Lexical idiosyncrasy in MWE extraction

Csaba Oravecz, Viktor Nagy and Károly Varasdi
Research Institute for Linguistics
Hungarian Academy of Sciences
{oravecz,nagyv,varasdi}@nytud.hu

1 Introduction

A wide scale of different NLP methods have been investigated for the extraction of Multiword Expressions from large corpora. While a good deal of recent research has been focusing on the development of reliable means to delineate different subclasses of MWEs with respect to the degree of their compositionality (Baldwin et al., 2003; McCarthy et al., 2003), it has been generally accepted that for the "simple" task of separating MWEs from fully productive word combinations, the substitutability of component words in a multiword unit with semantic neighbours could be a good indicative measure (Bannard et al., 2003). The underlying assumption is that MWEs do not generally tolerate the replacement of their components with semantically similar items. (Let us call this phenomenon lexical idiosyncrasy.) If we could represent this substitutability by some ranking measure, we will have reliable information whether a word combination could be considered a multiword unit or not (Pearce, 2001).

In our paper we will investigate the usability of lexical idiosyncrasy in MWE detection/extraction in Hungarian, and try to demonstrate that contrary to intuition, while the above hypothesis based on the lexical idiosyncrasy of MWEs might well be true, it can be problematic to detect it in large corpora with reliability necessary for an efficient extraction method: those methods based on measuring differences in terms of substitutability might not perform as well as "good old" ones based on association measures like some variant of MI or *t*-score.

The remainder of the paper is structured as follows. In section 2 we will give a brief description of the idiosyncratic semantic behaviour of MWEs, which can be utilised (at least in theory) to demarcate MWEs from productive word combinations or to identify different MWE classes. Section 3 will discuss the extraction methods we experimented with, while in section 4 we will present the experiments carried out under several scenarios and evaluate how the different techniques perform. We will not devote a separate section to related work or past research, rather we will be making constant references to resources and methods we slavishly adopt throughout the paper. Conclusions and suggestions for further work will end the paper in section 5.

When referring to different types of MWEs we make use of the terminology in Nunberg et al. (1994) and Sag et al. (2002), and in the description of the data Evert and Krenn (2001).

2 Semantic diagnostics for MWEs

2.1 Thematic incongruence

One of the most important characteristics of most types of MWEs, beside their formal rigidity, is that their meaning is definitely not compositional. Still, many MWEs have a compositional, i.e. literal, interpretation beside the idiomatic one. This fact has an important consequence in the process of human communication which cannot be ignored by the interlocutors: *a MWE can only be used to convey its idiomatic meaning if there is sufficient* ***thematic incongruence*** *between the topic of the discourse and the compositional meaning of the MWE*. The idiomatic meaning only becomes relevant to the hearer after excluding the literal meaning as incongruent with the discourse so far for pragmatic or other reasons. To avoid the breakdown of communication in the face of the speaker's apparently inconsistent utterance, the hearer tries to invoke its *idiomatic* meaning, in accordance with Davidson's Principle of Charity (Davidson, 2001). For example, one can utter "I had to carry the can" in the context of talking about actual physical actions done on a farm without conveying the (secondary) meaning that he or she was willing to take the blame for some misdeed. In this case, there is not enough thematic incongruence between the compositional meaning of the expression and the overall topic of the discussion for the hearer to start looking for a secondary interpretation in order to avoid the breakdown of communication. To utilize this fact in a computational setting, it is of course necessary to determine the "semantic center of gravity" of the text, but this is both theoretically and practically possible by applying state-of-the-art vector-based text classifying methods. Adding incongruent, that is, probably idiomatic, expressions will change the vector of the text to such an extent that would otherwise indicate the closing of the discussion or starting a new topic.

Although the methods that utilize thematic incongruence in full generality may require complex resources to determine the topic of the discourse, the special case discussed below can be handled through more modest means.

2.2 Semantic incongruence

MWEs often have "semantically strange" meanings: e.g. in the case of *sweep the problem under the rug* the subject apparently applies a physical action (sweeping) to an abstract entity (the problem), which, on the face of it, is a category mistake. Similar examples could be *that argument doesn't hold water* or *he ate his words*. This type of incongruence can be detected easily if the corpus is semantically annotated at least to some extent. We note that such semantically anomalous idioms seem to allow for a greater extent of syntactic variations: they can often undergo topicalization and their elements can be modified fairly freely. This flexibility can be explained by the salient nature of the anomaly in question: the presence of an apparent category mistake is independent of the syntactic structure of the utterance and is too salient to be ignored by the hearer.

2.3 Lexical idiosyncrasy

We have seen that the components of MWEs have a lesser degree of variability than those of fully productive phrases. As a consequence, the parts of a MWE usually cannot be replaced by (near) synomyms without losing the idiomatic meaning. If the idiomatic meaning happens to be the only one that the expression has (as in *by and large*), the replacement of (near) equivalents can make the phrase devoid of any sense (**by and big*). We call this phenomenon, i.e. that in the set of semantically equivalent expressions (the synset) only one is appropriate, *lexical idiosyncrasy*, and we expect that non-idioms will not obey this constraint (or to a lesser degree). We have tried to utilize this theoretical expectation to separate MWEs from non-MWEs by applying a machine-readable synonym dictionary of Hungarian to our corpus.

