On Extracting Multiword NP Terminology for MT

Svetlana Sheremetyeva
LanA Consulting ApS
Møllekrog, 4, Vejby
3210 Copenhagen, Denmark
lanaconsult@mail.dk

Abstract

The paper addresses the issue of MT knowledge acquisition and describes a new hybrid methodology for automatic extraction of multi-word nominal terminology. The approach is based on statistical techniques merged into a strongly lexicalized Constraint Grammar paradigm. It is targeted at intelligent output and computationally attractive properties.

1 Introduction

The quality of machine translation output is to a large extent influenced by the comprehensiveness of multiword term dictionaries where noun phrases (noun phrase terms) are more frequent than any other types of multiword expressions.

Noun phrases (NPs) can often be translated into other languages irrespective of the context and contribute significantly to the robustness of MT systems by reducing the ambiguity inherent in word to word matching and text analysis.

Multiword phrase databases are relevant for both RBMT and SMT systems; - in state-of-the-art statistical translation systems structural relations between source and target sentences are captured by means of phrases instead of isolated words (Zens et al., 2002; Koehn et al., 2003).

To build a two level hierarchy of phrases phrase driven SMT systems more and more focus on the linguistic concept of NP as the unit of decomposition (Hewavitharana et al, 2007).

Creating multilingual MT resources is based on the acquisition of unilingual lexicons as the first and basic step (Pohl, 2006; Hewavitharana et al, 2007; Daille and Morin, 2008).

In many cases, especially for low resource languages, lexical acquisition starts from the English side independent of the translation direction. For example, (Hewavitharana et al, 2007) developing an Arabic-to-English MT first extract English NPs as the Arabic parsers available do not produce desired accuracy. The quality of monolingual (English, in particular) extraction, is thus of primary importance for the quality of MT.

Another issue which matters a lot for practical MT systems is the speed of NP extraction process. It directly affects the costs of MT system development and maintenance. Despite a lot of research on extracting different kinds of multiword phrases related to MT the problem still presents a tough challenge (Piao et al., 2006). Extraction of NP phrases, in particular, is especially problematic as it normally involves parsing and often very expensive computationally.

We suggest a new hybrid NP extraction methodology based on statistical techniques merged into a strongly lexicalized Constraint Grammar paradigm which features computationally attractive properties and intelligent output. We illustrate our approach on the example of the English language as its resources are most widely used in MT research and the quality of these resources is still to be improved.

While testing our approach we got practical results for the patent domain corpus and saw how generic extraction procedures can take advantage of the domain restrictions. For patent MT the issue of extraction quality and speed is one of the priorities as its terminology is being constantly renewed and requires operative maintenance.

In what follows we first overview related work, we then define our task and describe the extraction procedure followed by evaluation results. We conclude with discussion and future work.

Proceedings of the 13th Annual Conference of the EAMT, pages 205–212,
Barcelona, May 2009

2 Related work

The range of the work related to NP term extraction is very wide and covers NP, multiword expression, collocation and keyphrase extraction. Keyphrase implies two features: phraseness and informativeness (Tomokiyo and Hurst, 2003) and while the issue of informativeness is beyond the scope of the current article, techniques used to identify phraseness are of direct interest to our research.

NP describes objects and concepts. It is a grammatical notion and techniques used for detecting noun phrases in the text are normally NLP-oriented. The most correct results in NP extraction can be expected with full-fledged NLP (symbolic) procedures, which while unquestionable under the assumption of perfect NLP parsing in reality will immediately lead to the problems of coverage, hence robustness and correctness[1]. Pure NLP parsing can be very time consuming and normally not portable.

An ultimate example of symbolic approach to extraction is a semantic tagger which annotates English corpora with semantic category information and is capable of detecting and semantically classifying many multiword expressions but can suffer from low recall (Rayson et al., 2004).

Current approaches to NP extraction in an attempt to raise recall and extraction speed involve statistical techniques where phrases, collocations or multiword expressions are determined as word sequences with no intention to limit the meaning in a linguistic sense. In pure statistical methods phrase extraction is based on n-gram extraction and may include such preprocessing steps as stoplist words removal and stemming[2]. Phrases are further selected based on various statistical collocation/phrasenessh metrics, e.g., mean and variance (Smadja, 1993) and binomial log likelihood ratio test (BLRT) (Dunning, 1993), to mention just a few.

