First Steps in Term and Collocation Extraction from English-Croatian Corpus

[1]Sanja Seljan, [2]Angelina Gašpar

[1]Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lucica 3, 10 000 Zagreb, Croatia
sanja.seljan@ffzg.hr
[2] Ministry of the Interior
Ulica grada Vukovara 33, 10000 Zagreb, Croatia
angelina.gaspar@yahoo.com

Abstract: Term and collocation bases represent valuable additional resources covering specific domain and frequently expressions, which then can be used in further research. The paper presents possible model of building terminology and collocation base, using statistical and linguistic approaches in order to gain experience in building of such resources for the English - Croatian language pair. The aim of the paper is not to evaluate tools, but to give an insight into use of tools and to gain experience in building, training and testing of language resources. In the paper, two types of statistically-based term and collocation bases have been compared, created out of the legislative documentation and then filtered through language dependant linguistic patterns.

Key words: term extraction, collocation, terminology base, automatic, tools, resources, statistical, linguistic, English, Croatian.

1 Introduction

In the past decades, Croatia has been undergoing a process that many EU countries have already experienced through economic, cultural and educational transition that influences all aspects of written communication. New types of cooperation and preparation activities for EU membership have caused an increased need for the translation. During parallel work of translators, use of shared resources has become indispensable for consistent translation. Besides use of dictionaries, term and collocation bases represent valuable additional resources covering specific domain and frequently used expressions, which than can be used in further research: building of multilingual bases, glossaries, thesauri, in information retrieval, machine translation, computer-assisted translation, document indexing and creation of semantic networks. A need for up-to-date reference work is even more obvious for not widely spoken languages. In the paper, bilingual term entries have been extracted from English-Croatian legislation. Texts have been sentence aligned and then used as a source for the automatic creation of term and collocation bases based on statistical approach, then elaborated using NooJ linguistic environment.

2 Term Extraction

Term extraction is an operation which takes a document as input and produces a list of term candidates as output. Term candidates are words or phrases which are potential terms of the subject area represented by the input document. In the paper term extraction is based on bilingual corpus relating to English and Croatian legislative documentation, i.e. regulations and decisions. Term extraction process generally includes the following phases (Harris et al., 2003; Thurmair, 2003): term acquisition or term extraction including identification of term candidates and term recognition including verification with pre-defined list, created by an expert in order to identify (un)known terms.

3 Resources, tools

The aim of the paper is not to evaluate tools but to gain experience in the process of semi-automatic creation of linguistic resources and to use it in building of Croatian language resources. The research is made on 10 English-Croatian legislative documents relating to the EU activities which have been aligned and then used for the extraction. The final list of the extraction purpose was compared with the reference list, which was manually created for this purpose.
The corpus consists out of 10 English legislative documents relating to the EU activities, and their corresponding Croatian translations, which have been revised and publicly accessible at http://ccvista.taiex.be. The documents consist of 20,094 words in English and corresponding 17,583 words in Croatian language due to its flectional nature. The first step included alignment and saving in formats suitable for further research. Texts have been aligned using Robert C. Moore's Bilingual Sentence Aligner at http://research.microsoft.com/~bobmoore/, creating 784 translation units which have been then used for the automatic extraction.

Table 1.

Documents		TU	Paragraphs		Words	
Eng	Cro		Eng	Hr	Eng	Hr
10	10	784	776	788	20094	17583

4 Tools

In the research, two types of statistically-based term extraction tools have been used: SDL Multi Term Extract Lexterm by the Open University in Barcelona (http:www.linguoc.cat) for which the same stop-list was created and used. The list was then filtered using Linguistically-based environment NooJ (http://www.nooj4nlp.net/), developed by Max Silberztein at University Franche-Comté Paris, France. For the purpose of disambiguation the dictionary was compiled

and set up at high priority level. The final list was then analyzed from the syntactic and semantic point of view by two independent professionals, one involved in translation and the other involved in the domain of the computational linguistics.

MultiTerm Extract (MTE), with its large variety of extraction possibilities is a valuable tool in the term extraction process from monolingual or bilingual documents and translation memories. For each term candidate it offers probable translations, both shown in a term candidate list on a user-friendly graphic interface. After validating terms and their translations it is possible to export them to MultiTerm XML or a tab delimited format. MTE is interrelated with other programs in SDL Trados package designed to assist to the translators before and during translation. Also, it makes possible that the following parameters have been set up: min term length, max term length, min translation frequency, and max number of translations.

