

Mesures de similarité distributionnelle entre termes *

Sondes Bannour, Laurent Audibert, Adeline Nazarenko

LIPN, UMR 7030 CNRS, Université Paris 13
99 av. J.-B. Clément - F-93430 Villetaneuse, France
prenom.nom@lipn.univ-paris13.fr

Résumé :

Cet article présente une étude de mesures de similarité entre des termes tels qu'ils peuvent être extraits par un extracteur de termes, en s'appuyant sur l'hypothèse distributionnelle selon laquelle des termes sémantiquement proches tendent à apparaître dans des contextes similaires. Les mesures de similarités étudiées combinent des fonctions de poids et des mesures de similarité élémentaires. Nous redéfinissons, dans un premier temps, la notion d'analyse distributionnelle habituellement appliquée sur des mots pour la combiner à une analyse terminologique et prendre en compte les unités composées. Cela implique une adaptation des fonctions de poids et de la notion de contexte. Nous présentons ensuite une étude méthodique de différentes combinaisons de fonctions de poids et de mesures de similarité. Certaines produisent des résultats convaincants. Intégrées dans une plateforme d'aide à la construction d'ontologies à partir de textes, de telles mesures ont vocation à faciliter le travail d'acquisition.

Mots-clés : Construction d'ontologies à partir de textes, approche distributionnelle, termes, contexte, poids, mesures de similarité

1 Introduction

Depuis son émergence, au début des années 1990, dans les recherches liées à l'ingénierie des connaissances, la notion d'ontologie est devenue le centre d'intérêt de plusieurs domaines en Informatique. Définie comme étant la représentation explicite, formelle et consensuelle des concepts propres à un domaine et des relations qui les relient (Gruber, 1993), la notion d'ontologie

*. Ce travail a été réalisé dans le cadre du programme Quaero, financé par OSEO, agence nationale de valorisation de la recherche ; il bénéficie également des résultats du projet Ontorule.

fournit une structure conceptuelle à partir de laquelle il est possible de développer des systèmes à base de connaissances qui soient partageables et réutilisables, notamment dans des domaines comme le Traitement Automatique des Langues (TAL), la Recherche d'Information (RI) ou le web sémantique.

De nombreux travaux ont cherché à s'appuyer sur des textes pour construire des ontologies (Cimiano, 2006). Deux grandes méthodes ont été proposées. La première consiste à regrouper les mots du corpus en classes sémantiques, avec l'idée que ces classes sémantiques représentent les concepts centraux du domaine reflétés par le corpus. Cette analyse distributionnelle « à la Harris » est une technique exploitée depuis longtemps pour la construction de classes sémantiques de mots. Elle consiste à rapprocher les mots sur la base de contextes qu'ils partagent, en faisant l'hypothèse que les mots les plus proches sémantiquement sont employés de manière similaire et tendent à apparaître dans les mêmes contextes. Malheureusement les classes sémantiques obtenues sont généralement assez bruitées et difficiles à traduire directement en concepts. L'autre approche repose sur l'analyse terminologique du corpus, la liste des termes du domaine donnant elle-même un aperçu de son « vocabulaire conceptuel ». Travailler sur des termes extraits par un extracteur de termes plutôt que sur des mots se justifie dans la mesure où les mots composant un corpus ne sont pas tous pertinents (mots vides, mots très fréquents, etc.) et où les termes complexes perdent de leur intérêt si on les découpe (par exemple, dans le domaine de l'astronomie, le terme « trou noir » a beaucoup plus de sens que les mots « trou » et « noir » considérés séparément). Cette deuxième approche a été mise en œuvre dans des outils d'acquisition d'ontologies comme TERMINAE (Szulman *et al.*, 2002) mais elle n'offre qu'une aide limitée à l'ingénieur de la connaissance qui doit souvent analyser et explorer de longues listes de termes pour identifier ceux qui renvoient à des concepts clefs du domaine.

Notre objectif ici consiste à combiner ces approches terminologique et distributionnelle pour assister davantage la construction d'ontologies. Nous cherchons à structurer la liste des termes proposés à l'ingénieur de la connaissance pour lui permettre d'explorer les termes les plus similaires de manière groupée.

Nous cherchons donc à mettre en œuvre une analyse distributionnelle sur des termes. De nombreuses mesures de similarité ont été proposées mais elles ont été testées pour des mots simples (section 2). Après avoir brièvement présenté l'extracteur que nous utilisons (section 3), nous proposons une définition de la notion de contexte adaptée au traitement des termes qui peuvent être des

unités polylexicales (section 4). Nous comparons ensuite dans la section 5 les résultats obtenus pour les différentes méthodes de calcul de similarité, ce qui permet de déterminer celle(s) qui s'applique(nt) le mieux aux termes, qu'ils soient simples ou complexes. Nous proposons enfin en section 6 une expérience sur une ontologie réelle pour valider notre approche.

