Une approche multi-vue pour l'extraction terminologique bilingue

Raphaël Rubino — Georges Linarès

Laboratoire Informatique d'Avignon
339, chemin des Meinajaries
84911 AVIGNON
{raphael.rubino, georges.linares}@univ-avignon.fr

RÉSUMÉ. *Ce papier présente une approche multi-vue pour la traduction de termes de spécialité, basée sur un lexique bilingue et un corpus comparable. Nous proposons d'étudier différents niveaux de représentation pour un terme : le contexte, le thème et la graphie. Ces trois approches sont tout d'abord étudiées individuellement, puis combinées afin de sélectionner les meilleures traductions. Des expériences menées sur la traduction de termes médicaux du français vers l'anglais montrent une amélioration de l'approche classique par contexte, atteignant une précision de 80,4% de bonnes traductions au rang 1.*

ABSTRACT. *This paper presents a multi-view approach for term translation spotting, based on a bilingual lexicon and comparable corpora. We propose to study different levels of representation for a term: the context, the theme and the orthography. These three approaches are studied individually and combined in order to rank translation candidates. We focus our task on French-English medical terms. Experiments on our new model show a significant improvement of the classical context-based approach, with a precision score of 80.4% for the first ranked translation candidates.*

MOTS-CLÉS : *extraction terminologique, lexique bilingue, corpus comparable*

KEYWORDS: *terminology extraction, bilingual lexicon, comparable corpora*

1. Introduction

Les lexiques bilingues sont des ressources particulièrement utiles pour la traduction automatique, notamment pour accroître la couverture du vocabulaire en traduction automatique statistique (TAS), généralement basée sur des corpus bilingues parallèles (Koehn, 2005). Cependant, ces corpus faisant défaut pour des domaines de spécialités, certains auteurs ont étudié les possibilités de construire des lexiques bilingues à partir d'autres sources de données : les corpus comparables (Fung, 1995, Rapp, 1995).

Repérer des traductions à partir de corpus comparables est une tâche populaire depuis une quinzaine d'années. Une des principales approches s'appuie sur l'hypothèse qu'un terme et sa traduction partagent des similarités contextuelles. Aidé d'un lexique bilingue, il est alors possible de représenter un terme par un vecteur de contexte, où chaque composante est une valeur d'association entre ce terme et un élément du lexique. Les vecteurs de la langue source sont ensuite comparés aux vecteurs de la langue cible afin de repérer les traductions parmi les candidats. Cette méthode repose sur une hypothèse de relative invariance des vecteurs de contexte d'une langue à l'autre.

Cette approche est souvent combinée à des heuristiques basées sur la graphie des mots. En effet, dans des domaines de spécialités, beaucoup de termes sont portés d'une langue à l'autre sans subir de modifications : les translittérations (Knight *et al.,* 1998). Cette caractéristique a motivé l'utilisation de la graphie pour l'extraction de terminologies bilingues.

Cependant, cette approche *classique* ne permet pas de prendre en compte la polysémie ou la synonymie (Gaussier *et al.,* 2004). L'introduction d'information sémantique est alors nécessaire à la désambiguïsation. C'est ce que nous allons présenter dans ce papier : la combinaison du contexte, du thème et de la graphie pour l'extraction de terminologie bilingue à partir de corpus comparables. Chaque paramètre est d'abord étudié individuellement, puis ils sont combinés afin d'accroître la précision générale du système, et de ce fait, la pertinence des traductions candidates. Nous concentrons nos efforts sur la traduction de termes médicaux du français vers l'anglais.

Nos études sur les vecteurs de contexte portent sur la taille de la fenêtre permettant de limiter, en nombre de mots, l'environnement d'un terme. Nous voulons capturer l'information se trouvant dans le contexte local d'un terme, mais aussi dans un contexte plus global. Nous pensons que certains termes sont caractérisés par leur environnement proche, alors que d'autres termes le sont plus par un environnement distant (Rubino, 2009).

Puis, nous abordons l'approche par thème en émettant l'hypothèse qu'un terme et sa traduction partagent des similarités d'un point de vue thématique. Nous voulons représenter un terme par un vecteur ordonné où chaque composante est un thème. La première position est occupée par le thème le plus proche d'un terme, la dernière par celui le plus éloigné. Afin d'obtenir une liste de thèmes généraux, nous modélisons

le corpus comparable dans un espace sémantique à l'aide de l'Allocation Latente de Dirichlet (LDA) (Blei *et al.,* 2003). Cette approche est adaptée à nos besoins : une représentation sémantique par sacs de mots et des dimensions (des thèmes) indépendantes.

Finalement, pour notre étude sur les paramètres graphiques de termes, nous présentons une série d'expériences d'extraction terminologique par minimisation de la distance de Levenshtein (Levenshtein, 1966) entre des termes de la langue source et de la langue cible.

