Vers l'exploitation d'analyse de dépendance en recherche d'information précise

Loic Maisonnasse

Vers l'exploitation d'analyse de dépendance en recherche d'information précise.

Loïc Maisonnasse

Laboratoire CLIPS-IMAG
BP 53
38 041 Grenoble cedex 9

RÉSUMÉ. *Nous présentons une étude sur l'utilisation des dépendances syntaxiques pour la tâche de Recherche d'Information (RI). Ces dépendances sont extraites automatiquement des textes par des processus du Traitement Automatique des Langues Naturelles (TALN). Nous cherchons à déterminer si leur utilisation peut être bénéfique en recherche d'information. Pour cela, nous étudions les différentes utilisations des dépendances dans les systèmes de RI. Après avoir décrit ces systèmes, nous présentons et expérimentons une indexation par dépendances syntaxiques. Nous étudions la répartition des dépendances sur le corpus, sélectionnons les dépendances intéressantes pour la RI et testons leurs utilisation en RI sur différentes langues. Les différentes expérimentations montrent l'utilisabilité des dépendances mais aussi la difficulté de leur intégration dans une tâche de RI.*

ABSTRACT. *This paper presents a study concerning the way of using the syntactic dependencies that are extracted from text with natural language processing tools for an information retrieval task,. We investigate the different uses of syntactic dependencies on information retrieval and how they are used to lead to a more precise information retrieval system. Then a dependency-based information retrieval method is presented. We inspect the dependencies repartition on a corpus, select dependencies that are useful for information retrieval, and then use dependencies in information retrieval task on different languages. The different experimentations show the usability of dependencies and the possibility of improving information retrieval result with them.*

MOTS-CLÉS : *Recherche d'information, traitement de la langue, dépendance syntaxique*

KEYWORDS: *Information retrieval, natural language processing, syntactic dependencies*

1. Introduction

La quantité de documents disponibles pour les systèmes de recherche d'information étant de plus en plus grande, il est nécessaire de s'orienter vers des systèmes plus en plus précis[1]. Les systèmes de RI actuels se basent sur le principe de la correspondance de mots : un document est pertinent s'il contient un ou plusieurs mots de la requête. Cependant, le mot ne représente qu'une partie de la sémantique de la phrase (Jaquemin, 1997), pour progresser vers des systèmes plus précis, l'extraction de connaissances à partir du texte semble nécessaire.

Des outils de TALN sont ainsi utilisés pour enrichir la représentation des documents. Des recherches utilisant des représentations plus sémantiques des documents ont été menées. C'est notamment le cas des logiques terminologiques (Meghini *et al.* 1998), des dépendances de Shank (Berrut *et al.* 1998) ou des graphes conceptuels (Chevallet, 1992). De nouvelles fonctions de similarités spécifiques à ces représentations ont été introduites et ces modèles ont fourni des résultats encourageants notamment pour leur adaptabilité à différents médias. L'utilisation de représentations sémantiques de niveau élevé a cependant introduit un certain nombre de problèmes. Ces représentations ne peuvent pas être extraites directement du texte et requièrent une indexation manuelle. Les fonctions de similarité proposées pour ces représentations sont souvent complexes et ne peuvent être utilisées directement sur de grands corpus.

Notre objectif pour aller vers plus de précision est d'utiliser une représentation intermédiaire du texte se situant à un niveau syntaxico-sémantique. On s'éloigne donc du "signal" de la suite des mots d'un texte, tout en restant à un niveau de complexité acceptable pour réaliser le processus de RI sur de larges corpus. Nous exploitons pour cela les données extraites des textes en langage naturel à l'aide d'analyseurs syntaxiques. De telles données ont déjà été exploitées en RI, (Strzalkowski *et al.*, 1996), (Metzler *et al.*, 1989). Nous nous basons ici sur une analyse en dépendance, il semble en effet que cette structure soit plus à même de représenter le thème qu'une analyse en constituant (Koster, 2004). Une dépendance représente le lien syntaxico-sémantique entre deux mots : le dépendant et son recteur (i.e. le mot dont il dépend). Certains mots d'une phrase sont liés à un recteur et cette dépendance est étiquetée par la relation syntaxique qu'entretiennent ces deux mots. Par exemple, dans la phrase *'le chat mange la souris'*, le nom *'chat'* dépend de son recteur *'manger'* par une relation de sujet. Dans une phrase, ces dépendances forment une structure d'arbre : l'arbre de dépendance.

Du résultat de l'analyse de dépendance nous extrayons des termes d'indexations semi-complexes qui consistent en des n-uplets composés par des mots et par le type

1. La précision d'un système de recherche d'information correspond à la proportion de document pertinents parmi tous les documents retrouvés par le système

de relation qui relie ces mots. Des dépendances syntaxiques sous forme de liens dépendant recteur, sont ainsi utilisées. Par ce formalisme, la phrase précédente est représentée par l'ensemble de relations suivantes : *'{SUBJ(chat,manger), OBJ(souris,manger)}'* où la première dépendance signifie que le nom *'chat'* est le sujet du verbe *'manger'* et la deuxième que le verbe *'manger'* a comme objet *'souris'*.

