UFRGS@CLEF2008: Indexing Multiword Expressions for Information Retrieval

Otavio Costa Acosta, André Pinto Geraldo, Viviane Moreira Orengo, Aline Villavicencio
Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
[ocacosta, apgeraldo, vmorengo, avillavicencio]@inf.ufrgs.br

Abstract

For UFRGS's participation on CLEF's Robust task, our aim was to assess the benefits of identifying and indexing Multiword Expressions (MWEs) for Information Retrieval. The approach used for MWE identification was totally statistical, based association measures such as Mutual Information and Chi-square. Contradicting our results on the training topics, the results on the test topics did not show any significant improvements. However, for some queries, the identification of MWEs was very important. We have also performed bilingual experiments which achieved 84% of their monolingual counterparts.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Linguistic processing. H.3.4 [Systems and Software]: Performance evaluation

Free Keywords

experimentation, performance measurement, multiword expression

1 Introduction

The identification and treatment of multiword expressions (MWEs) are means to improve the capabilities of Natural Language Processing (NLP) in solving problems. MWEs are sequences of words that act as a single unit for the purpose of linguistic analysis. The meaning of the expression is different from the meaning of its composing terms analysed individually, e.g. the terms "escape" and "goat" have a totally different meaning when used as a MWE.

The nature of MWEs is varied, which makes it difficult to devise a mechanism to identify and treat them in a uniform manner. Some estimates say that the number of MWEs in a language is in the same order of magnitude as the number of individual words used by a native speaker of that language.

The correct identification and treatment of MWEs is also important for Information Retrieval (IR). In an ideal IR system, the entries in the index should represent the concepts present in the documents. Indexing a MWE as separate terms will mean loss in semantics.

Our aim in this paper it to test the validity of applying statistical methods for the identification of MWEs applied to IR. This paper reports on the experiments we performed for CLEF's Robust task.

The remainder of this paper is organised as follows: Section 2 discusses some methods for the identification of MWEs; Section 3 describes our experimental runs and results; Section 4 presents our conclusions.

2 Identifying Multiword Expressions

The automatic identification of MWEs has been the focus of many investigations on NLP (Baldwin & Villavicencio, 2002; Nicholson & Baldwin, 2006; Pearce, 2002; Villavicencio et al., 2007). Most methods are based solely on statistics of the data collection, and this was the approach adopted here.

For the experiments described here, we implemented two measures of association that compare the joint probability of occurrence of a certain group of events: Mutual Information and Chi-Square. This probability $p(ab)$ is calculated based on the null hypothesis of statistical independence between these events $p_0(ab)$(Press et al., 1992). In our case, the events are the occurrences of words in a given position. For each pair of adjacent words, known as *bigram*, we use these measures to calculate the strength of the association between them. The stronger the association, the more likely the bigram will compose a MWE.

- **Mutual Information (MI)** measures the mutual dependence of the terms composing the bigram. The MI for the bigram w_1w_2 is calculated as shown in Eq 1.

$$MI = \sum_{a,b} \frac{n(ab)}{N} \log_2 \left[\frac{n(ab)}{n_0(ab)} \right] \quad (1)$$

- **Chi-Square** (χ^2) is based on a comparison of the observed frequencies with the expected frequencies. It is calculated according to Eq 2.

$$\chi^2 = \sum_{a,b} \frac{[n(ab) - n_0(ab)]^2}{n_0(ab)} \quad (2)$$

where: a corresponds either to the word w_1 or to ¬ w_1 (all but the word w_1), and b corresponds to the word w_2 or to ¬ w_2.

$n(ab)$ is the number of bigrams ab in the corpus
$n_0(ab) = n(a)n(b)/N^2$ is the predicted number or null hypothesis
$n(a)$ is the number of words a
N is the number of words in the corpus

The approach taken for MWE identification was as follows: first, we computed the co-occurrences for all bigrams in the collection. The second step was to collect from the web the frequencies of each single word and each bigram. Next, MI and χ^2 were computed for all bigrams. Then, the bigrams were ranked decreasing order of MI and χ^2. Finally, the two rankings were merged, and the top k bigrams were kept.