2 Mesures de similarité existantes

2.1 Généralités

L'étude de la similarité entre les mots a fait l'objet de nombreuses recherches. Lee (1999) analyse certaines mesures de similarité en prenant en considération leurs propriétés formelles. Lin (1998a) décrit et compare un ensemble de mesures de similarité. Weeds & Weir (2003) évaluent les mesures de similarité proposées par Lee (1999) et Lin (1998a) en termes de rappel et de précision. Strehl (2002) effectue une comparaison détaillée de mesures et étudie leur influence sur la classification ultérieure des mots (clustering).

Curran (2004) décompose les mesures de similarité existantes en deux composantes : les mesures et les poids. En effet, les approches de similarité distributionnelle sont fondées sur l'hypothèse que la similarité sémantique est reflétée par la similarité des contextes. Or, dans un contexte, tous les mots n'ont pas la même importance. Il paraît donc judicieux de pondérer leur impact avant l'application d'une mesure de similarité.

TF-IDF	$\frac{f(m, r, m')}{n(*, r, m')}$
GRAF94	$\frac{\log_2(f(m, r, m') + 1)}{\log_2(n(*, r, m') + 1)}$
MI	$\log_2\left(\frac{p(m, r, m')}{p(m, *, *)p(*, r, m')}\right)$
TTEST	$\frac{p(m, r, m') - p(*, r, m')p(m, *, *)}{\sqrt{p(*, r, m')p(m, *, *)}}$

TABLE 1: Fonctions de poids calculées pour un mot m

Bourigault (2002) a déjà proposé une analyse distributionnelle sur des unités complexes à l'aide de son outil UPERY mais son approche repose sur une analyse syntaxique du corpus. Nous cherchons, quant à nous, à faire une analyse distributionnelle sans prendre en compte d'autres informations syntaxiques que celles qui sont éventuellement fournies par des extracteurs de

COSINUS	$\frac{\sum \text{poids}(m_1, *r, *m') \times \text{poids}(m_2, *r, *m')}{\sqrt{\sum \text{poids}(m_1, *, *)^2 \times \sum \text{poids}(m_2, *, *)^2}}$
DICE†	$\frac{2 \sum \min(\text{poids}(m_1, *r, *m'), \text{poids}(m_2, *r, *m'))}{\sum \text{poids}(m_1, *r, *m') + \text{poids}(m_2, *r, *m')}$
JACCARD	$\frac{\sum \min(\text{poids}(m_1, *r, *m'), \text{poids}(m_2, *r, *m'))}{\sum \max(\text{poids}(m_1, *r, *m'), \text{poids}(m_2, *r, *m'))}$
où $\sum \text{poids}(m_i, *r, *m') \times \text{poids}(m_j, *r, *m')$ $\equiv \sum_{(r, m') \in ((m_i, *, *) \cap (m_j, *, *))} \text{poids}(m_i, r, m') \times \text{poids}(m_j, r, m')$	

TABLE 2: Mesures de similarité entre deux mots m_1 et m_2

termes. La question est de savoir si on peut obtenir de bons résultats à partir de la simple observation des cooccurrences de termes en corpus.

Les fonctions de poids et les mesures que nous évaluons dans cet article sont extraites des travaux de Curran (2004) qui reprend les notations de Lin (1998b). Elles sont résumées dans les tableaux 1 et 2 où :

- (m, r, m') est un élément de contexte où un mot m' entretient une relation r avec le mot m
- $p(m, r, m')$ est la probabilité de trouver les mots m et m' en relation r
- $f(m, r, m')$ est le nombre d'occurrences des mots m et m' se trouvant dans une relation r
- $f(*, r, m')$ est le nombre d'occurrences de mots qui apparaissent avec le mot m' dans une relation r
- $n(*, r, m')$ est le nombre de mots différents qui apparaissent avec le mot m' dans une relation r

2.2 Fonctions de poids TF-IDF et GREF94

Les fonctions de poids TF-IDF et GREF94 sont des fonctions issues du domaine de la recherche d'information qui reposent sur le principe de TF-IDF (Fréquence du terme - Inverse de la Fréquence du Document). Dans le cas de la similarité basée sur les contextes, TF représente la fréquence des éléments de contexte de m ($f(m, r, m')$) et IDF représente le nombre de mots différents avec lesquels m' apparait ($n(*, r, m')$). Une fréquence élevée des éléments de contexte indique que l'attribut (r, m') est pertinent alors qu'un nombre élevé d'attributs indique que l'attribut est peu discriminant. TF-IDF équilibre donc ces deux facteurs.