Dans cet article, nous présentons dans une première partie différentes utilisations des analyses en dépendances en RI. Nous exposons ensuite le processus mis en place pour tester l'intérêt des dépendances en RI et dans la partie suivante nous présentons les résultats obtenus à l'aide de ce processus sur différentes langues, nous concluons et discutons enfin des travaux futurs.

2. Les dépendances en RI

La plupart des modèles de RI se basent sur l'hypothèse que les termes sont indépendants. Cette hypothèse est évidemment fausse du fait que certains mots co-occurrent et qu'il existe des liens syntaxiques entre les mots d'une même phrase. Des recherches ont été menées dans le but de résoudre ce problème pour améliorer la précision des systèmes de RI. (Losee, 1994) utilise l'*'expected mutual information'* pour trouver les dépendances entre les termes des documents et étend le modèle probabiliste pour prendre en compte ces dépendances. Ses résultats montrent que plus le nombre de dépendances utilisées pour déterminer la probabilité d'un terme est grand, plus la performance du système est améliorée.

Des recherches similaires ont été effectués par Lee (Lee *et al.*, 2002). L'auteur utilise les dépendances syntaxiques extraites par son propre analyseur pour étendre le modèle probabiliste. Les résultats montrent une augmentation d'environ 5% de la précision moyenne.

D'autres travaux de RI s'appuient sur l'analyse de dépendances pour prendre en compte les dépendances dans la représentation des documents. Ces travaux s'articulent majoritairement autour de deux axes : soit les dépendances sont utilisées pour extraire des syntagmes, soit la structure de dépendance est utilisée comme index et une fonction de correspondance adaptée à cette structure est employée.

2.1. *Extraction de syntagmes*

Dans (Strzalkowski *et al.*, 1994), les auteurs extraient une représentation proche d'un arbre de dépendance à l'aide d'un analyseur basé sur des grammaires. Les auteurs sélectionnent ensuite un certain nombre de paires candidates à la formation de termes composés à l'aide de patron sur les dépendances. Ces termes sont ensuite ajoutés dans l'index des documents. Un schéma de pondération sous forme de tf-idf, pondérant les termes par rapport à leur fréquence à l'intérieur du document (tf) et

par rapport à leur fréquence documentaire inverse (idf)[2], est adapté pour donner plus d'importance à l'idf des termes composés. Les auteurs notent une augmentation de l'ordre de 20% de la précision moyenne avec l'utilisation des mots composés. Il n'est cependant pas possible de savoir si ce résultat est directement relié à l'utilisation des dépendances.

2.2. *Correspondance de structure*

Partant de l'hypothèse que la conversion des structures de dépendance en syntagmes entraîne une perte d'information, des recherches ont portées sur l'utilisation directe de ces structures.

Dans (Matsumura *et al.*, 2000) l'auteur extrait des arbres de dépendance à partir de phrases en japonais. Dans sa structure, les nœuds terminaux représentent des 'mots concepts' comme les noms, les adjectifs ou les adverbes, et les autres nœuds représentent des 'mots relations' comme les verbes ou les particules post-positionnelles. Les arbres sont extraits par une analyse de dépendance sur les titres des documents d'un corpus. La fonction de correspondance entre la requête et les documents est basée sur des découpages de l'arbre et la fonction de pondération se base sur le poids des nœuds calculé par leur fréquence documentaire inverse.

Pour leur part (Metzler *et al.*, 1989) extraient des arbres de dépendances binaires sur des phrases en anglais. Ces arbres sont extraits des documents à l'aide de l'analyseur COP (Constituent Object Parser). Dans ce système, l'utilisateur doit déterminer les termes pertinent pour sa requête et indiquer les dépendances entre ces termes. Le système évalue alors les documents pertinents pour la requête en effectuant plusieurs types de correspondances entre les dépendances de la requête et celles contenues dans les arbres des documents.

Les deux analyses précédentes proposent une unique structure de la phrase. Les ambiguïtés syntaxiques ne sont donc pas prises en compte. (Smeaton, 1999) propose un modèle appliqué sur des syntagmes dans lequel les ambiguïtés les plus courantes sont représentées dans l'arbre de dépendance. La fonction de similarité s'effectue à l'aide d'une correspondance sur les arbres. Les résultats obtenus par cette correspondance sont cependant inférieurs à ceux obtenus en appliquant une pondération sur les syntagmes représentés par les arbres.

