Mining the Web for Idiomatic Expressions Using Metalinguistic Markers

Filip Graliński

Faculty of Mathematics and Computer Science,
Adam Mickiewicz University,
ul. Umultowska 87, 60–687 Poznań, Poland,
filipg@amu.edu.pl

Abstract. In this paper, methods for identification and delimitation of idiomatic expressions in large Web corpora are presented. The proposed methods are based on the observation that idiomatic expressions are sometimes accompanied by metalinguistic expressions, e.g. the word "proverbial", the expression "as they say" or quotation marks. Even though the frequency of such idiom-related metalinguistic markers is not very high, it is possible to identify new idiomatic expressions with a sufficiently large corpus (only type identification of idiomatic expressions is discussed here, not the token identification). In this paper, we propose to combine infrequent but reliable idiom-related markers (such as the word "proverbial") with frequent but unreliable markers (such as quotation marks). The former could be used for the identification of idiom candidates, the latter – for their delimitation. The experiments for the estimation of recall upper bound of the proposed methods are also presented in this paper. Even though the paper is concerned with identification and delimitations of Polish idiomatic expressions, the approaches proposed here should also be feasible for other languages with sufficiently large web corpora, English in particular.

Key words: idiomatic expressions, Web mining

1 Introduction

Identification of idiomatic expressions (idioms) of a given language is of importance from both theoretical and practical perspectives. First, idiomatic expressions can provide insight into the history and culture of their users [1]. Second, idioms, due to their semantic (and sometimes even syntactic) idiosyncrasy, present a challenge in many NLP applications, especially in machine translation – on the one hand, idiomatic expressions are abundant in informal and semi-formal Internet texts, such as blog entries, message board threads, Facebook posts, on the other hand, idioms are not that frequent in typical parallel corpora[1]. Therefore it would make sense to try to enlist idiomatic expressions and

[1] Parallel corpora derived from film subtitles do contain substantial number of informal idiomatic expressions, the quality of their translations is, however, very poor. Idioms are very often mistranslated or rendered literally (!) by incompetent translators.

annotate them manually (e.g. in the context of machine translation – translate them manually and create a bilingual lexicon).

We focus here on *type-based identification*, i.e. the aim is to collect as many idiom types as possible rather than to decide whether a given expression is used idiomatically or literally in a specific context (*token-based classification*). In order to find idiomatic expressions in large Web corpora, idiom-related metalinguistic markers can be used. *Idiom-related metalinguistic markers* are expressions used by speakers of a given language to metatextually mark idiomatic expressions, e.g. words and expressions such as *proverbial*, *as it is said*, *as they say* [2]. Obviously, occurrences of idioms marked with idiom-related metalinguistic markers are rather infrequent. Although this makes them rather useless in token-based classification context, such markers can be used in type-based identification provided that a corpus is large enough to comprise a significant number of idiom-related markers.

This paper is concerned mainly with Polish idiomatic expressions (usually referred to as *frazeologizmy* or *związki frazeologiczne* in Polish linguistic literature).

2 Related Work

There is a rich research literature on the automatic identification of multiword expressions (MWEs), non-compositional expressions, collocations etc. (see e.g. [3]), idiomatic expressions represent, however, only a subset of all MWEs. [4] was specifically concerned with (both type- and token-based) identification of English idiomatic expressions. The method described in [4] was based on distribution of syntactic patterns of an idiom, i.e. variants of an idiom with respect to passivisation, determiners used, pluralisation. Such method is not fully applicable for all languages, for instance in Polish no determiners are used and passivisation is less frequent than in English. Furthermore, [4] focused only on some types of idioms (of the form V+NP).

Using metalinguistic markers for the identification of idiomatic expressions was proposed in [2]. The following Polish markers were considered:

- the adjective *przysłowiowy* (*proverbial*),
- the adverb *przysłowiowo* (*proverbially*),
- quotation marks,
- the phrase *tak zwany* (*so-called*) along with its abbreviated form *tzw.*,
- the phrase *jak to mówi* (*as they say*),
- the phrase *jak to się mówi* (*as it is said*),
- the adverb *dosłownie* (*literally*).