On the one hand, statistical techniques offer some clear advantages, such as speed, robustness and portability, over linguistically-informed methods. On the other hand, the results obtained statistically are not always "good" phrases, and the basic statistical systems may suffer from combinatorial explosion if calculations are made over a large search space.

To overcome the limitations of "pure" approaches a use of statistics supplemented by heuristics and linguistic techniques is more and more popular in the research community. In hybrid systems extraction often involves morphological normalization, so that each word (lexical item) can be identified regardless its actual morphological form. Two basic approaches to morphological normalization are stemming, where a word is transformed (usually heuristically) into its stem, and lemmatization, where a word is transformed into its base form by morphological analysis (Pecina, 2008).

In general, the process of multiword unit (NP including) extraction follows the steps of a) identification of candidates from the text and b) filtering the candidates. Particular hybrid extraction techniques differ in the amount and order in which linguistics and statistics are used.

(Smadja, 1993) creates a set of collocation candidates applying statistical co-occurrence information on a pretagged corpus and after extraction uses parsing for filtering out invalid results.

(Daille et al. 1994) make use of linguistic knowledge at the first stage of extraction to identify two-word noun phrases which correspond to a limited number of syntactic patterns on the previously tagged corpora. At the second stage statistical scores based on the number of occurrences of the pairs are used to select the "good" ones among the candidates.

(Seretan and Wehrli, 2006) use a syntactic parser in the first extraction stage for identifying two-word collocation candidates. The pairs are then partitioned according to their syntactic configuration. Finally, the log likelihood ratios test (Dunning, 1993) is applied to filter "good" NPs.

(Pecina, 2008) describes the extraction of two-word collocation candidates performed on morphologically normalized texts and filtered by a frequency filter and a part-of speech filter.

(Piao et al., 2005) suggest augmenting the power of multiword expression extraction by combining a statistical tool for searching and identifying English multiword expressions with a lexicon-based English semantic tagger (Rayson et al., 2004). The authors emphasize that training the tools on specific domains is essential for good results. Domain restrictions, such as strict structuring and sublanguage specificity is normally taken into consideration by various data and text mining tools applied to patent texts (Hull et al., 2001; Fattori et al., 2003).

[1] It is impossible to acquire knowledge including all words in all senses, a priori defined syntactic configurations and disambiguation rules.

[2] See, e.g., (Porter, 1980) for stemming algorithm.

3 Task definition

Our ultimate goal is to develop a methodology for extracting multiword NP terminology targeted to intelligent results and computationally attractive properties for facilitating and speeding up the development and maintenance of high-quality real-world MT systems.

The target of our extraction effort is thus defined by the intersection of five criteria: (i) multiword expression, (ii) noun, (iii) terminology, (iv) increase of recall and precision, (v) reduction of computational cost. For this work, we considered a string composed of several words to be a multiword expression if its meaning cannot be computed from its elements (Gross, 1986). However, in this definition, we, similar to (Laporte et al.,2008) consider that the possibility of computing the meaning of phrases from their elements is of any interest only if it is a better solution than storing the same phrases in lexicons.

We extracted only expressions belonging to the noun part of speech. We recognized them through the usual criteria regarding their morphosyntactic context.

We assumed that multiword noun expressions in a technical text are terms, see, e.g., (Daille,1994) for similar approach.

We aim to extract an NP candidate which is included into a larger NP candidate, only if the shorter NP functions individually in the processed corpus or text meant for MT. For example, if we have NP candidates such as,

1. antenna port selection method
2. antenna port selection
3. antenna selection method
4. port selection method
5. antenna port
6. selection method
7. port selection

then if candidates 2, 4, 6, 7 do not function individually in the corpus, only candidates 1, 3 and 5 will be included in the final output.

Such approach has obvious advantages in domain-tuned MT, e.g., for saving multilingual lexicon acquisition effort. This restriction, however, can always be lifted if necessary.

Our ambition is not to loose low frequency and unique NP terms.

We experimented with different proportions of statistical and linguistic knowledge in an attempt to increase recall and precision and reduce computational cost.

