An Evaluation of Methods for the Extraction of Multiword Expressions

Carlos Ramisch[♣♢], Paulo Schreiner[♣], Marco Idiart[♡] and Aline Villavicencio[♣♠]

[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♢]GETALP Laboratory, Joseph Fourier University - Grenoble INP (France)
[♡]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
[♠]Department of Computer Sciences, Bath University (UK)
{ceramisch,pschreiner}@inf.ufrgs.br, idiart@if.ufrgs.br, avillavicencio@inf.ufrgs.br

Abstract

This paper focuses on the evaluation of some methods for the automatic acquisition of Multiword Expressions (MWEs). First we investigate the hypothesis that MWEs can be detected solely by the distinct statistical properties of their component words, regardless of their type, comparing 3 statistical measures: Mutual Information, χ^2 and Permutation Entropy. Moreover, we also look at the impact that the addition of type-specific linguistic information has on the performance of these methods.

1. Introduction

The task of automatically identifying Multiword Expressions (MWEs) like phrasal verbs (*sell out*) and compound nouns (*science fiction*) using statistical measures has been the focus of considerable investigative effort, (e.g. Pearce (2002), Evert and Krenn (2005) and Zhang et al. (2006)). Among these, some research has focused on the development of methods for dealing with specific types of MWEs (e.g. Pearce (2002) on collocations and Villavicencio (2005) on verb-particle constructions), and some work on dealing with MWEs in general (e.g. Zhang et al. (2006)). These works tend to focus on one language (e.g. Pearce (2002) and Zhang et al. (2006) for English and Evert and Krenn (2005) for German).

As basis for helping to determine whether a given sequence of words is in fact an MWE some of them employ (language and/or MWE-type dependent) linguistic knowledge for the task, while others employ (language and type independent) statistical methods, such as Mutual Information and Log-likelihood (e.g. Pearce (2002) and Zhang et al. (2006)), or a combination of both (e.g. Baldwin (2005) and Sharoff (2004)). Given the heterogeneousness of the different phenomena that are considered to be MWEs, there is no consensus about which method is best suited for which type of MWE, and if there is a single method that can be successfully used for any kind of MWE. Therefore, it would be of great value to know if a given MWE extraction approach could be successfully applied to other MWE types and/or languages (or families of languages), and if so, how good their performance would be.

In this paper we use three distinct MWE types from two different languages to evaluate some association measures: Mutual Information (MI), χ^2 and Permutation Entropy (PE) (Zhang et al., 2006). We also investigate the effect of adding some language and MWE-type specific information to the identification task, proposing a new measure, Entropy of Permutation and Insertion (EPI).

This paper starts with a brief description of the data sets (§ 2.). We then present the two different approaches used for identifying MWEs: a language and MWE-type independent set of association measures (§ 3.), and a language and type dependent set (§ 4.). We finish with a discussion of the overall results (§ 5.).

2. The Data

The evaluation of these association measures was performed over three distinct data sets: a list of 3,078 English Verb-Particle Constructions (VPCs) with manual annotation of idiomatic verb-particle pairs (which we refer to as EN-VPC); a manually annotated list of 1,252 German adjective-noun pairs (DE-AN); and a manually annotated list of 21,796 German combinations of prepositional phrase and governing verb (DE-PNV).[1]

These data sets are used as gold standard for the evaluation, as they are annotated with information about positive and negative instances of each of these MWE types. In addition the two German sets also contain frequency information, based on which the association measures are computed. The only pre-processing done in the data sets was that for DE-PNV we filtered out all candidates that appear less than 30 times in the Frankfurter Rundschau (FR) German corpus, to obtain a cleaner data set. The frequencies for the English set were collected from two different sources: the Web, using Yahoo APIs, which return the number of pages indexed for each search (henceforth referred to as *Yahoo*) and a fragment of the British National Corpus (BNC - Burnard (2000)) of 1.8M sentences (the same employed by Zhang et al. (2006), henceforth BNC_f).

3. A Language and Type Independent Approach

For each data set we compute three type independent statistical measures for MWE identification: MI, χ^2 and PE. The first two are typical measures of association while PE is a measure of order association. PE was proposed by Zhang et al. (2006) as a possible measure to detect MWEs, under the hypothesis that MWEs are more rigid to permutations and therefore present smaller PEs. Even though it is quite

[1] The data sets were provided by: Timothy Baldwin for EN-VPC; Dictionary editors of Langenscheidt KG and Stefan Evert for DE-AN; Brigitte Krenn and Stefan Evert for DE-PNV. All data sets are available from multiword.sf.net/mwe2008/shared_task.html.

different from MI and χ^2, PE can also be thought as an indirect measure of statistical independence, since the more independent the words are the closer PE is to its maximal value ($\ln 2$, for bigrams).

