Original scientific paper

Comparative Analysis of Automatic Term and Collocation Extraction

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

Bojana Dalbelo Bašić, Jan Šnajder, Davor Delač
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
bojana.dalbelo@fer.hr, jan.snajder@fer.hr, davor.delac@fer.hr

Matija Šamec-Gjurin, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
msamecgj@ffzg.hr

Dina Crnec
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
dcrnec@ffzg.hr

Summary

Monolingual and multilingual terminology and collocation bases, covering a specific domain, used independently or integrated with other resources, have become a valuable electronic resource. Building of such resources could be assisted by automatic term extraction tools, combining statistical and linguistic approaches.

In this paper, the research on term extraction from monolingual corpus is presented. The corpus consists of publicly accessible English legislative documents. In the paper, results of two hybrid approaches are compared: extraction using the TermeX tool and an automatic statistical extraction procedure followed by linguistic filtering through the open source linguistic engineering tool. The results have been elaborated through statistical measures of precision, recall, and F-measure.

Key words: automatic extraction, term and collocation base, English language, evaluation metrics

Introduction

Term and collocation resources have become useful tool in business, education, and research. Building of such resources, both monolingual and multilingual, can be greatly facilitated by using existing tools for automatic term extraction. Although such tools – especially those combining statistical and linguistic approaches – certainly do require human intervention, they nonetheless can substantially speed up the process. At present times, when Croatia is approaching the EU and undergoing a period of intensive international written communication, use of electronic resources (monolingual and multilingual dictionaries, terminology and collocation bases) could be of a considerable help in the translation work. As the most frequently translated language pair is English-Croatian and vice versa, this paper presents the pilot project of the monolingual term extraction from the English legislative documents, in future to be followed by the Croatian language counter pair. Such resources covering a specific domain may be used in machine translation and computer-assisted translation, information retrieval, building of multilingual bases, glossaries, thesauri, document indexing, and in creation of semantic networks.

According to the Sager’s list of requirements (Love, 2000), the *term* should relate directly to the concept and express it clearly. It should be lexically systematic, not overlap in meaning with other terms, and independent from the context. It should not convey unnecessary information and it should not be pleonastic. The terms should conform to the general rules of word-formation, be capable of providing derivatives, and preserve the original transcription.

On the other hand, according to (Manning and Schütze, 2002) *collocations* contain two or more consecutive words expressing a conventional way of saying things and therefore appear more frequently near each other (e.g., *in general, King of England, freeze up*). Collocations are characterized by non-compositionality (meaning of the collocation can not be predicted from the meaning of the parts), non-substitutability (components can not be substituted), and non-modifiability (not modified through additional lexical material of grammatical transformations). Collocations are considered to be a subset of *multi-word expressions* that constitute arbitrary conventional associations of words within a particular syntactic configuration (Wehrli et. al, 2009). In *multi-word units* the component words include meaningful units (e.g., *Knight of the Round Table*). Although multi-word units are composed of two or more orthographic words (linked by dash, conjunction, or blank), they are treated as a single grammatical unit. Multi-word (MW) units can include foreign expressions (e.g., *ad hoc*), prepositions (e.g. *freeze up, depend on*), adverbs (e.g., *of course*), idiomatic noun constructions (e.g., *know how, per cent*), expressions (e.g., *well being*), as on BNC (British National Corpus) web-page.

The term extraction process, by which a list of term candidates is generated, generally includes two phases (Harris et al., 2003; Thurmair, 2003):

(i) *term extraction* (*term acquisition*), which amounts to identification of term candidates in a corpus, and

(ii) *term recognition*, which refers to verification with a pre-defined list created by an expert in order to identify the (un)known terms.

In this paper, the research on term and collocation extraction from monolingual corpus is presented. The corpus consists of English legislative documents publicly accessible at the EUR-Lex web page[1] providing direct free access to European Union law and the appropriate Croatian translation available at TAIEX–CCVista[2] – Technical Assistance and Information Exchange, containing translations of legal acts of the EU. In the paper the results of two approaches to term extraction are compared: (i) extraction using the TermeX tool, developed at Faculty of Electrical Engineering and Computing, Knowledge Technologies Laboratory and (ii) the automatic statistical extraction procedure followed by linguistic filtering through the open source linguistic engineering tool. The results have been elaborated through statistical measures of precision, recall, and the F-measure.

