International Conference on Control, Engineering & Information Technology (CEIT'13)
International Journal of Business & Economic Strategy (IJBES)

A quality study of noun phrases as document keywords for information retrieval

Chedi Bechikh Ali
ISG, Tunis University
LISI laboratory, INSAT, Carthage University
Email: chedi.bechikh@gmail.com

Hatem Haddad
ESSTHS, Sousse University
LIPAH laboratory, Faculty of science Tunis
Email: hatem.haddad@gmail.com

***Abstract*—This paper presents a study that we conducted on the quality of noun phrases (NPs) that we extracted and that will be used in our future work in the process of information retrieval. For this, we extract noun phrases from the corpus DEFT 2012 and compare them with noun phrases keywords chosen by the articles authors . An accuracy measure is defined to evaluate the extracted noun phrases.**

***Index Terms*—noun phrases, terms dependency, information retrieval, natural langage processing.**

I. Introduction

With the increasing volume of electronic information stored and exchanged, the number of retrieved documents by search engines is becoming very important. They are facing the challenge to be effective and also to order the relevant retrieved documents.

To achieve this precision, a system must use an expressive documents and queries representation. The representation expressiveness should be based on the keywords quality so the keywords should best represent the document content. Earlier works proved that the use of simple terms as keywords is not accurate enough to represent the documents contents due to the words ambiguity.

A solution to this problem is to use complex terms instead of simple terms [1]. The assumption is that complex terms are more likely to identify semantic entities that simple words and are then a better representation of the semantic documents content [2].

The study that we propose in this paper concerns the evaluation of complex terms quality. Indeed, we propose to use complex terms to represent document, and more precisely noun phrases (NPs). These noun phrases are extracted using a linguistic analysis.

We rely on syntactic information to extract noun phrases, which allows integrating dependencies between words and overcomes the paradigm of "bag of words". The remainder of the paper is organized as follows. In section 2 we review some existing work that model the dependencies between word for information retrieval. Section 3 describes our approch for extracting noun phrases. Section 4 presents the DEFT compaign and shows the relevance of NPs as keywords of the DEFT corpora empirically. Finally we offer concluding remarks.

II. Integration of dependency between words to represent meaning in information retrieval

Most IR models use keywords as "bags of words", then documents are represented as an unordered terms set. For example, in this representation, the text "the bear eats man '" and "man eats bears" are identical. However, these sentences clearly have different meanings.

On the other hand, terms dependencies and relations ships exist in a text collection of. For example, some couples occurrences of terms are correlated, such as the terms "information" and "retrieval." That one occurs is strong evidence that the other is also likely to occur.

So modeling dependencies between terms to better represent the meaning of documents and queries and to better represent the links between the terms. Better representation of the data allows a better understanding of the meaning of the text, which can improve the performance of SRI.

Most work on modeling term dependencies in the past have focused on the phrases, the proximity [3] or co-occurring term [4]. Most of these models take into account only the dependencies between pairs of terms. In [5], Fagan discusses how to identify and use non-syntactic phrases (statistics).

It identifies the sentences using factors such as the number of times the word occurs in the collection and the proximity of the terms of the sentence. For many collections, significant improvements in relevance are obtained when the phrases are defined as both terms in a query or in a document with a near unlimited. However, for other collections, this method of identifying phrases helped make marginal improvements or negative.

Gao et al. [6] have proposed to extend the IR language model with a dependency structure, called linekage that is inspired from a grammartical link [7], [8]. In the proposed model, a link between a words pair across a sentence is created taking into account several linguistically constraints. This model showed consistent improvements on a number of TREC collections. Unfortunately, the model requires to build a linekage information for each query, which requires a lot of time.

Other works use techniques for detecting noun phrases and syntactic sentences [9], [10]. However, all these tech-

niques have shown that it provides little or no improvements.

In [11], the author useds a linguistic analysis to extract noun phrases. These noun phrases are then integrated within the structure of documents indexing. The author used a measure called "information quantity" to calculate the Wight of the added noun phrases. Experiments results showed that using noun phrases provides better performance compared to the use of simple terms.