For a bigram with words $w_1 w_2$, χ^2 and MI are calculated respectively as:

$$\chi^2 \;=\; \sum_{a,b} \frac{[\; n(ab) - n_{\emptyset}(ab)\;]^2}{n_{\emptyset}(ab)}$$

$$\text{MI} \;=\; \sum_{a,b} \frac{n(ab)}{N} \;\log_2 \left[\, \frac{n(ab)}{n_{\emptyset}(ab)} \,\right]$$

where a corresponds either to the word w_1 or to $\neg w_1$ (all but the word w_1) and so on. $n(ab)$ is the number of bigrams ab in the corpus, $n_{\emptyset}(ab) = n(a)n(b)/N^2$ is the predicted number from the *null* hypothesis, $n(a)$ is the number of unigrams a, and N the number of bigrams in the corpus. For these two measures we only use the FR and BNC_f corpora, since for them the size of the corpus is known (the value of N). PE is calculated as:

$$\text{PE} = -\sum_{(i,j)} p(w_i w_j) \ln \left[\; p(w_i w_j) \;\right]$$

where the sum runs over all the permutations of the indices and, therefore, over all possible positions of the selected words in the bigram. The probabilities are estimated from the number of occurrences of each permutation (e.g. *computer science* and *science computer*) as:

$$p(w_1 w_2) = \frac{n(w_1 w_2)}{\sum\limits_{(i,j)} n(w_i w_j)}$$

For calculating PE we used the Yahoo corpus and for each of the data sets we restricted the search to return only pages in that language (English or German). The Yahoo corpus can be used for PE, since, unlike MI and χ_2, PE is calculated independently of the size of the corpus, and the use of Yahoo as a corpus can minimize the problem of data sparseness.

The results of these three evaluations can be seen in figures 1 to 3 and in table 1. In all these cases the statistical measures perform better than the baseline, with the expected trade-off between precision and recall. The exception is PE. When this measure is calculated on the basis of varying the order of the words, it provides a stronger contribution when there is no underlying grammatical constraint preventing the combination of the constituents in the permuted orders. If, as in the case of English VPCs, the particle is only expected after the verb (but not before), PE does not add much information, since due to grammatical constraints the permuted orders are not going to be often found.

For both EN-VPC and DE-AN, MI and χ^2 have very similar performances. However, for DE-PNV MI seems to have a much better predictive power than χ^2 and any of the other measures.

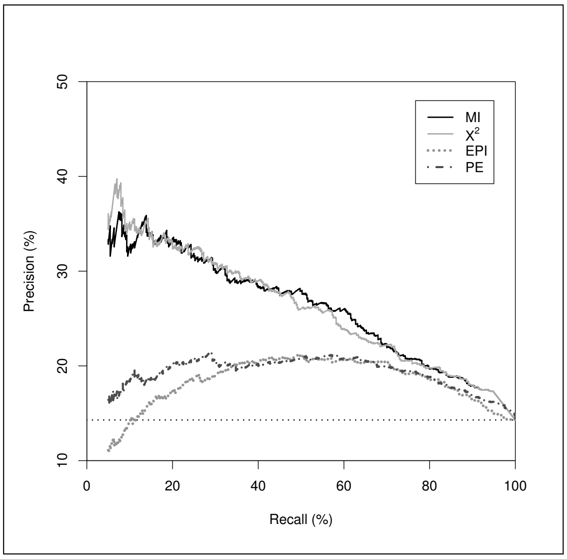

Figure 1: Precision-recall graphic for EN-VPC data set

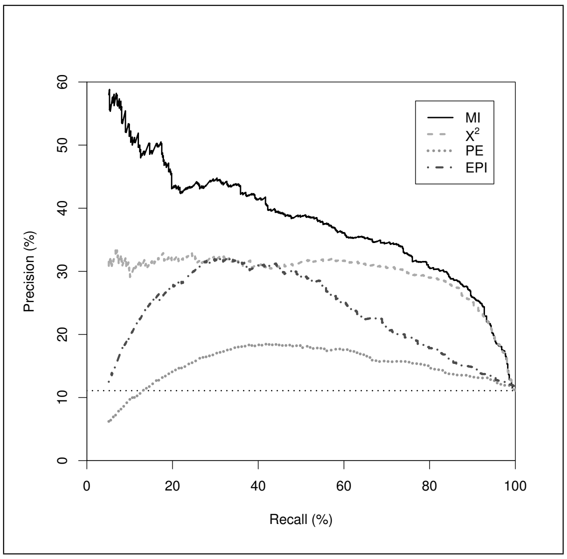

Figure 2: Precision-recall graphic for DE-PNV data set

4. A Language and Type Dependent Approach

In order to evaluate whether the addition of linguistic information can further improve MWE identification compared to the use of purely frequency-based measures, we performer further tests with two of the data sets: the German DE-PNV and the English EN-VPC. For that we introduce an entropy measure, the Entropy of Permutation and Insertion (EPI), that takes into account linguistic information about the MWE type. EPI is calculated as follows:

$$\text{EPI} = -\sum_{a=0}^{m} p(ngram_a) \ln \left[\; p(ngram_a) \;\right]$$

where $ngram_0$ is the original expression, and $ngram_a$ for $a = 1...m$, are m syntactic variants of the original expression. As before we calculate the probability of occurrences