Related work

There are several collocation extraction tools available today. One of the first collocation extraction tools is the Xtract (Smadja, 1991; Smadja, 1993). Xtract tries to detect collocations based on association measures (AMs), predictive relations, and phrasal templates. Collocate (Barlow, 2004) is a commercial tool that offers collocation extraction based on PMI and Log Likelihood association measures. A span of up to twelve words (12-gram) can be extracted using PMI, whereas Log Likelihood can be used only to extract 2-grams. Collocate does not use morphological normalization such as lemmatization, but is capable of processing previously POS-tagged corpora.

Another tool, presented in (Seretan, Nerima and Wehrli, 2004), is an advanced collocation extractor designed for computer aided translation. It differs from the aforementioned tools in that it focuses on syntactic analysis combined with AMs. The TermeX tool used in this work differs from the above-mentioned tools in that it provides a much wider range of AMs to choose from: as much as fourteen different AMs for extraction of 2-, 3-, 4-grams are provided. To improve extraction performance, TermeX uses morphological normalization, POS filtering, and filtering by frequencies.

Much of the work on the usability of extraction tools, hybrid approaches, and their integration into machine translation and information retrieval systems has been discussed by Thurmair (2003) Building of bilingual lexicons by extracting bilingual entries from aligned bilingual text using bidirectional transfer has been

[1] http://eur-lex.europa.eu/en/index.htm

[2] http://ccvista.taiex.be

discussed by Turcato(1998). According to Wehrli et al. (2009), "collocations could present a particular problem for machine translation, because of their frequency, their different morpho-syntactic properties, and long-distance dependencies." The ITS-2 system presented by Wehrli et al. (2009) is a large-scale translation system relying on a detailed linguistic analysis provided by the parser, which exploits monolingual lexicons. In their research, the transfer system is used to produce information-rich phrase-structure representation related to the predicate-argument structure, identifying multi-word expressions such as idioms and collocations. Extraction of collocations is made by a hybrid method, combining syntactic information from the parser with statistical methods for detection of typical constructions in the corpus. Another two hybrid models (Daille 1996; Izuha 2001) for term extraction from parallel bilingual text used linguistic various statistical scores for ranking. Extraction of multi-word expressions for the Croatian language have been presented in (Bekavac and Tadić, 2008).

According to the previous research, best results are obtained using hybrid approaches. In this research two types of hybrid approaches will be presented and compared. The first model uses statistical lexical association measures (AMs) combined with POS filtering and morphological normalization.

The second approach extracts the terms in two steps. It first uses statistical extraction regardless of the length of n-grams, filtered by a predefined frequency threshold and a stop-words list. In the second step, the list of potential candidates is fed through language dependant local grammars in the NooJ engineering tool, combined with its high-priority dictionary for disambiguation.

Research

Resources

This research includes a selection of ten different types of EU legislation documents related to the EU activities: three Council Decisions, one Commission Decision, one Decision of the European Central Bank, three Council Regulations, and two Commission Regulations – in total amounting to about 20,000 words. The documents have been translated from the original Croatian legislation. The texts have been revised and used for creation of a term and collocation base. Extraction process was made by two independent groups of researchers using:

- TermeX tool (Delač, 2009) developed at the Faculty of Electrical Engineering and Computing in Zagreb, Knowledge Technologies Laboratory;
- a statistically-based term extraction tool SDL Multi Term Extract and a linguistically-based environment NooJ (Silberztein, 2004) developed at the University Franche-Comté Paris, France.

For the purpose of evaluation, a reference list has been created by the human experts, representing the gold standard of terms typical for EU legislation vocabulary.

Tools

Approach A

The first approach uses the TermeX tool (Delač, 2009), a tool for construction of terminology lexica with possible applications in NLP. Collocation extraction in TermeX is based on association measures (AMs), statistical measures that provide information on how likely it is for an n-gram (the sequence of *n* words) to be a collocation. Extraction is done by creating ranked lists of n-grams based on their AM value. This way terms that are most likely to be a collocation become top ranked. TermeX implements fourteen AMs, based on Pointwise Mutual Information (PMI), Dice, and Chi-square. Implemented measures are extensions of the corresponding bigram measures for n-grams spanning up to four words as described in (Petrovic, 2009)0. In order to improve collocation extraction, TermeX implements POS filtering and morphological normalization to better cope with morphological complexity of natural languages.

