Automatic Identification of Lexical Units

Vidas Daudaravicius
Vytautas Magnus University, Faculty of Informatics
Vileikos 8, Kaunas, Lithuania
E-mail: vidas@donelaitis.vdu.lt

Keywords: lexical unit, lexical unit identification, token/type ratio, dice score, corpus size, average minimum law

Received: March 3, 2009

Lexical unit is a word or collocation. Extracting lexical knowledge is an essential and difficult task in NLP. The methods of extracting of lexical units are discussed. We present a method for the identification of lexical boundaries. The problem of necessity of large corpora for training is discussed. The advantage of identification of lexical boundaries within a text over traditional window method or full parsing approach allows to reduce human judgment significantly.

Povzetek: Opisana je metoda za avtomatično identifikacijo leksikalnih enot.

1 Introduction

Identification of a lexical unit is an important problem in many natural language processing tasks and refers to the process of extracting of meaningful word chains. The Lexical unit is a fuzzy term embracing a great variety of notions. The definition of the lexical unit differs according to the researcherŠs interests and standpoint. It also depends on the methods of extraction that provide researchers with lists of lexical items. Most lexical units are usually single words or constructed as binary items consisting of a node and its collocates found within a previously selected span. The lexical unit can be: (1) a single word, (2) the habitual co–occurrence of two words and (3) also a frequent recurrent uninterrupted string of words. Second and third notion refers to the definition of a collocation or a multi–word unit. It is common to consider a single word as a lexical unit. A big variety of the definition of the collocation is presented in Violeta Seretan work [12]. Fragments of corpus or strings of words consisting of collocating words are called collocational chains [7]. For many years the final agreed definition of the collocation is not made. Many syntactical, statistical and hybrid methods have been proposed for collocation extraction [13], [1], [5], [4]. In [10], it is shown that MWEs are far more diverse and interesting than is standardly appreciated. MWEs constitute a key problem that must be resolved in order for linguistically precise NLP to succeed. Although traditionally seen as a language independent task, collocation extraction relies nowadays more and more on the linguistic preprocessing of texts prior to the application of statistical measures. In [14] it is provided a language-oriented review of the existing extraction work.

In our work we compare Dice and Gravity Counts methods for the identification of lexical units by applying them under the same conditions. The definition of what is a Lexical Unit in a linguistic sence is not discussed in this paper. New presented technique extracts collocations like 'in the' that do not have meaning and have functional purpose. A question of keeping such collocations as lexical units is left open. At the same time, it is interesting to see that the frequency lists of such lexical units for English and Lithuanian (memeber of Balto-Slavonic language group) are now comparable.

2 Extracting vs. abstracting

Most of the collocation definitions refer to the collocation, which is constructed in an abstracting way. The collocations are not gathered directly from the text but rather constructed using syntactic and statistical information. The abstracted collocation is constructed using statistical information extracted from the corpus. The extraction of statistical information from a corpus is only the first step for constructing collocations. The process of constructing the collocation is called as a collocation extraction in many research works. In this paper we make difference between the extraction of collocation and the abstraction of colocation. The major difference between abstracting and extracting of collocation is the use of lexical boudaries. The extractive technique for the identification of lexical units takes a linear approach of consecutive counts of words in a text and of all the texts in a corpus. Thus calculations of combinability are applied to the continuous chain of words. The first step is to detect the strength of combinability for pairs of words in the corpus, the second step is to detect the boundaries of the lexical units. The Extractive technique marks lexical boundaries in the text and a word or a word chain between these boundaries is a lexical unit. A clear idea about the boundaries of the lexical units allows to determine the exact size of a corpus lexicon. Abstractive technique uses a statistical information extracted from a corpus and a definition of a threshold for a good lexical unit. The threshold in many cases is frequency. Tagging and parsing are also used for filtering out invalid results [8]. Both, abstractive and extractive, techniques use associative measures to evaluate the combinability of two items. A new technique is presented to solve the problem of identification of (uni-)multiword lexical units without any linguistic knowlegde when full automatization is necessary. Extractive technique is very practical, easy to implement, and could improve quality of results in many IR and IE tasks. Nevertheless the results can be used for lexicografical tasks.

