Italian-Arabic domain terminology extraction from parallel corpora

Fathi Fawi **Rodolfo Delmonte**
Department of Linguistic Studies and Comparative Cultures
Ca' Bembo 1075 – Università Ca' Foscari – 30123 Venezia
Email: fathi_fawi@yahoo.com Email: delmonte@unive.it

Abstract

English. In this paper we present our approach to extract multi-word terms (MWTs) from an Italian-Arabic parallel corpus of legal texts. Our approach is a hybrid model which combines linguistic and statistical knowledge. The linguistic approach includes Part Of Speech (POS) tagging of the corpus texts in the two languages in order to formulate syntactic patterns to identify candidate terms. After that, the candidate terms will be ranked by statistical association measures which here represent the statistical knowledge. After the creation of two MWTs lists, one for each language, the parallel corpus will be used to validate and identify translation equivalents.

Italiano. *In questo lavoro presentiamo il nostro approccio all'estrazione di termini composti da un corpus giuridico parallelo italiano-arabo. In una prima fase vengono estratti termini composti dai corpora monolingui tramite un approccio ibrido che combina le annotazioni linguistiche fornite dal POS tagging con le informazioni statistiche offerte dalle misure di associazione lessicale. In una seconda fase viene utilizzato il corpus parallelo per estrarre equivalenti di traduzione.*

1 Introduction

The development of robust approaches aiming at terminology extraction from corpora plays a key role in a lot of applications related to NLP, such as information retrieval, ontology construction, machine translation, etc. The main approaches adopted to terms extractions are linguistic-based, statistical-based, and hybrid-based. While the linguistic approach tries to identify terms by capturing their syntactic properties, called *synaptic compositions* (Pazienza et al., 2005), the statistical one uses different association measures (Church et al., 1989) to determine the degree of association or cohesiveness between the multi-word terms (MWTs) components. There is no doubt that the use of a hybrid approach, which combines linguistic and statistical information to identify candidate terms, can guarantee best results rather than relying basically on one approach (Frantzi et al., 1999).

In this paper we present our approach to extract MWTs from an Italian-Arabic parallel corpus of legal texts. The rest of this paper is organized as follows: in Section 2 we present related works; Section 3 describes our proposed approach to extract MWTs from parallel corpora; Section 4 presents the experiments and the results; and Section 5 explains the Conclusion and future works.

2 Related works

There are a lot of efforts that have been done to extract MWTs from monolingual corpora both in Italian (Bonin et al., 2010, Basili et al., 2001) and Arabic (El Mahdaouy et al.,2013, Al Khatib et al., 2010, Abed et al, 2013). The literature of terms extraction from parallel corpora reveals a high dependence on the heuristic methods which calculate the translation probability of terms in the source and target languages. NATools (Simões et al., 2003) uses co-occurrences count of terms in the parallel corpus for building a sparse matrix which will be processed to create a probabilistic translation dictionary for the words of the corpus.

Regarding the domain terminology extraction from parallel texts including Arabic, we can find only rare works, and this may be because of two reasons: a) Arabic is one of those languages which lack specialized parallel corpora in electronic format; b) Arabic is a complex language and its morphosyntactic features affect the overall performance of NLP tasks, especially

the bitext word alignment. In (Lahbib et al. 2014) an approach to extract Arabic-English domain terminology from aligned corpora was presented. The approach consists of the following steps: 1) morphological analysis and disambiguation of the corpus words; 2) extraction of relevant Arabic terms using POS to filter some words, and TF-IDF (Term Frequency- Inverse Document Frequency) to measure the relevance toward one domain; 3) alignment of the texts at the word level, using GIZA++; 4) translations extraction, based on a translation matrix generated from the alignment process, which consists of extracting, for each Arabic word in the corpus, the most likely corresponding translation. To evaluate the approach, a vocalized version of hadith corpus[1] has been used, gaving accuracy rates close to 90%. Here we can note some observations: firstly the approach relies on a probabilistic tool to align the texts at word level. This does not give good results with languages like Arabic which has its own syntactic and morphological features. Secondly, the corpus of evaluation is an Islamic corpus which contains a lot of Islamic terminologies which do not have a translation in other languages, but just transliteration.