In TermeX, a terminology lexicon is created by selecting collocations from lists of automatically extracted collocation candidates. Building of a single lexicon is referred to as a *project;* multiple corpora can be processed simultaneously as parts of a single project. For the purpose of this experiment, TermeX was first run on the *Acquis communautaire* corpus to gather the complete statistical data, after which the terms from the ten selected documents were extracted. The AMs used in this experiment were PMI for 2-grams and heuristic measures described in (Petrovic, 2009)0 for 3-grams and 4-grams. It should be noted that the AMs and POS filters used by TermeX are optimized for the extraction of noun phrases rather than verb phrases.

Approach B

The second type of research started from the language independent statistically-based approach with a predefined frequency threshold using SDL Multi Term Extract. It offered a number of term candidates and probable translations, both presented in a term candidate list on a user-friendly graphic interface. After validating terms and their translations, it was possible to export them to MultiTerm XML or a tab-delimited format. This list was then filtered from stop-words.

In the next step, the specialized language dependent tool NooJ was used. NooJ is a linguistic engineering platform providing tools for the formalization of language phenomena at different levels: orthography, morphology, lexicon, syntax, and semantics. It therefore includes large-coverage dictionaries and grammars, and parses corpora in real time. Its linguistic engine is multilingual and there are a dozen of modules for different languages available, as well as a dozen more being prepared. NooJ processes texts and corpora in numerous file formats (varying from HTML, PDF, and MS Office to XML documents). NooJ issues sophisticated queries in order to produce various results (i.e., concordances, statistical analysis, or information extracts). In this research, statistically obtained lists were filtered by 36 types of regular expressions (local grammars) in order to identify word combinations that match certain POS patterns. For the purpose of disambiguation, an additional pre-compiled filter dictionary in NooJ was set up at high priority level, after which the linguistic analysis was performed.

Lists

Reference list

The reference list contains 470 terms and collocations, excluding unigrams. The terms in the list vary from bodies' titles, functions' titles, documentation and common phrases, introductory and operative clauses, etc. Creation of a reference list is a rather difficult task, aiming to cover a specific domain, but balancing between lexical coverage, adequacy for the domain, and inclusion of typical expressions.

The reference list in this case study contains

- terms as semantic units in canonical forms (*acquiring company, annual account, applicant country*),
- collocations chosen because of their frequency at the pragmatic level as "preferred ways of expressing things", according to Thurmair (2003) (*adopt a resolution, decided as follows, entry into force, for the purpose of, having regard to*), names and abbreviations (*Economic and Monetary Union EMU, European Union EU, European Central Bank ECB*), and embedded terms relevant for the domain (*crime prevention, crime prevention bodies, national crime prevention measures*).

While terms are mainly noun phrases (346 out of 470), collocations also contain many verbal phrases. Distribution of n-grams in the reference list is presented in Table 1.

Table 1: Number of n-grams in the reference list

N-grams	2-grams	3-grams	4-grams	5-grams and more
Total: 430	119	138	98	75

List A

The list extracted with TermeX consists of 1816 terms. TermeX uses POS filters tuned to extract noun phrases consisting of two, three, and four words. Of the 1816 extracted terms, 758 consist of two words, 679 terms consist of three words, and 379 consist of four words. Most of the extracted terms are indeed semantically full noun phrases, mostly named entities and compound nouns.

List B

Using a language-independent statistically-based SDL Multi Term Extract tool, a list of terms has been obtained. This list was filtered by the list of stop-words, eliminating words such as determiners, pronouns, prepositions, conjunctions, etc. that appear at the beginning or at the end positions of candidates. The number of extracted term candidates, with frequency threshold set to 4, was 369. This list included not only semantically full terms, but also meaningless sequences of words or unfinished terms, requiring for, e.g., a noun, past participle, or a prepositional phrase, but extracted because of their frequency. These lists also contain terms that embed a noun and a number (e.g., Directive 68/151/EEC), which should not be included in the term base. Therefore, these lists contain considerable number of meaningless candidates, which would not pass the linguistic test. In the next step, 36 local grammars have been applied on the statistical list, containing mostly <A><N> and <N><N> candidates, followed by <N><PREP><N>, <A><V>+<G><N> (G for gerundive, i.e., verb in gerundive form), <N><CONJ><N> and <N><A>, as presented in Table 2. Percentage of local grammars is presented as one example per match. Because many of the lexical items were polysemous meanings, a new dictionary in NooJ was compiled and set up at high priority level. After linguistic filtering, a list of 512 term candidates has been created. The reason for bigger number (512 after linguistic filtering comparing to 369 candidates after statistical analysis) lies in the extraction of embedded terms (e.g., after pure statistical approach the term *applicable to public limited-liability companies in the Member* was extracted while after linguistic approach the following candidates were identified: *public limited, limited-liability, liability companies*; the term *Counterfeit Analysis Centre* extracted in statistical analysis was identified after linguistic analysis as *Counterfeit Analysis, Counterfeit Analysis Centre, Analysis Centre, Analysis Centres*).