Witten *et al.* (1999) décrivent d'autres variations de TF-IDF. L'ajout de 1 permet lorsque $f(m, r, m') = 1$ d'avoir un poids égal à 1 après l'application du logarithme (Grefenstette, 1994).

2.3 Fonction de poids MI

La fonction de similarité probablement la plus utilisée dans le domaine du TAL est l'information mutuelle (Fano, 1963) généralement définie comme :

$$I(x, y) = \log\left(\frac{P(x, y)}{P(x)P(y)}\right)$$

Il s'agit d'une information mutuelle ponctuelle qui compare la probabilité d'observer deux événements aléatoires x et y ensemble (distribution jointe) aux probabilités de les observer indépendamment (distribution indépendante). Si l'association entre x et y est forte, la probabilité jointe $P(x, y)$ est plus importante que $P(x)P(y)$ et par conséquent $I(x, y) > 0$.

Hindle (1990) utilise l'information mutuelle comme une fonction de poids entre les mots et les attributs (les relations sujet et objet dans son cas) dans la mesure de la similarité vectorielle :

$$I(m, r, m') = \log\left(\frac{p(m, r, m')}{p(m, *, *)p(*, r, m')}\right) = \log\left(\frac{f(m, r, m')f(*, *, *)}{f(m, *, *)f(*, r, m')}\right)$$

Il est courant de restreindre la variation de l'information mutuelle aux valeurs non négatives, comme l'ont fait Lin (1998a) et Ido *et al.* (1993).

2.4 Fonction de poids TTEST

La mesure de poids TTEST compare une valeur x à une distribution normale définie par sa moyenne μ , sa variance d'échantillon s^2 et son effectif N :

$$\tau = \frac{x - \mu}{s} \sqrt{N}$$

ce qui se traduit dans notre cas par :
$$\frac{p(m, r, m') - p(*, r, m')p(m, *, *)}{\sqrt{p(*, r, m')p(m, *, *)}}$$

2.5 Mesures de similarité JACCARD, DICE† et COSINUS

Les mesures de similarité JACCARD, DICE† et COSINUS sont largement utilisées dans le domaine de la recherche d'information. Elles ont été étendues pour prendre en compte les poids.

Les mesures JACCARD et DICE sont issues d'analyses écologiques respectivement au début et au milieu du XX^{ème} siècle.

La mesure de JACCARD compare le nombre d'attributs communs avec le nombre d'attributs uniques pour une paire de mots. Elle a été généralisée par Grefenstette (1994) en remplaçant l'intersection avec le poids minimum et l'union avec le poids maximum.

La mesure de DICE, étant deux fois le rapport entre le nombre d'attributs communs et le nombre total des attributs de chaque mot, a été transformée de

la même manière.

Les généralisations des mesures de DICE et JACCARD peuvent être équivalentes suivant la méthode considérée.

La mesure de COSINUS, initialement issue de l'algèbre linéaire, s'étend également aux vecteurs pondérés et est devenue la mesure standard des vecteurs pondérés dans le domaine de la recherche d'information. (Witten *et al.*, 1999) préconise l'utilisation de la mesure COSINUS dans le domaine de la recherche d'information plutôt que le PRODUIT SCALAIRE, ou les DISTANCES DE MINKOWSKI, car COSINUS résout certains problèmes inhérents à ces deux dernières (favorisation des vecteurs longs, discrimination des vecteurs dont la différence entre les longueurs est significative, etc.).

3 Extraction des termes

Afin d'extraire les termes sur lesquels nous travaillons, nous utilisons l'outil YaTeA¹ développé au LIPN (Hamon & Aubin, 2006). YaTeA est un outil qui permet de générer une liste de termes candidats à partir d'un corpus segmenté, lemmatisé et étiqueté morpho-syntaxiquement. La figure 3 illustre le travail d'un tel extracteur de termes.

