Identification and Treatment of Multiword Expressions applied to Information Retrieval

Otavio Costa Acosta, Aline Villavicencio, Viviane P. Moreira
Institute of Informatics
Federal University of Rio Grande do Sul (Brazil)
{ocacosta,avillavicencio,viviane}@inf.ufrgs.br

Abstract

The extensive use of Multiword Expressions (MWE) in natural language texts prompts more detailed studies that aim for a more adequate treatment of these expressions. A MWE typically expresses concepts and ideas that usually cannot be expressed by a single word. Intuitively, with the appropriate treatment of MWEs, the results of an Information Retrieval (IR) system could be improved. The aim of this paper is to apply techniques for the automatic extraction of MWEs from corpora to index them as a single unit. Experimental results show improvements on the retrieval of relevant documents when identifying MWEs and treating them as a single indexing unit.

1 Introduction

One of the motivations of this work is to investigate if the identification and appropriate treatment of Multiword Expressions (MWEs) in an application contributes to improve results and ultimately lead to more precise man-machine interaction. The term "multiword expression" has been used to describe a large set of distinct constructions, for instance support verbs, noun compounds, institutionalized phrases and so on. Calzolari et al. (2002) defines MWEs as a sequence of words that acts as a single unit at some level of linguistic analysis.

The nature of MWEs can be quite heterogeneous and each of the different classes has specific characteristics, posing a challenge to the implementation of mechanisms that provide unified treatment for them. For instance, even if a standard system capable of identifying boundaries between words, i.e. a tokenizer, may nevertheless be incapable of recognizing a sequence of words as an MWEs and treating them as a single unit if necessary (e.g. *to kick the bucket* meaning *to die*). For an NLP application to be effective, it requires mechanisms that are able to identify MWEs, handle them and make use of them in a meaningful way (Sag et al., 2002; Baldwin et al., 2003). It is estimated that the number of MWEs in the lexicon of a native speaker of a language has the same order of magnitude as the number of single words (Jackendoff, 1997). However, these ratios are probably underestimated when considering domain-specific language, in which the specialized vocabulary and terminology are composed mostly by MWEs.

In this paper, we perform an application-oriented evaluation of the inclusion of MWE treatment into an Information Retrieval (IR) system. IR systems aim to provide users with quick access to data they are interested (Baeza-Yates and Ribeiro-Neto, 1999). Although language processing is not vital to modern IR systems, it may be convenient (Sparck Jones, 1997) and in this scenario, NLP techniques may contribute in the selection of MWEs for indexing as single units in the IR system. The selection of appropriate indexing terms is a key factor for the quality of IR systems. In an ideal system, the index terms should correspond to the concepts found in the documents. If indexing is performed only with the atomic terms, there may be a loss of semantic content of the documents. For example, if the query was *pop star* meaning *celebrity*, and the terms were indexed individually, the relevant documents may not be retrieved and the system would

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 101–109,
Portland, Oregon, USA, 23 June 2011. ©2011 Association for Computational Linguistics

return instead irrelevant documents about celestial bodies or carbonated drinks. In order to investigate the effects of indexing of MWEs for IR, the results of queries are analyzed using IR quality metrics.

This paper is structured as follows: in section 2 we discuss briefly MWEs and some of the challenges they represent. This is followed in section 3 by a discussion of the materials and methods employed in this paper, and in section 4 of the evaluation performed. We finish with some conclusions and future work.

2 Multiword Expressions

The concept of Multiword Expression has been widely viewed as *a sequence of words that acts as a single unit at some level of linguistic analysis* (Calzolari et al., 2002), or as *Idiosyncratic interpretations that cross word boundaries (or spaces)* (Sag et al., 2002).

