Désambiguïsation de textes arabes pour l'extraction des candidats termes

L'apport de la structure des documents

Ibrahim Bounhas, Yahya Slimani

Département des Sciences de l'Informatique,
Faculté des Sciences de Tunis,
Université de Tunis El Manar, 1060, Tunis,
Tunisie

Bounhas.ibrahim@yahoo.fr / yahya.slimani@fst.rnu.tn

RÉSUMÉ. Les syntagmes nominaux ont une grande importance dans le processus de représentation et de recherche d'information. D'une part, ils constituent des candidats termes qui relèvent de la sémantique du domaine. D'autre part, les relations syntaxiques qui lient les constituants des termes composés encodent des relations sémantiques. Nous proposons d'extraire les syntagmes nominaux à partir de corpus arabes semi-structurés en résolvant les ambiguïtés qui peuvent apparaître à plusieurs niveaux d'analyse. Nous proposons un nouvel algorithme de désambiguïsation qui utilise divers types de contexte. Nous montrons que les relations syntaxiques et les liens entre les fragments du document jouent un rôle important dans la désambiguïsation.

ABSTRACT. Noun phrases have a great importance in the process of information representation and retrieval. In the one hand, they constitute domain relevant candidate terms. On the other hand, syntactic relations which link constituents of compound nouns can be used to infer semantic relations. We propose to extract noun phrases from Arabic semi-structured corpora by resolving ambiguities which can appear in the different levels of analysis. We propose a new algorithm of disambiguation which uses various types of context. The experiments show that contextual relations mined from logical links between fragments of the document and from syntactic relations have a great importance in the process of disambiguation.

MOTS-CLÉS : Syntagmes nominaux, Désambiguïsation de textes arabes, Documents semi-structurés.

KEYWORDS : Noun phrases, Arabic text disambiguation, Semi-structured documents.

1. Introduction

Les propriétés de la langue arabe rendent son traitement automatique difficile. C'est une langue dérivationnelle, inflectionnelle et agglutinante. S'ajoute à cela, l'absence de voyellation dans la majorité des textes disponibles. Ainsi, les textes arabes sont ambigus au niveau morphologique, syntaxique, sémantique et pragmatique. Ces ambiguïtés influencent plusieurs étapes dans le processus de recherche d'information. En effet, l'indexation de documents ou la construction de ressources termino-ontologiques requièrent l'extraction de termes (ou concepts) clés. Dans cette phase une étape de désambiguïsation est essentielle.

Notons que les noms sont considérés comme les entités qui représentent le sujet d'un document. Cependant, nous distinguons deux types de noms : les termes simples et les noms composés ou syntagmes nominaux. Ainsi, la première étape consiste en une analyse morphologique qui permet d'identifier les noms simples. Ensuite, une analyse syntaxique permet de constituer des syntagmes conformément à la grammaire arabe. Ces deux étapes génèrent des ambiguïtés morphologiques et syntaxiques d'où la nécessité d'une étape de désambiguïsation. Notre objectif consiste à étudier et proposer des solutions aux problèmes d'ambiguïté des textes arabes non voyellés dans le processus d'extraction des candidats termes. La section 2 énumère les solutions proposées dans la littérature. L'architecture que nous proposons est présentée dans la section 3. Notre solution consiste à utiliser divers types de relations contextuelles entre les termes lors de la désambiguïsation. La section 4 définit les principes de construction du graphe contextuel et l'algorithme de désambiguïsation. Dans la section 5, nous présentons et interprétons résultats d'expérimentations. La section 6 clôture ce papier et propose quelques perspectives pour de futures recherches.

2. Etat de l'art

2.1. Extraction de termes composés

Parmi tous les types de termes composés, nous nous intéressons aux noms composés. Bounhas et Slimani [3] ont résumé les différentes catégories de syntagmes nominaux. Ils ont aussi identifié les types, les rôles et les propriétés morphologiques des constituants de chaque type de syntagmes. Nous pouvons distinguer deux types d'approches d'extraction de syntagmes nominaux : les approches statistiques [7] et les approches linguistiques [2]. Les approches statistiques se servent de mesures d'association pour décider de la validité des termes. De telles approches ignorent des candidats termes valides dont la fréquence est faible. Par contre, les approches linguistiques exploitent l'information morphologique, syntaxique ou sémantique mise en application dans des