Table 2. Local grammars

Regular expressions (local grammars)	
	1 example per match
Most common	<A><N> **31%**
	<N><N> **30%**
	<N><PREP><N> 17%
	<A><V>+<G><N> 12%
	<N><CONJ><N> 6%
	<N><A> 6%
Least common	<N><CONJ><A><N><N>
	<V>(<DET>)<N><N>
	<V><A><N>
	<A><N><A><N>
	<A><N><P><N>

Results

Statistical Analysis

Statistical analysis is performed via measures of precision, recall and the F-measure, by comparing the lists of terms extracted by the two tools against the terms form the reference list.

Recall is defined as the proportion between valid computer extracted terms and expert extracted terms (the reference list), although it is hard to define the relevant set in the reference list regarding the quality and the quantity. The perfect recall score of 100% indicates that all valid terms were extracted, but does not say anything about the fact how many irrelevant terms were also extracted.

Precision is defined as the proportion between valid computer extracted terms and all computer extracted terms. As precision reflects the noise, it is also possible to have certain amount of false positives, i.e., terms that are extracted by the tool, but not included in the reference list. The perfect precision score of 100% indicates that every extracted term was relevant, but does not at all indicate whether all relevant items were extracted.

F-measure (or F-score) allows adjusting the relationship between recall and precision. The F-measure is the weighted harmonic mean between precision and recall.

Table 3. Results of extraction evaluation

	List A	**List B**
No. of terms	1816	508
Valid terms	202	234
Precision (%)	11.56	47.37
Recall (%)	42.98	49.79
F1 (%)	18.22	48.55

True positives were calculated by taking into account the inflectional variants of terms: a simple suffix stripping procedure was applied to conflate the inflectional variants to a single canonical form. Note that a more sophisticated morphological normalisation procedure (such as lemmatisation) was not required in this case: suffix stripping did not introduce any ambiguity and linguistic validity of norms was not required. Moreover, when comparing two terms, the determiners were ignored, so that, for instance, *adopt a decision* and *adopt decision* would be considered as match.

The results are shown in Table 3. The results for list A are rather unsatisfactory, while for list B they are modest. The number of terms common to both lists is 355. The low recall for list A can be traced down to the fact that TermeX tool does not extract verb phrases nor does it extract terms consisting of more than four words. If such terms are removed from the reference list, recall reaches up to 77.47%.

Results of the list B could be improved by lemmatization in order to have expressions in canonical forms, definition of upper/lower cases, precision of determiner in collocations, and by more detailed local grammars.

In the lists, there are also a number of false positives, i.e., terms and collocations not found in the reference list. We plan to address this issue as part of future work.

Conclusion

In this paper the results of two hybrid approaches to automatic term extraction were evaluated and compared. Human-created term and collocation lists differ from automatically created lists, mostly because of human knowledge, experience, and intuition that is involved in deciding whether a certain candidate can or can not be a term or a collocation.

Results show that extracted terms cover the specific domain in question and may serve to complement the dictionary, but there is certainly space for improvement.

Automatic extraction combined with human intervention may give usable results. We believe that the direction that should be taken is the fine-tuning of human criteria (when compiling the reference list) and application of hybrid models for automatic extraction.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants No. 130-1300646-0909 and 036-1300646-1986.

References

Aoughlis, Farida. A Computer Science Electronic Dictionary for NooJ. Lecture Notes in Computer Science, 2007. pp. 341-351

Barlow, M.: Collocate 1.0: Locating collocations and terminology. TX:Athelstan, 2004.

Bekavac, B.; Tadić, M. (2008) A Generic Method for Multi Word Extraction from Wikipedia. Proceeings of the ITI 2008 30th Int. Conf. on Information Technology Interfaces, 2008.

