A Study of Association Measures and their Combination for Arabic MWT Extraction

Abdelkader El Mahdaouy
LIM, Univ. USMBA
Fès, Maroc
Univ. Grenoble Alpes
CNRS LIG/AMA
Grenoble, France
a.mahdaouy@hotmail.fr

Saïd EL Alaoui Ouatik
LIM, Univ. USMBA
Fès, Maroc
s_ouatik@yahoo.com

Eric Gaussier
Univ. Grenoble Alpes
CNRS LIG/AMA
Grenoble, France
eric.gaussier@imag.fr

Abstract

Automatic Multi-Word Term (MWT) extraction is a very important issue to many applications, such as information retrieval, question answering, and text categorization. Although many methods have been used for MWT extraction in English and other European languages, few studies have been applied to Arabic. In this paper, we propose a novel, hybrid method which combines linguistic and statistical approaches for Arabic Multi-Word Term extraction. The main contribution of our method is to consider contextual information and both termhood and unithood for association measures at the statistical filtering step. In addition, our technique takes into account the problem of MWT variation in the linguistic filtering step. The performance of the proposed statistical measure (NLC-value) is evaluated using an Arabic environment corpus by comparing it with some existing competitors. Experimental results show that our NLC-value measure outperforms the other ones in term of precision for both bi-grams and tri-grams.

1 Introduction

Automatic Multi-Word Term extraction is an important task in many Natural Language Processing (NLP) applications (Boulaknadel et al., 2008b; Wen et al., 2007). The aim of the MWT acquisition process is to extract specific domain terms from special language corpora (Korkontzelos et al., 2008). The extraction of MWTs is crucial for terminology acquisition, since they are less ambiguous and less polysemous than single word terms, and since their internal structure encodes useful semantic relations (Wen et al., 2008).

There are three main approaches to MWT extraction. The first one makes use of linguistic filters. The second one relies on statistical measures based on termhood and/or unithood. Termhood denotes *"the degree to which a linguistic unit is related to a specific domain concept"*, and unithood denotes *"the degree of strength or stability of syntagmatic combinations or collocations"* (Kageura et al., 1996). Lastly, the third approach is hybrid and combines the linguistic and the statistical approaches. Hybrid methods extract MWTs using linguistic filters and then rank the list of candidate MWTs according to statistical measures.

In this paper, we propose a novel, hybrid method for Arabic MWT extraction. Like other hybrid methods, it includes two main filters. In the first one, we use a part-of-speech (POS) tagger to extract candidate MWTs based on syntactic patterns. In the second one, we propose a novel statistical measure, the NLC-value, that unifies the contextual information and both termhood and unithood measures. We compare this measure to alternative ones in the task of MWT extraction : NTC-value (Vu et al., 2008), LLR+C-value (Al Khatib et al., 2010), C/NC-value and LLR.

The remainder of this paper is organized as follows. In the next section, Section 2, we present the related work. Section 3 describes the proposed method to extract MWTs. In Section 4, we present how MWT variation is handled in the proposed method. Section 5 describes the experimental validation and Section 6 concludes this work and presents some perspectives.

2 Related Work

Several studies have been conducted on MWT extraction for many languages. These studies have either used a linguistic approach, a statistical approach, or a combination of them (hybrid approach). Most recent MWT extraction methods rely on a hybrid approach to efficiently extract MWTs, due to its higher accuracy compared to the two other approaches (Tadic et al., 2003). The linguistic approach uses technical analysis on the current knowledge of the language and its structure. There are two subcategories : approaches based on morpho-syntactic patterns (Daille, 1994) and those based on MWT boundary detection (Bourigault, 1994).

The main purpose of applying statistical methods for MWT extraction is to rank candidate terms based on a particular measure that gives higher scores to "good" candidate terms. Candidate terms above a particular threshold are selected for further processing. The reliance on frequency is based on the simple assumption that a frequent expression indicates an important representation of the domain in question. Therefore, frequent expressions are assumed to represent important concepts. Given a candidate multi-word term, frequency only counts how often the candidate occurs in the text, but doesn't give any information on the strength of the relationship between words composing the candidate multi-word term. Statistical approaches aim at extracting candidate terms from text corpora by means of association measures (Church et al., 1991) that concentrate on termhood and/or unithood to assign a score to candidate MWTs. These measures are based on frequency and co-occurrence information such as the T-score (Church et al., 1991), the loglikelihood ratio (LLR) (Dunning, 1994), the C/NC-Value (Frantzi et al., 1998), etc.