La combinaison de ces trois approches est réalisée par un vote. A notre connaissance, il n'y a pas d'études sur la combinaison de ces trois vues pour l'extraction terminologique bilingue. Après avoir présenté l'approche par vecteur de contexte en section 2, nous détaillons l'approche basée sur les thèmes en section 3, puis celle basée sur la graphie en section 4. La section 5 est consacrée aux expériences, suivie d'une présentation des résultats obtenus. Finalement, une discussion est proposée dans la dernière section.

2. Approche par contexte

Un corpus comparable est un ensemble de textes non parallèles ayant des caractéristiques communes (thématique similaire, style rédactionel comparable, etc.) et ayant été écris de manière indépendante. Dans ce type de corpus, un terme et sa traduction partagent des similarités dans le vocabulaire les entourant. S'appuyant sur cette hypothèse, de nombreux travaux décrivent différentes techniques afin d'extraire des terminologies bilingues. Une des premières études porte sur la comparaison de modèles de cooccurrences entre un terme et sa traduction dans un contexte local (voisins directs) (Fung, 1995).

L'utilisation d'un contexte plus étendu est étudiée par (Rapp, 1999). Cette méthode s'appuie sur un corpus comparable volumineux (plus de 100 millions de mots) et sur un lexique bilingue. Cette ressource peut être disponible à priori ou être générée automatiquement à partir de ressources monolingues (Koehn *et al.,* 2002).

Chaque terme étant associé individuellement aux éléments du lexique bilingue afin de composer le vecteur de contexte (dans la langue source ou cible), il est primordial de calculer une valeur d'association pertinente. De ce fait, de nombreuses mesures ont été étudiées et présentées dans la littérature, comme la probabilité conditionnelle, l'information mutuelle, le log-vraisemblance, etc. (Church *et al.,* 1990, Dunning, 1993, Evert, 2004).

Les vecteurs de contexte sont construits dans la langue source et cible, puis comparés à l'aide d'une mesure de similarité vectorielle. La plus populaire est la distance cosinus, mais de nombreux auteurs ont étudié des métriques alternatives, comme la métrique *city-block* ou encore la distance de Jaccard (Fung *et al.,* 1997, Rapp, 1999).

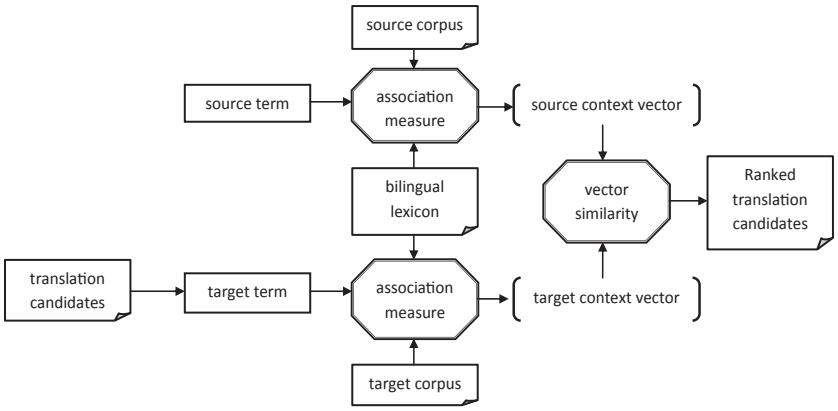

Figure 1. *Architecture de l'approche par contexte pour l'extraction terminologique bilingue.*

Plusieurs combinaisons entre les mesures d'association *terme-mot du lexique* et les métriques de similarité vectorielle ont été étudiées par (Laroche *et al.,* 2010). Ces travaux font office de référence (*baseline*) pour notre papier, car le domaine de spécialité est identique (domaine médical) et les ressources bilingues sont similaires. La configuration la plus efficace dans ces travaux est le rapport des chances (*odds-ratio*) combiné au cosinus. La formule du rapport des chances est détaillée en équation 1. Cette mesure se base sur une table de contingences regroupant les fréquences d'observations de deux termes dans une fenêtre donnée. Dans notre cas, quatre éléments sont dans la table de contingence : θ_{11} représente le nombre d'occurrences d'un terme et d'un mot du lexique, θ_{22} le nombre d'occurrences de deux éléments autres que le terme et le mot du lexique, θ_{12} les occurrences d'un terme et aucun mot du lexique, θ_{21} les occurrences d'un mot du lexique sans aucun terme.

$$odds(term_1, term_2) = log \frac{(\theta_{11} + 1/2)(\theta_{22} + 1/2)}{(\theta_{12} + 1/2)(\theta_{21} + 1/2)} \quad [1]$$

L'architecture générale de l'approche par vecteur de contexte est décrite sur la figure 1. Un des paramètres principaux dans cette approche est la taille de la fenêtre glissante permettant de compter les cooccurrences *terme-mot du lexique*. La taille de cette fenêtre, en nombre de mots, délimite le contexte d'un terme. Il peut être fixe, 10 ou 20 mots par exemple, ou encore dynamique, une phrase, un paragraphe, ... Dans ce papier, nous mettons particulièrement l'accent sur la taille du contexte. Nous utilisons des tailles de fenêtres différentes afin de capturer l'information contextuelle d'un terme de manière locale et globale. Une fenêtre de 20 mots, par exemple, signifie 20

mots en tout, sans compter le terme cherché. Ce dernier n'est pas forcément au centre de la fenêtre.