Extrait de corpus
En astronomie, un amas globulaire est un amas stellaire très dense, contenant typiquement une centaine de milliers d'étoiles distribuées dans une sphère dont la taille varie de 20 à quelques centaines d'années lumière.
Termes candidats extraits
amas globulaire, globulaire, amas stellaire, centaine, centaine de milliers, stellaire, amas, milliers

TABLE 3: Termes candidats identifiés par l'extracteur de termes **YaTeA** dans un extrait de corpus

4 Définition du contexte d'un terme

Pour exploiter l'hypothèse distributionnelle sur les termes, il faut définir formellement la notion de contexte d'un terme. En effet, étant donné que la fréquence des termes composés dans un corpus est bien plus faible que celle des mots qui le composent, utiliser des approches statistiques en considérant

1. Disponible en téléchargement depuis <http://search.cpan.org/~thhamon/Lingua-YaTeA/>.

ces termes comme des entités atomiques (après figement) s'avère problématique. Par ailleurs, des termes composés partageant la même tête (« étoile binaire » et « étoile massive ») ou imbriqués (« trou noir » et « trou noir supermassif ») risquent de se retrouver sans lien de proximité si on ne considère pas le fait qu'ils partagent des mots dans leur structure. Pour résoudre ces problèmes, la définition du contexte repose sur des mots (les termes composés présents dans le contexte sont donc décomposés) et comprend également les mots qui composent le terme cible. Ce choix permet de recueillir plus de preuves distributionnelles. Concrètement, le contexte d'un terme est défini comme l'ensemble des noms et verbes figurant dans la même phrase. Des fenêtrages plus petits ont été expérimentés (par exemple les deux ou trois mots qui suivent et précèdent le terme en question) mais les meilleurs résultats sont obtenus dans le cas de l'utilisation de la phrase comme fenêtre. Les mots vides (prépositions, articles, pronoms...) ainsi que les verbes d'état sont exclus de la définition du contexte car les expériences ont montré qu'ils sont peu informatifs. Les adjectifs sont également exclus car nous avons expérimentalement observé que les conserver dégrade les résultats sur nos corpus². Noms et verbes sont considérés sous leur forme lemmatisée.

Le terme, essentiellement le terme composé, est donc à la fois vu comme un élément atomique objet de notre étude, et comme un élément composite, puisque les mots qui le composent sont pris en compte dans la définition du contexte du terme. La figure 1 illustre cette notion de contexte d'un terme.

Cette définition de contexte, tout comme les méthodes de calcul de contexte à base de fenêtres glissantes, a l'avantage d'être pratiquement indépendante de la langue une fois le texte segmenté.

Phrase	
Une étoile est un <i>objet céleste</i> en rotation, de forme approximativement sphérique.	
Terme extrait	Contexte du terme
étoile	étoile, objet, rotation, forme
objet céleste	étoile, objet, rotation, forme

FIGURE 1: Illustration des contextes des termes « étoile » et « objet céleste » dans une phrase donnée

Formellement, un élément de contexte est représenté par un tuple (t, r, m) où t est le terme en question, m est un mot figurant dans le contexte du terme,

2. Nous envisageons d'étudier le rôle des adjectifs plus en détail par la suite.

et r est la relation « est dans la même phrase que ». Dans cette étude, nous ne considérons pas les relations syntaxiques. r devenant une relation invariable, le couple (t, m) suffit pour définir un élément de contexte d'un terme t .

Afin de pouvoir appliquer les fonctions de poids et les mesures sur les termes, nous devons les adapter à notre définition du contexte. Nous posons ainsi :

$N_{t,m}$: nombre de cooccurrences du terme t et du mot m dans une même phrase

$N_{t,.}$: nombre d'occurrences de mots qui cooccurrent avec le terme t dans une même phrase

$NT_{t,.}$: nombre de mots différents qui cooccurrent avec le terme t dans une même phrase

$N_{.,m}$: nombre d'occurrences de termes qui cooccurrent avec le mot m dans une même phrase

$NT_{.,m}$: nombre de termes différents qui cooccurrent avec le mot m dans une même phrase

$N_{.,.}$: nombre total d'occurrences de cooccurrences

Voici un exemple pour illustrer ces notations. Prenons l'extrait :

*Un **trou noir** possède une masse donnée, concentrée en un point appelé **singularité gravitationnelle**. Cette masse permet de définir une sphère appelée horizon du **trou noir***

Termes extraits : trou noir, singularité gravitationnelle

Mots du contexte dans la première phrase : trou, posséder, masse, point, singularité

Mots du contexte dans la deuxième phrase : masse, permettre, définir, sphère, horizon, trou

Dans cet exemple, nous avons :

$$- N_{\text{trou noir}, \text{masse}} = 2$$

$$- N_{\text{trou noir}, .} = 11$$

$$- NT_{\text{trou noir}, .} = 9$$

$$- N_{., \text{masse}} = 3$$

$$- NT_{., \text{masse}} = 2$$

$$- N_{.,.} = N_{\text{trou noir}, .} + N_{\text{Singularite gravitationnelle}, .} = 11 + 5 = 16$$

Les formules de poids pour les termes sont définies dans le tableau 4. Les fonctions de similarités restent, quant à elles, inchangées en appliquant les poids définis dans ce tableau.