While linguistic approaches focus on syntactic structures, statistical methods focus on the recurrent characteristics of MWTs. Both have their advantages and limitations. As mentioned by Boulaknadel et al. (2008), statistical approaches *"are unable to deal with low-frequency MWTs"* while pure linguistic approaches are *"language dependent and not flexible enough to cope with complex structures of MWTs"* . Hybrid methods try to combine linguistic and statistical techniques to extract MWTs in order to avoid the weaknesses of the two approaches.

Boulaknadel et al. (2008) have relied on a hybrid method to extract Arabic MWTs. As a first step, candidate terms that fit syntactic patterns are extracted from the output of the part-of-speech (POS) tagging tool proposed by Diab (2004). In the second step, the list of candidate terms is ranked according to one of the following association measures : log-likelihood ratio (LLR), Mutual Information (MI), FLR, and T-score. These measures have been evaluated on an Arabic corpus and he results obtained show that LLR outperforms the other association measures.

Bounhass et al.(2009) have followed the same approach (using again Diab's (2004) POS tagger and LLR) while focusing on compound nouns and thus using a more restricted set of syntactic patterns. For the bigrams, the obtained results outperform those obtained by Boulaknadel et al. (2008).

A similar study has been conducted by Al Khatib et al. (2010), based on the POS tagger proposed by (Al-Taani et al. , 2009) and an association measure that combines both termhood and unithood through a combination of the C-value and the LLR. Experimental results show promising results for the combined measure.

Most hybrid methods presented previously have been evaluated on 100 (best) candidate MWTs and deal with bi-grams (i.e. candidate MWTs of length 2). Moreover, they rely on LRR or a combination of LRR and C-value (Al Khatib et al., 2010) and ignore contextual information in the ranking step. To overcome this limitation, we introduce a new association measure that integrates contextual information and both termhood and unithood. Our overall approach is also hybrid and relies on the same linguistic filters as the ones used in the previous studies, based on syntactic patterns applied on the output of the POS tagger developed by (Diab, 2009).

3 Proposed Method

Our method for extracting MWT candidates comprises two major steps : the linguistic and the statistical filters.

3.1 Linguistic Filter

The proposed linguistic filter extracts candidate MWTs based on two core components ; the POS tagger and the sequence identifier. In the literature, several methods for Arabic POS tagging systems have ben developed. We have used the one proposed by (Diab, 2009) as it performs at over 96% accuracy and allows a number of variable user settings. The underlying system uses Support Vector Machine (SVM). Figure (1) illustrates the global schema of our linguistic filter. As a first step, our

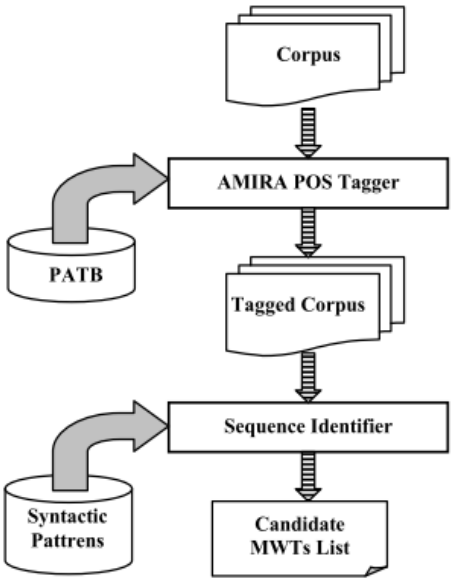

FIGURE 1: The global schema of the linguistic filter

method tags the corpus using the AMIRA toolkit (Diab POS Tagger) which is trained from the Penn Arabic TreeBank (PATB) to assign tags for each word in the corpus. Then, the sequence identifier tokenizes tagged files of the corpus and uses syntactic patterns in order to identify candidate terms that fit the rules of the grammar. We have extended the list of syntactic patterns used by Boulaknadel et al. (2008) as follows :

- $(Noun + (Noun|ADJ) + |(Noun|ADJ) + |(Noun|ADJ))$
- $Noun\ Prep\ Noun$

The second major step of the linguistic filter is handling the problem of MWTs variation to improve the effectiveness of extracted MWT candidates. Several categories of term variation are taken into account by this filter : graphical, inflectional, morpho-syntactic and syntactic variants, and are discussed in Section 4.

3.2 Statistical Filter

In this step, we apply a number of statistical measures to rank the list of candidate MWTs extracted by the linguistic filter. The main objective of our statistical filter is to consider both termhood and unithood measures.