3. Approche par thème

La modélisation d'un espace sémantique à partir d'un corpus de documents est une approche courante en indexation documentaire et en recherche d'information. L'analyse sémantique latente (LSA) (Deerwester *et al.,* 1990) est une méthode populaire, basée sur l'hypothèse que des documents peuvent être représentés dans un espace thématique. Dans cet espace, chaque dimension est un concept (ou thème). Chaque thème constitue une distribution de probabilités sur les mots contenus dans les documents. Cette méthode s'appuie sur une matrice *mots-documents* contenant les occurrences des mots dans les documents du corpus. Cette matrice creuse de grande dimension est réduite par une décomposition en valeurs singulières (SVD).

Une approche probabiliste à LSA (PLSA) a été introduite par (Hofmann, 1999), basée sur le principe de vraisemblance. Utilisée par (Gaussier *et al.,* 2004) afin de gérer la polysémie et la synonymie dans la tâche d'extraction terminologique, les auteurs introduisirent le modèle PLSA bilingue.

Dans ce papier, nous avons décidé d'utiliser l'allocation latente de Dirichlet (LDA) afin de modéliser notre corpus comparable dans un espace thématique. Introduite par (Blei *et al.,* 2003), le fonctionnement général de LDA est détaillé par la figure 2. Les travaux de (Boyd-Graber *et al.,* 2009) permettent de mettre en évidence que des thématiques sont partagées par des corpus non parallèles en allemand et en anglais.

Une approche multilingue basée sur LDA est introduite par (Ni *et al.,* 2009) pour extraire des thèmes à partir de Wikipédia. Des articles disponibles dans plusieurs langues (anglais et chinois) sont alignés grâce à la structure de Wikipédia. Suivant une approche similaire dans (Mimno *et al.,* 2009), les auteurs explorent une modélisation conceptuelle multilingue (plus de 10 langues) et mettent en évidence la possibilité d'aligner des concepts dans plusieurs langues. Leur approche est testée sur un corpus parallèle en premier lieu, puis sur des documents tirés de Wikipédia et regroupés par equivalence (un ensemble de documents contient les articles traitant du même sujet rédigés dans plusieurs langues). Ce corpus peut être considéré comme comparable. Cependant, le regroupement entre articles de plusieurs langues est encore une fois effectué selon les liens inter-langues de Wikipédia. C'est justement cet aspect que nous évitons dans nos travaux : nous voulons garder une indépendance à la structure des données. Nous n'utiliserons donc pas les liens entre les articles ni les liens entre les langues.

Nous voulons modéliser le corpus comparable dans un espace thématique, permettant ainsi d'obtenir la distribution du vocabulaire sur les dimensions (thèmes) de notre espace. Pour cela, nous réduisons le vocabulaire en langue source du corpus comparable au vocabulaire contenu dans le lexique bilingue. Nous modélisons ensuite

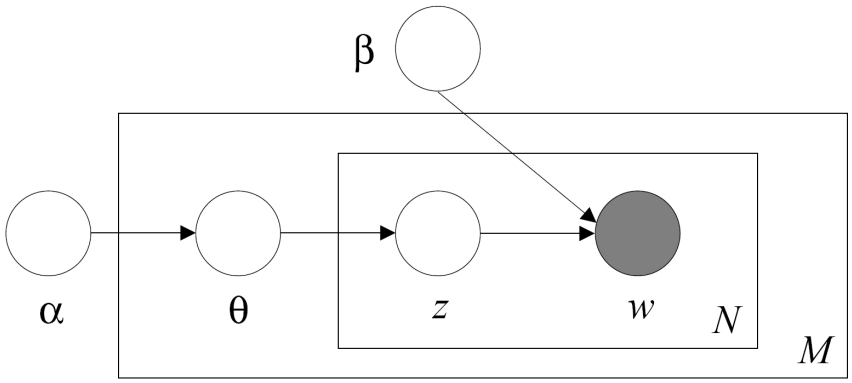

Figure 2. *Modèle graphique de LDA présenté par (Blei et al., 2003).*

ces données dans un espace thématique à l'aide de LDA (voir la figure 3). Voici le déroulement séquentiel de notre approche par thème :

– La partie du corpus dans la langue source est filtrée par le lexique bilingue : nous ne gardons que le vocabulaire que nous savons traduire.