One of the great challenges of NLP is the identification of such expressions, "hidden" in texts of various genres. The difficulties encountered for identifying Multiword Expressions arise for reasons like:

- the difficulty to find the boundaries of a multiword, because the number of component words may vary, or they may not always occur in a canonical sequence (e.g. *rock the boat, rock the seemingly intransigent boat* and *the bourgeois boat was rocked*);
- even some of the core components of an MWE may present some variation (e.g. *throw NP to the lions/wolves/dogs/?birds/?butterflies*);
- in a multilingual perspective, MWEs of a source language are often not equivalent to their word-by-word translation in the target language (e.g. *guarda-chuva* in Portuguese as *umbrella* in English and not as *?store rain*).

The automatic discovery of specific types of MWEs has attracted the attention of many researchers in NLP over the past years. With the recent increase in efficiency and accuracy of techniques for preprocessing texts, such as tagging and parsing, these can become an aid in improving the performance of MWE detection techniques. In terms of practical MWE identification systems, a well known approach is that of Smadja (1993), who uses a set of techniques based on statistical methods, calculated from word frequencies, to identify MWEs in corpora. This approach is implemented in a lexicographic tool called *Xtract*. More recently there has been the release of the *mwetoolkit* (Ramisch et al., 2010) for the automatic extraction of MWEs from monolingual corpora, that both generates and validates MWE candidates. As generation is based on surface forms, for the validation, a series of criteria for removing noise are provided, including some (language independent) association measures such as mutual information, dice coefficient and maximum likelihood. Several other researchers have proposed a number of computational techniques that deal with the discovery of MWEs: Baldwin and Villavicencio (2002) for verb-particle constructions, Pearce (2002) and Evert and Krenn (2005) for collocations, Nicholson and Baldwin (2006) for compound nouns and many others.

For our experiments, we used some standard statistical measures such as mutual information, pointwise mutual information, chi-square, permutation entropy (Zhang et al., 2006), dice coefficient, and t-test to extract MWEs from a collection of documents (i.e. we consider the collection of documents indexed by the IR system as our corpus).

3 Materials and Methods

Based on the hypothesis that the MWEs can improve the results of IR systems, we carried out an evaluation experiment. The goal of our evaluation is to detect differences between the quality of the standard IR system, without any treatment for MWEs, and the same system improved with the identification of MWEs in the queries and in the documents. In this section we describe the different resources and methods used in the experiments.

3.1 Resources and Tools

For this evaluation we used two large newspaper corpora, containing a high diversity of terms:

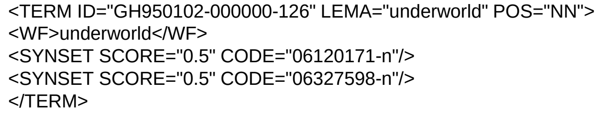
- Los Angeles Times (Los Angeles, USA - 1994)
- The Herald (Glasgow, Scotland - 1995)

Together, both corpora cover a large set of subjects present in the news published by these newspapers in the years listed. The language used is American English, in the case of the Los Angeles Times and British English, in the case of The Herald. Hereafter, the corpus of the Los Angeles Times will be referred as LA94 and The Herald as GH95. Together, they contain over 160,000 news articles (Table 1) and each news article is considered as a document.

Corpus	Documents
LA94	110.245
GH95	56.472
Total	166.717

Table 1: Total documents

The collection of documents, as well as the query topics and the list of relevance judgments (which will be discussed afterwards), were prepared in the context of the CLEF 2008 (*Cross Language Evaluation Forum*), for the task entitled *Robust-WSD* (Acosta et al., 2008). This task aimed to explore the contribution of the disambiguation of words to bilingual or monolingual IR. The task was to assess the validity of word-sense disambiguation for IR. Thus, the documents in the corpus have been annotated by a disambiguation system. The structure of a document contains information about the identifier of a term in a document (`TERM ID`), the lemma of a term (`LEMA`) and also its morphosyntactic tag (`POS`). In addition, it contains the form in which the term appeared in the text (`WF`) and information of the term in the WordNet (Miller, 1995; Fellbaum, 1998) as `SYNSET SCORE` and `CODE`, both not used for the experiment. An example of the representation of a term in the document is shown in Figure 1.



```
<TERM ID="GH950102-000000-126" LEMA="underworld" POS="NN">
<WF>underworld</WF>
<SYNSET SCORE="0.5" CODE="06120171-n"/>
<SYNSET SCORE="0.5" CODE="06327598-n"/>
</TERM>
```

