Using Statistical Features to Find Phrasal Terms in Text Collections

André Luiz da Costa Carvalho[1], Edleno Silva de Moura[1], Pável Calado[2]

[1] Universidade Federal do Amazonas, Brazil
andre@ufam.edu.br, edleno@dcc.ufam.edu.br
[2] Instituto Superior Técnico/INESC-ID, Portugal
pavel.calado@tagus.ist.utl.pt

Abstract. In this work we investigate alternatives to automatically detect phrasal terms, defined here as phrasal verbs, phrasal nouns, phrasal adjectives or phrasal adverbs found in a text. The automatic identification of phrasal terms may have several applications in text processing systems. We approach this problem and present a novel approach for detecting phrasal terms in a collection of documents. Our solution is based on machine learning and uses statistical features of the word n-grams found in the documents. We also investigate the particular impact of adding phrasal terms in the retrieval model of a search engine when processing queries on several data sets. Our results show that we are able to discover valid phrasal terms with a small error rate, achieving detection results ranging from 70% to 94% in terms of F1. Furthermore, the discovered phrasal terms, when used to enhance search tasks, allow improvements in retrieval performance of up to 11% in terms of MAP when considering all queries, and up to 36% in terms of MAP when considering only the queries that contained the detected phrasal terms.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords: phrasal terms, phrase queries

1. INTRODUCTION

In this work, we are interested in automatically detecting phrasal terms and using them to improve information retrieval tasks. We consider a *phrasal term* as a sequence of words that has a function of phrasal verb, phrasal noun, phrasal adjective or phrasal adverb in a text. Such phrasal terms are sequences of words that have a specific meaning, which can even be different from that of the individual words that compose the phrasal term. Examples of phrasal terms are "artificial intelligence", "comic book", "tax free", "legal issues", "formula one", "New York", and so on.

The task of finding phrasal terms may be associated with several applications related to text processing, such as text classification, search and keyword finding [tau Yih et al. 2006]. The objective of identifying and modeling phrasal terms is to produce a better representation of text documents when compared to the processing of only individual words as units. Individual words can, for instance, be composed together to build expressions with completely different meanings and such information may be missed when not modeling compositions. A prior knowledge of which word sequences have a specific meaning can be used for many practical purposes when processing textual sources of information, enriching the representation of the content found on those sources.

As a sample of application of phrasal term detection methods, we can mention the enrichment of traditional information retrieval models. Most popular textual information retrieval models represent

This work is partially supported by CNPq Ph.D. Scholarship (André Luiz da Costa Carvalho), CNPq individual grant 302209/2007-7 (Edleno Silva de Moura); SIRIAA project, CT-Amazônia/CNPq/553126/2005-9

documents as “bags of words”, i.e., sets of completely independent distinct terms [Baeza-Yates and Ribeiro-Neto 1999], disregarding the order in which the words appear in the document. The extraction of these independent words is done by the use of text parsers, which usually separate these terms in strings of alphanumeric characters separated by non-alphanumeric characters. After the parsing, each term is considered as an independent piece of information present in the document, with total disregard to its placing and the terms around it.

While this approach has yielded satisfactory results, it is clear that considering the words in a document as independent is not a faithful representation of the actual relationship between them. Words can be composed together to build expressions with completely different meanings (e.g., “artificial intelligence”, or “comic book”). Considering only individual words as terms can lead to loss of information, which can mislead the system and lead to erroneous results due to ambiguity. For instance, the word “gore” in a query might mean that the user is interested in horror movies; however if in the query before the word “gore” is the word “al”, it probably means that the user is interested in the politician Al Gore.

Here we argue that a prior knowledge of which word sequences have a specific meaning is important and can be used for many practical purposes in Information Retrieval, enriching the information available about documents. This is particularly important in web search applications, where the precision of results is more important than the recall; and where the most common type of queries, the informational queries, usually have larger numbers of words [Baeza-Yates et al. 2006]. This property raises the chance of informational queries containing phrasal terms. Further, informational queries rely mostly on the textual content as a source of information to improve the quality of search results, thus a better representation of texts in the information retrieval model may have higher impact in these web search engine queries.

Word sequences and word co-occurrence statistics have been shown useful in many different information retrieval tasks [Tan et al. 2002; Tesar et al. 2006; tau Yih et al. 2006; Zhang et al. 2007]. These examples of previous efforts to deal with word sequences in different information retrieval tasks represent a strong indication that indeed the detection and employment of meaningful word sequences can lead to improvements in the quality of various information retrieval tasks.