Regarding the domain terminology extraction from parallel corpora including the Italian language, we can mention the CLE project (Streiter et al., 2004), where a trilingual corpus with legal texts in Ladin, German and Italian has been created. CLE is stored in a relational database and is accessible via the Internet through BISTRO[2], the Juridical Terminology Information System of Bolzano. Furthermore, there is the LexALP project (Lyding et al., 2006), where sophisticated tools have been developed for the collection, description and harmonization of the legal terminology of spatial planning and sustainable development in four languages, namely French, German, Italian and Slovene.

3 The proposed approach

In this paper we propose a corpus-based approach to extract MWTs from bilingual corpora. It is a hybrid approach which combines statistical methods with linguistic knowledge. Providing the presence of a parallel corpus, the approach consists of the following phases:
1. using POS tagging to create candidate terms in each language;
2. applying statistical methods to rank the candidate terms in order to create a terminology list in each language;
3. using the parallel corpus for identifying translation equivalents of MWTs.

3.1 Morphological analysis

In this phase all the texts of the corpus are tagged at the POS level. The tagging task is done at monolingual level, given its dependency on the language. Regarding the Arabic texts we used the Amira tagger (Diab, 2009), which is based on a supervised learning approach. Amira system uses Support Vector Machine (SVM) for the processing of Modern Standard Arabic texts. In our case the POS tagging accuracy is close to 94%.

Regarding the Italian texts we used the VEST tagger (Delmonte, 2007). Vest is a symbolic rule tagger that uses little quantitative and statistical information. It is based on tagged lexical information and uses a morphological analyzer for derivational nouns, cliticized verbs and some adjectives. Vest has achieved around 95,7% of accuracy.

3.2 Create candidate terms

In this step we use the POS tagging and sequence identifier to form syntactic patterns in order to extract monolingual candidate terms which fit the rules of the grammar. For Arabic, we used the patterns proposed by El Mahdaouy et al.(2013):
–(Noun + (Noun|ADJ) + |(Noun|ADJ) + |(Noun|ADJ))
–Noun Prep Noun
For the Italian texts, we used the following set of POS patterns, proposed by Bonin et al. (2010):
Noun+(Prep+(Noun|ADJ)+|Noun|ADJ)+

3.3 Statistical filter

To rank the candidate MWTs and separate terms from non-terms, we used two statistical methods: Log-Likelihood Ratio (LLR) (Dunning, 1993) as *unithood* measure to rank the candidate terms extracted in the last phase; and C-NC value method as described in Frantzi et al., (1999) as the measure of *termhood*, i.e., for extracting relevant terms from those ranked by LLR.

1. http://library.islamweb.net/hadith/index.php
2. http://www.eurac.edu/bistro

3.3.1 Likelihood ratio

LLR is a widely used statistical test for hypothesis testing. LLR is a more suitable hypothesis testing method for low-frequency terms.
For bi-grams the LLR is defined as the following:

LLR $(w_1, w_2) = Nw_1;w_2 \log(Nw_1;w_2) + Nw_1;\text{-}w_2 \log(w_1;\text{-}w_2) + N\text{-}w_1;w_2 \log(N\text{-}w_1;w_2) + N\text{-}w_1;\text{-}w_2 \log(N\text{-}w_1;\text{-}w_2) - (Nw_1;w_2 + w_1;\text{-}w_2) \log(Nw_1;w_2 + w_1;\text{-}w_2) - (Nw_1;w_2 + N\text{-}w_1;w_2) \log(Nw_1;w_2 + N\text{-}w_1;w_2) - (w_1;\text{-}w_2 + N\text{-}w_1;\text{-}w_2) \log(w_1;\text{-}w_2 + N\text{-}w_1;\text{-}w_2) - (N\text{-}w_1;w_2 + N\text{-}w_1;\text{-}w_2) \log(N\text{-}w_1;w_2 + N\text{-}w_1;\text{-}w_2) + N \log(N)$,

where $Nw_1;w_2$ is the number of terms in which w_1 *and* w_2 co-occur; $Nw_1;\text{-}w_2$ is the number of terms in which only w_1 occurs; $N\text{-}w_1;w_2$ is the number of terms in which only w_2 occurs; $N\text{-}w_1;\text{-}w_2$ is the number of terms in which neither w_1 nor w_2 occurs; and N is the number of extracted terms.

