est grande, plus il faut remonter dans la hiérarchie, plus l'information partagée par les deux concepts est faible, et donc moins ces concepts sont similaires. Pour nos expérimentations, nous avons repris la mesure du plus court chemin entre deux nœuds, qui comprend la similarité comme l'inverse de la distance (1/distance), qui renvoie au nombre de nœuds présents sur le plus court chemin, les nœuds finaux compris (cfr. Pedersen *et al.*, 2004).

Les approches basées sur les nœuds s'appuient sur une estimation de l'information contenue dans chaque nœud. La plus utilisée est la mesure de Resnick (1995) qui part du principe que plus la probabilité de rencontrer un concept est faible, plus son contenu d'information (IC) est grand. Il quantifie l'information contenue par un nœud *c* comme le négatif du logarithme de sa probabilité :

$$IC(c) = -\log p\,(c)$$

Resnick part du principe que plus un concept est haut dans la hiérarchie, plus il est probable de le rencontrer (le nœud racine a une probabilité de 1), le contenu d'information sera donc d'autant plus élevé que l'on descend dans la hiérarchie. Par exemple, dans la hiérarchie représentée à la figure 1, *animal* contient moins d'information que *chien* ou *labrador*.

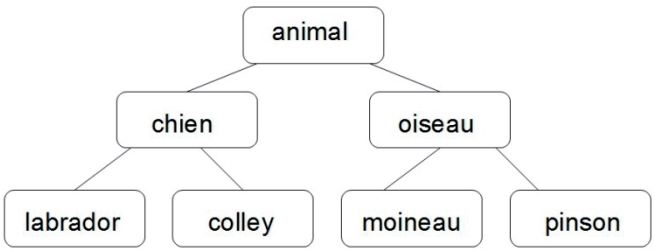

Figure 1 : exemple d'hiérarchie

La similarité entre deux concepts est basée sur l'information qu'ils partagent et elle correspond à l'IC maximum de l'ensemble des nœuds qui subsument les deux concepts. Ainsi, *labrador* et *colley* seront plus similaires que *chien* et *oiseau,* étant donné que le nœud parent des premiers ayant l'IC le plus élevé (*chien*) a un IC plus grand que le nœud parent des deuxièmes (*animal*).

3. Expérimentations

Les expérimentations visent à comparer l'utilisation d'une ressource lexicale et d'un corpus pour le calcul des similarités servant à prédire la valeur de mots pour différentes normes. L'emploi de plusieurs corpus permet également d'évaluer l'impact du choix du corpus pour de telles prédictions. Les sections qui suivent présentent les normes et les corpus utilisés, ainsi que la méthode d'évaluation, avant de présenter les résultats obtenus.

3.1. Normes

Les premières normes que nous avons reprises sont les normes ANEW (Bradley et Lang, 1999). Elles contiennent 1034 mots évalués par des groupes de 8 à 25 juges sur trois échelles en 9 points : la polarité (négatif, désagréable = 1 ; positif, agréable = 9), l'activation (calme = 1 ; excité = 9) et la dominance (impression d'être dominé = 1 ; impression d'être dominant = 9). Ces dimensions ont été obtenues à l'aide d'un système de notation affective basé sur des échelles non-verbales, où les réactions émotionnelles sont représentées sous forme d'images (Self-Assessment Manikin).

Les normes que nous avons utilisées pour les caractères concret et imagé ont été récoltées par Gilhooly et Logie (1980). Pour le caractère concret, 1944 mots ont été évalués par 35 participants, sur une échelle en 7 points allant de très concret (1) à très abstrait (7). Pour l'imagerie, ces mêmes mots ont été évalués par 37 participants, également sur une échelle en 7 points allant des mots qui ne suscitent pas ou difficilement des images mentales (1) aux mots qui en produisent facilement (7).