Figure 1: Structure of a term in the original documents

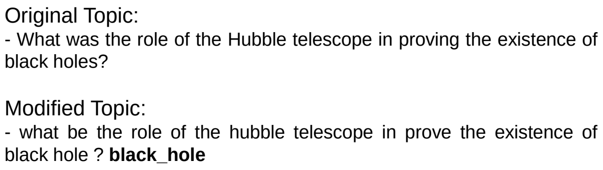
In this paper, we extracted the terms located in the `LEMA` attribute, in other words, in their canonical form (e.g. *letter bomb* for *letter bombs*). The use of lemmas and not the words (e.g. *write* for *wrote*, *written*, etc.) to the formation of the corpus, avoids linguistic variations that can affect the results of the experiments. As a results, our documents were formed only by lemmas and the next step is the indexing of documents using an IR system. For this task we used a tool called *Zettair* (Zettair, 2008), which is a compact textual search engine that can be used both for the indexing and for querying text collections. Porter's Stemmer (Porter, 1997) as implemented in *Zettair* was also used. Stemming can provide further conflation of related terms. For example, *bomb* and *bombing* were not merged in the lemmatized texts but after stemming they are conflated to a single representation.

After indexing, the next step is the preparation of the query topics. Just as the corpus, only the lemmas of the query topics were extracted and used. The test collection has a total of 310 query topics. The judgment of whether a document is relevant to a query was assigned according to a list of relevant documents, manually prepared and supplied with the material provided by CLEF. We used *Zettair* to generate the ranked list of documents retrieved in response to each query. For each query topic, the 1,000 top scoring documents were selected. We used the cosine metric to calculate the scores and rank the documents.

Finally, to calculate the retrieval evaluation metrics (detailed in Section 3.5) we used the tool *trec eval*. This tool compares the list of retrieved documents (obtained from *Zettair*) against the list of relevant documents (provided by CLEF).

3.2 Multiword Expression as Single Terms

In this work, we focused on MWEs composed of exactly two words (i.e. bigrams). In order to incorporate MWEs as units for the IR system to index, we adopted a very simple heuristics that concatenated together all terms composing an MWE using "_" (e.g. *letter bomb* as *letter_bomb*). Figure 2 exemplifies this concatenation. Each bigram present in a predefined dictionary and occurring in a document is treated as a single term, for indexing and retrieval purposes. The rationale was that documents containing specific MWEs can be indexed more adequately than those containing the words of the expression separately. As a result, retrieval quality should increase.



Original Topic:
- What was the role of the Hubble telescope in proving the existence of black holes?

Modified Topic:
- what be the role of the hubble telescope in prove the existence of black hole ? **black_hole**

Figure 2: Modified query.

3.3 Multiword Expressions Dictionaries

In order to determine the impact of the quality of the dictionary used in the performance of the IR system, we examined several different sources of MWE of varying quality. The dictionaries containing the MWEs to be inserted into the corpus as a single term, are created by a number of techniques involving automatic and manual extraction. Below we describe how these MWE dictionaries were created.

- **Compound Nouns (CN)** - for the creation of this dictionary, we extracted all bigrams contained in the corpus. Since the number of available bigrams was very large (99,744,811 bigrams) we filtered them using the information in the original documents, the morphosyntactic tags. Along with the `LEMA` field, extracted in the previous procedure, we also extracted the value of the field *POS* (*part-of-speech*). In order to make the experiment feasible, we used only bigrams formed by compound nouns, in other words, when the `POS` of both words was NN (*Noun*). Thus, with bigrams consisting of sequences of NN as a preprocessing step to eliminate noise that could affect the experiment, the number of bigrams with MWE candidates was reduced to 308,871. The next step was the selection of bigrams that had the highest frequency in the text, so we chose candidates occurring at least ten times in the whole corpus. As a result, the first list of MWEs was composed by 15,001 bigrams, called *D1*.