3. Processus d'évaluation des dépendances

Notre objectif est de déterminer comment utiliser, dans un système de RI, des dépendances extraites par des analyseurs syntaxiques existants et non spécifiques à

2. idf = log(N/n) où N est le nombre de documents dans le corpus, et n ceux qui contient le terme

la tâche de RI. Nous utilisons pour cela une structure moins complexe que l'arbre de dépendance mais où l'impact de la dépendance est quantifiable et où la fonction syntaxique liant les mots est présente.

Nous voulons prouver que l'utilisation des dépendances permet d'augmenter la performance en RI, notamment dans des cas de recherches orientées vers la précision des réponses, donc le cas de requêtes précises. Par exemple, si l'on interroge un système où les dépendances ne sont pas prises en compte avec la requête *'les mouvements pour la guerre en Irak'*, le mot *'pour'* pouvant être trouvé dans d'autres contextes au sein des documents ou bien supprimé par un anti-dictionnaire, des documents sur *'les mouvements contre la guerre en Irak'* peuvent être sélectionnés. L'utilisation des dépendances permet donc d'éviter cette imprécision par la prise en compte explicite au niveau de l'indexation des dépendances liant les mots '*mouvement*', *'pour'* et *'guerre'*.

Nous proposons donc une série d'expérimentations pour mettre en évidence ce phénomène et montrer expérimentalement l'intérêt de la prise en compte de ces relations de dépendances étiquetées lors de l'indexation.

3.1. *Enchaînement des expérimentations*

Nous extrayons les dépendances syntaxiques d'un texte dans le but de vérifier leur utilisabilité pour une tâche de RI et d'évaluer leur impact en RI par rapport à une indexation ordinaire à base de mots isolés.

Nous utilisons pour cela les dépendances comme termes d'indexation dans un modèle vectoriel ainsi que le montre la Figure 1. Dans cette expérimentation nous utilisons un analyseur morpho-syntaxique pour extraire deux types de termes d'indexations (descripteurs) : les lemmes et les dépendances. Nous filtrons ensuite ces descripteurs par rapport à des informations syntaxiques, notamment sur la fonction syntaxique des dépendances, car ces fonctions n'ont pas toutes le même potentiel pour représenter le contenu thématique. Après leur avoir affecté une pondération, nous stockons ces descripteurs dans deux vecteurs distincts, nous obtenons donc deux index par documents. L'interrogation des index s'effectue de manière symétrique, en constituant deux vecteurs requêtes à base des mêmes descripteurs et en appliquant une fonction de correspondance entre ces deux vecteurs et ceux contenus dans les deux index. Chacune des fonctions de correspondance retourne une liste de documents en réponse, classés par ordre de pertinence. Nous regroupons dans une dernière étape ces deux listes pour obtenir un unique classement de pertinence.

Ne connaissant pas le comportement des dépendances lorsqu'elles sont utilisées en RI, il nous faut savoir si ces descripteurs suivent les mêmes lois que les descripteurs habituellement employés en RI. Il restera ensuite à déterminer les types de dépendances réellement utiles à la tâche de RI (Khoo, 1997). Il faudra également

étudier les apports et les résultats d'une indexation à base de dépendance par rapport à une indexation à base de lemme et leur efficacité en fonction de la langue des documents. Finalement, le rapport entre les deux indexations, en particulier un modèle d'indexation explicitant leur complément mutuel, est également à étudier.