3.2. Corpus

Trois corpus ont été employés pour calculer les similarités entre les mots au moyen de l'ASL. Le premier, le corpus *TASA - General Reading up to 1st year college* est le corpus de référence pour les travaux psycholinguistiques qui s'appuient sur l'ASL (Landauer *et al.*, 1998 ; Landauer *et al.*, 2007). Ce corpus est composé d'extraits de textes, d'une longueur moyenne approximative de 250 mots, obtenus par un échantillonnage aléatoire de textes que lisent les élèves et les étudiants américains. La version à laquelle T.K. Landauer (Institute of Cognitive Science, University of Colorado, Boulder) nous a donné accès contient 44 486 documents et approximativement 12 millions de mots[2].

Le deuxième corpus, le BNC (British National Corpus, Aston et Burnard, 1998), est le corpus de référence en linguistique de corpus anglaise, conçu pour être représentatif de l'anglais britannique contemporain. Il est composé approximativement de 100 millions de mots (dont 10% issus de l'oral) et couvre de nombreux genres différents comme le langage académique, littéraire, journalistique. Comme les documents inclus dans ce corpus peuvent compter jusqu'à 45 000 mots, ils ont été subdivisés en segments de 250 mots, le dernier segment d'un texte étant supprimé s'il contient moins de 250 mots.

Le troisième corpus, WIKI (Wikipedia.org), connaît un succès de plus en plus important en traitement automatique des langues (Desgraupes, Loiseau et Habert, 2007; Strube et Ponzetto,

2 Un espace sémantique extrait d'une version légèrement différente de ce corpus est disponible sur le site http://LSA.colorado.edu.

2006). Il est construit au moyen de WikiExtractor.py, un script Python écrit par Antonio Fuschetto[3], qui extrait le texte brut de la base de données Wikipedia. Nous avons employé la base de données disponible en avril 2011 pour la version anglaise de Wikipedia. Comme pour le corpus BNC, les textes ont été découpés en segments de 250 mots et tout segment d'une taille inférieure à 250 mots a été supprimé.

Ces trois corpus ont été lemmatisés au moyen de TreeTagger (Schmid, 1994) et une série de mots outils (*and, be, the, that*...) ont été supprimés ainsi que tous les mots dont la fréquence totale est inférieure à 10. Les trois matrices de cooccurrences ont été soumises à une décomposition en valeurs singulières et les 300 premiers vecteurs propres ont été conservés, ce nombre étant habituellement considéré comme optimal (Landauer, Laham et Derr, 2004).

3.3. Méthode pour l'évaluation

Afin d'estimer l'efficacité de la méthode pour prédire de nouvelles données, nous avons employé la procédure de validation croisée *Leave-one-out* (LOOCV) dans laquelle chaque observation est prédite au moyen du modèle dérivé de l'analyse des toutes les autres observations disponibles, produisant ainsi une estimation non biaisée (Lachenbruch et Mickey, 1978 ; Stone, 1974). Toutefois, la combinaison de la régression de type *stepwise* et de la LOOCV pose un problème. Les procédures *stepwise* disponibles dans les logiciels statistiques effectuent la sélection des variables en s'appuyant sur l'ensemble des données. C'est seulement dans un second temps que la procédure LOOCV est appliquée en ne prenant pas en compte, à tour de rôle, chacune des observations. Il s'ensuit que les prédicteurs ont été sélectionnés sur la base de l'ensemble des données et que donc l'estimation LOOCV est favorablement biaisée (donne une estimation de l'efficacité de la prédiction supérieure à la valeur réelle) parce que chacune des observations a pu influencer en sa faveur la sélection initiale des prédicteurs.

Pour éviter ce problème, nous avons employé la procédure de double validation croisée qui consiste à retirer, à tour de rôle, chacune des observations *avant* d'effectuer la sélection des variables par la procédure *stepwise* et l'estimation du modèle qui est employé pour prédire cette observation (Stone, 1974 ; Schulerud et Albregtse, 2004). En d'autres mots, autant d'analyses *stepwise* qu'il y a de données dans le matériel d'apprentissage sont effectuées, laissant à chaque fois une observation de côté, et les variables sélectionnées sont employées pour construire le meilleur modèle pour ces observations. Ensuite, on emploie ce modèle pour prédire l'observation laissée de côté. De cette façon, l'estimation LOOCV est toujours non biaisée malgré l'emploi d'une procédure *stepwise*.

