On Operative Creation of Lexical Resources in Different Languages

Svetlana Sheremetyeva

LanA Consulting ApS, Møllekrog, 4, Vejby 3210 Copenhagen, Denmark

Abstract. Cognitive modeling is to a large extent mediated by lexicons thus bringing in focus operative creation of high quality lexical resources. This paper presents a methodology and tool for automatic extraction of lexical data from textual sources. The methodology combines n-gram extraction and a filtering algorithm, which operates blocs of shallow linguistic knowledge. The specificity of the approach is three-fold, - (i) it allows dynamic extraction of lexical resources and does not rely on a pre-constructed corpus; (ii) it does not miss low frequency units; (iii) it is portable between different lexical types, domains and languages. The methodology has been implemented into a tool that can be used in a wide range of text processing tasks useful for cognitive modeling from ontology acquisition, to automatic annotation, multilingual information retrieval, machine translation, etc.

1 Introduction

There is more and more evidence today that cognitive paradigms can complement computational models and boost natural language technology. NLP, in its turn, is one of the major vehicles that provides cognitive science with knowledge coded in the natural language, - indispensable source of all kinds of information [1]. Integration of NLP and cognitive modeling is to a large extent mediated by lexicons thus bringing in focus the issues of quality and operative creation of lexical data.

High quality linguistic resources are mostly handcrafted and their creation is costly and time-consuming. Much effort has been made to overcome this problem by automatically inducing grammars and lexicons from corpora. Among these efforts of great importance for cognitive modelling are the works that have concentrated on methodologies and techniques for automatic extraction of typed lexical units, such as noun, verbal, etc., phrases; they reflect different types of cognition processes.

There is an ongoing work to save development effort by suggesting language-independent extraction methodologies and tools [2], [3], [4], [5].

Another issue which matters a lot for using knowledge acquisition tools in practical applications (e.g., cognitive modelling) is the speed of extraction process as it directly affects the applicability potential and costs of system development.

The range of the work in automatic lexical acquisition is very wide and covers single- and multiword expression, collocation and keyphrase extraction, as, e.g. (in addition to those cited above) [6], [7].

Sheremetyeva S..
On Operative Creation of Lexical Resources in Different Languages.
DOI: 10.5220/0004089300440052
In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science* (NLPCS-2012), pages 44-52
ISBN: 978-989-8565-16-7

Typed lexical unit is a grammatical notion and the most correct extraction results 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. Pure NLP parsing can be very time consuming and is 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 [8].

In an attempt to raise recall and extraction speed current approaches to extraction involve statistical techniques where phrases are determined as word sequences with no intention to limit the meaning in a linguistic sense. Pure statistical methods are based on n-gram extraction and may include such preprocessing steps as stop list words removal and stemming. Phrases are further selected based on various statistical collocation/phraseness metrics, e.g., binomial log likelihood ratio test (BLRT) [9], or "unithood" [10] 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 basic statistical systems may suffer from combinatorial explosion if calculations are made on a large search space.

Most successful are hybrid approaches that in different proportions combine statistical metrics with linguistic knowledge, such as morphology, syntax and even semantics. Practically all researchers mentioned above use hybrid techniques.

While working quite well on English and other low inflecting languages, well developed extraction techniques are often rendered useless by rich morphology, case syncretism and relatively free word order of highly inflecting languages.

We attempt to contribute to the studies in the field and suggest a novel hybrid extraction methodology for typed lexical units, which could work well for both low and highly inflecting languages.

The specificity of our approach is three-fold, - (i) it allows operative dynamic extraction of lexical resources and does not rely on a pre-constructed corpus; (ii) it does not miss low frequency units; (iii) it is portable between different types of lexical units, domains and languages. The methodology was tested on English, French, Spanish and such highly inflecting language as Russian. It was implemented in a tool, which is intended to assist researches, translators, lexicographers, language teachers, professionals, analysts and system developers in dealing with different types of language processing. By means of the tool the end user can fulfill most of portability tasks her/himself.