3.3.2 C-NC value

The method C-NC value combines linguistic and statistical information (Frantzi et al.,1999). The first component, C-value measures the *termhood* of a candidate string using its statistical characteristics which are: number of occurrence; term nesting, which means the frequency of the candidate string as part of other longer candidate terms; the number of these longer candidate terms; and the length of the candidate string. It is defined as:

$$\text{C-value(a)} = \begin{cases} \log_2 (|a|) \cdot f(a) & \text{if } a \text{ is not nested,} \\ \log_2(|a|).\left(f(a) - \frac{1}{p(T_a)} \sum_{b \in T_a} f(b)\right) & \text{otherwise,} \end{cases}$$

where a is the candidate string; $|a|$ is the length in words of a; $f(a)$ is its frequency of occurrence in the corpus; T_a is the set of extracted candidate terms that contain a; $p(T_a)$ is the number of these candidate terms, and $\sum_{b \in T_a} f(b)$ are the sum of frequency by which a appears in longer strings. As we can see if the candidate string is not nested, its *termhood* score will be based on its total frequency in the corpus and its length. If it is nested, the *termhood* will consider its frequency as a nested string and the number of the longer strings into which it appears.

The NC-value component combines the C-value of a candidate string together with the contextual information. By *term context words* we mean the words which appear in vicinity of the extracted candidate terms in the text. A word can be defined as a term context word on the basis of the number of terms into which it appears. The criterion is that the higher the number of terms in which a word appears, the higher the likelihood that the word is a context word and that it will occur with other terms. So the weight of a context word will be calculated in this way: $weight(w) = \frac{a(w)}{n}$, where w is the context word; $a(w)$ is the number of terms into which w appears; n is the total number of candidate terms. So the N-value $(a) = \sum_{w \in C_a} f_a(w) \times weight(w)$, where $f_a(w)$ is the frequency of w as a context word of the term a and C_a is the set of context words of a. This measure is combined with the C-value to provide the C-NC value:

C-NC value(a) = $0.5 \times C\text{-}value(a) + 0.5 \times N\text{-}value(a)$

In our case the C-NC value receive as input the output of the *unithood* measures, namely LLR.

3.4 Identification of translation equivalents

The MWTs lists extracted by the C-NC-value in both languages will be recovered in the parallel corpus. The terms in their context will receive a marked format, using square brackets, to be distinguished from the rest of the words in the corpus. Then we used another algorithm to identify translation equivalents of terms from the parallel corpus. In every translation unit, which contains a source sentence with its target translation, created in TMX format, the system searches the terms between square brackets in both source and target languages. Primarily the system collects in a dictionary the bilingual terms for every translation unit present in the parallel corpus. Afterwards the system will validate the real translation equivalents in the dictionary. The relations types in the bilingual terms dictionary will be as follows:

- one2one
- many2many
- many2one
- one2many

} positive relations

- one2null
- many2null
- null2one
- null2many

} negative relations

After excluding the negative relations, since they will not produce translation equivalents, the system uses the following method for validating relevant equivalents of translation:
a) We use the LLR test, as described above, for estimating the association degree between the bilingual MWTs. In this case the system uses the statistical features of every bilingual MWTs pair in the parallel contest for calculating its LLR value.
b) As a second step the system uses a SMT, namely Google Translate: the idea here is that by means of the translation of the MWTs components the system can identify valid translation equivalents.
c) For the translation pairs which the LLR test and SMT system failed to identify, the system can use the MWTs index in the parallel context. This last choice relies on the idea that for our language pair the index of the words in the context can be considered a good indicator of translation relation. Within every translation unit, the code combines the words with the closest index in the bilingual context, with distance threshold value = 4.

4 Experiments and Results

4.1 The Corpus

We applied the approach to an Italian-Arabic parallel corpus specialized in the domain of international law (Fawi, 2015). The corpus comprises approximately one million words and is aligned at sentence level.

Italian MWTs	Arabic MWTs
1- camera d'appello 2- mandato d'arresto 3- responsabilità penale individuale 4- diritto internazionale umanitario 5- tenta di commettere il reato	1- دائرة الاستئناف 2- أمر بإلقاء القبض 3- المسؤولية الجنائية الفردية 4- القانون الإنساني الدولي 5- الشروع في ارتكاب الجريمة

Table 1. Italian-Arabic equivalent MWTs

4.2 Evaluation

The evaluation process of the term recognition system is a very complex task, not only because there is no specific gold standard for evaluating and comparing different MWTs extraction approaches, but also for the intrinsic nature of the *term* for which it is difficult to give a precise linguistic definition (Pazienza et al., 2005). Since there is no reference list against which we can measure the performance of our approach, we decided to carry out the evaluation mainly by manual validation. The approach validation consists of two parts: MWTs extraction from monolingual corpus (Table 2, 3) and MWTs extraction from parallel corpus (Table 4).