- **Best Compound Nouns** - after D1, we refined the list with the use of statistical methods. The methods used were the mutual information and chi-square. It was necessary to obtain frequency values from Web using the search tool *Yahoo!*, because despite the number of terms in the corpus, it was possible that the newspaper genre of our corpus would bias the counts. For this work we used the number of pages in which a term occurs as a measure of frequency. With the association measures based on web frequencies, we generated a ranking in decreasing order of score for each entry. We merged the rankings by calculating the average rank between the positions of each MWE; the first 7,500 entries composed the second dictionary, called *D2*.

- **Worst Compound Nouns** - this dictionary was created from bigrams that have between five and nine occurrences and are more likely to cooccur by chance. It was created in order to evaluate whether the choice of the potentially more noisy MWEs entailed a negative effect in the results of IR, compared to the previous dictionaries. The third dictionary, with 17,328 bigrams, is called *D3*.

- **Gold Standard** - this was created from a sublist of the Cambridge International Dictionary of English (Procter, 1995), containing MWEs. Since this list contains all types of MWEs, it was necessary to further filter these to obtain compound nouns only, using morphosyntactic information obtained by the TreeTagger (Schmid, 1994), which for English is reported to have an accuracy of 96.36%" (Schmid, 1994). Formed by 568 MWEs, the fourth dictionary will be called *D4*.

- **Decision Tree** - created from the use of the J48 algorithm (Witten and Frank, 2000) from *Weka* (Hall et al., 2009), a data mining tool. With this algorithm it is possible to make a MWE classifier in terms of a decision tree. This requires providing training data with true and false examples of MWE. The training set contained 1,136 instances, half true (*D4*) and half false MWEs (taken from *D3*). After combining several statistical methods, the best result for classification was obtained with the use of mutual information, chi-square, pointwise mutual information, and Dice. The model obtained from Weka was applied to test data containing 15,001 MWE candidates (*D1*). The 12,782 bigrams classified as true compose the fifth dictionary, called *D5*.

- **Manual** - for comparative purposes, we also created two dictionaries by manually evaluating the text of the 310 query topics. The first dictionary contained all bigrams which would achieve a different meaning if the words were concatenated (e.g. *space shuttle*). This dictionary, was called *D6* and contains 254 expressions. The other one was created by a specialist (linguist) who classified as true or false a list of MWE candidates from the query topics. The linguist selection of MWEs formed *D7* with 178 bigrams.

3.4 Creating Indices

For the experiments, we needed to manipulate the corpus in different ways, using previously built dictionaries. The MWEs from dictionaries have been inserted in the corpus as single terms, as described before. For each dictionary, an index was created in the IR system. These indices are described below:

1. **Baseline (BL)** - corpus without MWE.
2. **Compound Nouns (CN)** - with 15 MWEs of *D1*.
3. **Best CN (BCN)** - with 7,500 MWEs of *D2*.
4. **Worst CN (WCN)** - with 17,328 MWEs of *D3*.
5. **Gold Standard (GS)** - with 568 MWEs of *D4*.
6. **Decision Tree (DT)** - with 12,782 MWEs of *D5*.
7. **Manual 1 (M1)** - with 254 MWEs of *D6*.
8. **Manual 2 (M2)** - with 178 MWEs of *D7*.

3.5 Evaluation Metrics

To evaluate the results of the IR system, we need to use metrics that estimate how well a user's query was satisfied by the system. IR evaluation is based on recall and precision. Precision (Eq. 1) is the portion of the retrieved documents which is actually relevant to the query. Recall (Eq. 2) is the fraction of the relevant documents which is retrieved by the IRS.

$$Precision(P) = \frac{\#Relevant \bigcap \#Retrieved}{\#Retrieved} \quad (1)$$

$$Recall(R) = \frac{\#Relevant \bigcap \#Retrieved}{\#Relevant} \quad (2)$$

Precision and Recall are set-based measures, therefore, they do not take into consideration the ordering in which the relevant items were retrieved. In order to evaluate ranked retrieval results the most widely used measurement is the *average precision* (AvP). AvP emphasizes returning more relevant documents earlier in the ranking. For a set of queries, we calculate the *Mean Average Precision* (MAP) according to Equation 3 (Manning et al., 2008).