3.4. Analyses et résultats

La méthode ASG est utilisée pour prédire la valeur de mots sur cinq dimensions : la polarité (POL), l'activation (ACT), la dominance (DOM), le caractère abstrait (ABS) et l'imagerie (IMA). Dans le but de comparer l'emploi de corpus et de ressources linguistiques pour le calcul des similarités, deux variantes d'ASG ont été appliquées : une première qui utilise l'ASL et une seconde qui se base sur WordNet. Chacune fait l'objet d'une sous-section, la troisième étant consacrée à leur comparaison.

3 http://medialab.di.unipi.it/wiki/Wikipedia_Extractor (consulté le 1er décembre 2011)

3.4.1. Efficacité de ASG avec une ASL dans la prédiction des normes

ASG utilise l'ASL pour créer un espace sémantique d'où seront tirées les similarités entre les mots. Plusieurs corpus ont été employés pour construire cet espace sémantique, de façon à évaluer l'impact du choix des textes sur les prédictions. La qualité de ces prédictions a été évaluée au moyen de la corrélation (*r* de Bravais-Pearson) entre les valeurs estimées par la LOOCV et les valeurs fournies par les normes.

Le tableau 1 reprend les corrélations moyennes des LOOCV pour les différents ensembles de textes, N renvoyant au nombre de mots utilisés pour calculer ces corrélations. Un même seuil de signification de 0,05 pour la régression a été employé. La qualité des prédictions a été évaluée uniquement sur les mots des normes présents dans les trois espaces sémantiques et donc communs aux trois corpus, de façon à pouvoir directement les comparer.

	POL	**ACT**	**DOM**	**ABS**	**IMA**
N	948	948	948	1698	1698
TASA	0,75	0,57	0,61	0,76	0,69
BNC	0,77	0,59	0,64	0,78	0,67
WIKI	0,73	0,57	0,62	0,77	0,66

Tableau 1 : corrélations pour ASG avec ASL

Tout d'abord, pour un même corpus, nous pouvons observer des différences relativement importantes entre les normes (ex. différence de 0,2 entre l'activation et le caractère abstrait sur WIKI). Il est intéressant de remarquer que l'ordre d'efficacité pour les différentes dimensions est identique pour les trois corpus : les prédictions pour le caractère abstrait sont systématiquement meilleures, devant l'évaluation, l'imagerie, la dominance et l'activation. Cela laisse penser que certaines normes sont par essence plus difficiles à estimer. Deux explications peuvent être avancées. Une première met en cause la méthode qui serait moins efficace pour prédire certaines dimensions. Cela signifierait que ces dimensions ne peuvent pas être prédites sur la base des similarités sémantiques entre les mots. Une deuxième explication potentielle est que certaines dimensions sont plus difficilement estimables par des juges humains, ce qui a un impact sur la prédiction automatique. Nous manquons de données sur les variances des accords inter-juges pour les normes utilisées, ce qui ne nous permet pas de pencher vers une explication plutôt qu'une autre, bien que les deux soient sans doute valables.

Ensuite, les résultats sont plus ou moins constants au travers des différents corpus. Le plus grand écart est de 0,04 entre BNC et WIKI pour la polarité. Le choix du corpus ne semble donc pas être un élément déterminant pour le bon fonctionnement de la méthode ASG, en tout cas pour les corpus généraux non spécifiques à un domaine. La taille ne semble pas non plus importer, TASA comprenant plus ou moins 10 millions de mots, BNC 100 millions et WIKI 600 millions.

