Computing Idioms Frequency in Text Corpora

Jan Bušta

Faculty of Informatics, Masaryk University, Brno, Czech Republic
xbusta@fi.muni.cz

Abstract. The idioms are phrases which meaning is not composed from the meanings of each word in the phrase. This is one of the natural examples of violating the principle of compositionality that means that idioms are in area of natural language processing problem of meaning mining. To count the frequency of phrases such idioms in corpora has one big aim: To get to know which phrases we use often and which less. We do it to be able to start with getting the meaning of the whole phrases not just each word. This improves the understanding natural language.

1 Idioms

First step how be able to count the idiom frequency is to recognize the idiom. Although we have a dictionary (Slovník české frazeologie a idiomatiky), the idioms are in corpus in non-standard form. The question, what is and what is no more an idiom, is quite difficult to define. The word order in idioms is not fixed, there is for example an idiom *hrát druhé housle*[1] you find it in this form in the dictionary, but in the real text, the idiom could be *. . . já druhé housle hrát nehodlám*. It is shown that the word order is switched, the noun phrase is in the front of the verb phrase.

The structure of idioms in the dictionary are well primary divided in two categories:

– **the verb-based idioms and phrases** – the main word is a verb and the other words, which are optional, create only the context or/and further qualification,

  Example 1. hnout / pohnout někomu žlučí, dostat zaslouženou odměnu, být odkázán / bejt vodkázanej sám na sebe

– **the non-verb-based idioms and phrases** – the main meaning depends on the non-verbal word, very often on nouns (substantives), or is composed form the many words which are on the same meaning level.

  Example 2. bída s nouzí, brána pekla / do pekla / do pekel, bez velkých / dlouhých / zbytečných ceremonií

Note 1. Whether the phrase is verbal- or non-verbal-based depends on the language, the same meaning could be said in different forms in each language.

[1] All idioms in text are from SČFI [1]

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 71–74, 2008. © Masaryk University, Brno 2008

This division is very useful because the approach to this two groups is totally different. While for the verb-based idioms should the query take into account possible noun- and verb-phrase switching, by the non-verb-based idioms is the situation easier, because the word switching in this case is not common (there are some exception, but the idioms created this way sounds archaic, e. g. *země dary*).

Changing the case, gender or tense in idioms is not a real computing problem. Some idioms can occur in a specific positional form only, the phrase has an idiomatic meaning only if the words are next to each other, usually is that an adjective and a substantive which specify the base word, but in some idioms is also not fixed the number of words. You can start with the first part (whatever the first part is), insert some words which can specify the meaning, and finish the original phrase (see next example). This makes some problems with selecting the idiom in text.

Example 3. . . . , že by měla v budoucnu ***hrát*** *ve svém regionu* ***druhé housle****. . .*

1.1 Frege's principle

Idioms are the phrases which violating Frege's principle of compositionality. This fact makes this work meaningful, because we are not able to translate (or get the sense) from the sentence if it contains an idiom; first we have to define the meaning of the parts in sentence thereafter is the clear road to processing. Idioms can not be processed as it is but have to be preprocessed. The easiest way it to give them the fixed meaning/translation. It can be hard to work with all idioms therefore we will select the most frequented idioms.

2 Corpora

The primary corpus data come form the SYN2000c corpus made by Institute of the Czech National Corpus at Charles University in Prague. This corpus includes more than 100 millions words from complete texts sources. The SYN2000c is a synchronous corpus, which means that most the documents added into it have been published in the same time (1990–1999). The SYN200c corpus contains also some older documents from authors which were born after 1880. [2]

Using this corpus provides very good overview on the Czech language, the solution corresponds with today's language. Another advantage is that this corpus is tagged therefore is no problem with forms of words in any case, gender or tense.

3 Querying

The main work is in the querying. How to query the corpus if it contains the given idiom or not and if yes how many times. There is no problem, to get "some result", but it happens, that we have to think about the idiom structure and adapt the query to the idiom we are looking.