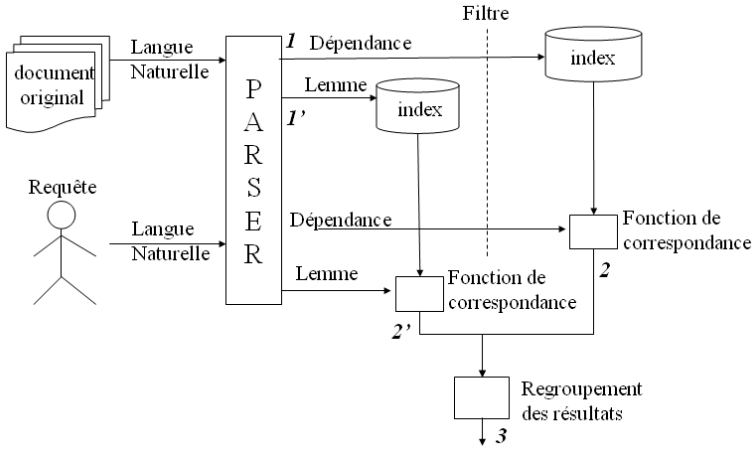

Figure 1. *Description de l'expérimentation*

Pour permettre de répondre à ces interrogations, nous mettons en place un certain nombre d'expérimentations. Pour tester le comportement des dépendances, nous étudions la répartition des lemmes et des dépendances extraites par l'analyseur syntaxique comme en 1 et 1' de la Figure 1. Nous analysons ensuite l'intérêt des différents types de dépendance, en variant les paramètres du filtre appliqué sur les dépendances extraites et en testant les résultats obtenus par la fonction de correspondance en 2 sur la Figure 1. Nous effectuons par la suite des indexations directement sur les deux descripteurs, en 2 et 2' sur la Figure 1, dans plusieurs langues et en variant le type de pondération, ceci dans le but de comparer les résultats obtenus à l'aide des lemmes et ceux obtenus à l'aide des dépendances. Nous testons, pour finir, les résultats obtenus à l'aide du regroupement des deux descripteurs, en 3 de la Figure 1, dans l'optique de tester la complémentarité de ces deux descripteurs au sein d'une tâche de RI. L'ensemble de ces expérimentations est décrit dans la partie 4 de cet article.

3.2. *Description du corpus*

Pour évaluer les dépendances, nous utilisons les corpus des campagnes CLEF 2002 et 2003 (Peters, 2003). Ces corpus sont constitués de collections de journaux

de différentes langues. Nous utiliserons seulement trois de ces langues : le français, le finnois et le russe. Le détail de ces collections est décrit dans le Tableau 1. La campagne CLEF 2002 contient un ensemble de 50 requêtes formées d'une phrase courte. La campagne 2003, quant à elle, contient 60 requêtes, chacune de ces requêtes est constituée de trois parties : un titre, une phrase courte et une description succincte du besoin d'information, un exemple de requête est donné dans le Tableau 2. Dans cet article nous utilisons seulement le titre et la phrase de chaque requête. Pour une utilisation multilingue, l'ensemble des ces requêtes est traduit manuellement dans les différentes langues.

Langue	Journal	Année	Nb. de document
français	Le monde	1994	44013
		1995	NA
	ATS	1994	43178
		1995	42615
finnois	Aamulehti	1994-1995	55344
russe	Izvestia	1995	16761

Tableau 1. *Description des collections*

Numéro de la requête	C146
Titre	Les Fast-Foods au Japon
Description	Quelles chaînes de fast-food nord-américaines ont un grand nombre de restaurants au Japon ?
Narration	Les documents pertinents doivent mentionner le nom des chaînes de fast-food américaines qui rencontrent le plus de succès au Japon, et peuvent contenir des informations supplémentaires concernant l'introduction de ce type d'alimentation dans la société japonaise.

Tableau 2. *Exemple de requête de CLEF 2003*

3.3. *Description des analyseurs syntaxiques*

A partir de ces corpus, nous extrayons automatiquement les lemmes et les dépendances. Pour cela nous utilisons différents analyseurs syntaxiques. L'analyseur 'Xerox Incremental Parser' (XIP) (Aït-Mokhtar et al., 2002) est employé pour le français, l'analyseur syntaxique Conexor (Tapanainen, 1999) pour le finnois, enfin l'analyseur ETAP (Boguslavsky *et al.*, 2003) pour le russe. Ces trois analyseurs extraient des informations syntaxiques à partir des textes. De ces informations, nous extrayons les lemmes et les dépendances. Les dépendances conservées sont constituées de deux ou trois lemmes et d'un type de dépendance qui représente le lien syntaxique entre les lemmes.

4. Résultats

Dans cette partie, nous présentons les résultats de ces expérimentations. Elles ont été réalisées avec le système XIOTA du laboratoire CLIPS de l'institut IMAG de Grenoble (Chevallet, 2004).

4.1. *Utilisabilité des dépendances*

Pour savoir si les dépendances peuvent être de bons descripteurs pour la RI nous vérifions en premier la loi de Zipf. Cette loi est l'une des bases de la RI par mots clefs. Elle a été empiriquement énoncée par G.K. Zipf (Zipf, 1949). Elle suppose que les symboles d'un ensemble organisé typologiquement s'organisent selon une loi de puissance. De ce fait, si on caractérise les mots par leur rang de fréquence, le mot le plus courant ayant le rang un, le deuxième mot le plus fréquent le rang deux etc., alors la fréquence du mot de rang 'i' est calculable par l'expression suivante :

$N_\sigma(i) = k \times i^{-\alpha}$, où k et α sont des constantes positives.

Du fait de cette loi de puissance, si on trace la distribution des fréquences par leur rang sur une échelle bi-logarithmique, la distribution a l'aspect d'une droite. Sur du texte naturel cette droite a un coefficient (α dans l'équation) proche de 1.

En RI, la loi de Zipf est considérée comme la base de la conjecture de Luhn. Cette conjecture suppose que les descripteurs qui sont intéressants pour la RI sont ceux de fréquences intermédiaires. Si un descripteur est trop fréquent alors il n'a pas assez de force de discrimination et s'il est très rare alors il est trop sélectif.

