Intégration de connaissances syntaxiques dans les modèles de langue pour la RI

Loïc Maisonnasse

Laboratoire CLIPS-IMAG
BP 53
38 041 Grenoble cedex 9
loic.maisonnasse@imag.fr

RÉSUMÉ. *En Recherche d'Information (RI) les méthodes purement statistiques basées sur des distributions de mots-clef ont actuellement atteint une limite. Cette limite n'est franchissable que par l'apport massif de connaissances extérieures au sein du système de RI. Nos travaux portent sur l'utilisation en RI des liens de niveaux syntaxiques entre les termes. Nous considérons ainsi les dépendances syntaxiques contenues dans l'arbre de dépendance produit par des analyseurs syntaxiques de surface. Pour intégrer ces informations en RI, le contexte des modèles de langue nous semble favorable. En effet, l'aspect théorique des modèles de langue est très intéressant, il est adaptable et permet l'intégration de nouvelles connaissances. Nous présentons ici, l'intégration des liens syntaxiques au sein d'un modèle de langue. Ce modèle est évalué sur une partie de la collection de CLEF. Les résultats montrent que l'intégration des dépendances syntaxiques abaisse les performances du système de RI. Face à ces résultats, nous souhaitons pour la suite de ces travaux nous orienter vers l'apport d'information de niveau plus sémantique.*

ABSTRACT. *In Information Retrieval (IR), statistic keyword based methods have reached a limit. This limit can only be cross by integrating, in large number outside source of knowledge, in IR system. Our work is based on the integration of the syntactic link between the terms produced by shallow parser. For that we consider the syntactic dependency of dependency tree produce by parser. For integrating these information in IR, we present the use of a language modeling approach. Language modeling approach theoretically framework is attractive as it can be adapted or in order to take into account new information. We present here the integration of the dependency relation in a language model. We evaluate this model on a part of the CLEF collection. The results show that the integration of dependency relation lowers the IR results. Consequently, knowing these results in the continuation of this work, we intend to integrate more semantic information instead of syntactic information.*

MOTS-CLÉS : *recherche d'information, analyse syntaxique, dépendance syntaxique, modèle de langue*

KEYWORDS: *information retrieval, shallow parsing, syntactic dependency, language model*

1. Introduction

Les performances des modèles de Recherche d'Information (RI) uniquement basés sur des méthodes statistiques ont actuellement atteint une limite. Seule l'utilisation de connaissances complémentaires à celles du corpus peut permettre de dépasser cette limite. Différents types d'informations sont utiles en RI. Leur impact est souvent orienté ; soit vers l'amélioration du rappel, utilisation d'ontologie pour l'extension de requête, soit dans le sens de l'amélioration de la précision, désambiguïsation des termes. Nous nous intéressons ici à l'amélioration de la précision en RI. Dans certains domaines, notamment professionnels, obtenir rapidement les bonnes réponses est crucial. Pour améliorer la précision, nous étudions les informations produites par des analyseurs syntaxiques et leur intégration en RI. Nous examinons, plus précisément, les relations syntaxiques entre les mots produites par des analyses en dépendance et nous intégrons ces dépendances linguistiques au sein d'un modèle de langue de RI. Nous introduisons les Modèles de Langue (ML) et les travaux sur l'utilisation de la syntaxe en RI. Nous décrivons ensuite les ML intégrant des relations. Enfin nous présenterons les résultats de l'utilisation de ces modèles sur des dépendances syntaxiques.

2. Les modèles de langues

En RI, les ML ont été introduits en 1998 par Ponte et Croft (Ponte *et al.*, 1998). Les auteurs proposent un modèle de RI où le score assigné à un document pour une requête est établi comme la probabilité que la requête soit générée par un modèle du document. La difficulté est d'établir ce modèle pour des documents de petites tailles. Pour résoudre le manque d'information lors de la construction du modèle, différentes méthodes de lissage (Hiemstra, 98) (Song *et al.*, 99) et différents modèles (Berger *et al.*, 1999) ont été proposés. Les résultats obtenus par ces modèles ont montré des performances équivalentes voire supérieures aux modèles classiques. Pourtant, ils se basent sur l'hypothèse d'indépendance des termes (unigramme) ou des multi-termes (bigramme, trigramme). Dans les ML, comme généralement en RI, cette hypothèse est plus une facilité mathématique qu'une réalité, une partie de la sémantique des documents s'exprimant à travers les relations qu'entretiennent les mots, en particulier les relations syntaxiques.