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \quad (3)$$

where $|Q|$ is the number of queries, R_{jk} is the set of ranked retrieval results from the top result until document d_k, and m_j is the number of relevant documents for query j.

4 Experiment and Evaluations

The experiments performed evaluate the insertion of MWEs in results obtained in the IR system. The analysis is divided into two evaluations: (A) total set of query topics, where an overview is given of the MWE insertion effects and (B) topics modified by MWEs, where we evaluate only the query topics that contain MWEs.

4.1 Evaluation A

This evaluation investigates the effects of inserting MWEs in documents and queries. After each type of index was generated, MWEs were also included in the query topics, in accordance to the dictionaries used for each index (for Baseline BL, the query topics had no modifications).

With eight corpus variations, we obtained individual results for each one of them. The results presented in Table 2 were summarized by the absolute number of relevant documents retrieved and the MAP for the entire set of query topics. In total, 6,379 relevant documents are returned for the 310 query topics.

Index	Rel. Retrieved	MAP
BL	3,967	0.1170
CN	4,007	0.1179
BCN	3,972	0.1156
WCN	3,982	0.1150
GS	3,980	0.1193
DT	4,002	0.1178
M1	4,064	0.1217
M2	4,044	0.1205

Table 2: Results — Evaluation A.

It is possible to see a small improvement in the results for the indices M1 and M2 in relation to the baseline (BL). This happens because the choice of candidate MWEs was made from the contents of the document topics and not, as with other indices, from the whole corpus. Considering the indices built with MWEs extracted from the corpus, the best result is index GS.In second place, comes the CN index, with a subtle improvement over the Baseline. BL surprisingly got a better result than the Best and Worst CN. The loss in retrieval quality as a result from MWE identification for BCN was not expected.

When comparing the gain or loss in MAP of individual query topics, we can see how the index BCN compares to the Baseline: BCN had better MAP in 149 and worse MAP in 108 cases. However, the average loss is higher than the average gain, this explains why BL obtains a better result overall. In order do decide if one run is indeed superior to another, instead of using the absolute MAP value, we chose to calculate a margin of 5%. The intuition behind this is that in IR, a difference of less than 5% between the results being compared is not considered significant (Buckley and Voorhees, 2000). To be considered as gain the difference between the values resulting from two different indices for the same query topic should be greater than 5%. Differences of less than 5% are considered ties. This way, MAP values of 0.1111 and 0.1122 are considered ties. Given this margin, we can see in Tables 3 and 4 that the indices BCN and WCN are better compared to the baseline. In the case of BCN, the gain is almost 20% of cases and the WCN, the difference between gain and loss is less than 2%.

Gain	60	19.35%
Loss	35	11.29%
Ties	215	69.35%
Total	310	100.00%
Difference between Gain and Loss		8,06%

Table 3: BCN x Baseline

Gain	26	8.39%
Loss	21	6.77%
Ties	263	84.84%
Total	310	100.00%
Difference between Gain and Loss		1.61%

Table 4: WCN x Baseline

Finally, this first experiment guided us toward a deeper evaluation of the query topics that have MWEs, because there is a possibility that the MWE insertions in documents can decrease the accuracy of the system on topics that have no MWE.

4.2 Evaluation B

This evaluation studies in detail the effects on the document retrieval in response to topics in which there were MWEs. For this purpose, we used the same indices used before and we performed an individual evaluation of the topics, to obtain a better understanding on where the identification of MWEs improves or degrades the results.

As each dictionary was created using a different methodology, the number of expressions contained in each dictionary is also different. Thus, for each method, the number of query topics considered as having MWEs varies according to the dictionary used. Table 5 shows the number of query topics containing MWEs for each dictionary used, and as a consequence, the percentage of modified query topics over the complete set of 310 topics.