TF-IDF	$\frac{f(t, r, m)}{n(*, r, m)} = \frac{N_{t,m}}{NT_{.,m}}$
GRAF94	$\frac{\log_2(f(t, r, m) + 1)}{\log_2(n(*, r, m) + 1)} = \frac{\log_2(N_{t,m} + 1)}{\log_2(NT_{.,m} + 1)}$
MI	$\log_2\left(\frac{p(t, r, m)}{p(t, *, *)p(*, r, m)}\right) = \log_2\left(\frac{N_{t,m} \times N_{.,*}}{N_{t,*} \times N_{.,m}}\right)$
TTEST	$\frac{p(m, r, m') - p(*, r, m')p(m, *, *)}{\sqrt{p(*, r, m')p(m, *, *)}} = \frac{N_{.,*} \times N_{t,m} - N_{.,m} \times N_{t,*}}{N_{.,*} \times \sqrt{N_{t,m} \times N_{t,*}}}$

TABLE 4: Adaptation des fonctions de poids

5 Évaluation des mesures sur un corpus d'astronomie

5.1 Corpus et protocole expérimental

Dans cette première expérience, nous considérons un corpus d'astronomie en langue française traitant de différents objets célestes naturels, et plus particulièrement du cycle de vie des étoiles. Ce corpus, de taille modérée, totalise 55 921 mots et a été constitué manuellement à partir de 23 articles de Wikipedia. Dans ces articles, les termes importants liés au terme vedette (titre de l'article) possèdent des liens hypertextes renvoyant vers d'autres articles de Wikipedia. Ces liens ont permis de constituer automatiquement une petite ressource terminologique. En effet, nous avons considéré que tous les termes d'un article de Wikipedia qui servent d'ancre à un renvoi hypertexte sont sémantiquement proches du terme vedette de cet article et nous avons sélectionné 12 termes vedettes de Wikipedia : *amas*, *astéroïde*, *comète*, *étoile*, *galaxie*, *nébuleuse*, *planète*, *pulsar*, *satellite naturel*, *soleil*, *supernova* et *trou noir*. Nous utilisons les 12 listes de termes correspondant à ces termes vedettes pour évaluer la pertinence de nos mesures de similarité.

Sur ce corpus, YaTeA produit une liste de 5 562 termes candidats représentés sous une forme canonique. Cette liste contient du bruit et est composée, pour près de 75%, de termes n'apparaissant qu'une seule fois dans le corpus. Ces termes dits « hapax », ne pouvant pas être pris en compte dans l'analyse distributionnelle, sont exclus de fait de cette étude³ pour aboutir à une liste de 1 413 termes.

3. Le traitement des hapax nécessiterait de prendre en compte d'autres critères de similarité, notamment ceux qui reposent sur la structure interne des termes.

5.2 Résultats

En utilisant la définition de contexte d'un terme présentée dans la section 4, nous avons mesuré la similarité entre les différents termes en mesurant la similarité de leurs contextes, conformément à l'hypothèse distributionnelle selon laquelle des termes sémantiquement proches apparaissent dans des contextes similaires.

L'approche de similarité poids-mesure pour laquelle nous avons opté utilise les poids pour mesurer l'importance d'un mot dans la définition du contexte d'un terme, et les mesures pour comparer les représentations pondérées des contextes des différents termes. Les douze combinaisons poids-mesure résultantes des différentes fonctions de poids présentées dans le tableau 4 et des mesures présentées dans le tableau 2 ont été implémentées et testées.

Pour évaluer ces mesures de similarité, nous avons comparé, pour chacun des 12 termes vedettes présentés ci-dessus, la liste des termes jugés les plus similaires selon la mesure de similarité avec la liste de référence du terme établie à partir des liens hypertexte de Wikipedia (section 5.1). Nous avons calculé la précision de l'ensemble des 5, 10, 40 et 80 termes jugés les plus similaires, selon les douze combinaisons poids-mesure. La moyenne des précisions obtenues pour les douze termes vedettes nous permet de comparer ces différentes combinaisons comme le montre le tableau 5.