3. La syntaxe en RI

En RI, l'utilisation d'informations syntaxiques n'est pas nouvelle, plusieurs méthodes utilisant ces informations ont été proposées. La plus répandue consiste à utiliser ces informations pour extraire des syntagmes. Dans (Tong *et al.*, 1996) et (Strzalkowski *et al.*, 1994) les auteurs se basent sur l'hypothèse qu'un syntagme est produit par une structure de dépendances. Les syntagmes sont extraits après analyse et ajoutés à l'index. Les auteurs obtiennent ainsi une augmentation de l'ordre de

20% de la précision moyenne. Partant de l'hypothèse que la conversion des structures en syntagmes entraîne une perte d'information, des recherches ont porté sur la structure. Dans (Matsumura *et al.*, 2000) (Metzler *et al.*, 1989), les auteurs extraient des arbres de dépendances à partir de phrases. Ils évaluent ensuite les documents pertinents par plusieurs correspondances entre les dépendances de la requête et des documents. Ces méthodes n'intègrent pas les ambiguïtés syntaxiques, Smeaton (Smeaton, 1999) propose un modèle, appliqué aux syntagmes, où ces ambiguïtés sont représentées. Cependant, les résultats du modèle sont inférieurs à ceux obtenus par les syntagmes. La syntaxe est aussi utilisée dans les systèmes de questions réponses, notamment lors de la sélection des phrases pouvant contenir la réponse. Punyakanok (Punyakanok *et al.*, 2004) établi une distance basée sur des transformations entre l'arbre d'une phrase et celui d'une requête. Hang (Hang *et al.*, 2005) apprend la probabilité de transformer un lien syntaxique en un autre.

4. Les dépendances dans les modèles de langue

Des travaux proposent d'intégrer la structure dans les ML. Ces travaux se basent sur la réduction du calcul de la probabilité des termes. Ce calcul, par expansion, est complexe ; chaque mot dépendant de ceux déjà apparus. Les auteurs limitent donc le nombre des dépendances aux plus fortes, la probabilité d'un terme w_i est calculée en fonction de son antécédent le plus probable w_j, celui qui maximise $P(w_i \mid w_j)$.

Dans (Nallapati *et al.*, 2002), les auteurs obtiennent les antécédents par le calcul d'un arbre de couverture maximum (MST) sur le degré de dépendance des termes de la requête. Ce degré est estimé par le calcul du cœficient de Jaccard. La probabilité de générer une phrase sachant un modèle de document M_D est alors le produit des probabilités de générer chaque terme sachant son antécédent dans le MST.

$P(S \mid M_D) = \prod_{w \in S} P(w \mid A(w), M_D)$ où $A(w)$ est l'antécédent de w.

La probabilité finale est obtenue par un lissage de la probabilité initiale avec un modèle de l'anglais. En parallèle, ce modèle est mixé avec un modèle unigramme. L'évaluation du modèle, sur une tâche de détection d'histoire, montre qu'un modèle unigramme donne de meilleurs résultats que le modèle basé sur les dépendances. Cependant, les performances sont améliorées en utilisant le modèle hybride.

Dans une autre approche (Gao *et al.*, 2004), les auteurs utilisent une structure de dépendance pour limiter les calculs. La structure est produite par un analyseur statistique dont les relations ne sont pas nécessairement grammaticales. Le modèle de document créé prend en compte ces dépendances. Les auteurs considèrent les liens entre les termes comme une variable cachée qui permet d'exprimer les dépendances. La génération de la requête est alors un processus en deux étapes. La structure de dépendance *L* est produite selon une probabilité *P(L|D)*. Puis la requête *Q* est générée selon *P(Q|L,D)*, les termes de la requête étant choisis en fonction des éléments liés dans L. La probabilité de produire la requête *P(Q|D)* sachant toutes les structures de dépendances possibles *Ls* est alors :

$$P(Q|D) = \sum_{Ls} P(Q, L|D) = \sum_{Ls} P(L|D)P(Q|L, D)$$

Les auteurs supposent ensuite que l'ensemble des structures possibles Ls est dominé par une unique structure L, la structure la plus probable. On obtient alors :

$$P(Q|D) = \log(P(L|D)) + \sum_{i=1..m} P(q_i|D) + \sum_{(i,j)\in L} MIq_i,(q_j|L,D) \text{ Où } MI(q_i, q_j|L, D) = \log \frac{P(q_i, q_j|L, D)}{P(q_i|D)P(q_j|D)}$$

Le modèle est évalué sur les collections de TREC. Les résultats obtenus sont meilleurs que ceux du modèle probabiliste ou du modèle unigramme. Cependant la différence avec ce dernier est faible. La prise en compte des dépendances entre les termes ne donne pas de bons résultats car il est difficile d'estimer les dépendances.

5. Modèles de langues sur les dépendances syntaxiques

Contrairement aux approches présentées précédemment, nous proposons ici d'intégrer les dépendances produites par un analyseur syntaxique au sein d'un ML. Nous utilisons l'analyseur XIP de Xerox. A partir de ces résultats, nous sélectionnons les lemmes porteurs de sens et les dépendances qui les relient. Pour chaque phrase, nous obtenons un graphe (voir Figure 1) où les nœuds sont des lemmes reliés par des dépendances linguistiques.