– Ce corpus langue source filtré est utilisé afin de modéliser un espace conceptuel grâce à LDA.

– Cet ensemble de concepts est *projeté* dans la langue cible par traduction des mots de chaque concepts, en gardant les probabilités $p(w_n|z)$.

– Pour chaque terme (langue source et cible) :

- La distance détaillée par l'équation 2 est calculée pour chaque concept (langue source et cible).

– Chaque terme est donc associé à une liste de concepts, ordonnés selon la distance obtenue.

$$d(term, z) = \sum_n p(w_n|z)\ odds(term, w_n) \quad [2]$$

Le fait d'avoir réduit le corpus au vocabulaire dont nous connaissons la traduction permet de projeter notre espace sémantique vers la langue cible. Cette projection ne reflète pas la réalité des distributions de probabilités de la langue cible, cependant cette méthode nous permet d'obtenir des dimensions alignées entre nos deux modèles. Notre objectif étant de mesurer une valeur d'association entre un terme et une dimension de l'espace, la mesure du odds-ratio entre un mot de cette dimension et un terme est pondéré par la probabilité d'observation du mot dans la dimension (voir

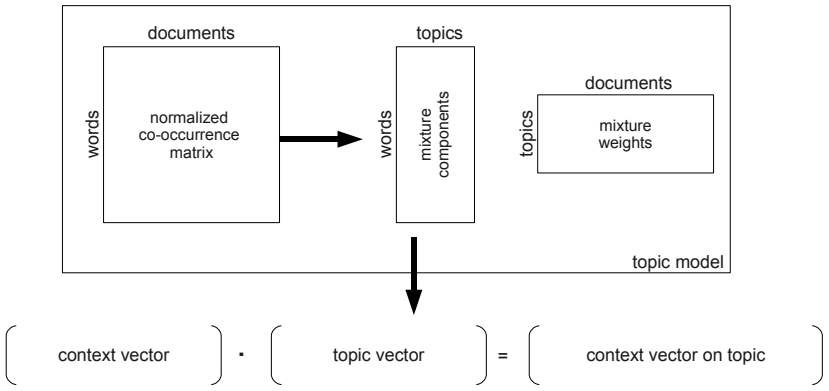

Figure 3. *Modélisation du contexte d'un terme selon différentes thématiques.*

l'équation 2). Pour chaque mot w du lexique bilingue, une probabilité de l'observer associé à un thème z est donnée par LDA. Cette probabilité est pondérée par la valeur d'association *odds-ratio* entre un terme et le mot du lexique. Ainsi, pour chaque terme de la langue source ou de la langue cible, nous estimons une valeur d'association à chaque thème en sommant ces valeurs. L'hypothèse étant qu'un terme et sa traduction partagent des similarités thématiques. Techniquement, à chaque terme est associé un vecteur ordonné contenant les thèmes du plus proche au plus éloigné. Les termes de la langue cible partageant des similarités conceptuelles avec un terme de la langue source sont plus enclins à être de bonnes traductions.

4. Approche par graphie

Comme il a été décrit dans (Koehn *et al.,* 2002), des langues liées comme l'anglais et l'allemand partagent des similarités orthographiques pouvant être utilisées pour l'extraction de traductions. Une méthode populaire pour la comparaison orthographique entre termes est la distance d'édition (ou distance de Levenshtein) (Levenshtein, 1966). Des auteurs ont effectué des expériences menées sur des langues latines, montrant qu'il est possible d'établir des règles de traduction portant sur les préfixes et les suffixes. Ainsi, comparer des termes entre leurs racines (ou leurs lemmes) permet de repérer des traductions possibles.

Un modèle introduit par (Shao *et al.,* 2004) est basé sur un système de *romanisation* pour le chinois, permettant d'effectuer un comparaison au niveau des lettres avec l'anglais. Cette technique permet de comparer des termes de langues non liés.

Dans notre étude sur l'approche par graphie, nous utilisons la distance de Levenshtein afin de comparer des termes médicaux français et anglais. Cette comparaison est effectuée entre les quatres premières lettres des termes ou entre les termes entiers.