3 Techniques for MWE extraction

As an initial experiment we will follow Pearce Pearce (2002) method since it is intuitively very appealing and relatively simple to implement, that is it can be readily adoptable for a new resource. In the lack of a Hungarian WordNet, this new resource is the electronic version of the Hungarian Synonym Dictionary (Kiss, 2001), containing 25.000 entries with 25.700 synonym sets consisting of around 280.000 synonyms (without sense information in the corpus all sub-senses of a headword are merged into one synonym set). In the following we briefly present the association measure applied in the method, adopted to the use of the synonym dictionary.

Let $\mathcal{D}$ denote the dictionary database, from which we can assign a set of synonyms to a particular word:

$$\mathcal{D} = \{S_1, S_2, S_3 \ldots\} \tag{1}$$

A possible lexical realisation of a concept C is E multiword expression as a sample point, for which: $E = \langle w_1 \ldots w_n \rangle$. The sample space is defined as all possible realizations of the concept C, which is the following in terms of the synonym sets:

$$\Omega(E) = \{w'_{1,n} : w'_i \in S_i, 1 \leq i \leq n\} \tag{2}$$

where S_i is the synonym set for w_i.

If we assume that E is a fully productive expression with its component words selected independently from each other then the probability of E_i can be approximated as the joint probability:

$$\hat{p}(E_i) = \prod_{i=1}^{n} p_i(w_i) \tag{3}$$

$p_i(w_i)$ is the probability that we select w_i from the synonym set S_i:

$$p_i(w_i) = \frac{f(w_i|S_i)}{\sum_{w \in S_i} f(w|S_i)} \tag{4}$$

This probability can then be compared to the maximum likelihood estimate of E_i:

$$p(E_i) = \frac{f(E_i)}{\sum_{E \in \Omega(E)} f(E)} \tag{5}$$

The z-score of the difference between the two probabilities then indicates how the given expression tolerates substitutability:

$$z_i = \frac{d_i}{\sigma(d)}, \quad \text{where} \quad d_i = p(E_i) - \hat{p}(E_i) \quad \text{and} \quad \sigma(d) = \sqrt{\frac{\sum_i (d_i - \mu)^2}{n}} \tag{6}$$

A high z-score would provide evidence for the presence of a MWE.

4 Evaluation

We evaluate on two different candidate lists (CL) compiled from the 150 million word POS disambiguated Hungarian National Corpus (Váradi, 2002):

1. adjective+noun combinations (L1)

2. transitive verb+noun in accusative case within one sentence (without a syntactic parse, as a possible approximation for a head-argument relation) (L2)

With Hungarian being a morphologically rich language it would be interesting to test an extraction method based on the morphological idiosyncrasy of certain word combinations. However, this is out of the scope of the present investigation so we use the lemmatised version of the candidate list to get more reliable statistics. As a threshold for the minimum frequency of occurrence of candidates we select the value of 5 (a setting of 10 gave practically the same results). The model based on the above z-score is referred to as M_z in the evaluation.

Results are compared to a baseline model in which MI values filtered by t-test are calculated in the usual manner (Church et al., 1994), i. e. occurrence counts are not restricted to semantically related substitutes (M_b). For a better assessment of the substitution based methods, a crude ranking measure similar to the one proposed in Pearce (2001) is also evaluated: the ratio (denoted s) calculated from the most frequent (with a count of f') and second most frequent (f'') expressions constructed from the synonym sets ($s = \frac{f'-f''}{f'}$; *model* M_s).

In the fourth model we attempt to construct a rudimentary generalisation of the synonym sets serving as a basis for the competing lexical realisations of a concept C, as a first step in the extension towards a thesaurus like resource. In vein similar to Bárdosi et al. (2004) we define the semantic similarity of two words in terms of the ratio of their shared synonyms in the following way. Let $Syn(w)$ denote the unified synsets of the headword w in the synonym dictionary. The semantic similarity of w_1 and w_2 is

$$Sim(w_1, w_2) = \frac{|Syn(w_1) \cap Syn(w_2)|}{|Syn(w_1) \cap Syn(w_2)| + |Syn(w_1) \setminus Syn(w_2)| + |Syn(w_2) \setminus Syn(w_1)|} \quad (7)$$

The similarity matrix is a sparse matrix as most headword pairs has no common synonyms.

The headwords are brought together using a hierarchical complete link clustering algorithm. The clustering process is stopped when the similarity of each pair of clusters became 0. The result is a set of maximally connected clusters of semantically related headwords. In order to produce the extended sets, the synsets assigned to the headwords in each cluster are joined together. With this method, about 4500 extended sets are derived from 12500 noun synsets, 1500 from 4500

with adjectives and 2700 from 7500 with verbs. This model is referred to as M_{ze} in the evaluation.