		Précision moyenne en considérant les n premiers termes				
Poids	Mesure	n=5	n=10	n=20	n=40	n=80
MI	JACCARD	15%	13%	9%	12%	12%
MI	DICE†	15%	13%	9%	12%	12%
MI	COSINUS	38%	35%	25%	20%	17%
TFIDF	JACCARD	2%	1%	2%	2%	4%
TFIDF	DICE†	2%	1%	2%	2%	4%
TFIDF	COSINUS	32%	32%	26%	20%	17%
TTEST	JACCARD	20%	13%	10%	11%	9%
TTEST	DICE†	20%	13%	10%	11%	9%
TTEST	COSINUS	47%	36%	25%	23%	18%
GRAF94	JACCARD	3%	3%	3%	3%	5%
GRAF94	DICE†	5%	4%	4%	4%	5%
GRAF94	COSINUS	38%	36%	25%	23%	19%

TABLE 5: Comparaison des différentes combinaisons poids-mesure implémentées

Comme l'on pouvait s'y attendre, les meilleures précisions sont obtenues pour les 5 premiers termes, et cette précision va décroissant lorsque l'on considère des termes de plus en plus éloignés sachant que la taille moyenne des listes de référence est de 95 termes.

Les combinaisons TFIDF_JACCARD, TFIDF_DICE†, GREF94_JACCARD et GREF94_DICE† produisent des valeurs de précision très faibles. En effet, l'analyse des formules permet de comprendre que ces combinaisons sont incompatibles et font remonter artificiellement des termes non pertinents avec une mesure de similarité maximale (égale à 1).

La mesure COSINUS est celle qui donne les meilleurs résultats, indépendamment de la fonction de poids avec laquelle elle est combinée. Les meilleures combinaisons sont TTEST_COSINUS et GREF94_COSINUS, avec une meilleure précision pour TTEST_COSINUS sur les 5 premiers termes.

étoile		trou noir		galaxie		amas		planète	
Terme	S	Terme	S	Terme	S	Terme	S	Terme	S
massive	0,27	trou	0,99	active	0,40	amas globulaire	0,76	planète naine	0,53
binaire	0,27	noir	0,98	galaxie spirale	0,37	globulaire	0,76	mars	0,38
étoile massive	0,26	supermassif	0,36	galaxie active	0,37	amas ouvert	0,51	planète géante	0,35
faible	0,26	trou noir supermassif	0,34	galaxie naine	0,35	ouvert	0,44	dixième	0,33
autre étoile	0,26	trou noir stellaire	0,33	spirale	0,35	amas stellaire	0,41	pluton	0,32
faible masse	0,26	horizon	0,31	galaxie elliptique	0,33	même âge	0,25	planète tellurique	0,32
étoile jeune	0,25	supermassifs	0,30	supermassif	0,31	amas de galaxie	0,24	notion	0,31
étoile binaire	0,25	trous noirs supermassifs	0,29	trou noir supermassif	0,30	omega centauri	0,24	notion de planète	0,31
filante	0,25	charge	0,29	elliptique	0,30	numéro	0,24	autre planète	0,29
étoile filante	0,25	entropie	0,28	centre de la galaxie	0,26	plupart des amas globulaires	0,23	couleur rouge	0,26

TABLE 6: Les dix termes les plus similaires selon TTEST_COSINUS pour une sélection de cinq termes vedettes. La colonne **S** désigne la mesure de similarité (comprise entre 0 et 1). Les cellules grisées correspondent aux termes appartenant aux listes de référence.

Les valeurs de précision présentées dans le tableau 5 sont utiles pour com-

parer les différentes mesures, mais prises individuellement, leur valeur absolue n'est pas très informative puisque de nombreux termes pertinents ne figurent pas dans les listes de termes de référence (par exemple « étoile massive », « étoile jeune » et « étoile filante » dans le cas de « étoile »).

Prenons, par exemple, la combinaison TTEST_COSINUS. Le tableau 6 présente les dix termes les plus similaires pour une sélection de cinq termes vedettes. Pour le terme vedette « galaxie », par exemple, nous remarquons que les termes « active », « spirale » et « elliptique » ne sont pas pris en compte dans le calcul de la précision, parce qu'ils sont absents des listes de référence qui sont nécessairement partielles du fait de leur mode de construction automatisé. Ces termes sont pourtant pertinents : ils décrivent les différents types de galaxies selon la classification de Hubble. Le tableau 7 propose une correction manuelle de la précision.

	Précision sur les 5 premiers termes		Précision sur les 10 premiers termes	
	PRef	PEstim	PRef	PEstim
amas	60%	100%	40%	70%
étoile	0%	60%	10%	70%
galaxie	60%	100%	50%	90%
planète	80%	80%	50%	50%
trou noir	60%	100%	50%	90%

TABLE 7: Précisions basées sur les listes de référence (PRef) et précision estimée manuellement (PEstim) pour les termes du tableau 6