Dans la suite nous allons donc tester si l'ensemble des dépendances extraites à partir de documents français suit la loi de Zipf. Pour cela nous utilisons une partie du corpus constituée de la collection française 'Le monde' de 1994, sur laquelle nous extrayons un certain nombre de descripteurs :

- les mots qui constituent les documents sans appliquer d'autres traitements;
- les lemmes extraits par l'analyseur XIP;
- les dépendances extraites par l'analyseur XIP.

	lemme	mot	dépendance
descripteur différent	185848	215751	5839732
Nombre d'occurrences des descripteurs	15299359	21859214	26264464

Tableau 3. ***Données sur les descripteurs utilisés***

Comme on peut le voir sur le Tableau 3, si le nombre de dépendances extraites de la collection est proche du double du nombre de lemmes, le nombre de

dépendances différentes est environ 30 fois supérieur au nombre de lemmes différents. Même si l'importance de ce nombre n'est pas négligeable, notamment pour la taille des index, il reste acceptable par rapport au nombre de dépendances théoriquement possibles.

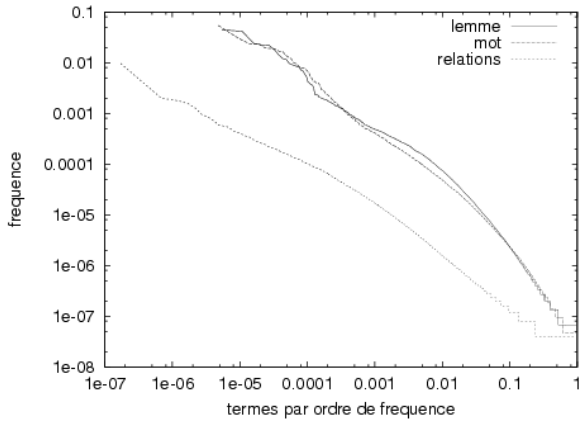

Figure 2. *Répartitions de la fréquence des termes par leur rang*

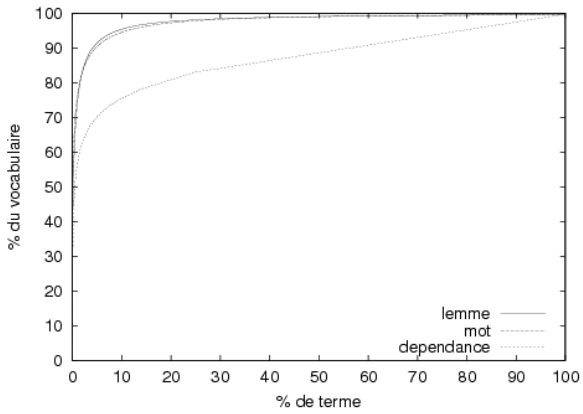

Figure 3. *Répartition du vocabulaire sur la collection*

Nous étudions ensuite la répartition des termes sur la collection. La Figure 2 montre que la répartition des dépendances semble bien suivre une loi de Zipf mais que le coefficient directeur de cette répartition est plus faible que pour celui des lemmes ; en effet, le calcul du coefficient directeur de la droite de régression pour les dépendances donne un résultat proche de 0.7 alors que pour les lemmes le

coefficient est proche de 1.5. On en conclut que la répartition des dépendances est plus uniforme que celle des lemmes. Ce phénomène est confirmé sur la Figure 3 où 10 % des lemmes représentent plus de 90% du vocabulaire de la collection alors que 10% des dépendances n'en représentent que 75%.

Il est aussi important de souligner sur la Figure 2 l'importance du nombre de dépendances n'apparaissant qu'une seule fois dans le corpus. Un processus de RI utilisant les dépendances va ainsi être fortement sélectif.

A condition de prendre en compte certaines des spécificités des dépendances lors de l'indexation (tel que le nombre de relations n'apparaissant qu'une fois), il semble donc possible d'utiliser ces dépendances pour une tâche de RI.

4.2. *Choix des dépendances*

En nous basant sur les résultats précédents, nous expérimentons maintenant l'utilisation des dépendances dans un model vectoriel.

Chaque analyseur syntaxique extrait différents types de dépendances, tous ces types n'ont certainement pas le même intérêt pour la RI. Dans cette partie nous essayons de déterminer par un test à l'aveugle quels sont les types de dépendances qui peuvent améliorer la RI. Ce test consiste à effectuer différentes indexations en variant les types de dépendances utilisés.

Nous effectuons ici le test pour les dépendances extraites par XIP sur des textes en français. Nous utilisons ici la collection du 'monde 1994' et les requêtes de CLEF 2002, les résultats obtenus sont résumés sur le Tableau 4. On peut noter que l'utilisation de certains types de dépendance a tendance à diminuer la précision des résultats. Ce sont principalement les dépendances sémantiquement pauvres telles que celles de type DETERM qui lient un déterminant avec le nom qu'il détermine et celles de type PREPOBJ qui lient une préposition avec le nom dont elle dépend. De plus, dans le cas de PREPOBJ cette dépendance est en partie redondante avec le type NMOD qui peut lier un nom, son modificateur et la préposition qui l'introduit.