3.2.1 The C-value

The C-value measures the termhood of a candidate string on the basis of several characteristics : number of occurrences, term nesting, and term length. It is defined as :

$$C\text{-Value}(a) = \begin{cases} \log_2(|a|) \cdot f(a) & \text{if } a \text{ is not nested,} \\ \log_2(|a|) \cdot (f(a) - g(a)) & \text{otherwise} \end{cases} \quad (1)$$

where $|a|$ denotes the length in words of candidate term a, $f(a)$ is the number of occurrences of a and :

$$g(a) = \frac{1}{|T_a|} \sum_{b \in T_a} f(b)$$

where $T(a)$ denotes the set of longer candidate terms into which a appears ($|T(a)|$ is the cardinality of this set).

As one can note, if the candidate term is not nested, its score is solely based on its number of occurrences and length. If it is nested, then its number of occurrences is corrected by the number of occurrences of the terms into which it appears.

3.2.2 The NC-value

The NC-value combines the contextual information of a term together with the C-Value. The contextual information is calculated based on the Nvalue which provides a measure of the terminological status of the context of a given candidate term. It is defined as :

$$N\text{value (a)} = \sum_{b \in C_a} f_a(b) \cdot \frac{|T(b)|}{n} \quad (2)$$

where C_a denotes the set of distinct context words of a, $f_a(b)$ corresponds to the number of times b occurs in the context of a and n is the total number of terms considered. This measure is then simply combined with the C-value to provide the overall NC-value measure :

$$NC\text{-value(a)} = 0.8 \cdot C\text{-value(a)} + 0.2 \cdot N\text{value(a)} \quad (3)$$

3.2.3 The NTC-value

The aim of the NTC-value (Vu et al., 2008) is to incorporate a unithood feature, through the T-score, to the C/NC-value to improve its performance. The T-score measures the adhesion or differences between two words in a corpus of N words as follows :

$$Ts(w_i, w_j) = \frac{p(w_i, w_j) - p(w_i) \cdot p(w_j)}{\sqrt{\frac{p(w_i, w_j)}{N}}} \quad (4)$$

where $p(w_i, w_j)$ corresponds to the probability of observing the bi-gram w_i, w_j in the corpus ; $p(w_i)$ is the probability of word w_i in the corpus and corresponds to the marginal probability of $p(w_i, w)$. The T-score is integrated in the C/NC measures through a re-weighting of the number of occurrences that privileges terms with a positive T-score :

$$F(a) = \begin{cases} f(a) & \text{if } \min(Ts(a)) \leq 0 \\ f(a) \ln(2 + \min(Ts(a)) & \text{otherwise} \end{cases} \quad (5)$$

where $\min(Ts(a))$ corresponds to the minimum T-score obtained from all the word pairs in a. Substituting $F(a)$ to $f(a)$ in Equation 1 yields the TC-value, which is then combined with the Nvalue as before, leading to the NTC-value :

$$NTC\text{-value(a)} = 0.8 \cdot TC\text{value(a)} + 0.2 \cdot N\text{value(a)} \quad (6)$$

The resulting metric (6) thus takes into account both contextual information and termhood and unithood measures.

3.2.4 The NLC-value

We follow here the same development as before but rely this time on the more accurate unithood feature LLR (Dunning, 1994), instead of the T-score, for the combination with the C/NC-value (Frantzi et al., 1998). LLR is a "goodness of fit" statistics that determines if the words in an observed n-gram come from a sample that is independently distributed (meaning they co-occur by chance) or not. The underlying measure is calculated for bi-grams by the following formula :

$$\begin{aligned} LLR(w_j, w_j) &= a \log(a) + b \log(b) + c \log(c) \\ &+ d \log(d) - (a + b) \log(a + b) \\ &- (a + c) \log(a + c) - (b + d) \log(b + d) \\ &- (c + d) \log(c + d) + N \log(N) \end{aligned}$$

with :
a : number of terms in which w_i and w_j co-occur ;
b : number of terms in which only w_i occurs ;
c : number of terms in which only w_j occurs ;
d : number of terms in which neither w_i nor w_j appear ;
N : total number of extracted terms.