The paper is structured as follows. Section 2 gives an overview of the extraction methodology. Section 3 describes its portability potential. Section 4 outlines possible shortcomings and ways to overcome them. Section 5 presents evaluation results. The paper concludes with methodology summarization and outlines its possible fields of application.

2 Methodology Overview

2.1 Approach

Our ultimate goal was to develop a possibly universal methodology for extracting typed lexical resources targeted to

- Intelligent results accounting for the specificity of input texts/corpora
- Reliable extraction of both high and low frequency typed lexical units
- No preconstructed corpus
- Multilingual and multipurpose use
- Computationally attractive properties.

By intelligibility we mean that our extraction results should be grammatically correct (not truncated) and thus can be comfortable both for human use, and for porting into other applications. To account for input specificity we aim to avoid extracting lexical units, which are included into longer candidates but do not function individually in the input document/corpus. We further provide a mechanism for relevancy calculation based of statistical distribution. Relevancy is to be user-defined depending upon the extraction purpose. We pay special attention to the reliable extraction of not only high frequency units but low frequency units as well. By multipurpose use we mean that the extraction results can be output in different "shapes" as exhaustive lemmatized/not lemmatized lists or user-defined-relevancy keywords. The last but not least goal is to let the end user be able to fulfill most of portability tasks her/himself.

We will describe our methodology on the example of single- and multiword noun phrase (NP) extraction. NPs describe objects and concepts and are considered closely connected to the content of utterances. They are most frequent in a text while extraction of NPs is especially problematic. It normally involves parsing and is often very expensive computationally [11]. Our extraction methodology combines statistical techniques, linguistic knowledge and heuristics. It starts with n-gram calculation. But unlike many extraction procedures ours skips stop words removal at the preprocessing stage. We discovered that it might "spoil" extraction results. For example, if the task is to extract NPs, the removal of traditionally used stop words (boldfaced) at the preprocessing stage from the fragment below:

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

will lead to the extraction of phrases, which do not exist in the input, such as

**table wireless location system*

**copy status transmitters.*

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. Proper NLP lemmatization is very expensive computationally. We therefore postponed lemmatization to the last stage of processing.

We further avoided using regular statistical metrics for extraction proper and used linguistic filters only. This was done to be able to reliably extract low frequency units and make the methodology portable to highly inflecting languages where exact matches of n-grams are much less frequent than in English. For the same reason linguistic filters in our methodology use very shallow linguistic knowledge.

2.2 Knowledge

The linguistic knowledge is very shallow and includes

(i) partial knowledge about the constraints on a typed unit structure in terms of parts-of speech (POS) for a particular language

(ii) a number of shallow lexicons each referenced to a particular (first, middle[1] or last) position in the typed unit to be extracted

(iii) specific rules of a strongly lexicalized constraint grammar. The rules are very simple and are as follows:

Rule 1
IF the first word in an n-gram belongs to Lexicon 1
THEN delete n-gram

Rule 2
IF the last word in an n-gram belongs to Lexicon 2
THEN delete n-gram

Rule 3
IF the middle word belongs to Lexicon 3
THEN delete n-gram

The rules are language–independent. They find and delete those n-grams which cannot be NPs, without determining n-gram full part-of-speech structures. The lexicons are of course language-dependent and contain lists of wordforms of relevant parts-of-speech in a particular language. For example, for the English language in the NP extractor Lexicon 1 contains explicitly listed wordforms of verbs, determiners, wh-words and prepositions. The specificity of these lexicons is that they only include POS-unambiguous wordforms. The advantage of using such lexicons is in avoiding a computationally (and resource) expensive procedure of POS disambiguation. The optional part of knowledge includes language-dependent lemmatizers.

2.3 Workflow

The extraction workflow is divided into a basic procedure which outputs a list of grammatically correct typed lexical units, and optional procedures that can be run on the output of the basic procedure and do not influence the extraction quality.