5. Expériences

Afin de mener à bien nos expériences, nous avons besoin de trois ressources bilingues : un corpus comparable, un lexique et une liste de candidats à traduire avec une traduction référence. Notre cadre d'expérimentation comprend un corpus comparable composé de la totalité des articles français et anglais de Wikipédia. Ces données sont disponibles sur la page destinée aux téléchargement des *dumps* de Wikipédia [1]. Un *dump* est un fichier XML contenant l'ensemble des articles dans une langue. Il est constitué de texte, mais aussi de balises spécifiques à Wikipédia (images, liens, etc.) qui ne sont pas utilisées dans nos expériences. Nous formatons les fichiers au format *trectext*, après avoir ôté les mots outils. L'outil d'indexation Lemur [2] nous permet de manipuler aisément la quantité de texte provenant de Wikipédia. Les candidats à traduire sont issus du thésaurus *Medical Subject Heading* (MeSH) [3], accompagnés d'une traduction de référence (Langlais *et al.,* 2008). Le lexique bilingue est tiré du glossaire multilingue de termes médicaux de l'Institut Heymans de Pharmacologie [4]. Ces deux ressources sont identiques à celles utilisées dans les travaux de (Laroche *et al.,* 2010). Les détails sur les ressources bilingues utilisées sont présentés dans le tableau 1.

Afin de modéliser les articles Wikipédia dans un espace thématique, nous utilisons une implémentation de LDA basée sur l'échantillonnage de Gibbs [5]. Nous construisons plusieurs modèles, chacun possédant un nombre différent de dimensions (de 20 à 200). Chaque modèle est estimé avec 2 000 itérations. Pour chaque série d'expériences, 3 000 termes anglais (1 598 termes composés d'un seul mot, 1402 termes multi-mots) sont à repérer, correspondant aux 3 000 termes français (1 635 termes composés d'un seul mot, 1 365 termes multi-mots). Nous étudions d'abord chaque approche (ou vue) individuellement. Le résultat pour chacune des vues est une liste ordonnées de termes anglais, le premier terme étant la traduction la plus probable selon le système. Nous observons ainsi la position de la traduction référence qui permet de déterminer la pertinence de l'approche. Puis, les résultats de chaque vue sont combinés à l'aide d'un vote. Nous pouvons, de ce fait, observer si une bonne traduction est placée au premier rang par plusieurs juges. Nous pensons que ce type de combinaison nous permet d'accroître la précision de notre approche, tout en gardant un rappel correct grâce à la complémentarité des vues.

1. http ://download.wikimedia.org/backup-index.html
2. http ://www.lemurproject.org
3. http ://www.nlm.nih.gov/mesh/
4. http ://users.ugent.be/˜rvdstich/eugloss/welcome.html
5. http ://gibbslda.sourceforge.net

corpus	documents	mots	mots uniques
candidats	-	-	3 000
lexique	-	-	9 000
Wikipédia FR	872 111	118 019 979	3 994 040
Wikipédia EN	3 223 790	409 792 870	14 059 292

Tableau 1. *Détails des ressources bilingues utilisées dans nos expériences.*

5.1. *Contexte*

Pour chaque terme, un vecteur de contexte est construit. Nous considérons un terme multi-mots de la même manière qu'un terme composé d'un seul mot : comme une entité terminologique. Nous faisons varier la taille du contexte afin de capturer différents types d'information. Puis, chaque vecteur de la langue cible est ordonné selon la similarité cosinus avec le vecteur de la langue source. Nous comparons uniquement des vecteurs construits avec les même paramètres. Le rappel est mesuré sur les 100 premiers termes de la langue cible ordonnés par le système. Nous avons aussi étudié les candidats dont les traductions sont placées en première position en utilisant un contexte d'une certaine taille uniquement. Il nous est alors possible de voir si certains termes sont mieux caractérisés grâce à un contexte proche, ou au contraire, par un contexte plus distant. Le tableau 2 contient les résultats de ces expériences, pour chaque taille de contexte individuellement.

	10 mots	20 mots	30 mots	40 mots	document
1	31,1	32,9	**33,7**	32,4	15,6
10	57,6	59,6	60,6	58,6	37,7
50	69,3	71,8	72,6	71,8	54,6
100	73,4	76,0	76,9	76,6	61,5
unique	**4,5**	1,7	1,5	1,7	3,1

Tableau 2. *Rappel pour les 100 premiers termes ordonnés par le système, pour des tailles de contexte allant de 10 mots au document complet. La ligne unique indique le pourcentage de bonnes traductions au rang 1 retrouvées uniquement par une taille de contexte.*

Nous pouvons remarquer que les résultats de rappel sont relativement proches pour des contextes allant de 20 à 40 mots. Un contexte de 30 mots permet de repérer le plus de bonnes traductions, quelque soit le rang entre 1 et 100. Cependant, c'est un contexte de 10 mots qui permet de placer au premier rang un nombre de bonnes traductions qui ne le sont pas avec les autres tailles de contexte. Certains termes, et leurs traductions, sont donc plus *attachés* à un vocabulaire les environnant. D'autres, par contre, sont retrouvés par le système grâce à un contexte plus grand, jusqu'au document complet (un article Wikipédia). Nous envisageons alors une combinaison de l'ensemble des ré-

sultats par vote à l'unanimité, présentée dans le tableau 3. Si la condition d'unanimité est remplie, le système retourne une traduction candidate, sinon aucun résultat n'est retourné.