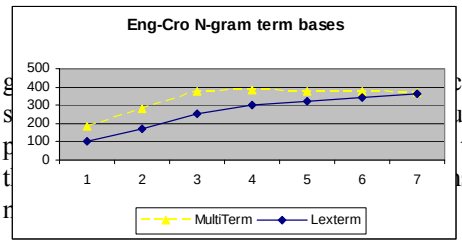
Lexterm (LT) tool, created at Open University Catalunya is an open source tool, using also statistical approach and extracting from monolingual or bilingual translation memories. It offers the list of most probable terms and their translations through one or more candidates. After validating terms and translations, it is possible to export them for further use. The term then could be rejected or selected and then one translation candidate chosen and if necessary, post-edited.

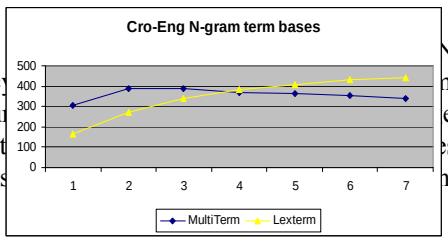
In both types the same stop-list has been used. Unigrams have not been included in this research. Low frequency terms were not identified as term candidates, as they didn't pass the statistical threshold set up at 4. Automatically created bases have been filtered by manually created list of stop words containing functional words, such as articles, conjunctions, prepositions, etc. in order to refine the suggested term bases. The term extraction conducted in this case-study is considered to be «bottom-up» approach having no preconceived terminology structure.

In the first step of this case-study the following lists have been created: a) English lists consisting of N-grams (from 2- to 8-grams, excluding unigrams) by MTE. b) Croatian lists of N-grams (from 2- to 8-grams, excluding unigrams) by LT. In both cases the 8-gram lists were used for further analysis

Table 2.

	Tool	**2gr**	**2-3gr**	**2-4gr**	**2-5gr**	**2-6gr**	**2-7gr**	**2-8gr**
Eng	MtEx	185	279	379	385	376	378	369
	Lext	105	173	253	301	321	340	362
Cro	MtEx	307	388	388	369	364	352	307
	Lext	164	271	339	382	410	431	164





5 Linguistic filtering

Automatically created lists include not only semantically full terms, but also meaningless sequence of words or unfinished terms, requiring certain complement, but chosen because of frequency. These lists also include terms containing noun and number (e.g. *Directive 68/151/EEC*) which would not be included in the term base, a number of candidates which would not pass the linguistic test, number of candidates differentiating in singular and plural (e.g. *adopted by Member State - adopted by Member States*), candidates differentiating in determiners (e.g. *law of a Member, law of a Member State, law of that Member)*, candidates starting with the same expression and differentiating in one word (e.g. *accordance with Article/ Directive/ paragraph/ Regulation, etc.),* candidates containing abbreviations (e.g. *Council of the ECB, development of SIS*) and a number of candidates that meet linguistic needs, but semantically do not form the clear term.

Statistically created lists have undergone through further linguistic filtering by local regular grammars within NooJ linguistic environment. Therefore, statistically created lists were analyzed through language dependant specific POS-patterns. A list of acceptable multi-word POS-patterns was created, partly in table 4., where A is used for adjective, N for noun, P for preposition, Conj for conjunction, V for verb. For the Croatian language the most frequent combination is AN (72%), followed by NN, NPN and AAN, and at the lower level ANN, NNN, etc.

The list of acceptable English terms created out of English-Croatian legislative parallel pilot-corpus using local grammar looks as follows:

Table 3.

	Frequency of syntactic patterns by MTExtract	Frequency of syntactic patterns by Lexterm
AN	**70 (25%)**	**36 (30%)**
NN	**42(15%)**	**22(18%)**
NPN	**33 (11,8%9**	**16 (13,3%)**
ANN	12	4
APN	11	9
V+ppN ….	10	8
TOTAL	**278**	**120**
Det	67 (25%)	38(31,6%)

The frequency of syntactic patterns extracted from both statistical lists show that the most represented patterns are ANs and NNs, followed by NPNs. In both lists there is significant proportion (25% and 31,6%) of expressions containing specific determiner, and therefore, possibly not identified in the exact search pattern.