3 Combinability measures

Two different statistical calculations of collocability counts are applied (Dice and Gravity Counts)in this work. A good overview of combinability methods is presented in [3].

3.1 Dice score

The Dice coefficient can be used to calculate the co-occurrence of words or word groups. This ratio is used, for instance, in the collocation compiler XTract [11] and in the lexicon extraction system Champollion [6]. It is defined as follows [11]:

$$Dice(x,y) = \frac{2 * f(x,y)}{f(x) + f(y)}$$

$f(x,y)$ being the frequency of co-occurrence of x and y, and $f(x)$ and $f(y)$ the frequencies of occurrence of x and y anywhere in the text. If x and y tend to occur in conjunction, their Dice score will be high. The logarithm is added in order to discern small numbers. Thus the formula is slightly modified. The combinability of the each pair of words using this method was measured on the basis of the formula:

$$Dice'(x,y) = log_2\left(\frac{2 * f(x,y)}{f(x) + f(y)}\right)$$

The human have to set the level of collocability manualy. We set the level of collocability at the Dice minus 8 in our experiments. This decision was based on the shape of the curve found in [3].

3.2 Gravity counts

Gravity Counts are based on the evaluation of the combinability of two words in a text that takes into account a variety of frequency features, such as individual frequencies of words, the frequency of pairs of words and the number of types in the selected span. Token/type ratio is used slightly different. Usually this ratio is used for the whole document. The difference is that the token/type ratio is calculated not for a document or a corpus but for a word within a selected span only. In our experiments we used the span equal to 1. The expression of Gravity Counts is as follows:

$$G(x,y) = log\left(\frac{f(x,y) * n(x)}{f(x)}\right) +$$

$$+log\left(\frac{f(x,y) * n'(y)}{f(y)}\right)$$

(x,y) is the frequency of the pair of words x and y in the corpus; $n(x)$ is a number of types to the right of x; $f(x)$ is the frequency of x in the corpus; $n'(y)$ is the number of types to the left of y; $f(y)$ is the frequency of y in the corpus. We set the level of collocability at the Gravity Counts 1 in our experiments. This decision was based on the shape of the curve found in [3].

4 Identifying the boundaries of lexical units

There were attempts to extract recurrent uninterrupted strings of unlemmatized word-forms [7]. The chains were identified purely on the ground of row frequency and consisted of chains up to five words in length. However, the applied method did not reveal whether the extracted chains are made of collocating words. In our case, the detection of the boundaries of a lexical unit is based on a full text approach. The idea behind this approach is that the corpus is used as a very long chain of words to calculate the combinability of adjacent word pairs. The counts starts with the first and ends with the last word of the corpus. Thus, the corpus is seen as a changing curve of the lexical combinability. Peaks, appearing above the point of a selected value are taken as collocability points that can form lexical units (see Figure 1 for an example of a sentence). Using a text as the basis for the identification of lexical units with the help of the collocability points allows detecting the boundaries of each lexical unit. A lexical unit is defined as a segment of text where the combinability of constituent adjacent word pairs is above the arbitrarily chosen point of collocability. The lower combinability of word pairs preceding and following the segment marks the boundaries of a lexical unit. The list of all such segments from the corpus is the list of its lexical units. Moreover, we introduce two new additional definitions of the boundary of lexical unit. We call them absolute minimum and average minimum laws.

4.1 Average minimum law

In addition to the collocability requirement the average minimum law can be applied. This law is applied to the three adjacent collocability points. The law can be expressed as follows: if $\frac{x_{-2}+x_0}{2} > x_{-1}$ then the boundary of a lexical unit is set at x_{-1} . The boundary of a lexical item is set, if the average of values of collocability points on both sides are higher. This law allows making additional boundaries of lexical units when collocability points are set. The identified lexical units are shorter and more clearcut (see Figure 2 for an example of a sentence).