	30 mots	**20+30 mots**	**10+30+40 mots**	**toutes**
rappel	**33,7**	27,8	19,2	6,2
précision	33,7	50,5	75,9	**83,7**
f-mesure	33,7	**35,8**	30,6	11,5

Tableau 3. *Combinaison des résultats au rang 1 obtenus par variation de taille de contexte. Un contexte de 30 mots est comparé aux combinaisons 20 et 30 mots, de 10, 30 et 40 mots, puis toutes les tailles (de 10 mots à l'ensemble du document).*

Pour ces expériences, nous mesurons le rappel, la précision, puis calculons la f-mesure, détaillée par l'équation 3. Nous observons une nette dégradation du rappel lors de la combinaison des tailles de contexte. Seulement 6,2% de bonnes traductions sont placées au premier rang par l'ensemble des contextes. Le score de précision de 83,7% est atteint par cette même configuration, et dépasse tous les autres résultats mesurés. Nous obtenons une f-mesure de 35,8% en combinant les résultats des contextes de 20 et 30 mots, en atteignant une précision de 50,5%.

$$\text{f-mesure} = \frac{2 \times (\text{précision} \times \text{rappel})}{\text{précision} + \text{rappel}} \quad [3]$$

5.2. *Thème*

L'idée principale de notre approche par thème est de voir s'il est possible de filtrer les 3 000 traductions candidates. Nous sommes conscients que cette méthode ne peut permettre, dans notre configuration, d'isoler un seul candidat, car le corpus utilisé afin de modéliser les thèmes est général (Wikipédia), et les candidats sont plus ou moins des termes techniques du domaine médical (*agranulocytose* et *agranulocytosis* ; *analyse des aliments* et *food analysis* ; etc.).

Nous estimons que la granularité de nos modèles (entre 20 et 200 thèmes) peut permettre de séparer des candidats trop éloignés des thèmes du terme à traduire. Cette information thématique peut nous permettre de valider une traduction candidate provenant d'une autre approche (basée sur les contextes par exemple). Pour chaque candidat, de langue source ou cible, nous calculons un vecteur ordonné contenant les valeurs d'association aux thèmes d'un modèle. Nous voulons ainsi observer si un terme en français et sa traduction en anglais se situent à la même position dans un espace sémantique.

Le thème le plus pertinent pour un terme en français est sélectionné. S'il se trouve parmi les trois premiers thèmes d'un candidat en anglais, ce candidat est retenu. Si

		20	50	100	200
1	**rappel**	23,6	33,4	33,0	10,4
	précision	0,12	0,06	0,07	0,18
	f-mesure	0,23	0,12	0,14	0,35
2	**rappel**	35,9	42,2	38,3	18,8
	précision	0,1	0,06	0,07	0,15
	f-mesure	0,19	0,12	0,14	0,31
3	**rappel**	44,1	45,9	41,2	24,1
	précision	0,08	0,06	0,07	0,14
	f-mesure	0,16	0,12	0,13	0,28

Tableau 4. *Résultats de l'approche par thème. Les trois premiers thèmes de la langue cible sont étudiés. Nous testons des modèles de 20, 50, 100 et 200 dimensions.*

ce candidat est la bonne traduction, le rappel est incrémenté. Nous présentons les résultats de ces expériences dans le tableau 4. Les modèles étant différents par leur nombre de dimensions, nous les avons testés individuellement. Nous mesurons aussi la précision de cette approche. Si un terme anglais n'est pas la bonne traduction (selon notre référence) mais partage des similarités thématiques avec un terme en français, la précision décroît. Ainsi, une précision de 0 indique que 3 000 candidats ont été retournés par le système, pour un terme à traduire.

Les résultats concernant la précision de notre approche par thèmes ne sont pas surprenants, quelque soit le modèle utilisé. En effet, de nombreux candidats sont placés au même endroit dans l'espace sémantique que la bonne traduction. De ce fait, pour un terme à traduire (langue source), plusieurs termes de la langue cible peuvent être rapportés par le système. C'est pour cela que l'approche par thème ne peut permettre seule de déterminer une bonne traduction, mais elle peut permettre de filtrer l'ensemble des traductions candidates, car un tiers de bonnes traductions partagent un thème similaire aux termes de la langue source, dans un espace de 50 dimensions. Ce rappel nous permet de penser à une combinaison entre l'approche par vecteur contexte et l'approche par thèmes.

5.3. *Graphie*

La distance de Levenshtein est utilisée afin d'ordonner les traductions candidates selon leur proximité orthographique. Soit sur les quatre premières lettre des termes sources et cibles, soit sur les termes entiers. Les résultats sont présentés dans le tableau 5. Nous étudions plusieurs rangs (du rang 1 au rang 10). Pour chaque rang pris en compte, si la bonne traduction est observée, le rappel augmente. Nous calculons un score de précision d'après le nombre de termes au rang donné n'étant pas la bonne traduction.