3.4.2. Efficacité de ASG avec WordNet dans la prédiction des normes

Les similarités entre les mots sont calculées à partir de WordNet, à l'aide du module Perl *WordNet-Similarity-2.05* de Pedersen *et al.* (2004). Deux similarités ont été testées : une basée uniquement sur la distance entre les mots (PATH) et une basée sur le contenu d'information des

mots (RES). Comme toutes ces mesures sont basées sur la structure en hiérarchie de WordNet, elles n'ont pu être appliquées que sur les noms. De plus, nous n'avons pu calculer des similarités qu'entre les noms des normes présents dans WordNet.

Le tableau 2 reprend les corrélations moyennes des LOOCV pour les différentes mesures de similarités testées. Tout comme pour ASG avec une ASL, un seuil de signification de 0,05 a été utilisé pour la régression. Une transformation logarithmique des données, log(sim + 1), a été effectuée, étant donné que la distribution des similarités se caractérise par une asymétrie positive[4].

	POL	**ACT**	**DOM**	**ABS**	**IMA**
N	731	731	731	1921	1921
PATH	0,46	0,44	0,44	0,78	0,48
RES	0,50	0,37	0,38	0,77	0,49

Tableau 2 : corrélations pour ASG avec WordNet

Force est de constater que les résultats sont globalement moins bons qu'avec l'utilisation d'une ASL à partir d'un corpus. On va même jusqu'à 0,33 de corrélation en moins pour la polarité. Il y a tout de même une dimension pour laquelle l'utilisation de WordNet semble pertinente : la dimension abstrait – concret. On obtient des résultats identiques à ceux obtenus à l'aide d'une ASL. Il semble donc que les relations d'hyperonymie entre les mots – qui forment la hiérarchie des noms dans WordNet – transmettent leur caractère concret ou abstrait.

Il convient toutefois d'être prudent lorsque l'on compare ASG basé sur WordNet et ASG basé sur une ASL. En effet, les résultats rapportés pour les deux méthodes n'ont pas été calculés sur le même nombre de mots. WordNet attribue un score à 217 mots de moins pour les normes ANEW en raison de la présence dans ces normes de mots appartenant à d'autres catégories grammaticales que nominale. Par contre, il attribue un score à 223 mots de plus pour les normes de Gilhooly et Logie (1980) parce que celles-ci ne contiennent que des noms et que WordNet couvre un lexique plus grand que les corpus en raison du seuil de fréquence minimale fixé à 10 pour ceux-ci. La section suivante opère une comparaison plus rigoureuse.

3.4.3. Comparaison des deux approches

Le tableau 3 reprend les corrélations moyennes des LOOCV pour les différentes mesures de similarités testées, qu'elles soient basées sur l'ASL ou sur WordNet, pour les mots communs à ces deux approches. Globalement, on observe très peu de différences avec les tableaux précédents, les corrélations obtenues étant le plus souvent légèrement plus élevées pour les mots communs. Il s'ensuit que cette analyse confirme la supériorité de l'ASL sur WordNet pour estimer des normes sauf pour la dimension abstrait – concret. On note toutefois un écart beaucoup plus élevé pour l'estimation de l'imagerie par WordNet puisque la corrélation s'accroît de 0,07 lorsqu'elle est calculée sur les 1690 mots communs par rapport à celle obtenue pour les

4 Cette transformation logarithmique des données a faiblement amélioré les prédictions pour la similarité PATH et très faiblement diminué celles pour l'activation et la dominance avec la similarité RES, mais sans réel impact.

1921 mots estimables par WordNet. L'origine de cette différence se trouve probablement dans le fait que les juges humains éprouvent des difficultés pour évaluer le degré d'imagerie pour les mots rares (Desrocher et Thompson, 2009). En raison du seuil de fréquence minimale employé pour l'estimation par ASL, on peut penser que les mots spécifiques à l'estimation par WordNet sont relativement rares.