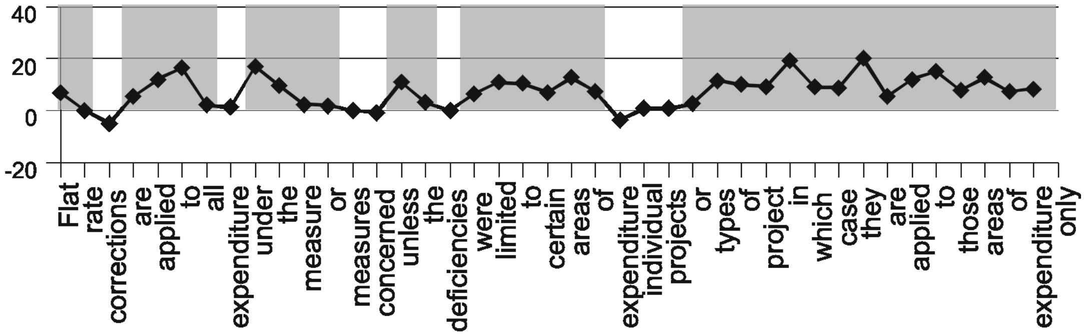

Figure 1: Identified lexical units of an example sentence, combinability values and collocability level at value 1. / Flat rate / corrections / are applied to all expenditure / under / the measure or measures / concerned / unless the deficiencies / were limited to certain areas of expenditure / individual / projects / or types of project in which case they are applied to those areas of expenditure only/

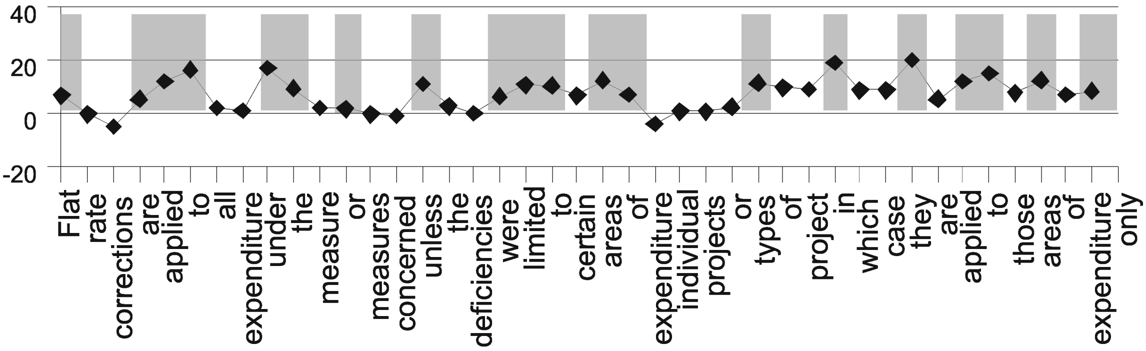

Figure 2: Identified lexical units of an example sentence, combinability values, collocability level at value 1 and average minimum law applied. / Flat rate / corrections / are applied to / all / expenditure / under the measure / or measures / concerned / unless the / deficiencies / were limited to certain / areas of expenditure / individual / projects / or / types of / project / in which / case / they are / applied to those / areas of / expenditure only /