Measure	Arabic		Italian	
	precision	recall	precision	recall
LLR	84%	74%	89%	80%

Table 2. Evaluation of the *unithood* measure

Measure	Arabic			Italian		
	n-best 100	n-best 300	n-best 500	n-best 100	n-best 300	n-best 500
C-NC value	84%	75%	69%	85%	80%	77%

Table 3. Precision of the C-NC value applied on the output of LLR with n-best = 100, 300, 500

measures	recall	precision
LLR	70 %	86 %
SMT system	51 %	88 %
Context Index	50 %	70 %

Table 4. Evaluation of the translation equivalents extraction

5 Conclusion

In this paper we presented our proposed approach to extract multi-word terms from parallel corpora in the legal domain. Regarding the monolingual extraction, we can observe that the results in Italian are a little higher than those in Arabic and this is due to the morphological complexity of the Arabic language which has an impact on the POS tagging performance and therefore on the MWTs extraction. Regarding the bilingual extraction we note that the mediocre recall in SMT system is due to the legal peculiarity of the corpus terms which do not always correspond to the Google translation, while the low recall in the method based on the MWTs index can be attributable to the limited reordering between the two languages. We believe that our attempt can be considered the first one of its type in the Arabic-Italian bilingual domain terminology extraction, and that the results are encouraging. Future work will focus on improving the performance of the approach.

References

Abed, A. M, Tiun, S., and Albared, M., 2013. Arabic Term Extraction Using Combined Approach On Islamic Document. In *Journal of Theoretical & Applied Information Technology*, vol. 58, no. 3, pp. 601 – 608.

Al Khatib, K., 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.

Attia M., Tounsi L., Pecina P., van Genabith J., Toral A. Automatic Extraction of Arabic Multiword Expressions. In *COLING 2010 Workshop on Multiword Expressions: from Theory to Applications*. Beijing, China

Basili, R., Moschitti, A., Pazienza, M., Zanzotto, F. 2001. A contrastive approach to term extraction. In *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)*, France, pp.119-128

Bonin, F., Dell'Orletta, F., Venturi, G., and Montemagni, S. 2010. A contrastive approach to multi word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, La Valletta,Malta, 19–21 May, pp. 3222–3229

Boulaknadel, S., Daille, B., Aboutajdine, D. 2008. A multi-word term extraction program for arabic language. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC, pp. 1485-1488.

Church K.W., Hanks P. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics*, pp.76–83

Delmonte R. 2007. VEST - Venice Symbolic Tagger. In *Intelligenza Artificiale*, Anno IV, N° 2, pp. 26-27

Diab, M. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt

Dunning T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. In *Computational Linguistics,* vol.19, No.1, pp. 61-74.

El Mahdaouy A., Ouatik S., Gaussier E. 2013. A Study of Association Measures and their Combination for Arabic MWT Extraction. In *10th International Conference on Terminology and Artificial Intelligence*, Paris, France

Fawi, F, 2015. Costituzione di un corpus giuridico parallelo italiano-arabo. To appear in *Second Italian Conference on computational Linguistics CliC-it 2015*, 3-4 December 2015, Trento.

Frantzi K., Ananiadou S. 1999. The C–value / NC Value domain independent method for multi–word term extraction. In *Journal of Natural Language Processing*, 6(3), pp.145–179

Lahbib W., Bounhasm I., Elayed, B. 2014. Arabic -English domain terminology extraction from aligned corpora. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences. Lecture Notes in Computer Science*, Vol. 8841, Springer, pp. 745-759

Lyding, V., Chiocchetti, E., Sérasset, G., Brunet-Manquat, F. 2006. The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, pp. 25-31

Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge Mining*, Springer Verlag, pp.255-279

Simões, A. and Almeida, J.J. 2003. NATools: A Statistical Word Aligner Workbench. In *Procesamiento del Lenguaje Natural*, 31, pp.217-224

Streiter, O., Stuflesser M., Ties, I. 2004: CLE, an aligned. Tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface. In *LREC 2004, Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*, May 24