Due to limited resources, a standard but rightly criticised (Evert and Krenn, 2001) manual n-best list evaluation is presented here ($n = 250$). We only present precision values since within this type of evaluation framework recall values carry no additional information about performance. Consider the calculation of *precision* in n-best lists: $p(n) = \frac{TP(n)}{n}$, where TP(n) is the number of true positives in the list containing n members. The *recall* here: $r(n) = \frac{TP(n)}{C}$, where C is the number of TPs in the whole significance list, which is constant. Then $r(n) = \frac{p(n) \times n}{C}$, so here *recall* is in effect a model independent transform of *precision* carrying no additional information. (If $n = |\textit{full significance list}|$ then naturally $r(n) = \frac{\frac{C}{n} \times n}{C} = 1 \rightarrow 100\%$. The information that *recall* values can bear independent of *precision* is the lowest $n < C$ value that can be assigned to a particular method, for which $r(n) = 1$, that is, the minimal threshold number in the significance list above which all TPs are present. This cannot be identified from *precision* graphs.) For this reason Table 1 contains only the *precision* values.

Type of CL	Number of candidates $f(E_i) > 5$	Precision $p(n = 250)$			
		M_b	M_s	M_z	M_{ze}
L1	191454	54.4%	15.2%	17.2%	17.5%
L2	100559	56.8%	9.2%	38.0%	39%

Table 1: Performance of different models.

4.1 Error analysis

We propose a tentative explanation for the disappointing performance of the substitution based extraction methods, which is conspicuous in the presence of a high number of non MWEs in the significance list: by altering the space of competitors many non MWEs comes out as statistically idiosyncratic simply because no frequent use of their synonym substitutes can be attested in corpus. By contrast, in the classic method (M_b), when candidates are selected on the basis of free (non synonym controlled) variations of the individual components this undesirable statistical idiosyncrasy disappears, and the unit comes out as "stochastically compositional" so will not qualify for a MWE. Thus, for general MWE identification this alteration of sample space does not seem to be a workable alternative for Hungarian.

5 Conclusion

We would be very cautious to jump into any rush conclusion. What we have tried is only to demonstrate that lexical idiosyncrasy, although being an intuitively very attractive phenomenon to utilise in MWE identification, does not appear to be an efficient diagnostics on large Hungarian corpora. For Hungarian, methods based on this characteristic of MWEs do not seem to perform well when supported by either a synonym dictionary or automatically constructed resources.

We offer the following explanation for the poor performance of a substitution based method, their primary problem being not so much the inability to make the distinction between compositional and non-compositional MWEs (emphasised as a main problem in e.g. Lin's work (1999) by Baldwin et al. (2003) and Bannard (2002)), but making the primary distinction between MWEs (including semantically compositional institutionalised expressions as well), and fully productive word combinations. When we want to make use of the lexical idiosyncrasy of MWEs we basically try to capture their exceptional behaviour with relatively basic statistics, therefore we rely here on their statistical idiosyncrasy. It seems, however, when we examine whether it is attestable in large corpora, we find that this is better manifest in the variations of the multiword components with respect to the whole vocabulary, and not just with respect to the reduction of the variations to ones where there is a synonymy (or some other semantic) relation with one or the other component of the multiword. It might well be the case that in the end not only a wide generalisation of the concept sets is necessary but the whole vocabulary must be taken into account, and we will asymptotically arrive at the good old stochastic extraction methods operating on the whole candidate list not constrained to semantically related substitutes.

There is ample room for further work with respect to testing the usability of the lexically idiosyncratic behaviour of MWEs in different extraction methods. A first step is to examine how results are influenced if a Hungarian WordNet becomes available, which is just in preparation in the time of the writing of this paper. It is also necessary to support our findings with more extensive testing including automatically constructed synonym or concept sets (thesauri) to examine whether the above assumption about the asymptotic behaviour can be empirically justified.

References

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL*

2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pages 89–96, Sapporo, Japan.

Bannard, C. (2002). Statistical techniques for automatically inferring the semantics of verb-particle constructions. Technical report, LinGO Working Paper No. 2002-06.

Bannard, C., Baldwin, T., and Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Bárdosi, V., Kiss, G., Kiss, M., and Rapcsák, T. (2004). Kísérlet magyar szavak jelentéshasonlóságának meghatározására a magyar szókincstár segítségével. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, pages 27–37, Szeged, Hungary.

Church, K. W., Gale, W., Hanks, P., Hindle, D., and Moon, R. (1994). Lexical substitutability. In Atkins, B. T. S. and Zampolli, A., editors, *Computational Approaches to the Lexicon*, pages 153–180. Oxford University Press.

Davidson, D. (2001). Radical interpretation. In *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford.

Evert, S. and Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Kiss, G., editor (2001). *Magyar Szókincstár*. Tinta Könyvkiadó, Budapest.

Lin, D. (1999). Automatic identification of noncompositional phrases. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 317–24, College Park, USA.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Pearce, D. (2001). Using conceptual similarity for collocation extraction. In *Proceedings of the 4th UK Special Interest Group for Computational Linguistics*.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Poceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (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)*, pages 1–15, Mexico City, Mexico.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 385–389, Las Palmas.