The basic extraction procedure consists of the following 3 steps:

- *Calculation of n-grams* (n = 1, 2, 3, 4) from *a raw* corpus/document.
- *Selecting the initial set of NP candidates* (singular and plural) using a strongly lexicalized constraint-based grammar as the major filtering mechanism.

[1] Here “middle” means “not first and not last”.

- *Filtering out* partial n-grams, i.e. n-grams which do not occur individually but only as parts of longer n-grams.

To select the initial set of NP candidates the grammar rules try to match the first, last and middle word of every n-gram against its position referenced lexicons. In case a lexical match is found the n-gram is discarded; otherwise it is added to a candidate set.

Every rule taken separately might let pass some of the ill-formed NP candidates for which no match in the lexicon was found. This happens because a lot of words which are POS-ambiguous are simply not included in the application lexicons. 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 that was not identified by one rule will be identified by another and thus discarded.

The “good” candidates are then checked on functioning isolated in the input text. This is done using a count-based criterion “Uniqueness” (U) which we define as the difference between an n-gram frequency and the sum of the frequencies of its (n+1)-gram extensions. 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. Extraction can stop at this point if no lemmatization or scoring are needed like, e.g., in case our extraction procedure is to be included in a concordance or parser.

The optional procedures include

- *Lemmatization* and count summing
- *Wordform merge* and count summing
- *Ranking and filtering* based on statistical criteria

If cumulative counts are needed to calculate a lexical unit relevancy, e.g., to identify keywords, these counts can be obtained after lemmatization or wordforms merge, - a much simpler procedure than lemmatization. Word merge differs from stemming in that it outputs grammatically correct wordforms rather than truncated strings of characters. A heuristic algorithm automatically merges inflected forms of the same lexical unit into one (not necessarily a lemma) with cumulative counts. This procedure is based on character match and is language-independent. In general our extraction methodology works equally well for high frequency and low frequency units (see Figure 1).

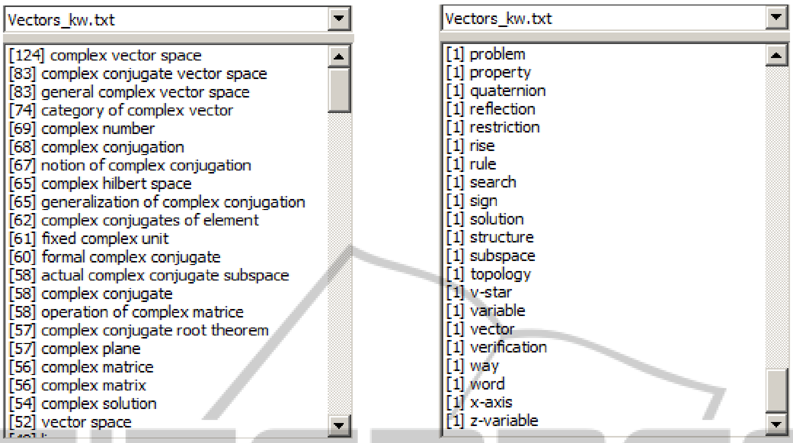
3 Portability

A universal algorithm realizing deletion rules trigged by cascade matches of the lexicons against n-grams and extremely shallow linguistic knowledge give our methodology a high portability potential.

The processing algorithm being universal, a generic portability procedure consists in creating shallow language- and application-dependent linguistic knowledge, which includes.