4 Approach

We tried and discarded the idea of starting extraction with stoplist words removal as it may lead to "bad" combinations of words. For example, removal of stoplist words (boldfaced) at the preprocessing stage from the patent fragment:

...**a** table **in which the** wireless location system continuously maintains **a** copy **of the** status **of** transmitters…

will lead to extraction of such "NPs" as

*table wireless location system
*copy status transmitters

Such combinations will not be filtered out automatically as they satisfy our grammar and they could have a high frequency due to the specificity of patent texts, where text fragments as above can be highly repetitive, e.g., in patent claims[3]. Even if unique they will not be discarded as our intention is to extract all NP terms.

We also decided against morphological normalization as preprocessing. Heuristic stemming algorithms, may fail to identify inflectional variants and lead to the extraction of wrongly combined and/or truncated character strings which are impossible to understand thus lowering precision. Proper NLP lemmatization is very expensive computationally. For these reasons we postponed lemmatization to the very last stage of processing.

We first calculate n-grams ($0<n<5$)[4] on a raw text, and then select NP candidates (singular and plural) with a strongly lexicalized constraint-based grammar as the major filtering mechanism. This initial candidate set is filtered by a count-based criterion. The key proposals here are:

- to apply shallow parsing based on grammar rules related to NP word order constraints to the raw text n-grams, and
- to apply these constraint rules through direct lexical (word string) match of an n-gram component against a lexicon rather than through POS tagging.

[3] A patent claim is a part of a patent where all essential features of an invention are formulated. A patent may include more than a hundred of claims.

[4] This is the most widely used limit for the number of words in n-gram extraction, but in our system "n" can be set to any number.

4.1 Resources

Our multiword NP extraction is based on the following resources:

- a shallow patent domain lexicon of part-of speech-unambiguous wordforms.
- application specific rules of a strongly lexicalized constraint grammar
- a heuristic noun lemmatizer

The shallow lexicon is a patent domain corpus-based list of wordforms with their part-of-speech information. The specificity of this lexicon is that it only includes part-of-speech-unambiguous wordforms. The advantage of using such a lexicon is in avoiding a computationally (and resource) expensive procedure of part-of-speech disambiguation. To build the lexicon we extracted part-of-speech-unambiguous lists of English wordforms from the lexicon of the patent MT system (Sheremetyeva, 2007).

Constraint-based grammar formalism is formalism in which a class of constraints is used to reduce a class of potential representations to the representations, which are well formed, or grammatical (Daniels and Meurers, 2004). In our approach we use the Phrase Structure Grammar constraints on the NP word order to select the initial set of NP candidates.

The specificity of our rules is that they are not the usual part-of-speech NP patterns to find n-grams which match these patterns. Our rules find those n-grams which cannot be NPs, without determining their full part-of-speech structures.

The rules are only applied to the first, last or middle (in case of 3- and 4-grams) words of an n-gram performing shallow (hence less expensive) n-gram parsing rather than complete parsing. The rules are as follows:

Rule 1
IF the first word in an n-gram is
Determiner/verb/preposition/wh-word/
THEN delete n-gram

Rule 2
IF last word in an n-gram is
adjective/verb/preposition/wh-word/article/
THEN delete n-gram

Rule 3.
IF the word, which is neither the first word,
nor the last word in a 3-gram,
is determiner/verb/wh-word
THEN delete the 3-gram

Rule 4
IF the word, which is neither the first word,
nor the last word in a 4-gram,
is /verb/wh-word
THEN delete the 4-gram

As can be seen, the rules do not exactly allow for all linguistically legal NP patterns. For example, the "determiner" constraint in Rule 1 is included because we do not want to extract NP candidates starting with articles or other determiners ("this", "that", etc.) as they are not included in MT lexicons.

Such application-motivated constraints are to some extent equivalent to stop words in traditional statistical approaches but in our case they are applied selectively and filter out inappropriate phrases rather than output non-existing NPs.

Due to the postponing of morphological normalization to the very last stage of processing, when an NP candidate set is supposed to consist of NPs only (plural and/or singular) we can afford to use a restricted noun lemmatizer rather than a full-fledged morphological lemmatizer. This again contributes a lot to processing robustness and resource/computation savings.