3 Experiments

This section describes our experiments submitted to the CLEF-2008 Robust Task. Sections 3.1 and 3.2 describe our monolingual experiments and their results, and Sections 4.3 and 4.4 refer to our bilingual runs and their results.

3.1 Description of Runs and Resources for Monolingual Experiments

We worked on the English news collections composed by LA Times 94 and Glasgow Herald 95. There are 169,477 documents in total. Two versions of the collection were available: a "plain" version, and a version with word-sense disambiguation (WSD) data.

Using the WSD documents (UBC version), we created a document collection composed by the lemmas in the texts. This collection was used as the basis for all our WSD runs.

For the runs in which we used MWE identification, we computed MI and χ^2 only for bigrams that had nouns (NN). In order to further reduce the number of bigrams, we discarded all word pairs with fewer than ten occurrences. After the rankings for MI and χ^2 were merged, we kept the top 7,500 bigrams. Having this list of MWE candidates, we searched for their occurrence in the text collection. Each time a MWE candidate was found in a document, we added the MWE candidate to the document joined by an underscore. For example, suppose the bigram "home page" was part of the MWE candidate list, all documents that had this bigram would have the term "home_page" appended to them. The underscore is to force the IR system to index the MWE as a unity rather than as two separate terms. Notice that we did not remove the original bigram from the text, we just added the compound form, joined by the underscore.

We also tested the opposite approach, i.e. removing all compounds from the texts. Since our collections were composed by lemmas, some terms were joined by an underscore, e.g. "to_have". For the two final runs, we removed all underscores joining word forms.

We used the Porter stemming algorithm (Porter, 1980) in three runs. Stop words were not removed.

The IR system we used was Zettair (Zettair), which is a compact and fast search engine developed by RMIT University (Australia) distributed under a BSD-style license. Zettair implements a series of IR metrics for comparing queries and documents. We used Okapi BM25 as some preliminary tests we performed on other data collections showed it achieved the best results.

We have submitted two baseline runs indexing the plain collection and five runs using the WSD-annotated documents. The details of the monolingual runs are shown in Table 1.

Table 1 - Details of the test collections for the monolingual runs

RunID	Description	Number of unique terms
Mono1	baseline run (plain collection)	595,025
Mono2	baseline run (plain collection) and stemming	494,861
WSD1	lemmas	592,459
WSD2	lemmas and 7500 MWE	606,938
WSD3	lemmas, 7500 MWE, and stemming	512,896
WSD4	lemmas, removing compounds	577,508
WSD5	lemmas, removing compounds and stemming	487,979

3.2 Results for Monolingual Experiments

The results for our monolingual runs are summarised in Table 2 and Figure 1. All performances were extremely similar both in terms of MAP and Pr@10. This similarity can be easily seen by the overlap in the recall-precision curves in Figure 1. Statistically, the only significant differences were found when comparing the runs in which some kind of linguistic processing was used (Mono2, and all WSD runs) to the baseline run Mono1. A T-test between Mono1 and Mono2, for example, resulted in a *p*-value of 0.005, showing that stemming yields significant improvements.

The best run overall was WSD5, however, the superiority to the other runs is only marginal. These results disagree with the results we obtained on the training topics, as for those indexing MWEs significantly improved overall performance. The reasons for this discrepancy still need to be evaluated. We did find great improvements for some individual queries. For example the identification of "oil price" as a MWE in topic 290 and "student fees" in topic 325 led to improvements of 22% in MAP.

Comparing our results with other participant's, we came in 5th place for both baseline and WSD tasks. These results are encouraging since the methods we used for MWE identification are very simple and can still be greatly improved.