For terms that consist of more than two terms, we calculate the LLR for each big-ram and then consider the minimum value obtained. The number of occurrences of a term is now re-weighted by this minimum value : $FL(a) = f(a) \cdot \ln(2 + \min(LLR(a)))$ which is used instead of $f(a)$ in the C-value, leading to the LC-value :

$$L\text{C-value(a)} = \begin{cases} \log_2(|a|) \cdot FL(a) & \text{if } a \text{ is not nested,} \\ \log_2(|a|) \cdot (FL(a) - GL(a)) & \text{else} \end{cases} \quad (7)$$

with $GL(a) = \frac{1}{|T_a|} \sum_{b \in T_a} FL(b)$

This measure is then combined with the Nvalue as before, leading to the NLC-value that integrates contextual information and both termhood and unithood :

$$NLC\text{-value(a)} = 0.8 \cdot LC\text{-value(a)} + 0.2 \cdot N\text{value(a)} \quad (8)$$

4 Term variation

As mentioned in the previous section, we have handled the problem of term variation at the linguistic step. Our method takes into account four types of variations : graphical variants, inflectional variants, morpho-syntactic variants and syntactic variants. Graphical variants concern orthographic errors occurred in writing a particular letters ("أ", "ي" and "ة") which are very common in Arabic. Furthermore, some letters go through a slight modification in writing, that doesn't necessarily change the meaning of the word. For example, the letter "ي" is replaced by another letter "ى" at the end of a MWT, as for "التلوث الكيميائي" which leads to "التلوث الكيميائى" meaning "chemical pollution". Inflectional variants are due to the use of different forms for the words constituting a MWT ; these different forms are related to gender and number of adjectives, as in "تلوث المحيط" (ocean pollution) and "تلوث المحيطات" (pollution of the oceans) and to the presence/absence

of a definite article, as in "تلوث مياه" (water pollution) and "تلوث المياه" (the water pollution). Morpho-syntactic variants affect the internal structure of term as the words it contains are related through derivational morphology. Two patterns control this type of variation in Arabic MWTs :

- $Noun1\,Noun2 \Leftrightarrow Noun1\,Adj$: "تلوث الهواء and "التلوث الهوائي" ("air pollution").
- $Noun1\,Adj \Leftrightarrow Noun1\,Prep\,Noun$: "برميل نفطي" and "برميل من النفط" ("barrel of oil").

We treat these three types of variations by using normalization method and the light stemming algorithm described in (Larkey et al., 2007) on each word of each MWT candidate.

Syntactic variants modify the internal structure of the MWT candidate by adding one or more words (as adjectives) but do not affect the grammatical categories of the content words of the original MWT candidate. Such variants can be identified, for a given MWT candidate, by searching for all the stemmed MWT candidates that contain it. All the elements that constitute an addition to the original MWT candidate are then considered as context terms.

5 Experiments and Results

5.1 The Corpus

Since there is no standard domain-specific Arabic corpus, we have built, in order to evaluate our approach, a new corpus specialized on the environmental domain with similar properties as the ones described (Boulaknadel et al. , 2008; Bounhas et al., 2009; Al Khatib et al., 2010).

The corpus built contain 1666 files comprising 53569 different tokens (without stop words) extracted from the Web site "Al-Khat Alakhdar" [1]. It covers various environmental topics such as pollution, noise effects, water purification, soil degradation, forest preservation, climate change and natural disasters.

5.2 Evaluation and Results

The evaluation of automatic MWTs extraction is a complex process and is usually performed by comparing each MWT candidate extracted to a domain-specific reference list. When there is no reference list available in the language retained, one can first translate the MWT candidates (using a machine translation system or a bilingual dictionary) and use a reference list available in another language. For the evaluation purpose, we have constituted automatically a reference list of all Arabic MWTs available in the latest version of AGROVOC [2] thesaurus and then use the stemming algorithm to remove prefixes and suffixes for each MWT in the reference list and the extracted MWT list. The next step consists of using an algorithm that considers a MWT candidate as correct if it is included in this list, noting that the MWT candidate and the term in the reference list should have the same number of stemmed words. Otherwise, we translate it and consider it as relevant whether its translation is contained in the European terminological database IATE [3]. Finally, the precision is calculated using the number of attested MWTs and the number of considred terms.

We computed the association scores (LLR, C-value, NC-value, NTC-value, LLR+C-value, NLC-value) for the MWT candidates and retain from each produced ranking for each statistical measure the k-best candidates, with k ranging from $100 - 300$ at intervals of 100. The experimental results illustrated in table 1 show that our method (NLC-value) outperforms the previous methods in term of the quality of the extracted MWTs.