4.2 Procedure

The multiword NP terminology extraction starts by simple calculation of raw text n-grams, $0<n<5$[5]. Note, that though our goal is to extract multiword NPs we do not discard the list of 1-grams at this stage.

We then apply the rules of our grammar to filter out n-grams which cannot be NPs and build an initial set of NP candidates. The matching procedure in rule application is reversed. It starts with trying to match the first, last and middle word of every n-gram against the lexicon.

In case a lexical match is found the morphological description of the matching word is checked. If the matching word in the lexicon has a part-of-speech forbidden by the rules, the n-gram is discarded; otherwise it is added to a candidate set. If no lexicon match is found for any of the n-gram components the n-gram is also assigned an NP candidate status thus making the rules absolutely robust. Another advantage of the grammar rules is that they are computationally simple.

[5] This is the most widely used limit in n-gram calculation, but actually "n" can be set to any number.

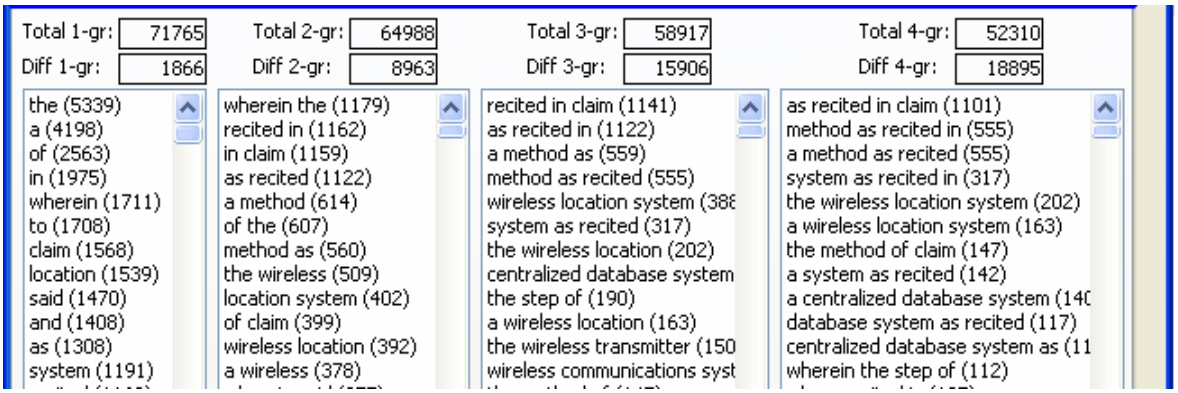

Figure 1. A fragment of top 1- to 4-gram lists before the application of the lexicalized constraint grammar rules. Numbers in brackets show frequencies.

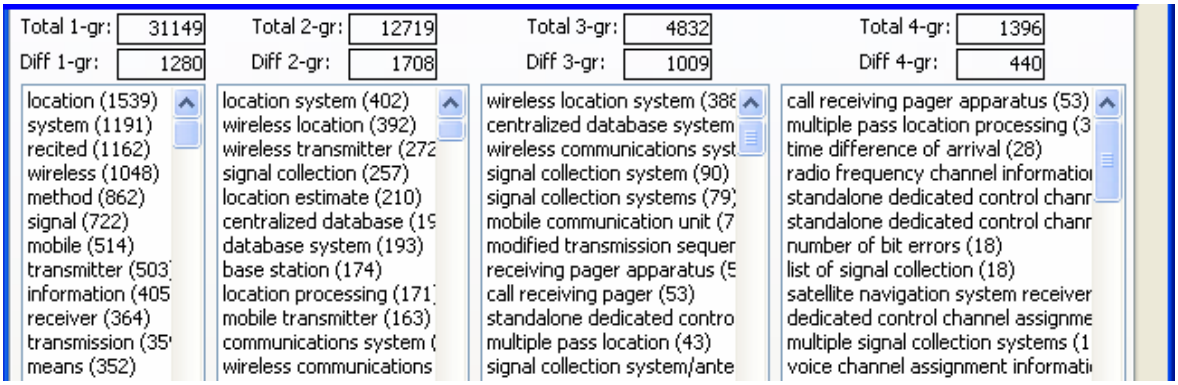

Figure 2. A fragment of top 1- to 4-gram lists after the application of the lexicalized constraint grammar rules. Numbers in brackets show frequencies.