Measure	Corpus	Data set EN-VPC	DE-PNV	DE-AN	EN-VPC-DICT
MI	BNC$_f$–FR	26.09%	39.05%	56.09%	39.59%
χ^2	BNC$_f$–FR	26.41%	29.85%	56.91%	41.46%
PE	Yahoo	17.96%	14.64%	40.35%	35.74%
EPI	Yahoo	19.33%	22.74%	–	39.23%
Baseline	–	14.29%	11.09%	41.53%	30.15%

Table 1: Average precisions for the studied measures

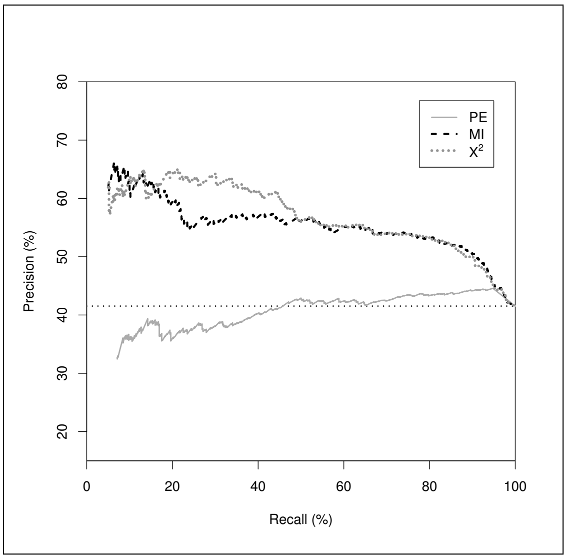

Figure 3: Precision-recall graphic for DE-AN data set

of any one of the variants as:

$$p(ngram_a) = \frac{n(ngram_a)}{\sum_{a=0}^{m} n(ngram_a)}$$

EPI is an extension to PE based on the idea that not all types of MWEs have the same behaviour. Therefore, if we know what kinds of modification an MWE type accepts or refuses in a particular language, we may be able to obtain more accurate entropies that may improve the identification task.

VPCs in English, for instance, have a very strict word order in that the verb comes before the particle, and accept little intervening material between them, but appear in a number of different syntactic configurations (e.g. the split and joint configuration for transitive VPCs). Thus, to identify VPCs contained in the EN-VPC data set we performed further Yahoo searches to account for some linguistic features that distinguish them from prepositional verbs or free verb-preposition combinations (e.g. *walk up the hill*).[2] The following patterns were used:

- Intransitive VPC: VERB + PARTICLE + DELIMITER
- Split Transitive VPC: VERB + NP + PARTICLE + DELIMITER
- Joint Transitive VPC: VERB + PARTICLE + NP + DELIMITER

We searched for exact matches of these patterns, where VERB corresponds to the verb element of a VPC candidate, PARTICLE, to the particle of the VPC, NP is either '*this*' or '*the* *' (with the Yahoo wildcard standing for one word), and DELIMITER is the preposition *with*. The delimiter is used to avoid retrieving pages where the particle is followed by an NP, which would also be ambiguous with prepositional verbs and free verb-preposition combinations, following Villavicencio (2005). Two distinct transitive VPC configurations were used: the *split* for when the verb is separated from the particle by an NP complement, and the *joint*, for when the verb and particle are adjacent to each other. Note that in the joint configuration pattern, there may be some false positive cases (e.g. prepositional verbs) since the delimiter is not immediatly following the particle anymore, which will introduce some noise in the frequencies obtained. However, since this is one of three configurations that are combined in EPI, even if there is some noise, it will be counterbalanced by the other configurations.

For VPCs an EPI closer to 1 indicates a VPC since variations are more characteristic of a genuine VPCs while non-VPCs will show a peak for the canonical form (*verb particle/preposition*). Figure 1, also shows the results for EPI, which has a higher precision-recall rate than PE, and therefore a higher average precision (19.33% vs 17.96%), but still lower than MI and χ_2.