6 Ordonnement d'une liste de termes dans une perspective d'aide à la construction d'ontologies

6.1 Corpus et protocole expérimental

Dans cette deuxième expérience, nous travaillons sur un corpus en langue anglaise qui traite du programme de fidélisation de la compagnie Aérienne « American Airlines ». Ce corpus de petite taille, fourni dans le cadre du projet européen Ontorule, totalise 5 915 mots. Une ontologie de 211 classes a été construite à partir de ce corpus à l'aide de la plateforme d'aide à la construction d'ontologies TERMINAE (Szulman *et al.*, 2002). À partir de cette ontologie, 15 concepts centraux⁴ ont été sélectionnés. Pour chacun de ces concepts,

4. Concepts non abstraits pour posséder des occurrences dans le corpus et non terminaux dans la taxonomie pour posséder des enfants.

une liste de référence a été construite en considérant ses instances (occurrences du terme vedette et de ses synonymes⁵) et celles de ses fils.

Après avoir évalué les différentes combinaisons poids-mesure dans l'expérience précédente et en nous appuyant sur la même définition du contexte d'un terme, nous retenons les deux meilleures combinaisons TTEST_COSINUS et GREF94_COSINUS. Ces deux combinaisons sont appliquées aux termes candidats extraits par YaTeA et qui sont au nombre de 438 (après avoir éliminé les hapax) et évaluées sur les 15 termes vedettes correspondant aux 15 concepts retenus, de la même manière que l'expérience précédente. L'objectif de cette expérience est d'évaluer si une mesure de similarité permet d'ordonner de manière pertinente pour l'ontologue la liste de termes qui lui est présentée quand il travaille sur un terme donné.

6.2 Résultats

		Précision moyenne en considérant les n premiers termes		
Poids	Mesure	n=5	n=10	n=20
TTEST	COSINUS	39%	33%	26%
GREF94	COSINUS	41%	36%	24%

TABLE 8: Précisions mesurées pour les combinaisons TTEST_COSINUS et GREF94_COSINUS

Les résultats de l'expérience sont présentés dans le tableau 8. Étant donné que la taille moyenne des listes de référence est de 20 termes, nous avons mesuré la précision moyenne de l'ensemble des 5, 10 et 20 termes jugés les plus similaires à nos 15 termes vedettes. Les résultats sont déjà encourageants. Sur les 5 premiers termes, 40% correspondent soit directement au concept du terme vedette, soit à l'un de ses fils. Cette précision chute à 25% pour une fenêtre de 20 termes.

Qu'en est-il des termes jugés « non pertinents » ? S'agit-il de bruit, ou de termes tout de même rattachés à un concept de l'ontologie ? Afin de répondre à cette question, nous avons réunis tous les termes qui ont servi à la construction de l'ontologie (c.-à-d. toutes les instances de concepts) dans une unique liste de référence totalisant 257 termes. Nous avons ensuite recalculé les précisions par rapport à cette liste. Les résultats observés pour les 15 termes vedettes sont résumés dans le tableau 9 :

5. Pour être totalement exacts, nous considérons aussi les parties des ces termes dans cette expérience.

		Précision moyenne en considérant les n premiers termes		
Poids	Mesure	n=5	n=10	n=20
TTEST	COSINUS	88%	82%	77%
GRAF94	COSINUS	88%	81%	75%

TABLE 9: Précisions mesurées pour les combinaisons TTEST_COSINUS et GRAF94_COSINUS dans le cas d'une seule liste de référence

Ainsi, sur les cinq premiers termes présentés à l'ontologie travaillant sur un terme vedette, 40% sont directement pertinents (tableau 8) et pratiquement 90% sont pertinents au regard de l'ontologie en cours de construction (tableau 9). Sur les vingt premiers termes, ces chiffres sont respectivement de 25% et plus de 75% ce qui reste encore très appréciable. Ainsi en moyenne, sur les 20 premiers termes présentés à l'ontologie travaillant sur un terme vedette, un quart sont directement pertinents pour le concept auquel est rattaché le terme vedette, une moitié est pertinente au regard de l'ontologie en cours de construction et seulement un quart vient « parasiter » le travail de l'ontologie.

Terme vedette ticket			
1	award ticket	11	award
2	fare ticket	12	attorney fee
3	rate	13	travel
4	rate ticket	14	full-fare economy class ticket
5	agency	15	aadvantage member
6	industry	16	economy class ticket
7	cost	17	charter
8	consolidator fare	18	charter flight ticket
9	unpublished fare ticket	19	flight ticket
10	fare	20	flight number

TABLE 10: Les vingt termes les plus similaires au terme *Ticket* selon TTEST_COSINUS