règles ou des programmes spécifiques à une langue. Conséquemment, elles sont dépendantes de la langue et pas assez flexibles pour faire face aux structures complexes des syntagmes [3]. Pour éviter les faiblesses des deux approches, une solution communément adoptée est de les combiner. Les travaux qui se sont intéressés à la langue arabe soufrent du manque d'une étape d'analyse morphologique ou de mesures statistiques qui permettent de filtrer les candidats termes ou encore ignorent certains types de syntagmes [3]. En effet, une approche d'extraction de syntagmes nominaux requiert l'intégration d'un analyseur morphologique qui permet d'analyser les mots et de calculer les traits morphologiques. Ensuite, un analyseur syntaxique identifie les termes composés en respectant les règles grammaticales de la langue. Cette approche a été implémentée par Bounhas et Slimani [3].

2.2. Analyse et désambiguïsation linguistique

Pour réduire les ambiguïtés, deux solutions sont envisageables. La première consiste à utiliser le contexte. Etant donnée une entité qui a plusieurs interprétations possibles, il s'agit, dans une première étape, d'associer à chaque interprétation un ou plusieurs contextes par apprentissage dans un corpus étiqueté. Dans une deuxième étape, on essaie de désambiguïser les entités dans un corpus de test par comparaison des nouveaux contextes à ceux appris dans la première étape. La deuxième solution consiste à résoudre les ambiguïtés d'un niveau en passant au niveau suivant. Par exemple, un analyseur syntaxique peut filtrer les solutions proposées par un analyseur morphologique pour n'en garder que les solutions compatibles avec les règles de la grammaire [1]. La modélisation du contexte ou du niveau sémantique requiert la mise en relation des termes sémantiquement proches. Ces relations peuvent être utilisées non seulement dans la désambiguïsation mais aussi pour la construction d'une ressource termino-ontologique. Elles peuvent être extraites en exploitant la structure des termes ou leurs contextes. La méthode implémentée dans Upery [6] assume que les termes liés syntaxiquement partagent certains traits sémantiques. Cette méthode n'a pas été testée pour la langue arabe alors que les travaux qui se sont intéressés à la construction de terminologies arabes ou à des tâches de recherche d'information sont principalement basés sur la cooccurrence.

3. Architecture du système

Notre modèle de désambiguïsation combine les approches étudiées dans la section 2.2. L'architecture que nous proposons se distingue par deux aspects : (i) elle intègre les différents niveaux d'analyse linguistique; (ii) elle combine trois types de contextes:

– Le contexte morphologique : nous intégrons l'outil MADA [9] qui exploite le contexte du mot dans la phrase pour choisir la meilleure solution morphologique.

– Le contexte syntaxique : comme Bourigault [6], nous considérons que chaque relation syntaxique constitue un contexte. Chacun des deux termes qui constituent le syntagme représente pour l'autre un contexte paramétré par le type de la relation.

– Le contexte pragmatique : nous appliquons notre approche pour les documents semi-structurés. La structure de tels documents induit des relations pragmatiques qui correspondent aux relations logiques entre les différents fragments.

Notre objectif consiste à évaluer l'impact des trois types de contexte sur le processus de désambiguïsation. L'architecture proposée est schématisée par la Figure 1. Nous commençons par extraire la structure logique des documents du corpus en utilisant l'analyseur macro-logique développé par Bounhas et Slimani [4]. Le niveau morphologique est traité par l'outil MADA [9]. Notre choix est justifié par le fait que cet outil est le seul, à l'heure actuelle, qui effectue l'annotation grammaticale, l'analyse et la désambiguïsation morphologique dans une seule étape. Au niveau syntaxique, nous intégrons l'outil développé par Bounhas et Slimani [3]. A partir des arbres syntaxiques générés, nous identifions deux types d'éléments. D'une part, nous stockons pour chaque phrase les éléments non ambigus (lemmes des noms simples et les noms composés). D'autre part, pour chaque ambiguïté détectée dans la phrase, nous stockons les solutions possibles (éléments ambigus).