The original EN-VPC data set was manually marked for true positives for all VPC candidates that are idiomatic. A closer look at the data, however, revealed that many of the unmarked candidates are nonetheless present in machine-readable dictionaries. Therefore, in order to evaluate the measures in terms of their effectiveness in detecting VPCs, regardless of their idiomaticity we used a list of 3,156 VPCs contained in either the Alvey Natural Language Tools (ANLT) lexicon (Carroll and Grover, 1989), the Comlex lexicon (Macleod and Grishman, 1998), and the LinGO English Resource Grammar (ERG) (Copestake and Flickinger, 2000)[3]. Using this as a gold standard, we obtained a new baseline of 30.15%, and a considerable improvement in performance. Average precision of χ^2, for example, improves to 41.46% (vs 26.41% with manual annotation). These results suggest that these measures seem more adequate to detect VPCs in general rather than to detect id-

[2] These features are based on those used by Baldwin (2005).

[3] Version of November 2001.

iomaticity in them.

For the DE-PNV data set, the first attempt to include linguistic information in the identification task is done by means of capturing inflectional patterns of German prepositions, which in the data set are marked with a "+" symbol if they inflect (e.g. *in+:Bett* as *ins Bett* or *im Bett*). To account for this variability we use the boolean operators available in Yahoo and a search for a combination like *in+:Bett liegen* originates the exact search term *(in OR im OR ins) Bett liegen* that has the potential to return either of those three prepositional forms occurring as the first word.[4]

Besides prepositional inflection, the other source of language-dependent information for the identification of DE-PNV is based on the assumption that fixed and semi-fixed MWEs do not accept determiners being inserted into the expression. This behaviour is essentially different from English VPCs, where genuine candidates do accept some syntactic variation. In German, a verb may appear before or after the indirect complement, depending on the context (e.g. both *in Kontakt treten* and –the less frequent but possible– *treten in Kontakt* might occur). However, true MWEs accept less well the addition of a determiner (except eventually for an article) placed between the preposition and the noun (e.g. *in Kontakt treten* but not *in großen Kontakt treten* nor *in den Kontakt treten*). To capture that we searched the Web for four different combinations (the Yahoo wildcard stands for a word like a pronoun, an article, an adjective, etc.): (1) *in Kontakt treten*, (2) *treten in Kontakt*, (3) *in * Kontakt treten* and (4) *treten in * Kontakt*. For DE-PNVs a high EPI indicates a more homogeneous distribution (i.e. not an MWE), while a low EPI suggests that there is a peak with only one acceptable form (i.e. indicating an MWE). This change in EPI interpretation shows that the measure can be easily adapted from one language and/or MWE type to another with the addition of some linguistic information and the appropriate interpretation. These patterns can be easily obtained, for instance with a linguist, and verified in a corpus (or in the Web), independently of expensive resources like dictionaries, huge corpora and thesauri and easily refined online.

Although the new measure is fairly superior than conventional PE for DE-PNV (figure 2), the result is far from being optimal, and we believe that some additional variation tests should be performed in order to reach higher quality levels. In terms of average precision, we go from 14.64% with PE to 22.74% with EPI.

The addition of linguistic information to both EN-VPC and DE-PNV had indeed an effect when compared to the standard PE. However, both MI and χ^2 still perform better.

5. Conclusions

One of the important challenges for robust natural language processing systems is to be able to successfully deal with Multiword Expressions and related constructions. In this paper we presented a first step towards investigating whether MWE identification methods can be robustly and successfully applied to different types of MWEs and different languages. The results suggest that although statistical measures on their own can detect trends and preferences in the co-ocurrences and combinations of words, for different languages and MWE types, they also have limited success in capturing some specific linguistic features, such as compositionality (in the EN-VPC data), which would require more sophisticated measures. Moreover, even if measures like MI and χ^2 seem to often agree on their rankings (Villavicencio et al., 2007), they may also have different performances for different MWE-types (e.g. for the DE-PNV). Finally, the individual performances of these measures may well be improved if they are combined together, offering different insights into the problem, and this is planned for future work.

6. References

Timothy Baldwin. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language*, 19(4):398–414.

Lou Burnard. 2000. User reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

John Carroll and Claire Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. In B. Boguraev and E. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Catherine Macleod and Ralph Grishman. 1998. Comlex syntax reference manual, Proteus Project.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

Serge Sharoff. 2004. What is at stake: a case study of russian expressions starting with a preposition. pages 17–23, Barcelona, Spain.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043.

Aline Villavicencio. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? *Journal of Computer Speech and Language Processing*, 19(4):415–432.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.

[4] Some noun-verb combinations exclude some prepositional forms (like the impossible **ins Bett liegen* and **in Bett liegen*, and these will be reflected in the frequencies obtained, with any occasional noise being automatically corrected by the size of the Web.