Metzler and Croft in [12] developed a general framework to modeling dependency between term through Markov Random Fields (MRF) in a language model. In particular, three variants of MRF: full independence (FI), the sequential dependency (SD), and total dependency (FD), have been proposed. Experiments showed the effectiveness of MRF model on different search tasks. They make use the Associated Press and Wall Street Journal subcollections of TREC, which are small homogeneous collections, and two web collections, WT10g and GOV2, which are considerably larger and less homoge- neous.

In [13], the authors proposed to incorporate the dependency between the terms in the model Divergence From Ramdomness. This model assigns scores [c'est à dire pondération des termes simples de plus ponédration ds paires de termes] to each pairs of query terms, in addition to the single query terms. This approach is shown to be robust as retrieval effectiveness is enhanced on the TREC .GOV2 Terabyte test collection, and its associated TREC 2005 and 2006 adhoc title-only topics.

Representation of dependencies between terms in these models is expressed by using NPs, that can be extracted statistically, linguistically, or by combining the two [14].

We note that the keywords extracted statistically to represent the meaning of the text may contain noise, which may affect negatively the performance of SRI, in the following example

The student will be probably attending a special reading on software engineering on Monday

the statistical methods extract compounds keywords such as " will probably ", " student will ", " be attending ", these keywords have no input in IR and can degrade the performance of IRS. On the other hand linguistic methods based on NPs will simply used compound keywords such as: "special reading", "reading is special software", "software engineering", etc. We note that these keywords are more likely to represent the content or topic of the sentence than those extracted by statistical methods.

We propose to represent the contents of documents and queries using a linguistic method based on the extraction of SNs.

III. Our noun phrases extracting approche

The main objective of extracting keywords from textual corpora is the acquisition of useful knowledge for Information Retrieval System that we use to index the documents content.

Light Natural Language Processing already showed that it can improve matching results by combining the methods of classical Information Retrieval with noun phrases recognition [9], [15], [16]. The aim is to extract noun phrases based on the syntactic dependencies between words.

Our approach is based on two parts:

1) We conduct a linguistic analysis with a tagger, which generates a tagged collection. Each word is associated to a tag corresponding to the syntactic category of the word, example: noun, adjective, preposition...
2) Then, we use the tagged collection to extract a set of noun phrases. Candidate noun phrases are extracted by the identification of syntactic patterns.

We adopt the definition of syntactic patterns in [11], [17], where a pattern is a syntactic rule on the order of concatenation of grammatical categories which form a noun phrase:

- V: the vocabulary extracted from the corpus
- C: a set of lexical categories
- L : the lexicon $\subset V \times C$

A pattern is a syntactic rule of the form::

$$X := Y_1 Y_2 Y_k ... Y_{k+1} Y_n$$

where Yi $\in$ C and X is a noun phrase.
Examples :
Adjectiv Noun: "next month", "white hair", etc.
Noun Prepostion Noun: "picture of animals", "weapon of destruction", etc.

IV. Experimental study:

To evalute the quality of the extracted noun phrases and their ability to index documents, we use the DEFT[1] collection and we compare the keywords proposed by the authors to the keywords we extract. Our work focus on tow and three words keywords.

A. Campaign DEFT 2012:

Started in 2005, the aim of DEFT (Défi fouilles de textes) is to evaluate, using the same corpora, methods and systems of different research teams. DEFT 2012 campaign challenge is to evaluate the ability of research systems to extract keywords and use them to index the content of scientific papers published in journals of Humanities and Social Sciences [18]. So given a scientific paper with a keyword set proposed manually by experts (mainly the papers authors), can a system identify the same keywords set? Experts proposed simple terms and noun phrases as keywords. In our sudy, we focus on noun phrases (NPs). We use four text collections from DEFT 2012.