The list is probably exhaustive, with the exception of the markers *frazeologizm*, *związek frazeologiczny*, *idiom* (all denoting "idiomatic expression" in Polish) – such markers are used in texts showing more linguistic awareness and are more likely to accompany well-known (and less interesting) idioms, already recorded in dictionaries.

The adjective *przysłowiowy* (along with its adverbial form *przysłowiowo*) appears to be the most distinctive idiom-related marker, other markers were not used in [2]. It should be noted, however, that quotations marks are much more frequent than *przysłowiowy*. The problem is that they are used for many other purposes. Results reported in this paper concern combining *przysłowiowy* (reliable but relatively rare) with quotation marks (frequent but unreliable). In this way, we extend work reported in [2]. Also, some estimations of recall upper bound of methods proposed in [2] and in this paper will be given here.

Exploiting expressions indicative of idiomatic usages (such as *proverbially*, *metaphorically speaking*) and quotation marks was mentioned in [5]. That paper was, however, focusing on token-based classification and, as expected, such indicative terms were not a useful feature in the task reported there.

3 Corpus

The same corpus as in [2] were used (732M words). The corpus contained 29737 sentences with *przysłowiowy/-o* marker. This sentences were pre-processed in the following way:

- a given sentence was tokenised (PSI-Toolkit[2] was used for the segmentation),
- *przysłowiowy/-o* token was found,
- at most 8 tokens to the left and to the right of the *przysłowiowy/-o* token were taken.

This way, a set of snippets containing the word *przysłowiowy/-o* was obtained.

3.1 Test set

A random sample of 2000 snippets was selected. Snippets belonging to the development and test set used in [2] were excluded. (The test set of [2] was used as the development set now.)

In each snippet an idiomatic expression referred to by *przysłowiowy/-o* was manually tagged and delimited (if present at all). The number of all idioms marked in the corpus sample was 1008 (50.4%) – the *przysłowiowy/-o* marker is used for other (not idiom-related) purposes by Polish speakers as well.

The following criteria for idiomacity have been used [6]:

- semantic non-compositionality,
- lexicosyntactic fixedness,
- prevalence (in order to reject one-time, ad hoc expressions, phrases not listed in traditional paper dictionaries were checked with Web search engines – 5 independent occurrences were required for an expression to be classified as an idiom),

[2] `http://psi-toolkit.wmi.amu.edu.pl`

– "graphicalness" (*obrazowość* in Polish, a criterion traditionally put forth in Polish linguistic literature).

A substantial number (188) of idioms marked in the test set were discontinuous, i.e. split into two or more parts separated by some intervening material (mostly pronouns). For instance, in the snippet (the *przysłowiowy/-o* token is replaced with #):

Początkowo prowadzę uczniów wręcz za # rączkę
= At the beginning I lead the students just by the # hand

the idiom *prowadzić za rączkę* (= *lead by the hand*) was split into two parts (not counting discontinuity introduced by the idiom-related marker). Such discontinuous idioms are not likely to be properly delimited by the procedure described in Section 4, they were not, however, removed from the test set.

4 Identification and Delimitation Procedure

The identification and delimitation proposed in [2] was used as a baseline and starting point. Now we are going to describe briefly this procedure.

The procedure starts with locating the *przysłowiowy/-o* token in a given snippet. Then we attach tokens on the right until some condition is true. This way, the right boundary of an idiom is determined. Next, we switch to the left side and, in a similar manner, attach tokens until some condition is true. The alternative of the following conditions yielded the best results for the right side (as reported in [2]):

– the phrase gathered so far combined with the current token occurs only once in the corpus,
– the association strength between the phrase gathered so far and the current token (measured using pointwise mutual information, PMI [7], on the whole corpus) is below a threshold,
– the current token is a punctuation mark (excluding quotation marks),
– the current token is the second quotation mark.

The conditions for the left side are the same as those for the right side with one exception – the first quotation mark on the left stops the procedure as well.

After the main part of the procedure is finished, an extra filter is applied: words which are unlikely to occur at the beginning (mainly conjunctions) or the end (conjunctions and prepositions) of an idiom are stripped from the delimited expression.

Identification is based on the delimitation: a delimited expression is classified as an idiom if and only if it is composed of two or more tokens. (Note that idioms are, by definition, multi-word expressions).