D'autres types de dépendances ont peu d'impact sur le résultat de la RI, ce sont les dépendances qui représentent plus la structure de la phrase que son contenu informationnel telle que PRECOMMA, CONNECT, et NEGAT. Les dépendances qui donnent les meilleurs résultats pour la RI sont alors celles tel que NMOD, NARG, NN car elles se composent des éléments informatifs de la langue (noms, verbes …).

Type de dépendances utilisées	précision moyenne
Toutes les dépendances	0,1522
CLOSEDNP	0,1351
CLOSEDNP NMOD	0,173
CLOSEDNP NMOD PREPOBJ	0,1567
CLOSEDNP NMOD DETERM	0,1363
CLOSEDNP NMOD VARG	0,1709
CLOSEDNP NMOD VMOD	0,1774
CLOSEDNP NMOD VMOD SUBJ	0,1774
CLOSEDNP NMOD VMOD SUBJ PRECOMMA	0,1736
CLOSEDNP NMOD VMOD SUBJ MWEHEAD	0,1826
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG	0,1847
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN	0,1848
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN CONNECT	0,1838
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN AUXIL	0,1829
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN COREF	0,1841
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN NEGAT	0,1847
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP	0,1854
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP	0,1877
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP DEEPOBJ	0,1859
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP REFLEX	0,1867
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP COORDITEMS	0,1863
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP DEEPSUBJ	0,1874
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP ADJARG	0,1872
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP INTERROG	0,1877
CLOSEDNP NMOD VMOD SUBJ MWEHEAD NARG NN STRAYNP SEQNP SUBJCLIT	0,1876

Tableau 4. *Résultats par type de dépendance*

4.3. *Comparaison lemmes et dépendances*

Dans cette partie nous utilisons les différents corpus de CLEF 2003 pour évaluer une indexation à base de dépendances. Nous testons dans un premier temps cette indexation sur le corpus français puis nous effectuons des tests similaires avec les deux autres langages.

Une expérience ayant une pondération, ne prenant en compte que le nombre d'occurrences d'un descripteur dans un document (tf), produit la courbe de rappel

précision de la Figure 4. Une deuxième expérimentation, en utilisant une pondération à base de tf-idf, produit la courbe de la Figure 5.

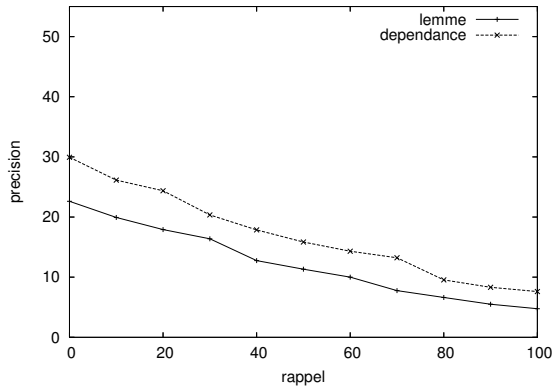

Figure 4. *Courbe de rappel précision sur le corpus français de CLEF 03 avec tf*

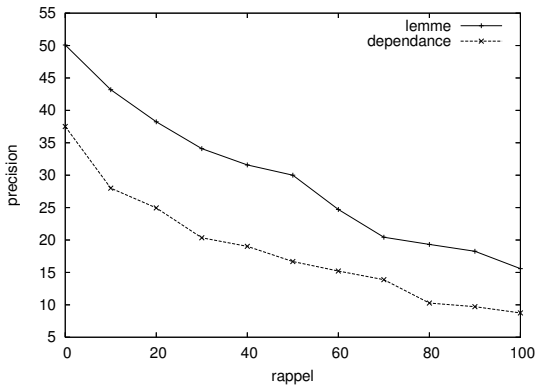

Figure 5. *Courbe de rappel précision sur le corpus français de CLEF 03 avec tf-idf*

À l'aide du tf, les résultats obtenus avec l'index basé sur les dépendances sont meilleurs que les résultats avec l'index basé sur les lemmes. L'utilisation d'une pondération, tel que tf-idf sur ces descripteurs, inverse les résultats. En effet avec l'utilisation de cette pondération, le résultat des lemmes est fortement amélioré alors que celui des dépendances n'est pas amélioré dans la même proportion. Cela signifie qu'une pondération à base de tf-idf n'est pas suffisamment adaptée aux dépendances et c'est une pondération pour les mots. Cette observation est renforcée par les résultats sur les deux autres langues (cf. Tableau 5) où l'utilisation du tf-idf sur les dépendances ne modifie pas les résultats et qui, de plus, dans le cas du finnois les

fait baisser. Le tableau montre aussi que l'utilisation d'autres pondérations telle que la 'divergence from randomness' DFR (Amati *et al.*, 2002) ont le même effet que le tf-idf, l'amélioration apportée aux lemmes est largement supérieure que celle apportée aux dépendances.