[1] www.deft2012.limsi.fr

TABLE I
CORPORA STATISTICS

	Corpus1	Corpus2	Corpus3	Corpus4
Size	1.9 Mo	1.9 Mo	1.2 Mo	1.2 Mo
Number of documents	140	140	94	93
Average terms per documents	6417	6350	6392	6114
Keywords per corpus	238	256	167	148
Average keywords per document	1.9	2.24	2.25	2

```
as_2007_015986ar.xml    revitalisation linguistique;jeux de sociétés;noms de lieux;utilisation du territoire
as_2007_018375ar.xml
as_2007_018382ar.xml    savoir ordinaire;savoir scientifique;
```

Fig. 1. Example of reference file

B. Text Collections description

Four corpora are designed to identify keywords. We present in the table 1 statistics about this corpora. We want to see if the NPs we extract correspond to NPs defined by the authors as index of documents. So our job is to extract NPs from documents and compared them with NPs proposed by the authors.
For every corpus there is a reference file that contain all the keywords, the Figure 1 shows an example of a reference file, we can see that there is four NPs for the document 'as_2007_015986ar.xml': revitalisation linguistique, jeux de sociétés, noms de lieux, utilisation du territoire. The Figure 2 shows an example of the document identified 'as_2007_015986ar.xml' of the corpus. NPs defined by the authors are between the tags <mots[2]> and </mots>. for the corpus 1 and corpus 2 that are the learning corpora , the keywords are not defined in the corpus 3 and the corpus 4 that are the test corpora.

C. Noun phrases extraction results

To study the quality of the NPs extracted we defined a precision measure adapted to our problem that calculates the extraction precision.

$$Accuracy = R/P$$

R: Number of relevant noun phrases extracted from a document.
P: Number of complex keywords proposed by the authors for a document.

We compute also the average accuracy for all the documents.

$$AP = \frac{\sum_{i=1}^{n} Pi}{N}$$

Pi: Is the accuracy for the document i.
N: Total number of documents that contain NPs.

Table 2 present statistics of the four corpuses. We note that we extract many NPs relatively to the number of NPs proposed by the specialists. The average of the extracted NPs for the first corpus is 702 per document, but the average number of the proposed NPs is 1.9. In this work we content with the study of the NPs that are composed of two or three terms.

For the corpus 1 the average precision is 87%, knowing that we retrieved 207 NPs by identification of syntactic patterns among 238 NPs proposed in the reference file by the expert. For the corpus 2 the average precision is 84% and we retrieved 215 NPs from 256 proposed.

The NPs that we don't retrieve are not equivalent to NPs, for example "sans papier", other NPs are not retrieved because they match with tagging mistakes like " Pensée du désordre" where "Pensée" is tagged as a verb and the pattern " verb+ preposition+ noun" isn't a NP.

For the corpuses 3 and 4 the results was lower than the results for the corpuses 1 and 2, because the keywords aren't included in the documents like the corpus 1 between the tags <mots> and many of them aren't composed by words of the document that is this NPs don't exist in the documents and are just proposed by the experts. For the corpus 3 we find 57 NPs from 167 proposed, wheras we find 40 NPs from 148 proposed.

V. CONCLUSION

We present in this paper different methods used in information retrieval to capture terms dependencies for representing the meaning of documents, which will improve the performances of the IRS. We introduce our method of extracting dependencies between the terms of the documents, this method is based on the extraction of nouns phrases (NPs) by linguistic analysis with the use of syntactic patterns.

The study of the DEFT corpus and the comparison of the NPs that we have extracted with the NPs proposed by the experts showed that the NPs we proposed are relevant as keywords of the studied documents. Evaluation proved that our method produces NPs with 87% and 84% precision respectively for corpus 1 and corpus 2, that we consider being significantly high.

REFERENCES

[1] S. Boulaknadel. Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe. In CORIA, pages 341-346, (2006)

[2] *mots* is the traduction of the term *words* in french

TABLE II
EXTRACTED NPS STATISTICS

	Corpus1	Corpus2	Corpus3	Corpus4
Average NPs per document	702	701	889	812
Number of retrieved NPs	207	215	57	40
Average precision	87%	84%	34%	27%