- Acquisition of syntactic configurations for a typed phrase of interest
- Acquisition of the stop lexicons

- Creation of a lemmatizer (optional).



```
Vectors_kw.txt
[124] complex vector space
[83] complex conjugate vector space
[83] general complex vector space
[74] category of complex vector
[69] complex number
[68] complex conjugation
[67] notion of complex conjugation
[65] complex hilbert space
[65] generalization of complex conjugation
[62] complex conjugates of element
[61] fixed complex unit
[60] formal complex conjugate
[58] actual complex conjugate subspace
[58] complex conjugate
[58] operation of complex matrice
[57] complex conjugate root theorem
[57] complex plane
[56] complex matrice
[56] complex matrix
[54] complex solution
[52] vector space
```

```
Vectors_kw.txt
[1] problem
[1] property
[1] quaternion
[1] reflection
[1] restriction
[1] rise
[1] rule
[1] search
[1] sign
[1] solution
[1] structure
[1] subspace
[1] topology
[1] v-star
[1] variable
[1] vector
[1] verification
[1] way
[1] word
[1] x-axis
[1] z-variable
```

Fig. 1. Top and bottom fragments of the NP extraction results from an English scientific paper on mathematical modeling.

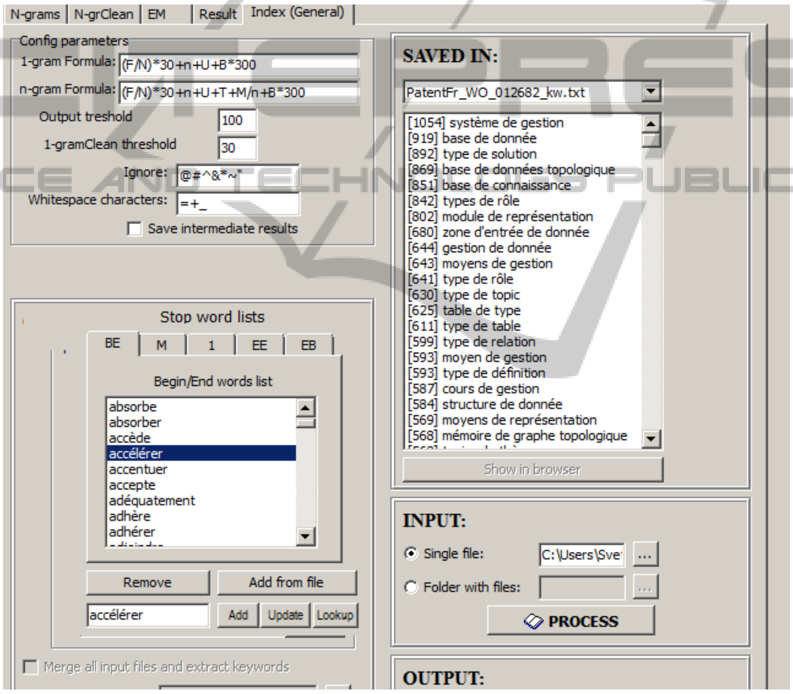
Acquisition of syntactic knowledge is reduced to the identification of those part-of-speech classes that are forbidden or not desirable in the first, middle and last positions in a phrase structure. Acquisition of the stop lexicons is reduced to the acquisition of wordforms sorted into POS classes and then combining these lists according to the syntactic configuration of a unit to be extracted. The fist and second tasks are pretty straightforward and can be fulfilled by the end user. The development of a merge module and lemmatizer is the responsibility of the developers.

4 Possible Shortcomings

The straightforward shortcoming is due to POS ambiguity, which can cause undesirable noise. To bypass this problem we restricted the content of the stop list lexicons to unambiguous wordforms only, which of course relaxes filtering. However, our experiments showed that the successive application of different stop lexicons to the same n-gram compensates for this drawback. Shortcomings can also be expected if the methodology is applied to a highly inflecting language, such as Russian, where a typical word has from 9 (for nouns) up to 50 forms (for verbs) [12] and exact n-grams are much less frequent than in English. This can lead to the U criterion failure in identifying NP extensions. For example, the Russian NP *“средства контроля колес» («wheel control means») will* not be recognized as an extension of the NP *«средствам контроля» («control means»)* because of the not matching case strings of *«средства» (means*_N_nom*)* and *«средствам»* (*means*_N_dat). Both short and long NPs will in such a case be included in the final output. Since such “leaked” candidates are still “good” this shortcoming can be neglected. The shallow filter can also pass some noun combinations forbidden by the Russian grammar. Evaluation (see Section 5) showed a small number of this kind of mistakes and they could be either neglected or filtered out by a human.