		1	2	3	4	5	10
4 lettres	**rappel**	34,0	39,0	45,9	65,4	100	100
	précision	15,9	3,9	0,5	0,1	0,03	0,03
	f-mesure	21,7	7,2	1,0	0,2	0,07	0,07
termes	**rappel**	50,7	54,8	59,6	67,4	77,3	99,3
	précision	83,5	29,6	5,6	1,4	0,5	0,2
	f-mesure	63,1	38,5	10,3	2,8	1,0	0,3

Tableau 5. *Résultats de l'approche par graphie, selon les rangs de 1 à 10. Deux méthodes sont testées : entre les 4 premières lettres des termes et entre les termes entiers.*

Nous pouvons voir qu'avec une distance de Levenshtein sur les termes entiers, 50,7% des bonnes traductions sont placées au premier rang, avec une précision de 83,5%. Ces résultats sont explicables par la proximité des langues française et anglaise, et par le domaine de spécialité. Le rappel atteint 99,3% au rang 10, cependant la précision est médiocre. Cette précision très faible s'explique par la mesure utilisée. En effet, la distance de Levenshtein place plusieurs candidats au même rang (par exemple, tous les termes ayant une distance de 1 sont placés au même rang). Nous pensons cependant qu'il est intéressant d'utiliser cette approche dans notre modèle multi-vue, la complémentarité des approches pouvant permettre de différencier les candidats à la traduction.

5.4. *Combinaison*

Nous présentons les résultats de la combinaison des différentes vues dans le tableau 6. Trois combinaisons sont étudiées : deux plus *classiques* entre le contexte et les thèmes ou le contexte et la graphie, et une nouvelle étant les trois approches ensembles.

L'approche par contexte combinée avec l'orthographe représente notre référence, car cette combinaison est la plus courante. Chaque approche (ou juge) participe au système de vote, et nous étudions si la majorité absolue ou l'unanimité des *juges* ont placé les bonnes traductions au premier rang. Si aucune de ces conditions n'est remplie (le vote n'a atteint ni la majorité absolue, ni l'unanimité), notre système est incapable de retourner une traduction au terme en français. Nous proposons deux manières d'inclure l'approche par contexte dans la combinaison : comme un seul juge (soit une seule taille de contexte) ou comme cinq juges (un par taille de contexte). Pour la taille de contexte fixe, nous avons choisi le nombre de mots permettant d'obtenir la meilleure f-mesure au rang 1.

L'approche par contexte combinée avec la distance de Levenshtein permet d'atteindre un rappel de 19% avec une précision de 99,1%. En utilisant toutes les tailles de contexte, le score de rappel est légèrement accru (21,1%) mais la précision est

		cont+thème	cont+graph	cont+thème+graph
30 mots	**rappel**	13,9	19,0	24,2 (7,6)
	précision	100	99,1	99,3 (100)
	f-mesure	24,4	31,9	39,0 (14,1)
toutes	**rappel**	20,8 (2,6)	21,1 (4,0)	26,9 (1,7)
	précision	76,2 (100)	76,6 (97,6)	80,4 (100)
	f-mesure	32,7 (5,0)	33,1 (7,7)	40,3 (3,4)

Tableau 6. *Résultats au rang 1 pour la combinaison des approches. L'unanimité des juges est indiquée entre parenthèses, si les juges sont plus de deux. L'approche par contexte est notée cont, l'approche par graphie est notée graph.*

diminuée, selon la majorité des juges. Si l'on regarde le rappel pour l'unanimité des juges (4%), nous pouvons remarquer que le recouvrement des termes anglais placés en première place pour chaque juge est très faible. Cependant, notre système ne propose jamais de mauvaises traductions (quand la précision est à 100%). Ces résultats mettent en évidence un aspect remarquable de notre approche multi-vue : la capacité du système à proposer une traduction dont il est certain, et à s'abstenir lorsque les vues divergent.

De plus, la combinaison des toutes les vues permet d'obtenir la meilleure f-mesure. C'est la configuration la mieux adaptée à notre tâche, ce qui montre la complémentarité des trois approches combinées.

6. Discussion

Nous présentons dans cet article une approche multi-vue pour l'extraction bilingue de terme médicaux, basée sur un corpus comparable et un lexique bilingue. Une approche basée sur des vecteurs de contexte est combinée à un modèle thématique afin de filtrer les candidats trop éloignés d'un point de vue sémantique. Nous incluons également des paramètres orthographique afin d'améliorer les résultats.

La combinaison des trois approches permet d'atteindre un rappel de 26,9% et une précision de 80,4% pour les candidats placés au premier rang. Ces résultats montrent la complémentarité des trois vues. Les travaux de (Laroche *et al.,* 2010) combinant le contexte et l'orthographe des termes sur les mêmes candidats (3 000) avec un lexique bilingue identique (9 000 entrées) permettent d'atteindre une précision de 55,2%. Nous améliorons ainsi de 25,2% la précision avec notre combinaison de vues.