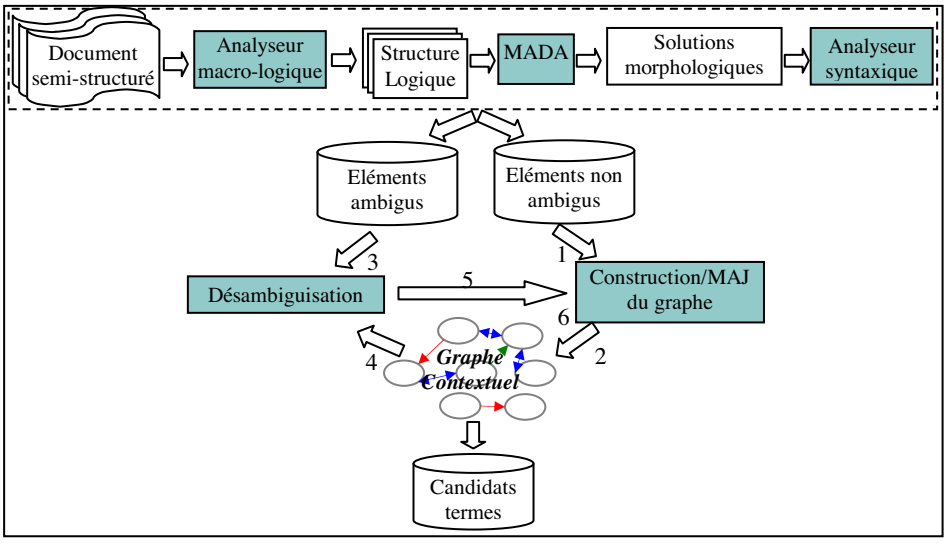

Figure 1 : *L'architecture proposée.*

Nous procédons ensuite à une étape d'apprentissage qui consiste à désambiguïser manuellement les termes qui apparaissent dans les titres et les sous-titres des documents. Ces derniers représentent à la fois un faible pourcentage en terme de quantité par rapport à la taille du corpus et les éléments les plus importants qui reflètent la sémantique des documents. A partir des éléments non ambigus et ceux désambiguïsés, un réseau contextuel de lemmes et de syntagmes est construit. Un algorithme de désambiguïsation traite les ambiguïtés en exploitant les connaissances de ce réseau qui est enrichi au fur et à mesure que les ambiguïtés sont traitées. A la fin, nous générons la liste des termes qui représentent l'index de chaque fragment du corpus.

4. L'algorithme de désambiguïsation

Cet algorithme utilise un graphe contextuel constitué par les termes (simples ou composés) et leurs relations. Chaque terme ajouté au graphe induit de nouveaux arcs dans le graphe ou renforce les arcs existants. Ainsi, chaque arc a un poids et un label. En effet, nous distinguons deux types de relations comme suit :

– Les relations syntaxiques : certaines relations sont symétriques (e.g. les syntagmes conjonctifs). Dans ce cas, les constituants du syntagme sont liés par un arc doublement orienté qui a comme étiquette le nom de la relation syntaxique. Pour d'autres relations non symétriques (e.g. le cas des syntagmes descriptifs), le syntagme est composé de deux éléments à savoir: la tête et l'expansion. L'expansion (respectivement la tête) est liée à la tête (respectivement l'expansion) par un arc dont le label est constitué du type de la relation concaténé à la chaîne « _tête » (respectivement « _expansion »).

– Les relations pragmatiques : Ces relations sont définies en utilisant le modèle d'indexation proposé par Bounhas et Slimani [5]. En effet, les termes qui apparaissent dans un noeud N_i de la structure du document sont liés par une relation étiquetée "Sup" aux termes qui apparaissent dans tous les pères hiérarchiques de N_i

Résoudre une ambiguïté revient à choisir une solution parmi une liste de solutions $LS = \{u_1,u_2,\ldots,u_n\}$. Dans le cas d'une ambiguïté morphologique, les u_j correspondent aux lemmes possibles d'un mot. Dans le cas d'une ambiguïté syntaxique l'ensemble LS contient des termes composés qui correspondent à différents regroupements. Dans les deux cas, nous évaluons la pertinence de chaque solution u_j par rapport aux connaissances actuelles du réseau contextuel. Il s'agit donc d'évaluer le degré de corrélation d'une solution avec son contexte. En effet, l'ajout de u_j engendre l'ajout d'un ensemble d'arcs $S = \{c_1, c_2,\ldots,c_m\}$ où c_i met en relation le terme u_j avec le terme v_i avec une relation de type r_i. Nous modélisons le choix de la meilleure solution comme une tâche de recherche d'information. La requête est composée des contextes $c_i = (v_i, r_i)$ et les documents sont les éléments de LS. Décider de la validité de u_j, revient à évaluer sa pertinence par rapport à la requête composée des arcs qu'elle engendre ce qui revient à