5 Testing and Evaluation

For testing and evaluation we have implemented our methodology into a tool, which, the knowledge changed as described in section 3, can be used for different extraction tasks in multiple domains and languages. The tool has flexible settings and lets the user administer both the linguistic knowledge, and scoring parameters through the interface. We passed the tool with the knowledge tuned to the extraction of English patent domain key phrases (NPs) to the end-users (researchers, translators and teachers) at ESC SUSU (Educational and Scientific Center of Innovative Language Technology at South Ural State University, Russia).



Config parameters
1-gram Formula: (F/N)*30+n+U+B*300
n-gram Formula: (F/N)*30+n+U+T+M/n+B*300
Output treshold 100
1-gramClean threshold 30
Ignore: @#^&*~"
Whitespace characters: =+_
□ Save intermediate results

Begin/End words list
absorbe
absorber
accède
accélérer
accentuer
accepte
adéquatement
adhère
adhérer

accélérer

□ Merge all input files and extract keywords

SAVED IN:
PatentFr_WO_012682_kw.txt
[1054] système de gestion
[919] base de donnée
[892] type de solution
[869] base de données topologique
[851] base de connaissance
[842] types de rôle
[802] module de représentation
[680] zone d'entrée de donnée
[644] gestion de donnée
[643] moyens de gestion
[641] type de rôle
[630] type de topic
[625] table de type
[611] type de table
[599] type de relation
[593] moyen de gestion
[593] type de définition
[587] cours de gestion
[584] structure de donnée
[569] moyens de représentation
[568] mémoire de graphe topologique

INPUT:
● Single file: C:\Users\Sve
○ Folder with files:

OUTPUT:

Fig. 2. A screenshot of the keyword (NP) extractor interface for the French patent domain, which is created based on the described methodology and resources. Numbers in square brackets show relevancy scores which are set as a logical combination of statistical parameters such as n-gram frequency (F), average n-gram frequency (N), the number of most frequent 1-gram components of an n-gram (T), summed frequency of these components (M), n-gram length (n), uniqueness (U), etc. In the left bottom corner shown is the knowledge administration interface.

After a half-day training session the users were able to port the tool to English, French, Spanish and Russian. Figure 2 shows the interface of the tool ported to the French patent domain with a fragment of extraction results. The tool has been used at

ESC for more than two years. The tasks include mostly the extraction of noun and verbal phrases from both single documents and open corpora from patent, economy, mathematical modeling and programming domains. The purposes of the extraction tasks were collecting resources for bilingual dictionaries, selecting foreign language teaching material, on-the-fly terminology look-up in the original papers on the Internet (used by translators), and domain specificity analysis for different research.

The evaluation results presented in this paper are based both on developer testing, and end-users feedback. The quality evaluation method for every language consisted in comparing the results of the basic extraction procedure (see Section 2) with gold reference lists from randomly selected patent domain corpora. The test corpora consisted of 20 000 wordforms for each language. The gold lists were built semi-automatically by calculating n-grams (n = 1, 2, 3, 4) with an off-the-shelf tool and their subsequent manual cleaning. It was not feasible to account for the uniqueness factor with gold lists thus constructed. We therefore considered all units, which were grammatically correct to be relevant results. The gold and result lists were further automatically compared by means of another tool, which produced difference lists that, in turn, were analyzed by humans. We used recall and precision measures for quantitative estimates. Evaluation results for a major extraction tasks, - NPs are given in Table 1.

Table 1. Evaluation results for the major typed lexical unit - NP, showing the quality of extraction based on the described methodology.

