A multi-word term extraction program for Arabic language

Siham Boulaknadel, Beatrice Daille, Driss Aboutajdine

LINA FRE CNRS 2729 University of Nantes, GSCM_ LRIT University Med V
2 rue la Houssinire BP 92208 44322 Nantes, B.P. 1014 Faculte des Sciences de Rabat
siham.boulaknadel@univ-nantes.fr, beatrice.daille@univ-nantes.fr, aboutaj@fsr.ac.ma

Abstract

Terminology extraction commonly includes two steps: identification of term-like units in the texts, mostly multi-word phrases, and the ranking of the extracted term-like units according to their domain representativity. In this paper, we design a multi-word term extraction program for Arabic language. The linguistic filtering performs a morphosyntactic analysis and takes into account several types of variations. The domain representativity is measure thanks to statistical scores. We evalutate several association measures and show that the results we otained are consitent with those obtained for Romance languages.

1. Introduction

The development of a terminology extraction tool for Arabic language requires linguistic specifications of terms. Indeed, Arabic being an agglutinative language, statistical methods could not be applied straight and should be associated to linguistic treatments.

The aim of this work is to develop a multi-word term (MWT) extraction tool for Arabic. The identification of MWT is crucial for terminology extraction because MWTS are less polysemous and more numerous that SWTs. For intance, (Nakagawa and Mori, 2002) show that more than 85% of domain-specific terms are multi-word terms.

To extract MWTs from corpora, we adopt the standard approach that combined grammatical patterns and statistical score (Cabré et al., 2000). First, we defined the linguistic specification of MWTs for Arabic language. Then, we develop a term extraction program and evaluate several statistical measures in order to filter the extracted term-like units for keeping the most representative of domain specific corpus.

This paper is organised as follow: section 2 focuses on related work. In section 3, we present the linguistic specifications of Arabic multi-word terms and we detail in section 4 our methodology of multiword terms extraction. We presente our results in section 5 and we conclude in section 6.

2. Related work

Approaches to MWT extraction proposed so far can be divided into three categories: a) statistical approaches based on frequency and co-occurrence affinity, b) symbolic approaches using parsers, lexicons and language filters, and c) hybrid approaches combining different methods (Smadja, 1993) (Church and Hanks, 1990) (Daille, 1994) (Sag et al., 2002) (Maynard and Ananiadou, 2000).

In practice, most statistical approaches employ linguistic filters to extract candidate MWTs (Church and Hanks, 1990). One of the main problems confronting statistical approaches, however, is that they are unable to deal with low-frequency of MWTs. In fact, the majority of the words in most corpora have low frequencies, occurring only once or twice. This means that a major part of multiword terms are excluded by statistical approaches. Lexical resources and parsers are used to obtain better coverage of the lexicon in MWT extraction. For example, In their DEFI Project, (Michiels and Dufour, 1998) used dictionaries to identify English and French multiword terms and their translations in the other language. Like pure statistical approaches, purely knowledge based symbolic approaches also face problems. They are language dependent and not flexible enough to cope with complex structures of MWTs. As (Sag et al., 2002) suggest, it is important to find the right balance between symbolic and statistical approaches. Our main interest in this paper is the development of a hybrid MWT tool for identifying and extracting Arabic MWTs from corpora.

3. Linguistic specifications of Arabic multi-word terms

Arabic words are formed with root-pattern schemes. Mutiword terms (MWTs) defined as idiosyncratic interpretations cross word boundaries (or spaces) (Sag et al., 2002). The main property of MWTs is the morphosyntactic one : its structure belongs to well-known morphosyntactic structures such as N ADJ, N1 N2, etc. that have been studied by Roman (1990) for Arabic.

3.1. MWT patterns

They consist of series of local grammar rules to detect MWTs having two content words. (Grammatical function words such as prepositions and determiners are permitted to intervene between the two content words). Some examples met in our corpus are presented in Table 1. For the description of Arabic terms, we applied Buckwalter's transliteration system [1] which transliterates Arabic alphabet into Latin alphabet.

3.2. MWT variation

The module for automatic term acquisition takes into account term variations. We followed the typology suggested by (Daille, 2005).