Table 2 – Monolingual Results in terms of MAP and Pr@10

RunID	Mean Average Precision	Precision at 10
Mono1	0.3120	0.3400
Mono2	0.3395	0.3544
WSD1	0.3424	0.3550
WSD2	0.3391	**0.3587**
WSD3	0.3434	0.3531
WSD4	0.3432	0.3531
WSD5	**0.3465**	0.3537

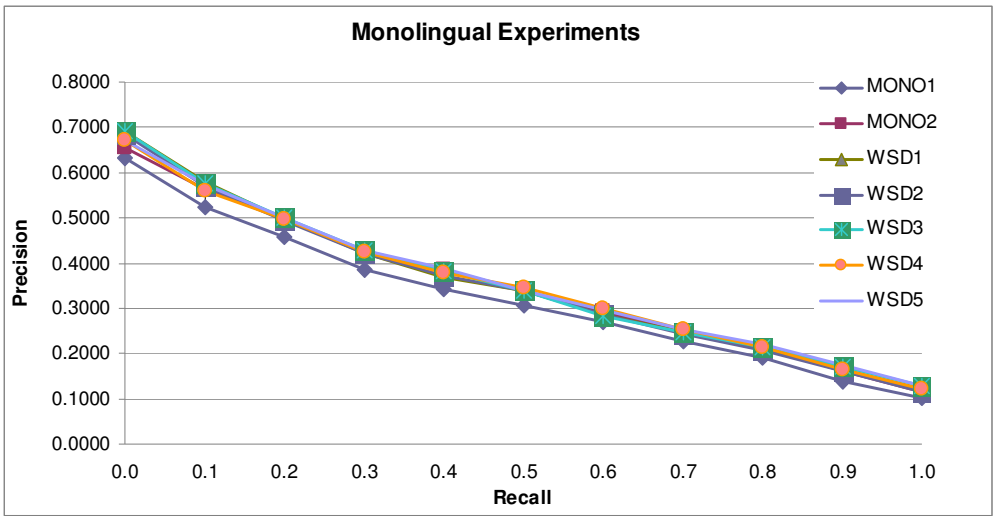

Figure 1 – Recall/Precision curves for Monolingual Runs

3.3 Description of Runs and Resources for Bilingual Experiments

In addition to the monolingual experiments, we also submitted four bilingual runs using Spanish topics to query English documents. Our approach used to map concepts between languages was the same as described in (Geraldo & Orengo, 2008). The idea is to use algorithms for mining association rules (ARs) to map concepts between languages.

Since the approach requires a sample of parallel documents and the document collections were in English only, 20% of the collection was automatically translated using Google Translator[1]. The Apriori algorithm (Agrawal et al., 1993) for mining ARs was applied over these simulated parallel documents. Each word in the original query was replaced by the words in the target language which remained after the filtering step. Table 3 shows the details of our bilingual runs. In one of the runs, Bi3, we also tested a modification to the BM25 algorithm that aims at giving more weight to rare terms. This modified version is also described in (Geraldo & Orengo, 2008).

Table 3 - Details of the test collections for the bilingual runs

RunID	Description	Number of unique terms
Bi1	baseline bilingual run, BM25	595,025
Bi2	stop-word removal, stemming, BM25	487,979
Bi3	stop-word removal, stemming, BM25+	487,979
Bi-WSD	stop-word removal, lemmas, BM25	592,459

3.4 Results for Bilingual Experiments

The results for our bilingual runs are summarised in Table 4 and Figure 2. The best performance was achieved by Bi3, which combines stop-word removal, stemming and our modification to BM25. This advantage is statistically significant in relation to all other runs, showing that our modification to BM25 does improve retrieval performance. Stemming was also found to yield significant improvements. The recall-precision curves clearly show the ranking of the runs. When comparing to other participants, our best run scored very well, being the best overall.

If we compare the performance of a monolingual run and its bilingual counterpart (Mono2 and Bi2), we find that the bilingual version achieves 84% of the monolingual performance. The superiority of the monolingual run is statistically significant.

[1] http://www.google.com/translate_t

The run in which we used WSD information (i.e. the lemmas) was the worst. This is because the sample of parallel documents used as a basis for mining ARs had the original word forms and not the lemmas.