4.1 Enhancements to the procedure

The procedure described in [2] is based on the *przysłowiowy/-o* metalinguistic marker. We decided to combine this with the most frequent (though less reliable) idiom-related marker – quotation marks. Note that in the original procedure proposed in [2] quotation marks are used *locally* (i.e. they are taken into account only if they occur in the same snippet as *przysłowiowy/-o*), now we are going to use them *globally*, searching for a given expression enclosed in quotation marks in the *whole corpus*. This enhancement will be applied in both the delimitation and identification step.

Delimitation As the delimitation step is concerned, we add a new stop condition, namely we stop attaching tokens if:

$$\frac{\Phi(\text{“ } t_{k+1} \dots t_{i-1} \text{ ”})}{\Phi(t_{k+1} \dots t_{i-1})} > \theta,$$

where $\Phi(t_{k+1} \dots t_{i-1})$ is the corpus frequency of the phrase gathered so far ($t_{k+1} \dots t_{i-1}$ using the notation as in [2]), $\Phi(\text{“}t_{k+1} \dots t_{i-1}\text{”})$ – the corpus frequency of the same phrase enclosed in quotation marks, θ – threshold. In other words, we stop if the phrase gathered so far is used with quotation marks with a sufficient frequency (as evidenced by the corpus). This condition is applied for both the left and the right side of *przysłowiowy/-o* marker. The threshold value of θ (0.07) was tuned on the development set.

For instance, let us consider the following snippet:

że bezpieczeństwo na większości placów budów zaczęło być # oczkiem w głowie wielu pracodawców
= that safety on most constructions sites started to be # the apple of the eye [for] many employers

The snippet contains an idiomatic expression *być oczkiem w głowie* (= *be the apple of the eye*). The original procedure, as presented in [2], would erroneously delimit *oczkiem w głowie wielu* – PMI would be not enough to stop before *wielu* (in turn, *być* would not be attached as the frequency of *być oczkiem w głowie wielu* is zero). With the enhancement proposed here, the idiom would be correctly limited (for the corpus used: Φ(oczkiem w glowie) = 234, Φ(“oczkiem w glowie”) = 27).

Identification The only criterion for idiom identification used in [2] was the length of the candidate phrase (at least two tokens after the delimitation step). In order to improve the precision, we decided to add another condition: a phrase will be classified as an idiom only if it is found in the corpus with quotation marks. More precisely, the following expressions are tried for a candidate phrase of the form $t_i \dots t_{k-1}$ # $t_{k+1} \dots t_j$:

- “ $t_i \dots t_{k-1} t_{k+1} \dots t_j$ ”
- $t_i \dots t_{k-1}$ “ $t_{k+1} \dots t_j$ ”

– $t_i \ldots t_{k-1}$przyslowiowy " $t_{k+1} \ldots t_j$ "

For instance, let us consider the snippet:

I czuje się jak po # pobycie na wsi i juz teraz rozumiem dlaczego
= And I am feeling like after # a stay in the countryside and I understand why

There is no idiomatic expression in this snippet (it is puzzling why the writer used *przysłowiowy* here at all). The original procedure would, however, return *po pobycie na wsi* (= *after a stay in the countryside*) as an idiomatic expression. As none of the following sequences of tokens:

– *" po pobycie na wsi "*
– *po " pobycie na wsi "*
– *po przysłowiowym " pobycie na wsi "*

are to be found in the corpus, the phrase will be rejected by the modified procedure.

5 Evaluation

The results are given in Table 1. The method denoted as "PMI-based + filter" in [2] was used as baseline. Precision, recall and F-measure are calculated by taking as true positives only those expressions that were correctly identified *and* delimited. "Quotes/delim" denotes using quotation marks in the delimitation step, whereas "quotes/ident" – in the identification step (see Section 4).

Table 1. Results for the test set. C – number of correctly recognised and delimited idioms, T – number of all reported idiomatic expression, P – precision, R – recall, F – F-measure.

	C	T	P	R	F
baseline	312	1006	0.310	0.307	0.308
baseline + quotes/delim	314	999	0.314	0.309	0.312
baseline + quotes/ident	253	577	0.438	0.249	0.317
baseline + quotes/delim + quotes/ident	258	597	0.432	0.254	**0.320**

Using quotation marks in the delimitation step did not improve the results much. Better results were obtained for quotation marks used in the identification step, the precision was increased without hurting recall too much.