[1] http://www.qamus.org/transliteration.htm

Pattern	Sub-pattern	Arabic MWT	English translation
N ADJ		**Al**tlwv **Al**kmyAAy	chemical pollution
N1 N2		tlwv **Al**mAA	water pollution
	N1 b N2	Altlwv b Alrsas	pollution with lead
N1 PREP N2	N1 l N2	AltErD l AlAmrAD	exposure to diseases
	N1 mn N2	Altxls mn Alnfayat	waste disposal

Table 1: MWT patterns

3.2.1. Graphical variants

By graphical variants, we mean the graphic alternations between the letters p and h. Table 2 shows some examples of graphic alternations.

Graphic alternation	Arabic MWT	English translation
p/h	tlwv Altrb**p**/tlwv Altrb**h**	ground pollution

Table 2: *Graphical variants*

3.2.2. Inflectional variants

In Arabic inflectional variant is a central issue in language processing as Arabic is an agglutinative language, with a very rich inflectional system. The amount of inflectional forms in which a given lemma or the 'canonical form' of a given term can appear in texts is extensive.

Inflectional variants include the number inflection of nouns, the number and gender inflections of adjectives (Jacquemin, 2001; Nenadic and Spasic, 2002), and the definite article that is carried out by the prefixed morpheme (Al).

Table 3 shows some examples of inflectional variants.

Type	Arabic MWT	Variant	English translation
Number	tlwv AlmA'	tlwv Alm**yAH**	water pollution**(s)**
Definitude	tlwv hwAAy	**Al**tlwv **Al**hwAAy	**(the)** air pollution

Table 3: *Inflectional variants*

3.2.3. Morphosyntactic and syntactic variants

Morphosyntactic variants refer to the synonymy relationship between two MWTs of different structures. The example below shows synonymic terms of N1 ADJ and N1 PREP N2 structures.

The syntactic variants modify the internal structure of the base-term, without affecting the grammatical categories of the main item which remain identical. We distinguish modification and coordination variants. Table 5 shows some examples of syntactic variants.

Structure alternation	Arabic MWT	English translation
N1 ADJ ⇔ N1 PREP N2	b}r nfty ⇔ b}r mn alnft	oil wells

Table 4: *Morphosyntactic variants*

Type	Sub-type	Arabic MWT	Variant
Modification	insertion	Altkwyn l ltrbp composition of the soil	Altkwyn Almstmr l ltrbp permanent composition of the soil
Modification	postposition	drjp AlHrArp degree of temperature	drjp AlHrArp AlEAlyp high degree of temperature
Coordination	expansion	tlwv Altrbp pollution of soil	tlwv AlmyAh **w** Altrbp pollution of soil and water
Coordination	head	AlmkhAtr mn Altlwv Risks of pollution	AlmkhAtr **w** AlwqAyp mn Altlwv Risks and prevention of pollution

Table 5: Syntactic variants

4. Multi-word term extraction

The term extraction process is performed in two major steps: the selection of MWT-like units, using part-of-speech that has been assigned by the diab's tagger (Diab et al., 2004), and the ranking of MWT-like units by means of statistical techniques, log-likelihood ratio (LLR) (Dunning, 1994), FLR (Nakagawa and Mori, 2003), Mutual Information (MI^3) (Kenneth and Hanks, 1989) and t-scores (Church et al., 1991). To filter the MWT-like units, we used their part-of-speech that has been assigned by the diab's tagger (Diab et al., 2004). The MWT-like string patterns are described through morphosyntactic rules presented in Table 6.

Association measures are used in order To rank MWT like strings that have been collecting in the first step, The well-known association measures rely on different concepts. So, we compute several measures: log-likelihood ratio (LLR) (Dunning, 1994), FLR (Nakagawa and Mori, 2003), Mutual Information (MI^3) (Kenneth and Hanks, 1989) and t-scores (Church et al., 1991). Mutual Information (MI^3)

MWT Pattern	Part of speech pattern
N1 N2	NN[P]? | NNs[P]?
N1 ADJ	NN[P]? | NNs[P]? |JJ
N1 PREP N2	NN[P]? | NNs[P]? | IN | NN[P]? | NNs[P]?