	POL	**ACT**	**DOM**	**ABS**	**IMA**
N	701	701	701	1690	1690
TASA	0,76	0,59	0,63	0,76	0,69
BNC	0,76	0,60	0,64	0,78	0,68
WIKI	0,73	0,57	0,62	0,77	0,66
PATH	0,45	0,45	0,44	0,79	0,56
RES	0,50	0,38	0,37	0,78	0,55

Tableau 3 : corrélations pour ASG avec ASL et WordNet (mots communs)

4. Conclusion

Nous avons présenté une méthode automatique pour estimer des normes lexicales, qui peut reposer aussi bien sur un corpus de textes que sur une ressource linguistique permettant le calcul de similarités entre les mots. Les analyses effectuées, visant à comparer les deux types de support, montrent que le calcul des similarités à partir d'informations sur les cooccurrences des mots en contexte est plus efficace dans le cadre de la prédiction de normes que leur calcul à partir d'informations contenues dans WordNet. Ce résultat est inattendu en raison de la très haute qualité des informations présentes dans WordNet. Une dimension fait tout de même exception, à savoir la dimension concret – abstrait, qui obtient des résultats similaires dans les deux cas. Il serait intéressant de déterminer si, malgré la moindre efficacité des estimations basées sur Wordnet, combiner cette approche avec l'approche basée sur un corpus ne permettrait pas d'améliorer encore les prédictions.

La procédure basée sur un corpus montre également des résultats encourageants pour les dimensions d'imagerie, de polarité et de dominance. Par contre, elle est moins efficace pour l'activation. Des études complémentaires sont nécessaires pour déterminer si cette moindre efficacité trouve son origine dans la méthode, qui serait moins efficace pour prédire certaines normes, ou chez les juges, qui éprouveraient plus de difficultés pour les estimer.

Plusieurs corpus généraux (non spécifiques à un domaine) et de différentes tailles ont été testés afin de voir si l'efficacité de la méthode est fort dépendante du corpus. Tous obtiennent des résultats similaires, et ce pour toutes les dimensions étudiées. Le choix du corpus ne semble donc pas déterminant. Il en est de même pour la taille, un corpus de 10 millions de mots semble suffisant. Par contre, l'impact de la taille des normes n'est que partiellement évalué. Il semble qu'elle ne soit pas déterminante pour la qualité des prédictions (les normes ABS et POL obtenant des résultats similaires sur corpus avec une différence de plus de 700 mots), mais quelle est la taille minimale nécessaire au maintien d'une bonne prédiction ? La question de la spécificité du corpus à un domaine mériterait également d'être étudiée. Il serait intéressant de voir dans quelle

mesure cela affecte les prédictions et si ces dernières reflètent les spécificités du domaine. Dans ce cas précis, la méthode ASG permettrait de produire des normes en fonction de l'usage prévu et de répondre ainsi aux besoins d'applications spécifiques. Enfin, la question de l'origine de la faible efficacité d'ASG avec WordNet pour les dimensions autres que l'abstraction mériterait certainement d'être approfondie. Tous ces points présentent un certain nombre de limites de l'étude présentée et justifient néanmoins des recherches complémentaires.

Références

Aston G. and Burnard L. (1998). *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Bestgen Y. (1994). Can emotional valence in stories be determined from words. *Cognition and Emotion*, vol. 8: 21-36.

Bestgen Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. *Actes de CIFT'*02, pp. 81-94.

Bestgen Y. (2006). Déterminer automatiquement la valence affective de phrases : Amélioration de l'approche lexicale. *Actes des JADT 2006*, pp. 179-188.

Berry M.W. (1992). Large scale singular value computation. *International Journal of Supercomputer Application*, vol. 6: 13-49.

Biber, D. (1998). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Bradley M.M. and Lang P.J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings, Tech. Rep. No. C-1, Gainesville, FL: Center for Research in Psychophysiology, University of Florida.

Bradshaw J.L. (1984). A guide to norms, ratings, and lists. *Memory and cognition*, vol. 12: 202-206.