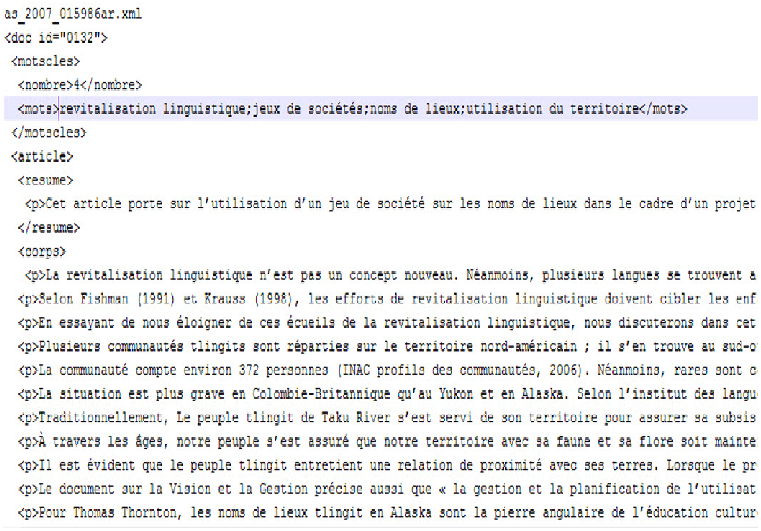

Fig. 2. The document 'as_2007_015986ar.xml' of corpus 1

[2] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In RIAO, pages 200-217, (1997)

[3] W. B. Croft, H. R. Turtle, and D. D. Lewis. The use of phrases and structured queries in information retrieval. In Proc. 14th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 32-45, (1991)

[4] C. J. van Rijsbergen. A theoretical basis for the use of cooccurrence data in information retrieval. Journal of Documentation, 33(2):106-119, (1977)

[5] J. L. Fagan. Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In Proc. tenth Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 91-101, (1987)

[6] J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In Proc. 27th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 170-177, (2004)

[7] Lafferty, J., Sleator, D. and Temperley, D. Grammatical trigrams: A probabilistic model of link grammar. in 'In Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language'. pages. 89-97.(1992)

[8] Pietra, S. D., Pietra, V. J. D., Gillet, J., Lafferty, J. D., Printz, H. and Ures, L.. Inference and estimation of a long-range trigram model. in 'ICGI '94: Proceedings of the Second International Colloquium on Grammatical Inference and Applications'. Springer-Verlag. pages 78Ű92,London, UK.(1994).

[9] Arampatzis, A., van der Weide, T., Koster, C. H. A. and van Bommel, P. . An evaluation of linguistically-motivated indexing schemes. in 'In Proceedings of the 22nd BCS-IRSG Colloquium on IR Research',(2000)

[10] Pickens, J. and Croft, W. B. An exploratory analysis of phrases in text retrieval. in 'In Proceedings of RIAO (Recherche dInformation assiste par Ordinateur'. pages 1179-1195, (2000)

[11] H. Haddad. Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information. Thèse de doctorat, Université Joseph Fourier, (2002)

[12] Metzler, D. and Croft, W. B. A markov random field model for term dependencies. in 'SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval'. ACM. New York, NY, USA. pages 472-479, (2005)

[13] Peng, J., Macdonald, C., He, B., Plachouras, V. and Ounis, I. . Incorporating term dependency in the dfr framework. in 'SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval'. ACM. New York, NY, USA. pages 843-844, (2007).

[14] Bechikh Ali, C. and Haddad, H. Approche hybride d'indexation pour la recherche d'information. EGC-M 2012 : 3ème Edition de la Conférence Internationale sur l'Extraction et la Gestion des Connaissances - Maghreb, pages 120-124.Hammamet, Tunisie, 13-15 Novembre (2012)

[15] Haddad, H. French noun phrase indexing and mining for an information retrieval system. In String Processing and Information Retrieval, 10th International Symposium, pages 277-286, Manaus, Brazil, October (2003)

[16] D. A. Hull, G. Grefenstette, B. M. Schulze, E. Gaussier, H. Schütze, and J. O. Pedersen. Xerox trec-5 site report : Routing, filtering, nlp, and spanish tracks. In Proceedings of the Fifth Text REtrieval Conference (TREC-5), pages 167-180, (1997)

[17] Amri, A., Mbarek, M., Bechikh Ali, C., Latiri, C. and Haddad, H. Indexation à base des syntagmes nominaux. JEP-TALN-RECITAL 2012, Atelier DEFT 2012: DÉfi Fouille de Textes. pages 33-39, Grenoble, France, 4-8 Juin (2012)

[18] Paroubek, P., Zweigenbaum, P., Forest, D. and Grouin, C.. Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012. JEP-TALN-RECITAL 2012, Atelier DEFT 2012: Défi Fouille de Textes. pages 1-13, Grenoble, France, 4-8 Juin (2012)