6 Manual list and comparison

For the purpose of this case study a reference term and collocation base has been manually created. The task was given to the translator unaware of the reasons for this research in order to obtain classic term base created by human.

Table 4.

	Manual list	Final list
AN	**96 (31%)**	**109 (29,5%)**
NN	**75 (24,3%)**	**84 (22,7%)**
NPN	41 (13,3%)	47 (12,7%)
V+gN	17	19
ANN	15	18
NConjN ….	12	13
TOTAL	**308**	**369**
Det	16	21 (5,7%)

The only information that was given was not to include single words, but compounds, i.e. two or more lexical items. According to the experiences, the main problem was to define lexical coverage and adequacy for the domain, balancing between granularity and generality, but also to decide whether certain term is a candidate for the standard list, especially collocations frequently used, but not belonging to the professional term base (e.g. *adopt provisions, approve draft terms, company being acquired, enter into force, formed in accordance with*, etc.). Table 5 presents the syntactic structure of the manually created list. Comparing the three lists, it can be seen that the advantage is also given to ANs and NNs (in total 55%) followed by NPN pattern.

Difference between lists lies also in semantics: in automatically created list terms such as *accordance with Article/ Directive* are found (8 term candidates), while in manual list the proposed term is *in accordance with*. Manually created term would be *having regard to* (1 term candidate) while statistically extracted terms would be *having regard to Council/ to the initiative/ to the opinion/ to the Treaty, etc.* In the manual list term candidates appear in number singular, whereas in automatically created lists appear in singular and in plural, embedding also prepositional phrases.

In this research, the term of false positives appeared, i.e. terms that are not extracted manually in the reference set, but appear in automatically derived sets (Harris et al., 2003) and could be used as terms. The terms identified by the automatic extraction and not included into reference set, were reviewed and if agreed, included then into the reference set. False positives include also terms that are not always semantically justified, but they appear frequently in the text (e.g. *in accordance with, adopted in implementation, applicable to public limited-liability companies, appointed or approved, binding in its entirety and directly applicable*.). Therefore, the final list contains reference list and include additional list of false primitives from both statistical lists, counting all together 369 terms and collocations, i.e. total of 52,2% of ANs and NNs, followed by 12,7% of NPNs.

7 Conclusion

In the paper the generic method for term and collocation extraction has been presented, relying on subset of parallel corpus of English-Croatian legislation for the purpose of creating, training and testing. Human-created term lists differ from automatically created lists, mostly because of human knowledge, experience and intuition when deciding whether certain candidate can or can not be a term. Although manual lists contain more meaningful candidates, they are rather time-consuming. Automatically created list contain a lot of duplication and meaningless candidates requiring human correction, but are still created at lower cost and time. The final list contains manually created list and the list of false positives, composed in the biggest proportion out of ANs and NNs, followed by NPNs. Statistically created list, filtered by language-dependant linguistic patterns, overlap with manually created list in 70 terms, while further overlapping is reduced because of the determiner in the middle of expression, differences in number, capitals, etc. The results would be considerably improved if bigger corpus is used, comparative language dependant patterns and lemmatization. Statistically created list, filtered by linguistic engineering tool, could help, though human post-editing and refinement are required. Extracted terms tend to cover specific domain and could serve as an additional base to the dictionary.

This work is an outcome of the research project (130-1300646-0909).

References

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

Drouin, P. (2002) Acquisition automatique de termes : l'utilisation des pivots lexicaux spécialisés. Doctorat Univ. de Montréal, http://www.olst.umontreal.ca/pdf/DrouinPhD2002.pdf

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

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.

L'Homme, M—C. Hee S.B. (2006) A Methodology for Developing Multilingual Resources for Terminology. In Proceeding of LREC 2006. Language Resources and Evaluation. http://www.olst.umontreal.ca/pdf/LREC-2006-Lhomme-bae.pdf)

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

Thurmair, G. (2003) 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 03.

Vintar, Š. (2000) Extracting terminological collocations from parallel corpus. 5th EAMT Workshop.

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