Table 6: *Pattern and Part-of-speech Mapping*

(Kenneth and Hanks, 1989) was taken from Information Theory. Other measures such as the log-likelihood ratio (LLR) (Dunning, 1994) and t-score (Church et al., 1991) are based on hypothesis testing. FLR (Nakagawa and Mori, 2003) make the hypothesis that MWTs are often built around a limited number of single words and measures how many distinct words are part of MWTs.

5. Experimentation and Evaluation

5.1. Corpora used

Document retrieval experiments for Arabic language are done on general language corpora gathering newspaper articles are available. As it does not exist specialised domain corpora, we built a specialised corpus: the texts are taken from the environment domain and are extracted from the Web sites "Al-Khat Alakhdar"[2] and "Akhbar Albiae"[3]. The environment domain covers various environmental topics such as pollution, noise effects, water purification, soil degradation, forest preservation, climate change and natural disasters. The corpus contains 1.062 documents, 475.148 words from which 54.705 are distinct. The length of these documents varies between a paragraph and 40 pages.

5.2. Reference list

We create a reference list to automatically annotate the results of the statistical measures in section 5.3.. The reference list is collected from a list of known Arabic terms from the environmental domain such as Agrovoc [4]. In total we compiled a list of 65,000 MWTs.

5.3. Comparing statistical measures

Before proceeding to the statistical analysis, however, one more step is required which is generating the bigrams and the corresponding frequency counts. In generating the bigrams, the program starts first by extracting strings using the pattern shown in the previous section.
We compare four commonly-used approaches (defined in Table 7) to measure the strength of association of bigram word strings. Formulae are defined in terms of the contingency table. In this table, n11 is the frequency of the bigram xy, n12 is the frequency of x followed by any word other than y, and n1p is the total frequency all bigrams with x as the first word. m11 is the expected value of the bigram xy ($m11 = \frac{np1n1p}{npp}$). And for FLR score, XY = N1 N2N L, where Ni (i= 1,.., L) is a simple word. Then a geometric mean: LR of XY is defined as follows.

$$LR(CN) = \prod_{i=1}^{L}((LN(N_i)+1)(RN(N_i)+1))^{\frac{1}{2L}} \quad (1)$$

Where

$$LN(N) = \sum_{i=1}^{\#LDN(N)} \#L_i \quad (2)$$

$$RN(N) = \sum_{j=1}^{\#RDN(N)} \#R_j \quad (3)$$

LDN(N) and # RDN(N) : These are the number of distinct simple words which directly precede or succeed N. LN(N) and RN(N) are the frequencies of nouns that directly precede or succeed N.

Method	**Formula**
LLR (Dunning, 1994)	$2(n11 \log \frac{n11}{m11} + n12 \log \frac{n12}{m12} + n21 \log \frac{n21}{m21} + n22 \log \frac{n22}{m22})$
T-score (Church et al., 1991)	$\frac{n11 - \frac{n1p}{np1npp}}{n11^2}$
FLR (Nakagawa and Mori, 2003)	$FLR(CN) = LR(CN) * f(CN)$
Mutual Information (MI^3) (Daille, 1994)	$\log_2 \frac{n11^3}{n12n21}$

Table 7: Statistical algorithms used to measure the association strength of a word pair xy.

5.4. Selecting the best measure

We evaluate the statistical algorithms against the environmental corpus (section 5.1.). We compute the association scores of the candidate multiword terms . We take from each produced ranking a set of 100 true terms, that match with our reference list, and calculate the precision as shown in equation :

$$precision = \frac{attested\ \ multiword\ \ terms}{all\ \ extracted\ \ sequences} \quad (4)$$

We attested that a term is relevant to the environment domain if it has already been listed in existing terminology database[5]. We note that the LLR, FLR and t-score measures, that are based on the significance of association measure, outperform the MI^3 measure. Note that LLR outperform other methods. These results are consistent with those reported in (Daille, 1994) (Hong et al., 2001) The results are shown in Table 8.