Every rule taken separately will let pass some of the ill-formed NP candidates for which no match in the lexicon was found, as a lot of words being part-of-speech ambiguous are simply not in the application lexicon. However, successive application of the grammar rules to different words of the same n-gram compensates for this lack of the lexicon coverage. A "bad" NP not identified by one rule will be identified by another and thus discarded.

For example, the 3-gram "change the system" will not be forbidden by Rule 1, as the ambiguous word "change" (it can be a verb or a noun) is excluded from our lexicon, but this 3-gram will still be discarded by Rule 3, which demands to discard 3-grams containing determiner ("the" in this case) in the middle.

We thus can handle NP word order constraints in a computational parsing, without invoking additional layers of representation (i.e., disambiguated tagging).

The quality of filtering with our application modified lexicalized PHSG can be judged by comparing the n-gram lists in Figures 1 and 2, which show the top of 1- to 4-gram lists before and after the application of the grammar rules.

At the next stage of processing we create an expansion matrix over the initial set of all grammar filtered candidates, 1-gram including. A fragment of an expansion matrix is shown in Figure 3.

The matrix is created to make a decision whether shorter NP candidates which are parts of longer NPs function individually and "have the right" to be included in the final output.

For this purpose we introduce the count-based criterion "Uniqueness" (U) which is defined as the difference between an n-gram frequency and the sum of frequencies of its (n+1)-gram expansions.

A low U-value shows that the candidate is unlikely to be used individually. We experimentally selected the U=0 or U< 0 values as thresholds for filtering out undesired candidates.

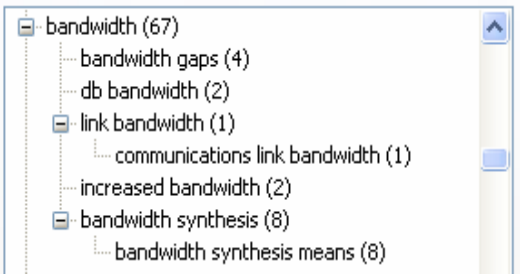

Figure 3. The expansion matrix of the 1-gram "bandwidth". Given in brackets are frequencies.

For example, in Figure 3 the candidates `link bandwidth` and `bandwidth synthesis` have U=0 and will be discarded; the rest 5 multiword candidates will be included in the candidate list for further filtering. After cleaning candidate duplicates in the expansion matrix [6] we once again run the grammar filter to discard the residue of "bad" candidates and then apply our noun lemmatizer. The duplicates[7] cleaned, the resulting set is output. Fragments of the output with top and low frequency NP terms are shown in Figures 4 and 5. Given in brackets are frequencies.

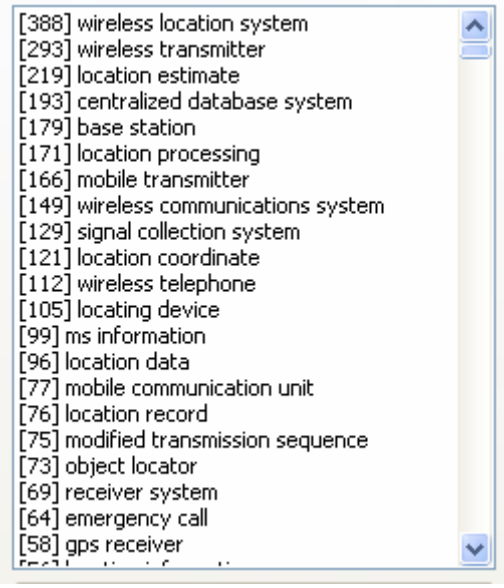

Figure 4. Top frequency multiword NP terms extracted over the evaluation corpus.

[6] One and the same candidate can appear as expansion in different n-gram nests.

[7] When a term in plural is lemmatized it may duplicate the term in singular which was already in the candidate set.