Langue	Type de pondération	Lemme	Dépendance
Français	Tf	0.1155	0.1614
	Tf-idf	0.2841	0.1757
	DFR	0.4130	0.2114
Finnois	Tf	0.0911	0.0906
	Tf-idf	0.2357	0.0841
	DFR	0.3846	0.1003
Russe	Tf	0.0601	0.0729
	Tf-idf	0.1433	0.1035
	DFR	0.2740	0.1701

Tableau 5. *Résultat en précision moyenne pour les trois langues*

Un certain nombre des dépendances extraites des requêtes ne sont pas présentes dans l'index des documents. Cela vient en partie du grand nombre de dépendances différentes et de la quantité de celles n'apparaissant qu'une fois. Ce phénomène est cependant accentué par la structure des requêtes utilisées dans la campagne de CLEF qui sont sous la forme de questions ou de phrases telles que : 'trouver des documents qui…'. Elles décrivent donc un besoin d'information et leur structure est différente de celle des documents à rechercher. Par conséquent, cela diminue les performances du système. Un traitement particulier des requêtes brutes, telles qu'elles apparaissent dans la collection de test, est donc nécessaire pour éviter cet artefact.

Au final, un certain nombre de requêtes sont améliorées et donnent de meilleurs résultats par l'utilisation des dépendances plutôt que des lemmes. Sur cette collection, 14 requêtes donnent de meilleurs résultats avec les dépendances, et pour la majorité de celles-ci, la précision à 5 documents est améliorée. Cela signifie donc que dans *certains cas*, l'utilisation de dépendances peut permettre d'améliorer la précision d'un système de RI.

4.4. *Regroupement des index*

Puisque les dépendances peuvent améliorer les résultats d'une recherche d'information, nous proposons de combiner les deux index dans le but de tirer profit des spécificités des deux descripteurs. Pour cela, les deux vecteurs index d'un document sont regroupés à l'interrogation et des pondérations leur sont affectées. Dans un premier temps, nous avons simplement regroupé les résultats à base de tf-idf. Par ce regroupement, les résultats sont améliorés (voir Tableau 6). Cependant

l'amélioration varie beaucoup en fonction des langues. En effet, les résultats en russe sont fortement améliorés (28%) alors que ceux en français ne le sont que beaucoup plus faiblement (3%).

	français	Russe	finnois
Résultat lemme	0.2841	0.1433	0.2357
Résultat dépendance	0.1757	0.1035	0.1003
Résultat regroupement	0.2939	0.1916	0.264

Tableau 6. *Résultat en précision moyenne du regroupement des résultats en tf-idf sur les trois langues*

Rappelons que les lemmes et les dépendances sont considérés comme indépendants. Or il est clair que ces deux descripteurs sont indéniablement liés, du fait même qu'une dépendance est constituée de deux lemmes. Pour prendre en compte ce lien entre ces deux types de descripteurs, nous proposons une pondération qui prenne en compte la répartition des lemmes pour pondérer les dépendances. Le poids d'une dépendance est ainsi obtenu par la multiplication du tf-idf des lemmes qui la composent. Cependant une telle pondération ne permet pas pour l'instant d'améliorer sensiblement les résultats précédents. Une modélisation de la relation entre dépendance et terme, reste donc à mettre au point.

5. Conclusion

Cet article montre d'une part, que la loi de Zipf est toujours respectée pour les dépendances et d'autre part, que l'utilisation des dépendances syntaxiques en tant que descripteurs dans une tâche de RI, donne des résultats encourageants. En effet, dans certains cas (qui restent à analyser), on a pu observer que les dépendances permettent d'améliorer la RI et d'atteindre une meilleure précision.

Du fait de leur répartition particulière face aux termes, et du fait qu'elles sont constituées de lemmes, les dépendances n'ont pas le même comportement que les autres descripteurs face aux modèles de pondération standard. Elles sont donc des descripteurs particuliers nécessitant une pondération adaptée, notamment par rapport aux spécificités de leur répartition sur le corpus.

En parallèle, un facteur essentiel à prendre en compte est la performance de l'analyseur syntaxique et la qualité des dépendances extraites. Les variations des performances sur les différentes langues peuvent provenir de la structure même des langues mais aussi de la qualité des informations extraites par les analyseurs syntaxiques. Etudier les variations de performance entre différents analyseurs sur une même langue pourrait être profitable, de même qu'intégrer dans une pondération

finale une mesure de la certitude de l'existance d'une dépendance (Brunet-Manquat, 2004).