For querying the corpus is the Corpus Query Language[3] which gives the power to create complex queries. It allows to specify the context or distance of each words in idioms. The result reflects very sensitively the used query.

Next example show how search the verb-based idiom *hrát druhé housle* in corpus.

Example 4.
```
[lemma="hrát"][word!="\."]{0,5}"druhé""housle"|
"druhé""housle"[word!="\."]{0,5}[lemma="hrát"]
```

The result of this query will be a list of phrases which begins with any from of the first part (in this this case the verb *hrát*, next word can be everything else except the dot sign, there should be 0–5 words this type (inserted words) and at the and is the other part of idiom which are consist from the two contiguous words *druhé* and *housle*. The other part implements the switching of the parts of the idiom. The count of found idioms using this query is 47, but if we searching only the occurrence of the word phrase *druhé housle*, the count will be 52. The difference between this two results is in the fact, that the phrase *druhé housle* can be used separately in the non-idiomatic meaning (music stuff). Knowing the right borders of idiom will decide, if the phrase is an idiom or not.

The situation in the field of non-verb-based idioms seems to be easier, but there are other things which could be solved. Many of non-verb-based idioms have more than one. The structure of them can consist from the static part and the part which could be changed. This second part is created from word (or words) which have the same/near meaning. The idioms *dar řeči*, *dar jazyka* and *dar výmluvnosti* are according to the dictionary the same. This idioms should be detect and searched as:

Example 5.
```
[lemma="dar"][word="řeči|jazyka|výmluvnosti"]
```

In this idiom example is also impossible to divide the idiom in two parts, this is the property of majority of non-verb-based idioms.

A special group of idioms are one-word idioms, e. g. *žízeň*. In this case is the frequency of idiom identical to the plain frequency of the word which is not exact. Many of occurrences are the words in his base meaning, to divide the base and the idiomatic meaning of the word is context dependent. There is no solution how to recognize this idioms without any other supporting method.

In the idioms written in SČFI are sometimes a word (words) of idiom in non-literary form. It would not be if the *lemma* will match the *lemma* of the literary equivalent, but there are not, it makes difficulties by searching in corpus.

3.1 Dividing the idioms

To find an automatic procedure of making the queries is the good classifying the idioms and do more specific groups of idioms than verb- and non-verb-based in which would be such of them which satisfy the prepared slots for making the final query.

It seems to by useful (in case of verb-based idioms) to divide them by their possibility to change the position of parts of the idioms, by possibility to accept the inserted words in the middle, by the alternatives parts (synonymic phrases inside the idiom). One of the important variable, which should be set, is the maximal distance of the parts of idioms. If this value will be to slow, some idioms will be not found; if it will be to high, some other phrases will be found which will be not idioms.

Non-verb-based idioms are easier to recognize also the dividing in groups linked with their structure. For the non-verb-based idiom groups is the main characteristic if there is a alternative part or words have to be side-by-side.

4 Conclusion

The processing the idioms, cleaning the data and preparing the form of the idioms is the fist step to be able to create the queries which will match the highest count of the idioms without matching any other phrases which are in the corpus but are not idioms.

The second step is prepare the skeleton of query, the slots have to be as much as possible specific (to accept only the right group of idioms).

After doing the previous steps we can start with the querying the corpus. The results are only very hard to evaluate in particular because there are no results done. We can sure compare results of some idiomatic searches but not all results.

Maybe will be some methods used also by computing the frequency of idioms in other language although the structure of idioms is language dependent.

Acknowledgments. This work has been partly supported by the Academy of Sciences of Czech Republic under the projects 1ET100300419 and 1ET200610406, by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 407/07/0679.

References

1. Čermák, F., collective: Slovník české frazeologie a idiomatiky. Academia (1994).
2. Institute of the Czech National Corpus: Structure of technical and other specialised literature according to the technical orientation (2008) [Online; accessed 12-November-2008].
3. Christ, O.: A modular and flexible architecture for an integrated corpus query system. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart (1994).