	Top MWT considred		
Stat. measures	**100**	**200**	**300**
LLR	75,0%	70,5%	64,3%
C-value	71,0%	69,0%	67,3%
NC-value	74,0%	70,0%	68,3%
NTC-value	80,0%	71,5%	69,7%
LLR+C-value	73,0%	72,0%	68,3%
NLC-Value	82,0%	75,5%	73,0%

TABLE 1: Results obtained for different statistical measures

Furthermore, the combination of the context information and the C-value improves the performance of the process of MWT extraction because the NC-value outperforms the C-value for each

1. http ://www.greenline.com.kw

2. www.fao.org/agrovoc/

3. http ://iate.europa.eu/iatediff

considered MWT list. The unithood feature LLR outperforms the C/NC-value as expected from previous studies. Figure 2 illustrates the precision obtained for the C/NC-value and the LLR.

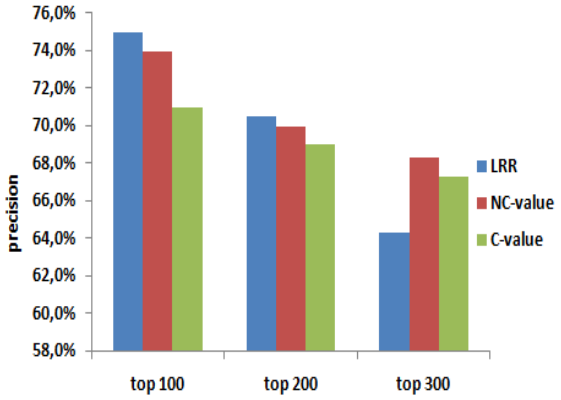

FIGURE 2: Precision obtained for the LLR and the C/NC-value

The integration of contextual information and the T-score unithood measure to the C-value improves the performance of MWT acquisition, since the NTC-value has better precision than the C/NC-value, as illustrated in Figure 3.

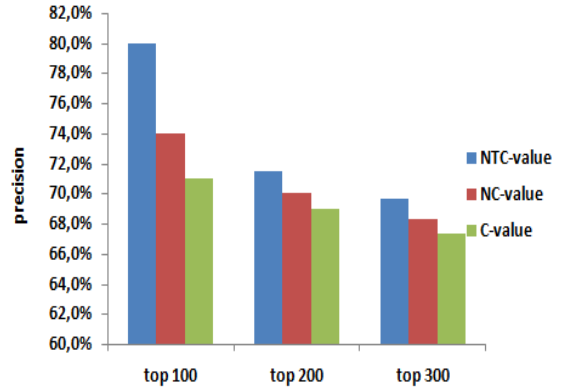

FIGURE 3: Precision obtained for the C/NC-value and the NTC-value

Lastly, the combination of termhood and unithood measures (NTC-value, $LLR + C$-value, NLC-value) is essential for MWT extraction, since all the measures based on this combination perform better than measures using only termhood or unithood (C-value, NC-value, LLR). We note that the statistical measure we have propose, NLC-value, outperforms all other measures. This measure is based on the accurate unithood feature LLR, combined with the NC-value. The NLC-value method takes advantages from previous works proposed in (Vu et al., 2008) and (Al Khatib et al., 2010) taken into account contextual information and both termhood and unithood association measures. Figure 4 presents a comparaison of the precision for different statistical measures that combine termhood and unithood.

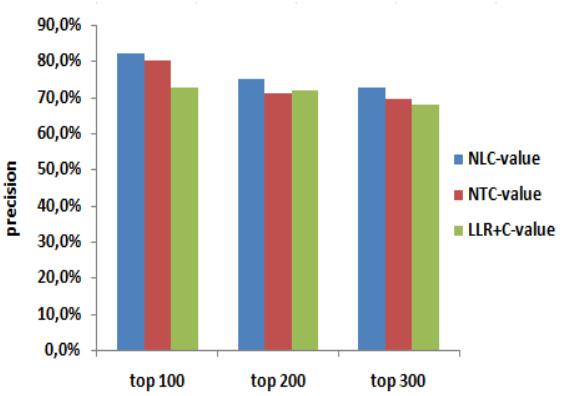

FIGURE 4: Precision obtained for different statistical measures that combine termhood and unithood

The number of different terms evaluated are 1095 amongst other 1800 terms, moreover the statical measures share 141 terms. The tables 2 and 3 represent the number of terms found in agrovoc and IATE respectively.