Daille, Béatrice. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pages 49–66. MIT Press, Cambridge, Massachusetts, 1996.

Delač, Davor; Krleža, Zoran; Dalbelo Bašić, Bojana; Šnajder, Jan; Šarić, Frane : TermeX: A Tool for Collocation Extraction. Lecture Notes in Computer Science (Computational Linguistics and Intelligent Text Processing, 2009; 149-157.

Dillinger, M. Dictionary Development Workflow for MT: Design and Management. MT Summit VIII, 2001.

Drouin, Patrick. *Acquisition automatique de termes : l'utilisation des pivots lexicaux spécialisés*. Thèse de doctorat présenté à l'Université de Montréal, Montréal, 2002. http://www.olst.umontreal.ca/pdf/DrouinPhD2002.pdf

Drouin, P. Term extraction using non-technical corpora as a point of leverage. In *Terminology*, 2003, vol. 9, no 1, p. 99-117. http://www.olst.umontreal.ca/pdf/ Terminology_2003.pdf, 9.6. 2008.)

Harris, M.R.; Savova, G. K.; Johnson, T.M.; Chute, C.G. (2003) A term extraction tool for expanding content in the domain of functioning, disability, and health: proof of concept. J Biomed Inform.; 36(4-5) 250-9, 2003.

Izuha, T. Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts (R). MT Summit VIII, 2001.

Keller, Frank. Evaluation: Connectionist and Statistical Language Processing http://homepages.inf.ed.ac.uk/keller/teaching/internet/lecture_evaluation.pdf, 9.6.2008.)

Love, Stacy. *Benchmarking the performance of Two Automated Term-extraction systems: LOGOS and ATAO*. Mémoire de maîtrise, Université de Montréal, 2000. http://www.olst.umontreal.ca/pdf/memoirelove.pdf, 6.6.2008.)

L’Homme, Marie-Claude and Hee Sook Bae. A Methodology for Developing Multilingual Resources for Terminology. In *Proceeding of LREC 2006. Language Resources and Evaluation, 2006*. http://www.olst.umontreal.ca/pdf/LREC-2006-Lhomme-bae.pdf, 9.6. 2008.)

Manning, Chiristopher. D.; Schütze, H. Foundations of Statistical Natural Language Processing. MIT, 2002.

NooJ http://www.nooj4nlp.net

Petrović, S., ·Šnajder, J., Dalbelo Bašić, B.: Extending lexical association measures for collocation extraction. Computer, Speech and Language, 2009 (doi:10.1016/j.csl.2009.06.001.).

SDL MultiTerm Extract http://www.sdl.com/en/products/products-index/multiterm.asp

Smadja, F.: Retrieving collocations from text: Xtract. In: Proceedings of 31th Annual Meeting of the Association for Computational Linguistics. Vol. 19, 1993, 143-177

Smadja, F.: From n-grams to collocations: An evaluation of Xtract. In: Proceedings of 29th Annual Meeting of the Association for Computational Linguistics, 1991. 279-284

Seretan, V., Nerima, L., Wehrli, E.: A tool for multi-word collocation extraction and visualization in multilingual corpora. Proceedings of EURALEX Congress, 2004.

Silberztein, M. NooJ: A Cooperative, Object-Oriented Architecture for NLP. In : INTEX pour la Linguistique et le traitement automatique des langues. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté, 2004.

Tadić, M., Šojat, K.: Finding multiword term candidates in Croatian. Proceedings of Information Extraction for Slavic Languages 2003 Workshop IESL, 2003, 102-107

Thurmair, Gregor. Making Term Extraction Tools Usable. Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, EAMT-CLAW 2003.

Turcato, Davide. Automatically Creating bilingual Lexicons for Machine Translation from Bilingual Text. Proceedings of the 17th international conference on Computational linguistics, vol. 2, 1998, 1299 - 1306

Vintar, Špela. Extracting terminological collocations from parallel corpus. 5th EAMT Workshop, 2000.

Wehrli, E.; Seretan, V.; Nerima, L.; Russo, L. Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy. Proceedings of the 13th Annual Conference of the EAMT, 2009, 128–135

Zienlinski, D.; Safar, Y.R. (2005) Research meets practice: t-survey 2005: An online survey on terminology extraction and terminology management.