Cohen, A.S., Minor, K.S., Najolia, G.M. and Hong, S.L. (2009). Laboratory-based procedure for measuring emotional expression from natural speech. *Behavior Research Methods*, vol.41: 204-212.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, vol. 41: 391-407.

Desgraupes B., Loiseau S. and Habert B. (2007). Wikipedia as corpus: the wiki2tei parser. 2007 Scientific Report, http://rs2007.limsi.fr/index.php/LIR:Page_1.

Desrochers A. and Saint-Aubin J. (2008). Sources de matériel en français pour l'élaboration d'épreuves de compétences en lecture et en écriture. *Canadian Journal of Education*, vol. 31: 305-326.

Desrochers, A. and Thompson, G.L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods,* vol. 41: 546-557.

Draper, N.R. and Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York : Wiley.

Esuli A. and Sebastiani F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC'06*, pp. 417-422.

Gilhooly K.J. and Logie R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1,944 words. *Behavior Research Methods and Instrumentation*, vol 12 : 395-427.

Hogenraad R., Bestgen Y. and Nysten J.L. (1995). Terrorist rhetoric: Texture and architecture. In E. Nissan and K.M. Schmidt (Eds.) *From information to knowledge. Conceptual and content analysis by computer*, Oxford, England: Intellect Books, pp. 54-67.

Hogenraad R. and Orianne E. (1981). Valences d'imagerie de 1.130 noms de la langue française parlée. *Psychologica Belgica*, vol. 21 : 21-30.

Hu M. and Liu B. (2004). Mining Opinion Features in Customer Reviews. *Proceedings of AAAI*, pp. 755-760.

Jiang J. and Conrath D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proc. on International Conference on Research in Computational Linguistics*, pp.19-33.

Kamps J. and Marx M. (2002). Words with Attitude. *Proceedings of the 1st Interational Conference on Global WordNet*, pp. 332-341.

Kim S.M. and Hovy E. (2004). Determining the sentiment of opinions. *Proceedings of COLING*, pp. 1367-1373.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998). An introduction to latent semantic analysis, *Discourse Processes*, vol.25: 259-284.

Landauer T.K., Laham D. and Derr M. (2004). From paragraph to graph: Latent Semantic Analysis for information visualization. *Proc. of the National Academy of Science* 101, pp. 5214-5219.

Landauer T.K., McNamura D., Simon D. and Kintsch W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Lavaur J-M. and Font N. (2000). Guide bibliométrique des études normatives et évaluatives pour les recherches en psycholinguistique. *Kabaro*, vol. 1: 141-158.

Maki W.S., McKinley L.N. and Thompson A.G. (2004). Semantic distance norms computed from and electronic dictionary (WordNet). *Behavior Research Methods , Instruments, & Computers*, vol. 36 (3) : 421-431.

Miller G.A., Beckwith R., Fellbaum C., Gross D. and Miller K. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, vol. 3: 235-244.

Pang B. and Lee L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, vol. 2: 1-135.

Pedersen T., Patwardhan S. and Michelizzi J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. *Proc. of HLT-NAACL*, pp. 38-41.

Pennebaker, J.W., Mehl, M.R. and Niederhoffer, K.G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, vol.54: 547-577.

Popping, R. (2000). *Computer-assisted Text Analysis*. London: Sage.

Resnick P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proc. of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44-49.

Schulerud H. and Albregtsen F. (2004). Many are called, but few are chosen: Feature selection and error estimation in high dimensional spaces. *Computer Methods and Programs in Biomedicine*, vol. 73: 91-99.

Stone, P.J. (1997). Thematic text analysis: New agendas for analyzing text content. In Roberts C.W. (Eds.). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts.* Mahwah, NJ: Erlbaum, pp. 35-54.

Turney P.D. and Littman M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report, National Research Council Canada, 2002.

Turney P.D. and Littman M. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, vol. 21: 315-346.

Vincze N. and Bestgen Y. (2011). Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. *Actes de TALN11*, vol. 1 : 223-234.