	Top MWT considred		
Stat. measures	**100**	**200**	**300**
LLR	35	60	80
C-value	27	59	82
NC-value	32	62	82
NTC-value	35	60	83
LLR+C-value	34	60	84
NLC-Value	41	65	86

TABLE 2: the number of terms found in agrovoc foreach measure

6 Conclusion

In this work, we have presented a hybrid method for Arabic MWT acquisition ; this method takes advantage of existing linguistic and statistical approaches. As a first step, we apply linguistic filters to extract MWT candidates based on syntactic patterns using a sequence identifier component. Then, MWT variants are identified through a morphological analysis of he extracted MWTs based on light stemming. In the statistical step, we

	Top MWT considred		
Stat. measures	**100**	**200**	**300**
LLR	40	81	113
C-value	44	79	120
NC-value	42	78	123
NTC-value	45	83	126
LLR+C-value	39	84	121
NLC-Value	41	86	133

TABLE 3: the number of terms found in IATE foreach measure

have proposed a novel statistical measure, NLC-value, that consists of ranking MWT candidates by considering contextual information and both termhood and unithood statistical measures.

Experiments are performed for bi-grams and tri-grams on an environment Arabic corpus. The experimental results show that our method outperforms the previous ones in term of quality of the extracted MWTs. In conclusion, the combination of the best association measures that integrate contextual information and both termhood and unithood statistical measures improves the performance of the MWT acquisition process.

In a near future, we plan on using the extracted MWTs in an information retrieval system as complex terms often constitute a better representation of the content of a document than single word terms.

References

Al Khatib K, and Badarneh A. 2010. *Automatic extraction of arabic multi-word terms*. In Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 411-418.

Al-Taani A, and Abu-Al-Rub S. 2009. *A rule-based approach for tagging non-vocalized Arabic words*, volume 6. The International Arab Journal of Information Technology, p 320.

Boulaknadel S, Daille B, and Aboutajdine D. 2008 a. *Multi-word term indexing for Arabic document retrieval*. In Proceedings of the The IEEE symposium on Computers and Communications, pp. 869-873.

Boulaknadel S, Daille B, and Aboutajdine D. 2008 b. *A multi-word term extraction program for Arabic language*. the 6th international Conference on Language Resources and Evaluation LREC.

Bounhas I, and Slimani Y. 2009. *A hybrid approach for Arabic multi-word term extinction*. International Conference on Language Processing and Knowledge Engineering, pp. 1-8.

Bourigault D. 1994. *LEXTER, un logiciel d'EXtraction de TERminologie, Application à l'acquisition des connaissances à partir de textes*, phd thesis. Ecole des Hautes études en Sciences Sociales, Paris.

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

Daille B. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, Phd thesis. University of Paris 7.

Diab M Hacioglu K, and Jurafski D. 2004. *Automatic tagging of Arabic text : From raw text to base phrase chunks*. In Proceedings of North American Association for Computational Linguistics NAACL, pp. 149-152.

Diab M. 2009. *Second Generation Tools (AMIRA 2.0) : Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking*. In Proceedings of International Conference on Arabic Language Resources and Tools.

Dunning T. 1994. *Accurate Methods for the Statistics of Surprise and Coincidence*, volume 19. Computational Linguistics, pp. 61-74.

Frantzi K. T, Ananiadou S, and Tsujii T. 1998. *The C-Value/NC-Value Method of Automatic Recognition for Multi-word terms*. Journal on Research and Advanced Technology for Digital Libraries, pp. 115-130.

Kageura K, and Umino B. 1996. *Methods of Automatic Term Recognition A Review*, volume 3. Terminology.

Korkontzelos I, Ioannis P. Klapaftis, and Manandhar S. 2008. *Reviewing and Evaluating Automatic Term Recognition Techniques*. In Procedings of the 6th international Conference on Advances in Natural Language Processing, pp. 248-259.

Larkey S. Leah , Ballesteros L, and Connell E. Margaret. 2007. *Light Stemming for Arabic Information Retrieval*, volume 38. Text, Speech and Language Technology, pp. 221-243.

Tadic M, and Sojat K. 2008. *Finding multiword term candidates in Croatian*. In the Proceedings of IESL2003 Workshop, pp. 102-107.

Vu T, Aw A. Ti, and Zhang M. 2008. *Term Extraction Through Unithood And Termhood Unification*. In Procedings of IJCNLP.

Wen Z, Yoshida T, and Xijin T. 2007. *Text classification using multi-word features*. In Proceedings of the The IEEE symposium on Computers and Communications, pp. 3519-3524.

Wen Z, Yoshida T, and Xijin T. 2008. *A Study on Multi-word Extraction form Chinese Documents*. Advanced Web and Network Technologies, pp. 42-53.