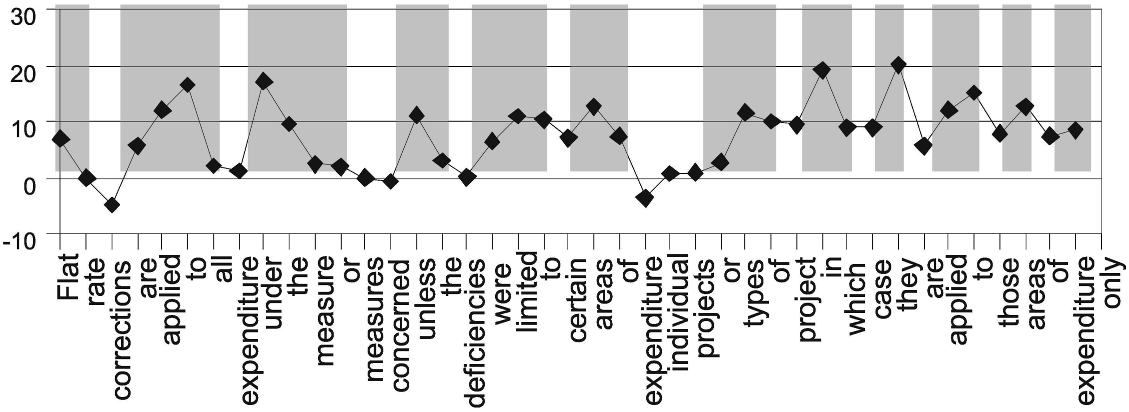

Figure 3: Identified lexical units of an example sentence, combinability values, collocability level at value 1 and absolute minimum law applied. / Flat rate / corrections / are applied to all expenditure / under the measure or measures / concerned / unless the / deficiencies / were limited to certain / areas of expenditure / individual / projects / or / types of project / in which case / they are / applied to those / areas of / expenditure only /

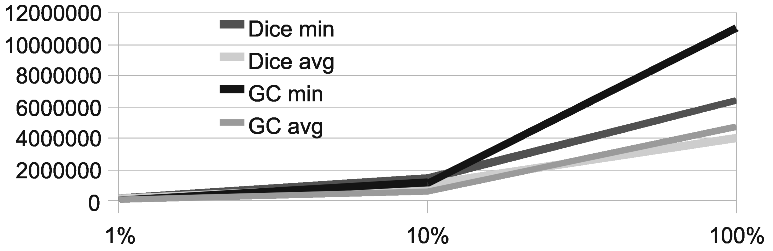

Figure 4: The number of lexical units (types) in the selected corpus (*x-axis has logarithmic scale*)

Gravity Counts average minimum law			Dice average minimum law		
100%	10%	1%	100%	10%	1%
of the	the	and	and	and	and
in the	and	the	the	the	the
and	in the	of	of the	of the	of
to the	the	of the	of	of	of the
on the	of	in the	in the	in the	in the
the	to the	to	in	in	to
at the	in	in	to	to	in
for the	on the	a	a	a	a
and the	to	for	to the	to the	to the
to be	or	’s	’s	on the	that
of	’s	to the	on the	’s	for
or	for	that	for	for	on the
in	a	is	that	that	to be
by the	that	by	or	or	and the
of a	and the	was	for the	for the	’s
from the	is	with	is	and the	or
with the	for the	or	and the	is	for the
that	with	on	to be	to be	was
to	at the	on the	was	at the	with
it was	by	from	at the	with	is

Table 1: The top 20 lexical units for different size of corpus and scores

100%	10%	1%
she might	she might	she might
have been	have been the	have been
the	headmistress	the
headmistress	of a	headmistress
of a	certain type of	of a
certain	girls ’	certain type of
type of	school	girls ’
girls ’	, now	school
school	almost extinct	, now
, now almost extinct	, or a	almost extinct
, or a	mother superior	, or a
mother superior	in an	mother superior
in an	enclosed order	in an
enclosed	.	enclosed order
order		.
.		
at any rate	at any rate	at any rate
there could be	there could be	there could be
no doubt	no doubt	no doubt
that she had	that she had	that she had found
found	found	the
the	the	temptation
temptation	temptation	of the
of the	of the	flesh resistible
flesh resistible	flesh resistible	.
.	.	