First, it is interesting to observe the values of MAP for all topics that have been altered by the identification of MWEs. These values are shown in Table 6.

As shown in Table 6 we verified that the GS index obtained the best result compared to others. This

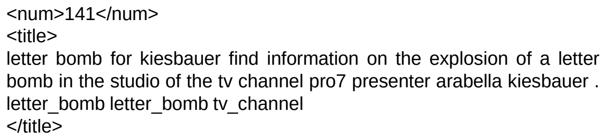
Index	Topics with MWEs	% Modified
BL	0	0.00%
CN	75	24.19%
BCN	41	13.23%
WCN	28	9.03%
GS	9	2.90%
DT	51	16.45%
M1	195	62.90%
M2	152	49.03%

Table 5: Topics with MWEs

Index	MAP
CN	0.1011
BCN	0.0939
WCN	0.1224
GS	0.2393
DT	0.1193
M1	0.1262
M2	0.1236

Table 6: Results - Evaluation B

was somewhat expected since the MWEs in that dictionary are considered “real” MWEs. After GS, best results were obtained from the manual indices M1 and M2. The index that we consider as containing the lowest confident MWEs (WCN), obtained better results than Decision Trees, Nominal Compounds and Best Nominal Compounds, in this order. One possible reason for this to happen is that the number of MWEs inserted is higher than in the other indices. Compared with the BL, all indices with MWE insertion have improved more than degraded the results, in quantitative terms. Our largest gain was with the index GS, where 55.56% of the topics have improved, but the same index showed the highest percentage of loss, 22.22%. Analyzing the WCN, we can identify that this index has the lowest gain compared to all other indices: 32.14%, although having also the lowest loss. But, 60.71 % of the topics modified had no significant differences compared to the Baseline. Thus, we can conclude that the WCN index is the one that modifies the least the result of a query. The indices CN and BCN had a similar result, and knowing that a dictionary used to create BCN is a subset of the dictionary CN, we can conclude that the gain values, choosing the best MWE candidates, does not affect the accuracy, which only improves subtly. But the computational cost for the insertion of these MWEs in the corpus was reduced by half. In terms of gain percentage, indices M1 and M2 were superior only to WCN, but they are close to other results, including the DT index, which obtained an intermediate result between manual dictionaries and CN. Analyzing some topics in depth, like topic 141 (Figure 3), the best the result among all the indices was obtained by the CN.



```
<num>141</num>
<title>
letter bomb for kiesbauer find information on the explosion of a letter
bomb in the studio of the tv channel pro7 presenter arabella kiesbauer .
letter_bomb letter_bomb tv_channel
</title>
```

Figure 3: Topic #141

Table 7 shows the top ten scoring documents retrieved for query topic 141 in the baseline. The relevant document (in bold) is the fourth position in the Baseline. After inserting the expression *letter bomb* twice (because it occurs twice in the original topic), and *tv channel* that were in dictionary D1 used by the CN index, the relevant document is scored higher and as a consequence is returned in the first position of the ranking(Table 8) . The MAP of this topic has increased 75 percentage points, from 0.2500 in Baseline to 1.000 in the CN index. We see also that the document that was in first position in the Baseline ranking, has its score decreased and was ranked in fourth position in the ranking given by the CN. This document contained information on a “small bomb located outside the of the Russian embassy” and has is not relevant to topic 141, being properly relegated to a lower position.

An interesting fact about this topic is that only the MWE *letter bomb* influences the result. This was verified as in the index BCN, whose dictionary does not have this MWE, the topic was changed only because of the MWE *tv channel* and there was no gain or loss for the result.

The second highest gain was of M1 index, in topic 173. The gain was of 28 percentage points. On the other hand, we found a downside in M1 and M2 indices, although they improved results on average, they have reached very high values of loss in some topics.