6 Estimations of Recall Upper Bound

What percentage of idiomatic expressions can be retrieved – at best – using idiom-related metalinguistic markers, assuming a corpus of texts indexed by

a general Web search engine (such as Google) is available? In order to estimate this[3], we selected a random sample of 301 Polish idioms from *Wielki słownik polsko-niemiecki* (*Great Polish-German dictionary*) [8], the total number of idioms recorded in this dictionary being 7292. Then, for each idiom from the sample a number of queries were issued to Google search engine. Each query was constructed by taking a given idiom and inserting a form of the adjective *przysłowiowy* or the adverb *przysłowiowo*. It turned out that for 118 idioms (39%) at least one query returned some Web pages containing a given idiomatic expression with the *przysłowiowy* marker.

Similar experiment was carried out for English idiomatic expressions – a random sample of 100 English expressions was selected from over 4800 idioms listed at English Wiktionary[4]. The queries were generated with words *proverbial*, *proverbially* and *literally*[5]. Web occurrences with metalinguistic markers were found for 88 out of 100 idioms.

Obviously, the bigger the textual haystack, the more proverbial needles – one can resort to other resources than Web texts indexed by a general search engine, for instance Web archives[6] or old newspapers and books digitised for digital libraries. This way, idiomatic expressions can be traced diachronically. For instance, there are 2834 occurrences of the word *przysłowiowy* in newspapers, magazines and books stored in Polish digital libraries[7] (the number is likely to be underestimated because of the poor quality of the OCR used).

7 Conclusions and Future Work

We showed that looking globally at all occurrences of an expression marked with idiom-related metalinguistic markers improves type-based identification. It would probably make sense to check all corpus occurrences of a candidate expression, not necessarily accompanied with idiom-related markers. Such methods look promising as was shown with estimations of recall upper bound for Polish and English.

Acknowledgements. This paper is based on research funded by the Polish Ministry of Science and Higher Education (Grant No. N N516 480540).

[3] Note that we estimate the recall against the set of all idiomatic expressions, not against the test set as in Section 5.

[4] `https://en.wiktionary.org/wiki/Appendix:English_idioms`

[5] Using *literally* might seem paradoxical, as it is supposed to mark non-idiomatic usages. This word *is*, however, sometimes used for idiomatic occurrences [5] (cf. `http://xkcd.com/725/`). Moreover, *literally* even for a non-idiomatic occurrence suggests that a given expression *can* be used idiomatically anyway, which is enough for type-based identification.

[6] E.g. `http://www.archive.org`

[7] Most of them not indexed by Google search engine.

References

1. Liontas, J.I.: Context and idiom understanding in second languages. EUROSLA Yearbook **Volume 2, Number 1** (2002) 155–185
2. Graliński, F.: Looking for proverbial needles in the proverbial haystack. In Kopotek, M.A., Marciniak, M., Mykowiecka, A., Penczek, W., Wierzcho, S.T., eds.: Intelligent Information Systems. New Approaches. Wydawnictwo Akademii Podlaskiej, Siedlce, Poland (2010) 101–111
3. Lin, D.: Automatic identification of non-compositional phrases. In: In Proceedings of ACL-99. (1999) 317–324
4. Fazly, A., Cook, P., Stevenson, S.: Unsupervised type and token identification of idiomatic expressions. Computational Linguistics **35**(1) (2009) 61–103
5. Li, L., Sporleder, C.: Linguistic cues for distinguishing literal and non-literal usages. In Huang, C.R., Jurafsky, D., eds.: COLING (Posters), Chinese Information Processing Society of China (2010) 683–691
6. Lewicki, A.M.: Aparat pojęiowy frazeologii. In Lech Ludorowski, W.M., ed.: Z badań nad literaturą i językiem. Panstwowe Wydawnictwo Naukowe (1974) 135–151
7. Gale, W., Church, K., Hanks, P., Hindle, D.: Using statistics in lexical analysis. In Zernik, U., ed.: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum Associates, Hillsdale, N.J. (1991) 115–164
8. Wiktorowicz, J., Frączek, A., eds.: Wielki słownik polsko-niemiecki. Wydawnictwo Naukowe PWN, Warszawa (2008)