Table 4 – Bilingual Results in terms of MAP and Pr@10

RunID	Mean Average Precision	Precision at 10
Bi1	0.2560	0.2838
Bi2	0.2860	0.2880
Bi3	**0.3639**	**0.3575**
Bi-WSD	0.2177	0.2469

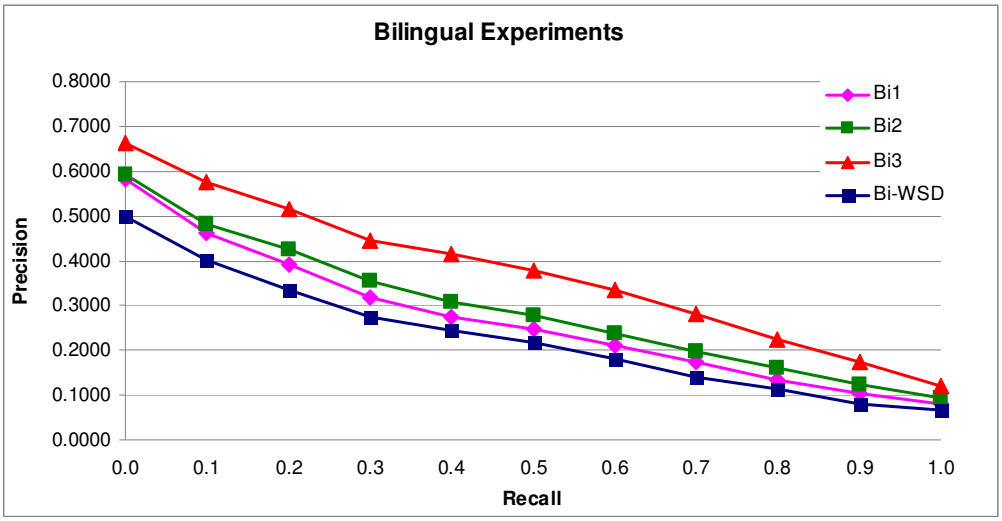

Figure 2 – Recall/Precision curves for Bilingual Runs

4 Conclusions

This paper reports on UFRGS's experiments for the Robust task at CLEF 2008. Our aim was to assess the benefits of identifying and indexing MWEs for CLIR. The methods we employed for MWE identification, Mutual Information and Chi-square, are totally automatic. They were based on associations between adjacent words.

The results of the experiments have shown no significant improvements overall. However, for the training topics we found that indexing MWE enhanced the performance. The methods were used are very simple, and further work will concentrate in improving their results.

In addition to the monolingual experiments, we also submitted four bilingual runs using Spanish topics to query English documents The method used to map concepts between languages employed algorithms for mining association rules. Our bilingual experiments achieved 84% of their monolingual counterparts.

Acknowledgements

This work was partially supported by CNPq Universal 484585/2007-0. Otavio Acosta is funded by a studentship from CAPES.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*. Washington, D.C.

Baldwin, T., & Villavicencio, A. (2002). Extracting the Unextractable: A Case Study on Verb-particles. In *Sixth Conference on Computational Natural Language Learning - CoNLL 2002*.

Geraldo, A. P., & Orengo, V. M. (2008). UFRGS@CLEF2008: Using Association Rules for Cross-Language Information Retrieval. In F. Borri, A. Nardi & C. Peters (Eds.), *Working Notes of CLEF 2008*. Aarhus, Denmark.

Nicholson, J., & Baldwin, T. (2006). Interpretation of Compound Nominalisations Using Corpus and Web Statistic. In *Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

Pearce, D. (2002). A Comparative Evaluation of Collocation Extraction Techniques. In: Third International Conference on Language Resources and Evaluation.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program, 14*(3), 130-137.

Press, W. H., Teukolsky, S. A., Veterling, W. T., & Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*: Cambridge University Press.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., & Ramisch, C. (2007). Validations and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1034-10343). Prague.

Zettair. Retrieved 11/06/07, 2007, from www.seg.rmit.edu.au/zettair/