Position	Document	Score
P1	LA043094-0230	0.470900
P2	GH950823-000105	0.459994
P3	GH951120-000182	0.439536
P4	**GH950610-000164**	**0.430784**
P5	GH950614-000122	0.428766
P6	LA091894-0425	0.428429
P7	GH950829-000082	0.422941
P8	GH950220-000162	0.411968
P9	GH950318-000131	0.406006
P10	GH950829-000037	0.402806

Table 7: Ranking for Topic #141 - Baseline

Position	Document	Score
P1	**GH950610-000164**	**0.457950**
P2	GH950614-000122	0.436753
P3	GH950823-000105	0.423938
P4	LA043094-0230	0.421757
P5	GH951120-000182	0.400123
P6	GH950829-000082	0.393195
P7	LA091894-0425	0.386613
P8	GH950705-000100	0.384116
P9	GH950220-000162	0.382157
P10	GH950318-000131	0.380471

Table 8: Ranking for Topic #141 - CN

In sum, the MWEs insertion seems to improve retrieval bringing more relevant documents, due to a more precise indexing of specific terms. However, the use of these expressions also brought a negative impact for some cases, because some topics require a semantic analysis to return relevant documents (as for example topic 130, which requires relevant documents to mention the causes of the death of Kurt Cobain — documents which mention his death without mentioning the causes were not considered relevant).

5 Conclusions and Future Work

This work consists in investigating the impact of Multiword Expressions on applications, focusing on compound nouns in Information Retrieval systems, and whether a more adequate treatment for these expressions can bring possible improvements in the indexing these expressions. MWEs are found in all genres of texts and their appropriate use is being targeted for study, both in linguistics and computing, due to the different characteristic variations of this type of expression, which ends up causing problems for the success of computational methods that aim their processing.

In this work we aimed at achieving a better understanding of several important points associated with the use of Multiword Expressions in IR systems. In general, the MWEs insertion improves the results of retrieval for relevant documents, because the indexing of specific terms makes it easier to retrieve specific documents related to these terms. Nevertheless, the use of these expressions made the results worse in some c]ases, because some topics require a semantic analysis to return relevant documents. Some of these documents are related to the query, but do not satisfy all criteria in the query topic. We conclude also that the quality of MWEs used directly influenced the results.

For future work, we would like to use other MWE types and not just compound nouns as used in this work. Other methods of extraction and a further study in Named Entities are good themes to complement this subject. A variation of corpora, different from newspaper articles, because each domain has a specific terminology, can also be an interesting subject for further evaluation.

References

Otavio Acosta, Andre Geraldo, Viviane Moreira Orengo, and Aline Villavicencio. 2008. Ufrgs@clef2008: Indexing multiword expressions for information retrieval. Aarhus, Denmark. Working Notes of the Workshop of the Cross-Language Evaluation Forum - CLEF.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. Sixth Conference on Computational Natural Language Learning - CoNLL 2002.

Timothy Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of multiword expression decomposability. ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA. ACM.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. Third International Conference on Language Resources and Evaluation - LREC.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech & Language - Special Issue on Multiword Expression - Volume 19, Issue 4, p. 450-466.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update.

Ray Jackendoff. 1997. The architecture of the language faculty. MIT Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press. 1394399.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41, November.

Jeremy Nicholson and Timothy Baldwin. 2006. Interpretation of compound nominalisations using corpus and web statistic. Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. Third International Conference on Language Resources and Evaluation.

Martin F. Porter. 1997. An algorithm for suffix stripping. pages 313–316, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Paul Procter. 1995. *Cambridge international dictionary of English*. Cambridge University Press, Cambridge, New York.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild? the mwetoolkit comes in handy. In *Coling 2010: Demonstrations*, pages 57–60, Beijing, China, August. Coling 2010 Organizing Committee.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickiger. 2002. Multiword expressions. a pain in the neck for nlp. Third International Conference on Computational Linguistics and intelligent Text Processing.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. Computational Linguistics.

Karen Sparck Jones. 1997. What is the role of nlp in text retrieval? University of Cambridge.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Zettair. 2008. The zettair search engine. (disponível via WWW em http://www.seg.rmit.edu.au/zettair).

Yi Zhang, Valia Kordoni, Aline. Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.