Table 2: The boundaries of lexical units identified by Dice

4.2 Absolute minimum law

In addition to the collocability requirement the average minimum law can be applied. This law is applied to the three adjacent collocability points. The law can be expressed as follows: if $x_{-2} > x_{-1} \wedge x_0 > x_{-1}$ then the boundary is set at x_{-1}. Informally, the boundary of a lexical item is set, if the values of collocability points on both sides are higher. The identified lexical units are wider compared to the average minimum low (see Figure 3 for an example of a sentence).

5 Experiments and results

We used whole British National Corpus for experiments. Three corpora sizes were used in experiments: whole, 10% and 1% of the corpus. We used row text without any tagged information.

5.1 Dictionary size

The number of lexical units identified in the corpus using the respective methods is presented in Figure 4. The number of lexical units extracted with the help of Dice and Gravity Counts scores using average minimum law is similar. The absolute minimum low yields to the different size of the dictionary. The result of the number of lexical units shows the trend line of possible total number of lexical units. We can expect maximum of about 5-6 million of lexical units in English using Dice score and average minimum law. This number is comparable to different languages, e. g., Lithuanian with rich morphology and almost free word order. In [9] the number of word types in corpus comparable to BNC is 1.8 million. In [8] the number of extracted collocations using similar method from Lithuanian

100%	10%	1%
she might have been the headmistress of a certain type of girls ’ school , now almost extinct , or a mother superior in an enclosed order .	she might have been the headmistress of a certain type of girls ’ school , now almost extinct , or a mother superior in an enclosed order .	she might have been the headmistress of a certain type of girls ’ school , now almost extinct , or a mother superior in an enclosed order .
at any rate there could be no doubt that she had found the temptation of the flesh resistible	at any rate there could be no doubt that she had found the temptation of the flesh resistible .	at any rate there could be no doubt that she had found the temptation of the flesh resistible .

Table 3: The boundaries of lexical units identified by Gravity Counts

corpus is 20 millions. We used new laws of minimum in our experiments. It is obvious that if the law of average minimum would be applied in [8] work then the number of collocations would drop to 6-7 millions or more for Lithuanian. Thus we are able to speak about the similar number of lexical units which could be applied for any language. For instance, the machine translation system ATLAS-II v.13 by Fujitsu has 5.44M terminological entries and about 1M to 1.5M general entries [2].

5.2 Top 20 lexical units

Another goal of our research is to discover which score is less sensitive to the corpus size. The size of corpus differs in applications. In [3] is shown that Mutual Information score heavily depends on the corpus size and it is very difficult to set the level of collocability. Dice and Gravity Counts scores do not consider corpus size. We performed several experiments to compare method dependability on the size of the corpus. We used Dice and Gravity Counts score together with the average minimum law on the different corpus sizes. We took 1%, 10% and full corpus of BNC. We built the dictionaries of lexical units and took top 20 lexical units for every case. The results are shown in Table 1. The list of top 20 lexical units identified using Dice score almost does not change. While the list of lexical units identified using Gravity Counts changes. This is sufficient to state that Dice score is stable, not sensitive to the corpus size and is reliable in many NLP applications. This statement is confirmed by the examples in Table 2 and Table 3. The same two sentences are taken and the boundaries of lexical units are identified. The law of average minimum is used. We can see that the identified boundaries of lexical units using Dice score do not change considerably. While in case of Gravity Counts the change of boundaries is observable often.

6 Conclusions

The numbers of lexical units in most languages is comparable and amounts to 6-7 millions. It should be applicable for the most of indoeuropean languages. The lexical unit is very important in NLP and is applied widely. But the notion of lexical unit is not clear and hard to define. We propose a definition of a lexical unit as a sequence of word-forms extracted from row text by using collocability feature and setting boundaries of lexical units. This approach is more clear compared to a widely used n-gram definition of a lexical unit. The boundaries are predictable and easier controlled compared to n-gram model. The result of setting lexical boundaries for the small and large corpora are stable using Dice score. Thus Dice score is reliable in many NLP applications. The average minimum law allows making additional boundaries of lexical units when collocability points are set. Identified lexical units are shorter and more clearcut. Human judgment on the boundaries of lexical unit is reduced considerably as the setting of collocability level is not so sensitive when the average minimum law is applied.