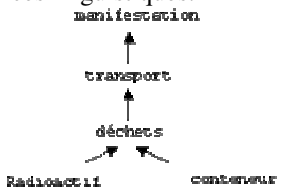

Figure 1: *Structure utilisée pour la phrase : "les manifestations contre le transport de déchets radioactifs par conteneurs en Allemagne."*

Le ML que nous utilisons est une version simplifiée du modèle de Gao. Nous sélectionnons la structure la plus probable comme celle produite par l'analyseur et nous utilisons des estimations simples. Nous estimons *P(L|D)*, comme la probabilité que deux termes (q_i,q_j) soit liés s'ils apparaissent dans la même phrase d'un document et nous interpolons (λ_l) cette probabilité avec celle obtenue sur la collection (F1). Nous testons, une deuxième estimation (F2) où *P(L|D)* est la probabilité qu'il existe une dépendance de q_j vers q_i sachant que q_i existe. Nous estimons *P(qi|D)* comme la probabilité qu'un lemme apparaisse dans un document et nous interpolons (λ_d) sur la collection. Pour éviter que les lemmes ou les dépendances absents de la collection rendent nuls les résultats d'une requête, nous éliminons des requêtes les éléments absents de la collection. Enfin, pour MI(qi, qj|L,D) nous utilisons l'estimation de Gao (Gao *et al.*, 2004).

5.1. *Evaluation du modèle*

Nous évaluons ce ML sur une partie des collections CLEF (Peters, 2004). Nous utilisons la collection 'Le Monde 95' avec les requêtes de CLEF 2004. Les résultats obtenus sont les suivants :

		F1	F2
λ_l	λ_d	MAP	MAP
0.1	0.5	37.6%	37.3%
0.1	0.9	38.0%	37.8%
0.1	0.9999	37.3%	37.3%
0.3	0.5	38.9%	38.1%
0.3	0.9	40.2%	39.2%
0.3	0.9999	40.1%	**40.1%**
0.5	0.5	38.4%	37.2%
0.5	0.9	39.1%	38.2%
0.5	0.9999	**40.7%**	39.7%

Tableau 1. *Résultats en précision moyenne en fonction des interpolations*

pondération	Précision moyenne
Tf-idf (lemme)	0.3058
DFR (lemme)	0.4067

Modèle vectoriel

λ_l	MAP
0.1	39.8%
0.3	40.7%
0.5	39.4%
0.9	35.0%

Modèle unigramme

Tableau 2. *Résultats en précision moyenne sur d'autres modèles à base de lemmes*

Les résultats à base de ML sont supérieurs aux résultats obtenus par la pondération tf-idf et égalisent ceux de la pondération DFR. Le modèle utilisant les dépendances égalise le modèle unigramme lorsque la prise en compte de celles-ci au sein des documents est faible. Les bons résultats sont obtenus lorsque l'influence des dépendances est limitée. Nous testons deux variations simples de l'estimation de *P(L|D)*, leurs résultats sont proches avec des variations similaires. Ces résultats semblent améliorables par l'utilisation de meilleures estimations, notamment pour *P(L|D)* qui fournit des probabilités élevées au sein des documents et des probabilités beaucoup plus faibles sur la collection.

6. Conclusion

Nous avons montré qu'il est possible d'intégrer les dépendances syntaxiques produites par un analyseur syntaxique au sein d'un modèle de langue. Cependant les résultats montrent que cette intégration n'est pas suffisante pour supplanter les autres approches, un ML plus adapté ou des estimations plus complexes prenant en compte les spécificités des dépendances syntaxiques pourraient améliorer les résultats. De même, l'ajout d'informations, telle qu'une estimation de la pertinence des structures syntaxiques, améliorerait le modèle. S'il est envisageable d'obtenir

une amélioration par l'utilisation des dépendances syntaxiques, ces relations seules semblent insuffisantes pour la RI. Nous souhaitons donc par la suite utiliser ces relations comme base pour extraire des relations de niveau plus sémantique qui s'intégrerons plus facilement dans un modèle de RI.

7. Bibliographie

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

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

Djoerd Hiemstra, *« Using Language Models for Information Retrieval »,* Thèse de doctorat, University of Twente, 1998

Gao, J., Nie, J.Y., Wu, G., Cao, G., « Dependence language model for information retrieval », *SIGIR-2004*. Sheffield, UK, juillet 25-29, 2004, p. 170-177.

Nallapati, R., Allan, J., « Capturing term dependencies using a language model based on sentence trees », C*onference on Information and knowledge management*, ACM Press, 2002, p.383-390.

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

Tong, X., Zhai, C., Milic-Frayling, N., Evans, D.A., « Evaluation of syntactic phrase indexing », *The Fifth Text Retrieval Conference (TREC-5)*. edition D. K. Harman, 1997.

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, 1989, p. 296-316.

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

Ponte, L.M., Croft W.B., « A language modeling approach to information retrieval », *ACM SIGIR*, 1998, p. 275-281.

Berger, D., Lafferty J., « Information retrieval as statistical translation », *ACM SIGIR*, Berkeley, California, United States, 1999, p. 222-229.

Song, F., Croft, W.B., « A General Language Model for Information Retrieval », I*nternational conference on Information and knowledge management*, 1999, p. 316-321.

Punyakanok., V., Roth, D., Yih, W., « Mapping Dependencies Trees: An Application to Question Answering », *AI & Math*, 2004

Hang, C., Renxu, S., Keya, L., Min-Yen, K., Tat-Seng, C., « Question answering passage retrieval using dependency relations », *ACM SIGIR*, Salvador, Brazil, pp. 400-407, 2005