Prenons un exemple concret : considérons les 20 termes les plus similaires au terme vedette « Ticket » selon la mesure TTEST_COSINUS (tableau 10). Les cases colorées en gris foncé correspondent à des termes qui ont servi à la construction de la classe « Ticket », les cases colorées en gris clair correspondent à des termes qui ont servi à la construction d'autres classes de l'ontologie et les cases non colorées correspondent à des termes considérés non

pertinents dans la construction de l'ontologie. Prenons par exemple les termes « *aadvantage member* » et « *flight number* », ces deux termes sont pertinents pour les classes « *Member* » et « *Flight_number* » respectivement mais leur apparition dans la liste des termes les plus similaires au terme « *Ticket* » est plutôt intéressante. En effet, dans l'ontologie, la classe « *Member* » entretient avec la classe « *Ticket* » la relation *purchase* et la classe « *Ticket* » entretient avec la classe « *Flight_number* » la relation *reflects*. Autrement dit, la mesure de similarité a permis non seulement de faire remonter des termes pertinents pour le terme vedette ou pour l'ontologie dans son ensemble mais aussi d'explorer d'éventuels liens entre certaines classes de l'ontologie, ce qui constitue une aide considérable dans le processus de construction de l'ontologie.

7 Conclusion

En nous appuyant sur les études existantes de similarité entre mots, nous avons proposé une approche permettant de mesurer la similarité entre des termes extraits par un extracteur de termes. Dans cette perspective, nous avons introduit une définition du contexte d'un terme composé qui tient compte à la fois de sa nature atomique, puisque les mesures de similarité s'appliquent dans notre cas sur des termes, et de son aspect composite, puisque les mots qui le composent sont pris en compte individuellement lors de l'extraction des contextes. Cette définition de contexte ne repose pas sur la syntaxe, ce qui facilite la mise en œuvre de notre approche. Nous avons ensuite présenté une étude de différentes combinaisons de fonctions de poids et de mesures de similarité dont certaines produisent des résultats tout à fait convaincants.

Comme nous l'avons montré dans une seconde expérience, ces mesures permettent de faire face au bruit important produit par un extracteur de termes. Par exemple, de telles mesures pourraient être intégrées dans une plateforme d'aide à la construction d'ontologies à partir de textes afin de mieux structurer la liste des termes traitée. Un autre exemple d'application est le cas où l'on dispose d'une ressource terminologique pertinente, mais incomplète, pour un domaine donné. Un bon moyen de compléter automatiquement cette ressource, sans introduire trop de bruit, serait d'extraire les termes les plus proches des termes constituant la ressource existante à partir d'un corpus du domaine.

La prochaine étape de notre travail consiste à utiliser notre module de calcul de similarité pour regrouper des termes en clusters et ainsi apporter une aide supplémentaire à la structuration d'une ontologie.

Remerciements

Nous tenons à remercier Thibault Mondary (LIPN) qui a constitué le corpus d'astronomie et la ressource terminologique associée, ainsi que Nouha Omrane (LIPN) et Sylvie Szulman (LIPN) pour la construction de l'ontologie associée au corpus « American Airlines ».

Références

- BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *TALN '02*, p. 75–84.
- CIMIANO P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation And Applications*. New York, USA : Springer.
- CURRAN J. R. (2004). *From Distributional to semantic Similarity*. PhD thesis, University of Edinburgh.
- FANO R. M. (1963). *Transmission of Information : a Statistical Theory of Communications*. MIT Press, Cambridge, MA USA.
- GREFENSTETTE G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA : Kluwer Academic Publishers.
- GRUBER T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**, 199–220.
- HAMON T. & AUBIN S. (2006). Improving term extraction with terminological resources. In *FinTAL '06*, p. 380–387.
- HINDLE D. (1990). Noun classification from predicate.argument structures. In *ACL '90*, p. 268–275.
- IDO D., MARCUS S. & MARKOVITCH S. (1993). Contextual word similarity and estimation from sparse data. In *ACL '93*, p. 164–171.
- LEE L. (1999). Measures of distributional similarity. In *ACL '99*, p. 25–32.
- LIN D. (1998a). Automatic retrieval and clustering of similar words. In *ACL '98*, p. 768–774.
- LIN D. (1998b). Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, p. 57–63.
- STREHL A. (2002). *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, The University of Texas.
- SZULMAN S., BIÉBOW B. & AUSSENAC-GILLES N. (2002). Structuration de terminologies à l'aide d'outils d'analyse de textes avec terminae. *TAL*, **43**(1), 103–128.
- WEEDS J. & WEIR D. (2003). A general framework for distributional similarity. In *EMNLP '03*, p. 81–88.
- WITTEN I. H., MOFFAT A. & BELL T. C. (1999). *Managing gigabytes : compressing and indexing documents and images*. Morgan Kaufmann.