Type-based Search of Idiomatic Expression

Jan Bušta

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
busta@fi.muni.cz
http://nlp.fi.muni.cz/

Abstract. This paper presents evaluation of different approaches to extract verb-noun idiomatic expressions in Czech. These approaches are based on the structure of the idiom and its behavior in language. PMI and syntactic and lexical fixedness modified using VerbaLex and generated thesaurus provide useful tool for choosing best idiomatic candidates for manual annotation and evaluation. Moreover we focused on general adapting the algorithms for Czech.

Key words: idioms, idiomatic candidates, syntactic fixedness, lexical fixedness, transitive verbs, thesaurus

1 Introduction

In any language there are always multi word expressions which are about to breake the Frege's Principle of compositionality. Let's call them idioms. But, it is not so easy to determine, if the compound word expression breaks this principle or not.

The need of determining such language segments is obvious: Every time we deal with machine translation from one language to another, we deal with this non word-by-word translation. The words of idioms cannot be translated by this basic approach, we have to mark them and handle separately. Lexisting lexicons of idiomatic phrases are always old and does not allow to search the today's language. Computer processing helps to build this kind of lexicons very fastly with the minimum amount of time needed by the lexicographers.

There are many approaches, how to define the idiom better, but there will be always the problem with the decision which is individual for every language user (or language user group). We based this work on the weakest presumption of Principle of compositionality and provide the language user more reliable data sources to determine the border line, therefore any time we speak about idiom, we just mean idiomatic candidates.

In further chapters we will present the basics of approaches to automatic idiom search, modified algorithms, the differences to approaches to English and evaluation of the methods.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 93–96, 2013. © Tribun EU 2013

2 Type-based: Verb-noun

In this article we present the approaches to type-based idioms. It defines our working area as the idiomatic phrases which consist from transitive verb and noun in accusative case (the transitive verb requires direct subject in accusative case). We can be sure, there is an object (in accusative case) for all transitive verbs we will handle.

Let's define the idiomatic phrase as a tuple $< v, n >$, where v is a transitive verb and n is the object.

3 The base: Fixedness

Talking about the Frege's principle we assume, it could represented by fixedness. The more fixed the phrase is, the more idiomatic behaviour we expect, so we are convinced that the idiomacity and the language fixedness are related and correlating.

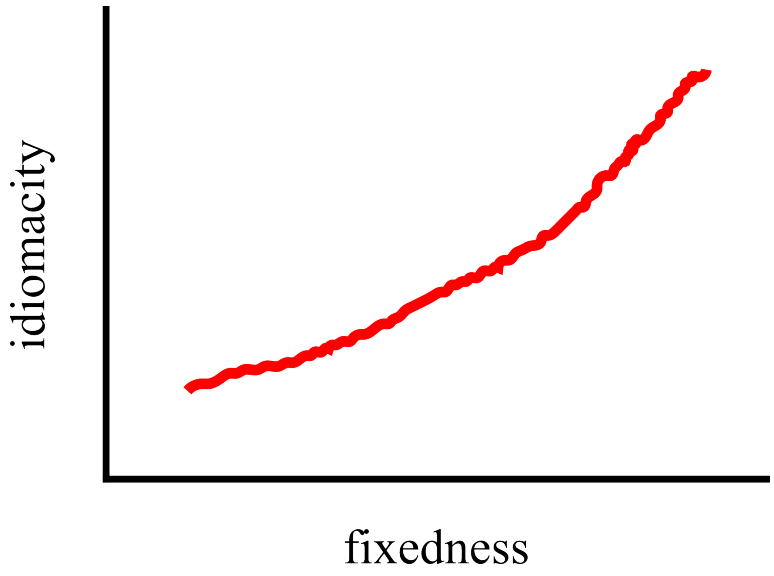

Fig. 1: Fixedness and idiomacity correlation.

The fixedness is much easier to implement as an algorithm for computer processing than the Frege's principle itself. There are multiple metrics available to show the fixedness of the phrase.

First we build $\mathcal{V}$, the set of transitive verbs acquired from *VerbaLex*. For all these verb we search the corpus for concordances in the form of $< verb, noun >$ or $< verb, intersegment, noun >$ with preserving the accusative case restriction. The intersegment can be adjectives or pronouns.

So we have all the candidates phrases and further algorithms show us the fixedness of each pair found in corpus.

We construct distributional thesaurus based on the word sketches for every noun n, let the set be $\mathcal{T}_n$, and let the $f(v, n_j)$ be frequency, where v is the transitive verb of candidate pair $< v, n >$ and n_j is element of $\mathcal{T}_n$. Let $f(v, *) = \sum_{n_j \in \mathcal{T}_n} f(v, n_j)$ and $f(*, n_j) = \sum_{v \in \mathcal{V}_n} f(v, n_j)$. $\mathcal{N}$ is set of all nouns found in corpus and being object of any transitive verb.