6. Conclusion

In this paper we presented our approach for the extraction of term candidates from Arabic technical texts. We have applied an a hybrid approach for the extraction of MWT in Arabic for environment domain, combining the detection of term candidates through linguistic techniques with

[2]http://www.greenline.com.kw
[3]http://www.4eco.com
[4]www.fao.org/agrovoc/
[5]www.fao.org/agrovoc/

Type	P(%)
FLR	60%
T-score	57%
LLR	85%
MI^3	26%

Table 8: *Precision, recall*

the subsequent ranking of candidates according to different statistical measures. We defined the linguistic specification of MWTs for Arabic language. Results obtained for Arabic are similar to that of Romance languages.

Further works have still to be made to evaluate this approach in differents domains and applications such as information retrieval or information extraction. Since MWTs have been useful for various applications of terminology processing (Ibekwe-SanJuan and Condamines, 2007) (Marcelline et al., 2003) and in IR (Haddad, 2002) (Ahlgren and Keklinen, 2006), we believe that our Arabic term extraction program will fill a gap in Arabic specialized language processing

7. References

P. Ahlgren and J. Keklinen. 2006. Swedish full text retrieval: Effectiveness of different combinations of indexing strategies with query terms. *Information Retrieval*, 9(6):681–697.

M.T. Cabré, R. ESTOP, and J. VIVALDI. 2000. Automatic term detection: a review of current systems. *Recent Advances in Computational Terminology*, 2(1):53–88.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C. Association for Computational Linguistics.

K. Church, W. Gale, P. Hanks, and D. Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. U. Zernik.

B. Daille. 1994. *Approche mixte pour l'extraction de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Universit de Paris 7, France.

B. Daille. 2005. Variations and application-oriented terminology engineering. *International journal of theoretical and applied issues in specialized communication*, 11(1):181–197.

M. Diab, K. Hacioglu, and D. Jurafsky. 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. In *In Proceedings of NAACL-HLT*, pages 149–152, Boston, USA.

T. Dunning. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

H. Haddad. 2002. *Extraction et impact des connaissances sur les performances des systmes de recherche d'information*. Ph.D. thesis, Universit Joseph Fourier, Grenoble, France.

M. Hong, S. Fissaha, and J. Haller. 2001. Hybrid filtering for extraction of term candidates from german technical texts. In *TIA 2001 : terminologie et intelligence artificielle*, pages 223–232.

F. Ibekwe-SanJuan and M. T. Condamines, A.and Cabr Castellv. 2007. Application-driven terminology engineering. *special issue of the Terminology Journal*, 2:1–17.

C. Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing Techniques*. MIT Press, Cambridge.

W. C Kenneth and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C.

R.H. Marcelline, K.S. Guergana, M.J. Thomas, and Christopher G.C. 2003. A term extraction tool for expanding content in the domain of functioning, disability, and health: proof of concept. *Journal of Biomedical Informatics*, 36(4/5):250–259.

G. Maynard and S. Ananiadou. 2000. Identifying terms by their family and friends. In *Proceedings of the 18th conference on Computational linguistics*, pages 530 – 536, Saarbrcken, Germany.

A. Michiels and N. Dufour. 1998. Defi, a tool for automatic multi-word unit rec-ognition, meaning assignment and translation selection. In *In Proceedings of the First Interna-tional Conference on Language Resources and Evaluation*, pages 1179–1186, Granada, Spain.

H. Nakagawa and T. Mori. 2002. Nested collocation and compound noun for term recognition. In *In Proceedings of the First Workshop on Computational Terminology COMPTERM'98,*, pages 64–70.

H. Nakagawa and T. Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

G. Nenadic and S. Spasic, I.and Ananiadou. 2002. Automatic acronym acquisition and term variation management within domain-specific texts. In *In: Proceedings of 3rd International Conference on Language, Resources and Evaluation, LREC-3*, pages 2155–2162, Las Palmas, Spain.

A. Roman. 1990. *La Grammaire de l'arabe*. PUF, Paris.

I.A. Sag, F. Baldwin, T.and Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *CICLing*, pages 1–15.

F. Smadja. 1993. Xtract : An overview. *Computers and the Humanities*, 26:399–413.