Dans nos futurs travaux nous devrons trouver une pondération qui prenne en compte les spécificités des dépendances. Nos devrons appliquer des modèles de conversion sur la requête de manière à les approcher de la syntaxe du texte à retrouver. Enfin il nous faudra approfondir l'étude des dépendances dans le but de savoir quelles sont celles qui reflètent le mieux le thème du texte et qui sont ainsi pertinentes pour la RI.

Une autre direction de recherche consiste à aller vers une structure d'indexation inter-lingue, pour cela nous envisageons de remplacer les lemmes des dépendances par des acceptions inter-lingues.

Je remercie Jean-Pierre Chevallet, Gilles Sérasset et Christophe Brouard pour leurs relectures et leurs conseils pour le contenu de cet article.

6. Bibliographie:

Aït-Mokhtar, S., Chanod, J.P., Roux, C., « Robustness beyond shallowness : Incremental Deep Parsing. » *Special Issue of the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data*, Cambridge University Press, p. 121-144, 2002.

Amati, G.; van Rijsbergen, C., « Probabilistic models of information retrieval based on measuring the divergence from randomness », *ACM Transaction on Information Systems*, Volume 20 Issue 4, p. 357-389, 2002.

Berrut, C., Chiaramella, Y., « Indexing medical reports in a multimedia environment: the rime experimental approach. », *proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, p. 187–197, 1989

Boguslavsky, I., Apresian, J., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., Tsinman, L., « ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT », *MTT 2003. First International Conference on Meaning-Text Theory*, Paris, Ecole Normale Superieure, p. 279–288, 2003

Brunet-Manquat, F., « Description et conception d´une plate-forme robuste combinant des analyseurs d´énoncés. » *journal en ligne ISDM (Informations, Savoirs, Décisions et Médiations)*. vol. 13, 12 pages, février 2004.

Chevallet, J.P., « Un Modèle Logique de Recherche d'Informations appliqué au formalisme des Graphes Conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels », PhD thesis, Université Joseph Fourier, Grenoble, 1992

Chevallet, J.P., « X-IOTA Une plateforme distribuée ouverte pour l'expérimentation en Recherche d'Information », *in COnférence en Recherche Information et Applications CORIA'2004*, Toulouse, 10 - 12 mars, 2004.

Jacquemin, C., « Variation terminologique : reconnaissance et acquisition automatique des termes et de leurs variantes en corpus », Habilitation à diriger des thèses, Université de Nantes, 1997

Khoo, S-G. C., « The Use of Relation Matching in Information Retrieval », *LIBRES: Library and Information Science Research Electronic Journal*, ISSN 1058-6768, Volume 7 Issue 2; 1997

Koster, C.H.A., « Head/Modifier Frames for Information Retrieval.», *CICLing 2004*, Springer LNCS 2945, Heidelberg, p. 420-432, 2004

Lee, C., Lee, G.G., « Probabilistic information retrieval model for dependency structured indexing system. », *ACM SIGIR 2002 workshop on mathematical/formal methods in information retrieval*, 2002

Losee, R.M, « Term Dependence: Truncating the Bahadur Lazarsfeld Expansion », *Information Processing and Management*, vol. 30, n°2, p. 293-303, 1994

Matsumura, A., Takasu, A., Adachi, J., « The effect of information retrieval method using dependency relationship between words. », *Proceedings of the RIAO 2000 Conference*., p. 1043–1058, 2000

Meghini, C., Sebastiani, F., Straccia, U., « Mirlog: a logic for multimedia information retrieval. », *information Retrieval: Uncertainty and Logics. Advanced models for the representation and retrieval of information,* Kluwer Academic Publishing, Dordrecht, NL, p. 151–185, 1998

Metzler, D.P., Haas S.W., « The constituent object parser: syntactic structure matching for information retrieval », *ACM Transactions on Information Systems*, vol. 7, n°3, p. 296-316, 1989

Peters, C., « Introduction to the CLEF 2003 Working Notes », *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway, 2003

Smeaton A.F., « Using NLP or NLP Resources for Information Retrieval Tasks *», Natural Language Information Retrieval*, T. Strzalkowski (Ed.), Kluwer Academic Publishers, p.99-111, 1999.

Strzalkowski, T., Carballo, J.P., Marinescu, M. , « Natural Language Information Retrieval: TREC-3 Report. » *Overview of the Third Text REtrieval Conference TREC 1994*, p. 39-54, 1994

Tapanainen, P., « Parsing in two frameworks: finite-state and functional dependency Grammar », PhD thesis, University of Helsinki, Language Technology, Department of General Linguistics, Faculty of Arts, 1999.

Zipf, G.K., *Human Behavior and the Principle of Least-Effort.* Addison-Wesley, Cambridge, MA, 1949