mesurer le degré d'appariement de la solution avec son contexte. Soit une requête $RQ=S=(c_1,c_2,\ldots,c_m)$. Nous proposons une approche possibiliste pour le calcul de la pertinence. Ce choix se justifie par le fait que ce modèle est adapté aux problèmes de l'imprécision et de l'incertitude tels que celui de la résolution de l'ambiguïté. La pertinence d'une solution u_j pour RQ, est calculée à travers les mesures de possibilité (Π) et de nécessité (N). L'expression $\Pi(u_j|RQ)$ est proportionnelle à [8]:

$$\Pi' (u_j|RQ) = \Pi(c_1| u_j) * \ldots * \Pi(c_m| u_j) = Freq_{1j} * \ldots * Freq_{mj}$$

Dans notre cas, $Freq_{ij}$ est égale au poids de l'arc ayant le label r_i qui relie u_j et v_i. La nécessité de u_j pour la requête RQ, notée $N(u_j|RQ)$, est calculée comme suit:

$$N(u_j|RQ) = 1- \Pi (\neg u_j|RQ) \text{ Avec } \Pi (\neg u_j|RQ) = (\Pi(RQ|\neg u_j) * \Pi(\neg u_j))/\Pi(RQ)$$

D'une manière analogique, $\Pi(\neg u_j|RQ)$ est proportionnelle à:

$$\Pi'(\neg u_j|RQ) = (1- \phi u_{1j}) * \ldots * (1- \phi u_{mj})$$

Avec: $\phi u_{ij} = \text{Log}_{10}(|LS|/nS_i) * (Freq_{ij})$

Dans cette formule, $|LS|$ est le cardinal de l'ensemble LS. nS_i est le nombre de solutions u_j dans LS qui sont corrélées avec le contexte c_i (i.e. $Freq_{ij} >0$). Le Degré de Pertinence Possibiliste (DPP) de u_j est égal à la somme des deux mesures: $DPP(u_i) = \Pi(u_j|RQ) + N(u_j|RQ)$. L'algorithme sélectionne la solution ayant le DPP les plus élevé et ajoute les arcs q'elle engendre au graphe.

5. Expérimentations et résultats

Les corpus utilisés dans nos expérimentations sont extraits à partir de six livres encyclopédiques des citations arabes groupées par thème[1]. En effet, nous avons construit trois corpus qui correspondent aux thèmes mariage, boissons et purification. Nous avons choisi les thèmes qui ne se chevauchent pas avec d'autres dans les six livres. Le Tableau 1 présente des statistiques sur chaque domaine en termes des différents types de fragments, du nombre de mots et du nombre de candidats termes simples et composés.

	Boissons	Mariage	Purification
Nombre des titres de niveau 1	1	1	10
Nombre des titres de niveau 2	200	444	745
Nombre des paragraphes	715	1091	2129
Nombre de mots	15780	30975	49335
Nombre de noms simples pertinents au domaine	163	290	271
Nombre de noms composés pertinents au domaine	540	574	822

Tableau 1. *Statistiques sur les fragments et les termes dans les trois corpus.*

[1] Il s'agit des livres: «صحيح البخارى», «صحيح مسلم», «سنن أبي داود», «سنن الترمذى», «سنن النسائي» et «سنن ماجة»

Nous testons notre algorithme de désambiguïsation en utilisant différentes combinaisons de types de contextes. Pour chaque combinaison, nous évaluons les résultats en termes du nombre des candidats termes extraits sachant que ces derniers sont filtrés en utilisant la mesure TF-IDF (Term Frequency–Inverse Document Frequency). Nous utilisons comme métriques d'évaluation, le rappel (R) et la précision (P). Comme liste de référence, nous nous sommes aidés d'un dictionnaire spécialisé [2]. Le Tableau 2 récapitule les résultats obtenus dans le cadre de deux expériences pour l'extraction des termes simples et composés. Dans la première, nous adoptons la solution choisie par MADA. Dans le deuxième, nous considérons toutes les solutions morphologiques. Il est clair que cette solution produit de meilleurs résultas, car MADA commet certaines erreurs et élimine plusieurs candidats termes corrects qui n'apparaissent pas en premier rang dans le classement. Ainsi, il est préférable de résoudre les ambiguïtés morphologiques en considérant d'autres types de contexte. Les meilleurs résultats sont obtenus en combinant le contexte syntaxique et le contexte pragmatique. L'élimination de l'un de ces deux types de contexte détériore les résultats en termes de rappel et de précision.