References

[1] Brigitte Orliac, Mike Dillinger (2003) Collocation extraction for machine translation, *MT Summit IX*, New Orleans, USA, 23-27 September 2003, pp.292-298

[2] Christian Boitet, Youcef Bey, Mutsuko Tomokio, Wenjie Cao, Hervé Blanchon (2006) IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations,ă *International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation [IWSLT 2006]*, November, 27-28, 2006, Kyoto, Japan, pp.23–30.

[3] Daudaravicius V., Marcinkeviciene R. (2004) Gravity Counts for the Boundaries of Collocations, *International Journal of Corpus Linguistics, 9(2)*, pp. 321Ű-348.

[4] Dekang Lin (1998) Extracting collocations from text corpora, *In First Workshop on Computational Terminology*, Montreal, 1998, pp. 57Ű-63.

[5] Gael Dias (2003) Multiword unit hybrid extraction, *In Proceedings of the ACL Workshop on Multiword Expressions*, Sapporo, Japan, pp. 41Ű-48.

[6] Marcinkeviciene R., Grigonyte G. (2005) Lexicogrammatical Patterns of Lithuanian Phrases, *The Second Baltic Conference on Human Language Technologies proceedings*, Tallinn, pp. 299Ű-305.

[7] Rimkute E., Daudaravicius V., Utka A. (2007) Morphological Annotation of the Lithuanian Corpus, *45th Annual Meeting of the Association for Computational Linguistics; Workshop Balto-Slavonic Natural Language Processing*, Praha, pp. 94–99.

[8] Smadja, F. (1993) Retrieving Collocations from Text: XTRACT, *Computational Linguistics, 19(1)*, pp. 143-177.

[9] Smadja, F., McKeown, K. R. and V. Hatzivassiloglou (1996) Translation Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics, 22(1)*, pp. 1-38.

[10] Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger (2002) Multiword Expressions: A Pain in the Neck for NLP, *In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico, pp. 1–15

[11] Stubbs, M. (2002) Two quantitative methods of studying phraseology in English, *International Journal of Corpus Linguistics, 7(2)*, pp. 215-244.

[12] Violeta Seretan (2008) Collocation Extraction Based on Syntactic Parsing, *Ph.D. thesis*, University of Geneva.

[13] Violeta Seretan and Eric Wehrli (2006) Accurate collocation extraction using a multilingual parser, *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, Australia, pp. 953-Ű960.

[14] Violeta Seretan and Eric Wehrli (2006) Multilingual collocation extraction: Issues and solutions. *In Proceedings of COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, 2006, Sydney, Australia, pp. 40-Ű49.

Appendix 1. *Risk* Lexical units extracted using Gravity Counts and average minimum law

1% of BNC (1 million words)	
Lexical unit	Frequency
increase the risk of	1
risk	51
risk of	8
the risk	9
the risk of	5
the risk of another	1

Appendix 2. *Risk* Lexical units extracted using Dice and average minimum law

1% of BNC (1 million words)	
Lexical unit	Frequency
at risk	9
calculated risk	1
currency risk	1
environmental risk	1
health risk	1
particular risk	2
primary risk	1
real risk	1
reducing risk	1
risk	13
risk being	1
risk being disappointed	1
risk being considered	1
risk losing	2
risk undermining	1
risk using	1
risk of	7
risk than being overweight	1
risk than an asset	1
risk factor	1
risk slipping	1
risk arguing	1
risk assessment	1
risk her	1
risk her rage	1
risk damage	1
risk missing	1
risk cream	1
risk element	1
serious potential risk	1
serious risk	1
safety risk	1
the risk	5
the risk of	6
the risk compared with	1
the great risk of	1
were at risk	1