L'approche par vote n'étant pas forcément la plus pertinente, il serait judicieux d'utiliser une autre manière de combiner les trois vues. Différentes combinaisons possibles sont à l'étude.

Nous envisageons d'améliorer le modèle thématique en modélisant le corpus dans la langue cible, afin d'évaluer les probabilités de distribution des thèmes sur le vocabulaire, et non plus de *projeter* le modèle dans la langue source. Une autre amélioration possible réside dans la sélection parmi les éléments du corpus de sous-parties relatives au domaine de spécialité. Ainsi, le modèle obtenu serait plus fin et les thèmes plus spécifiques.

Le lexique bilingue est utilisé dans sa totalité pour nos expériences, or l'avantage de l'approche par contexte réside dans l'utilisation de points d'ancrages non ambigus. Une étude possible peut concerner l'utilisation de mots les plus discriminant pour un terme à traduire, ce qui réduirait par la même occasion la taille des vecteurs de contexte.

Remerciements

Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (ANR), par l'intermédiaire du projet AVISON (ANR-007-014).

7. Bibliographie

Blei D., Ng A., Jordan M., « Latent Dirichlet Allocation », *The Journal of Machine Learning Research*, vol. 3, p. 993-1022, 2003.

Boyd-Graber J., Blei D. M., « Multilingual topic models for unaligned text », *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, p. 75-82, 2009.

Church K., Hanks P., « Word Association Norms, Mutual Information, and Lexicography », *Computational linguistics*, vol. 16, n° 1, p. 22-29, 1990.

Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, vol. 41, n° 6, p. 391-407, 1990.

Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Evert S., « The Statistics of Word Cooccurrences : Word Pairs and Collocations », *Ph.D. Thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart*, 2004.

Fung P., « Compiling Bilingual Lexicon Entries from a Non-parallel English-Chinese Corpus », *Proceedings of the 3rd Workshop on Very Large Corpora*, p. 173-183, 1995.

Fung P., McKeown K., « Finding Terminology Translations from Non-parallel Corpora », *Proceedings of the 5th Annual Workshop on Very Large Corpora*, p. 192-202, 1997.

Gaussier E., Renders J., Matveeva I., Goutte C., Dejean H., « A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora », *Proceedings of the 42nd ACL conference*, ACL, p. 526, 2004.

Hofmann T., « Probabilistic Latent Semantic Indexing », *Proceedings of the 22nd ACM SIGIR conference*, ACM, p. 50-57, 1999.

Knight K., Graehl J., « Machine Transliteration », *Computational Linguistics*, vol. 24, n° 4, p. 612, 1998.

Koehn P., « Europarl : A Parallel Corpus for Statistical Machine Translation », *MT summit*, vol. 5, Citeseer, 2005.

Koehn P., Knight K., « Learning a Translation Lexicon from Monolingual Corpora », *Proceedings of the ACL workshop on Unsupervised lexical acquisition*, vol. 9, ACL, p. 9-16, 2002.

Langlais P., Yvon F., Zweigenbaum P., « Translating Medical Words by Analogy », *Intelligent Data Analysis in Biomedicine and Pharmacology*, Washington, DC, USA, p. 51-56, 2008.

Laroche A., Langlais P., « Revisiting Context-based Projection Methods for Term-translation Spotting in Comparable Corpora », *Proceedings of the 23rd Coling conference*, Beijing, China, p. 617-625, August, 2010.

Levenshtein V., « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady*, vol. 10, p. 707-710, 1966.

Mimno D., Wallach H., Naradowsky J., Smith D., McCallum A., « Polylingual topic models », *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2-Volume 2*, Association for Computational Linguistics, p. 880-889, 2009.

Ni X., Sun J., Hu J., Chen Z., « Mining Multilingual Topics from Wikipedia », *Proceedings of the 18th international conference on World Wide Web*, ACM, p. 1155-1156, 2009.

Rapp R., « Identifying Word Translations in Non-parallel Texts », *Proceedings of the 33rd ACL Conference*, ACL, p. 320-322, 1995.

Rapp R., « Automatic Identification of Word Translations from Unrelated English and German Corpora », *Proceedings of the 37th ACL conference*, ACL, p. 519-526, 1999.

Rubino R., « Exploring Context Variation and Lexicon Coverage in Projection-based Approach for Term Translation », *Proceedings of the RANLP Student Research Workshop*, ACL, Borovets, Bulgaria, p. 66-70, September, 2009.

Shao L., Ng H., « Mining New Word Translations from Comparable Corpora », *Proceedings of the 20th ACL*, ACL, p. 618, 2004.