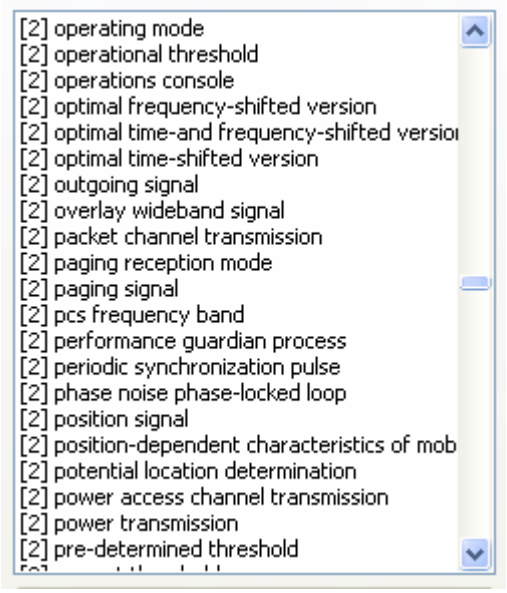

Figure 5. Low frequency multiword NP terms extracted over the evaluation corpus.

To summarize the extraction procedure is as follows:

1. IDENTIFICATION OF CANDIDATES
   a. Calculating raw text n-grams.
2. FILTERING
   a. Fist filtering of candidates with the use of the lexicon and constraint grammar rules.
   b. Calculation of an extension matrix
   c. Second filtering of candidates with the U-criterion
   d. Cleaning the resulting list (removing duplicates)
   e. Third filtering of candidates with the use of the lexicon and constraint grammar rules
   f. Lemmatization of resulting NPs
   g. Cleaning of the resulting list (removing duplicates)
3. OUTPUT

5 Evaluation

Our evaluation scheme covered the two basic demands: quality and speed. The quality evaluation method consisted in comparing our result list with a gold reference list.

The gold list was built manually by linguist students following the guidelines formulated in Section 3. The evaluation was performed over a patent corpus of 72000 words for which it was feasible to create a gold standard. The results of the evaluation are given in Table 1.

Total number of gold NPs	1425
Total extracted phrases	1476
Task correct NPs	1351
Gerundial phrases	43
Short NPs, not used individually	58
Missed NPs longer than 4 words	52
Missed NPs shorter than 4 words	16
Incorrect phrases	24

Table 1. Evaluation results.

Gerundial phrases are those like given below in bold face:

a server for **receiving tasking requests** from other applications

Such phrases if not actually NPs can still be attributed to nominal terminology as they normally mean processes and translationwise often correspond to regular NPs in other languages.

Most of short NPs, not used individually appeared in the final output due to "technical" reasons, namely, because we limited ourselves to a 4-gram window which does not allow for extracting NP terms containing more than 4 words. This makes it impossible to properly calculate the U value of shorter terms included in long ones. Examples of such terms are shown below (extracted NPs are in bold face).

multiple discrete frequency **elements of consistent amplitude**

outgoing **real time two-way communication**

caller **generated wireless local loop** communication system

One way to fix this problem is to widen the extraction window which might increase the computation time, but whether it really matters is left for further experiments. On the other hand, the shorter NPs, though not functioning individually, can still be included in an MT lexicon leaving translation of longer phrases to translation grammars. The number of such long terms is not very large, - 52 out of 1425 in our test.

The numbers of "bad" mistakes are shown in the last two rows on Table 1.

The speed of NP extraction is to a great extent increased due to the computational savings provided by our approach which removes a lot of n-grams from further computations at the early stage of extraction (compare numbers given on the top of Figures 1 and 2) and users shallow parsing and restricted lemmatizer.

In addition to that the extraction speed depends upon such factors as the load on the server, the speed of the network, and the size of the input text. Patents range in size from a few kilobytes to 1.5 megabytes. We can report that on a regular Hewlett-Packard X86-based PC it usually takes a fraction of a second to process a patent. An XML file of 8 megabytes containing 150 patents is processed in less that 2 min.

6 Conclusions

In this paper we described a methodology for extraction of multiword NP terms. The methodology provides for intelligent output and has computationally attractive properties due a specific combination of statistical, NLP and heuristic techniques. It includes n-gram calculation, shallow parsing based on strongly lexicalized constraint grammar. The grammar rules are applied to raw text n-gram components through direct lexical (word string) match against a non-ambiguous lexicon.