	1- Avec contexte morphologique						2- Sans contexte morphologique					
	Simples		Composés		**Total**		Simples		Composés		**Total**	
Types de contextes	R	P	R	P	R	P	R	P	R	P	R	P
Syntaxique et pragmatique	0,72	0,16	0,53	0,45	0,58	0,18	0,84	0,40	0,91	0,47	0,89	0,46
Pragmatique uniquement	0,74	0,31	0,44	0,44	0,51	0,38	0,77	0,33	0,55	0,42	0,60	0,38
Syntaxique uniquement	0,57	0,28	0,20	0,30	0,29	0,29	0,77	0,33	0,41	0,52	0,50	0,42

Tableau 2. *Résultats d'extraction de candidats termes*

6. Conclusion

Dans une perspective d'indexation de documents et de construction de ressources termino-ontologiques, nous avons présenté une nouvelle approche d'extraction de candidats termes simples et composés, à partir de documents semi-structurés arabes, en mettant l'accent sur le problème des ambiguïtés. Notre proposition combine des approches linguistiques et statistiques. En outre, nous proposons de nouvelles méthodes d'indexation et de désambiguïsation. Du point de vue expérimentation, nous avons testé et combiné trois types de contextes à savoir le contexte morphologique, le contexte syntaxique et le contexte pragmatique. Ce dernier exploite les liens logiques entre les fragments du document comme information contextuelle pour la désambiguïsation. Nous avons prouvé que ce type d'information améliore considérablement les performances de l'outil d'extraction. Nous pensons aussi que le réseau contextuel produit

[2] محمد رواس قلعه جي وحامد صادق قنيبي. 1988. معجم لغة الفقهاء. دار النفائس, الطبعة الثانية

constitue une base de connaissances linguistiques qui peut être exploitée pour construire une ressource termino-ontologique.

7. Bibliographie

[1] C. Aloulou, L. H. Belguith, A. H. Kacem, and A. Ben Hamadou, Conception et développement du système MASPAR d'analyse de l'arabe selon une approche agent, *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, 2004.

[2] M. A. Attia, M. Rashwan, A. Ragheb, M. Al-Badrashiny, H. Al-Basoumy and S. Abdou. A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields. *The 6th international conference on Advances in Natural Language Processing*, pp. 65 – 76, Gothenburg, Sweden, 2008.

[3] I. Bounhas, and Y. Slimani. A hybrid Approach for Arabic Multi-Word Term Extraction, *IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'09)*, pp. 429-436, Dalian, China, September 24-27, 2009.

[4] I. Bounhas, and Y. Slimani. A social approach for semi-structured document modeling and analysis. *International Conference on Knowledge Management and Information Sharing KMIS 09*, pp. 95-102, Madeira, Portugal, 6 - 8 October, 2009.

[5] I. Bounhas, and Y. Slimani. A hierarchical approach for semi-structured document indexing and terminology extraction", *International conference on information retrieval and knowledge management* (CAMP'2010), Shah-Alam, Malaysia, 2010, pp. 314-319.

[6] D. Bourigault. Upery: un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *The 9th Conference on Natural Language Processing TALN'2002*, pp.75-84 Nancy, France.

[7] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pp. 76–83, Vancouver, B.C. Association for Computational Linguistics.

[8] B. Elayeb, F. Evrard, M. Zaghdoud & M. Ben Ahmed. Towards An Intelligent Possibilistic Web Information Retrieval using Multiagent System. *The International Journal of Interactive Technology and Smart Education (ITSE)*, Special issue: New learning support systems, Emerald Group Publishing Limited, 2009, 6(1): 40-59.

[9] N. Habash, O. Rambow and R. Roth. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. *The 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, pp.102-109, Cairo, Egypt, 22-23 April, 2009.