One of the metrics we used is PMI, which is defined for all pairs as:

$$PMI(v, n_j) = log \frac{|\mathcal{V} \times \mathcal{N}| f(v, n_j)}{f(v, *) f(*, n_j)}$$

The lexical fixedness is based on PMI and defined as

$$Fix_{lex}(v, n) = \frac{PMI(v, n) - \overline{PMI}}{s}$$

The $\overline{PMI}$ is the mean of all $PMI(v, n_j)$ and s is the standard deviation. This approach is motivated by the fact, that idiomatic phrases consist of such a word collocation which is significantly different from the others. Fazly and Stevenson define the set $\mathcal{T}_n$ not as thesaurus, but as a set of synonyms to the noun n. We assume, that if the noun from idiom can be replaced by the noun from thesaurus, the fixedness is lower than by using the set of synonyms.

The last algorithm we present is measuring the syntactic fixedness. Fazly and Stevenson approach is based on syntactic changes of the phrase: passivation, pluralisation and type of article, but those features are not feasible for Czech. Handling the passivation in Czech was not selected because the corpus searching queries were not prepared for this purpose. The free word order makes this task difficult if one wants to be precise. The type of article is not significant even, the Czech language do not use it. The pluralisation is the only one of the syntactic fixedness computation parameters which could be used in Czech, but according to the fact, that this is the only one, another approach is presented: To measure the syntactic fixedness we observed the changes in intersegment length. Let the $f_i(v, n) = f(v, n)$ where i is the length of intersegment. The syntactic fixedness of the pair as follows:

$$Fix_{syn}(v, n) = \frac{max(f_0(v, n), f_1(v, n), f_2(v, n))}{\sum_{i=0}^{2} f_i(v, n)}$$

This shows us that the phrase is more fixed (idiomatic) if most of the occurrences are in just one form. The length of the intersegment is fixed and language does not allow the flexibility by changing the internal structure of the idiom.

4 Conclusions

All three approaches produce lot of candidate phrases, but it is much easier for human evaluation to read the candidate phrases instead of searching the whole text for the idioms.

Modifying the algorithms for Czech was beneficial so that we can use it also for other Slavic languages and generate the candidate dictionary.

The evaluation showed that with comparing to the existing lexicons of idiomatic phrases, there were the verb-noun pairs on the top which were not in the lexicon (this has been marked by annotators).

Distinguishing of idiomatic phrases is a complex task and our work shows that we can easily modify the current approaches and make them better for other languages.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Bannard, Colin. *A measure of syntactic flexibility for automatically identifying multiword expressions in corpora* In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, 2007.
2. Bušta, Jan. *Výpočet četnosti výskytů hesel SČFI v korpusu*. Bakalářská práce, Fakulta informatiky, Masarykova univerzita, Brno, 2009.
3. Čermák, František. *Frazeologie a idiomatika česká a obecná*. Praha: Nakladatelství Karolinum, 2007.
4. Čermák, F., Hronek, J. – editors. *Slovník české frazeologie a idiomatiky*. Praha: Academia, 1994.
5. Čermák, F. a kol. *Slovník české frazeologie a idiomatiky 3, Výrazy slovesné*. Praha: Leda, 2009.
6. Fazly, A. a Stevenson, S. *Automatically constructing a lexicon of verb phrase idiomatic combinations*. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*. 2006.
7. Fellbaum, Christiane. *The determiner in English idioms*. In *Idioms: Processing, structure, and interpretation*. Hillsdale: Lawrence Erlbaum Assiciates, 1993.
8. Hlaváčková, D. a Horák, A. *VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech*. In *Proceedings of the Computer Treatment of Slavic and East European Languages 2005*. Bratislava, 2005.
9. Horák, A., Rychlý, P., Kilgarriff, A. *Czech word sketch relations with full syntax parser*. In *After Half a Century of Slavonic Natural Language Processing*. Brno: Masaryk University, 2009.
10. Karlík, P., Nekula, M., Rusínová, Z. – editors. *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny, 2008.
11. Kilgarriff, A., Rychly, P., Smrz, P. a Tugwell, D. *The Sketch Engine*. In *EURALEX Proceedings 2004*. Lorient, 2004.
12. Lin, D. *Automatic identification of non-compositional phrases*. In *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 1999.
13. Rychlý, Pavel. *Korpusové manažery a jejich efektivní implementace*. Doktorská práce, Fakulta informatiky, Masarykova univerzita, Brno, 2000.