The methodology is robust as it does not depend on lexicon coverage and excludes such statistically or NLP expensive techniques as vast combinatorial computations or proper tagging and parsing.

We illustrated the approach on the example of patents in the English language but preliminary experiments show that it is portable to different domains and languages.

Different applications can benefit from the techniques proposed here, ranging from knowledge acquisition for RBMT systems or phrase-based SMT systems to machine-aided NLP tools.

We plan to extend this work in a number of ways. We are currently working on including the NP extractor into the analysis module of a patent MT system.

Another perspective is to extent the application to a multilingual keyphrase extraction tool for further use in multilingual search and information extraction.

References

Daille Béatrice, Éric Gaussier, & Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. *Coling 1994: the 15th International Conference on Computational Linguistics*: Proceedings, August 5-9, 1994, Kyoto, Japan; pp. 515-521.

Daille Béatrice and Emmanuel Morin. 2008. A effective compositional model for lexical alignment. *IJCNLP 2008: Third International Joint Conference on Natural Language Processing*, January 7-12, 2008, Hyderabad, India; pp 95-102.

Daniels, M. and Meurers, D. 2004. GIDLP: A grammar format for linearization-based HPSG. HPSG04 Conference proceedings.

Dunning Ted . 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74

Fattori, Michele., Pedrazzi Giorgio., Turra, Roberta. 2003. Text mining applied to patent mapping: a practical business case. World Patent Information,25

Gross, M. 1986. Lexicon-Grammar. The representation of compound words. In Proceedings of the 11th International Conference on Computational Linguistics,COLING'86, Bonn, West Germany, pp. 1-6.

Hewavitharana Sanjika, Alon Lavie, and Stephan Vogel. 2007. Experiments with a noun-phrase driven statistical machine translation system. *MT Summit XI*, 10-14 September 2007, Copenhagen, Denmark. *Proceedings*; pp.247-253.

Hull D., A i it-Mokhtar, S., Chuat, M., Eisele, A,, Gaussier, E., Grefenstette, G. 2001. Language technologies and patent search and classification. World Patent Inf 23.

Koehn P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In Proc. of the 41th Annual Meeting of the Association for Computational Linguistics.

Laporte, Eric, Takuya Nakamura and Stavroula Voyatzi. 2008. A French Corpus Annotated for Multiword Nouns. Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008) pp.27-30.

Pecan, Pavel. 2005. An extensive empirical study of collocation extraction methods. In Proceedings of the ACL Student Research Workshop, pp 13–18.

Pecina, Pavel .2008 Reference Data for Czech Collocation Extraction. Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008) pp.11-14..

Piao, S. L., Rayson, P., Archer, D. and McEnery, T. 2005. Comparing and Combining A Semantic Tagger and A Statistical Tool for MWE Extraction. *Computer Speech & Language* Volume 19, Issue 4, pp. 378-397

Pohl, Gábor. 2006. English-Hungarian NP alignment in MetaMorpho TM. *EAMT-2006: 11th Annual Conference of the European Association for Machine Translation*, June 19-20, 2006, Oslo, Norway. Proceedings; p.69-74

Rayson, P., Archer, D., Piao, S., and McEnery, T., 2004. The UCREL semantic analysis system. In: Proceedings of the LREC-04 Workshop, beyond Named Entity Recognition Semantic Labelling for NLP Tasks, Lisbon, Portugal, pp.7–12.

Seretan, Violeta and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. *Coling-ACL 2006: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, 17-21 July 2006; pp.953-960.

Sheremetyeva, Svetlana. 2007. On Portability of Resources for Quick Ramp-Up of Multilingual MT for Patent Claims. Proceedings of the workshop on Patent Translation in conjunction with MT Summit XI, Copenhagen, Denmark, September 10-14.

Smadja, F. 1993. Retrieving collocations from text. Xtract. *Computational Linguistics* 7(4):143–177.

Tomokiyo, T., and Hurst, M. 2003. A language model approach to keyphrase extraction. Proceedings of ACL Workshop on Multiword Expressions.

Zens, R., Och, F.J., Ney, H.: Phrase-based statistical of LNCS. Springer Verlag (September 2002) machine translation. In: Advances in artificial intelligence. 25. Annual German Conference on AI. Volume 2479 18–32.