	English	French	Spanish	Russian
Recall	97,5 %	96,3%	96,0%	98,4%
Precision	94,8%	92,5%	91,8%	93,4%

It was not feasible to compare the results and portability potential of our methodology to those of other reported works. Few exhaustive evaluations of the extraction results have been carried out and no evaluation reports on methodologies portability potential are available.

The end-users who participated in testing reported high satisfaction with the extraction results and the simplicity of knowledge administration/porting. On average one person/day was spent to acquire linguistic knowledge (as described in section 3) for a new language. It took even less time to port the tool to a new type of linguistic unit (e.g., from NP to VP) within one language. It required only a small brush up of stop lexicons when porting the tool to a new domain within the same language and unit type. Low frequency units are extracted as reliably as high frequency units.

6 Conclusions

The paper described a hybrid extraction methodology and illustrated its portability potential across domains, applications and languages. The key proposals are:

- to calculate n-grams from a *raw text*
- to use “deletion rules” rather than POS-pattern “search” rules
- to use the knowledge related to the typed unit *word order constraints* rather than

full-fledged POS patterns

- to apply constraint rules through *direct lexical (word string) match* of an n-gram component against position-referenced lexicons (*avoid tagging and syntactic parsing*)

The methodology features intelligent output and computationally attractive properties.

An overall testing and evaluation showed that the methodology stays robust for English, French, Spanish, as well as for Russian. The basic extraction procedure outputs all inflected forms of proper types of lexical units.

Different applications can benefit from the techniques proposed here, ranging from knowledge acquisition for cognitive modeling to indexing, unilingual and multilingual information retrieval, extraction, summarization, machine translation, language learning/teaching and the like.

References

1. Motivation in Grammar and the Lexicon (Human Cognitive Processing), .Ed. Panther KU., G. Radden. John Benjamin’s publishing Company (2011) 313.
2. Cholakov K, Kordoni, V., Zhang, Y.: Towards domain-independent deep linguistic processing: Ensuring portability and re-usability of lexicalized grammars. In: Proceedings of COLING 2008 Workshop on Grammar Engineering Across Frameworks (GEAF08), Manchester, UK (2008).
3. Lefever E., Macken, L., Hoste, V.: Language-independent bilingual terminology extraction from a multilingual parallel corpus. In: Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece (2009) 496–504.
4. Valderrabanos V. A. S., Belskis, A., Iraola L.: TExtractor: a multilingual terminology extraction tool. In: Proceedings of the second international conference on Human Language Technology Research, San Diego, California (2002) 393-398
5. Seretan, V., Wehrli, E. Multilingual collocation extraction with a syntactic parser. In: Language Resources and Evaluation, 43(1) (2009) 71–85.7.
6. Daille B., E. Morin. An effective compositional model for lexical alignment. IJCNLP 2008: Third International Joint Conference on Natural Language Processing, January 7-12, Hyderabad, India (2008) 95-102.
7. Michou A., Seretan, V.: Tool for Multi-Word Expression Extraction in Modern Greek Using Syntactic Parsing. In: Proceedings of the EACL Demonstrations Sessions. Athens, Greece (2009).
8. Rayson, P., Archer, D., Piao, S., and McEnery, T.The UCREL semantic analysis system. In: Proceedings of the LREC-04 Workshop, beyond Named Entity Recognition Semantic Labelling for NLP Tasks, Lisbon, Portugal, (2004) 7–12.
9. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1) (1993) 61–74.
10. Thuy, V., Aw, A., Zhang, Min.: Term extraction through unithood and termhood unification. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), Hyderabad, India (2008).
11. Piao, S. L., Rayson, P., Archer, D., McEnery, T.: Comparing and Combining A Semantic Tagger and A Statistical Tool for MWE Extraction. Computer Speech & Language Volume 19, Issue 4, (2005) 378-39715.
12. Sharoff, S.: What is at stake: a case study of Russian expressions starting with a preposition. In: Proceedings of the Second ACL Workshop on Multiword Expressions Integrating Processing, July (2004).