Our solution to detect phrasal terms uses statistical features of the word *n-grams* (sequences of n words) found in the documents and a classifier to determine which are valid phrasal terms. We present empirical evidence showing that this is an effective solution to the problem. The intuition behind this approach is that phrasal terms share similar statistical features that might be able to distinguish them from non-phrasal sequences. It provides an effective mean of discovering meaningful word sequences that is grammar independent (at least, in regard to western languages) and time and space efficient, since it does not require the use of complex Natural Language Processing methods. Experiments indicate that, as the database grows larger the effectiveness of the proposed also increases, achieving precision values up to 94,45%.

We also show that using these phrasal terms can indeed lead to better results in information retrieval tasks. To this effect, we propose two different approaches that are able to use this information to enhance results in document search tasks, and perform experiments demonstrating the usefulness of our approach. Experiments in four different collections show gains of up to 36% in Mean Average Precision, when compared to a traditional single-word search.

This paper is organized as follows: In Section 2 we present work related to phrasal term detection and its applications. In Section 3 we present the proposed approach to detect phrasal terms. In Section 4 we present how to use the phrasal terms to improve results in text document search tasks. In Section 5 we present the experiments performed to evaluate the quality of the phrasal term detection approach and also the gains obtained by its use when searching for documents. Finally, in Section 6, conclusions and future work are presented.

2. RELATED WORK

Zhang et al. [Zhang et al. 2007] tackled the problem of noun phrase detection and its possible application to enhancing search engine queries. Noun phrases can be considered a subset of phrasal terms, since phrasal terms do not necessarily have to be nouns. The method proposed by Zhang uses a number of different datasets, such as Wikipedia[1] and WordNet [Miller et al. 1990], Natural Language parsers and Google search[2] as sources of evidence, combined in order to detect noun phrases. This work provides good evidence that a subset of phrasal terms (noun phrases), can lead to improvements in text search. However, while the objective of this work is similar to ours (besides the broader spectrum of phrasal terms in comparison to noun phrases), the use of external databases and natural language processing techniques make the method unsuitable to many information retrieval scenarios, due to its high computational cost and dependency from external sources. In contrast, the method we here propose relies only on the text collection itself to find phrasal terms, and this processing can be done offline at indexing time, i.e., prior to the actual use of the text collection.

Mladenik et al. [Mladenic and Grobelnik 1998] made experiments using n-grams (up to 5-grams) to improve the performance of a naive Bayes classifier. The authors found out that the highest improvement was achieved when adding 2-grams to the classification, and also that n-grams with n larger than 3 actually had no positive influence in the classification.

Tesar et al. [Tesar et al. 2006] proposed the use of 2-itemsets (selected sets of two words co-occurring in documents) instead of *bigrams* (sequences of two words) to enhance classification, and made experiments comparing the use of both with a number of feature selection techniques. While the results achieved for both bigrams and 2-itemsets were superior to the ones achieved when considering only unigrams, the results obtained by adding bigrams were consistently superior to those with 2-itemsets, indicating that bigrams are a better option for classification purposes since computing 2-itemsets is also more expensive.

Based on the assumption that the use of all the bigrams present in the pages would add noise to the categorization, Tan et al. [Tan et al. 2002] applied a feature selection method to select only bigrams that are likely to be useful for the classification. The authors proposed an algorithm based on the information gain metric, combined with some frequency thresholds to select these bigrams, achieving significant gains in classification. This approach, however, is focused on classification and can not be trivially adapted to be used in other information retrieval tasks, in contrast with our method. Also, it is orthogonal to our method, since it is not interested in finding meaningful bigrams , but simply bigrams that are good class discriminators.

Finally, it is interesting to mention that in [Ekkerman and Allan 2003], Ekkerman et al. claim that, up to that point, many efforts have been made to improve text categorization by the use of bigrams, with little to no observable improvement in the classification results. He also proposed a method to include bigrams information in the classification, based on distributional clusters, but this method yielded statistically insignificant improvements in the quality of results. As discussed above, subsequent works have had more success in this task.

3. DETECTING PHRASAL TERMS

We define a *phrasal term* as a sequence of words that has a function of a phrasal verb, phrasal noun, phrasal adjective or phrasal adverb in a text. It is important to notice that, in our definition of phrasal term, it is crucial that the set of words has a specific, understandable, meaning. For instance, while “Carnegie Mellon University” is a phrasal term, “Mellon University” is not since it loses its meaning without the term “Carnegie”.

[1] www.wikipedia.com
[2] www.google.com

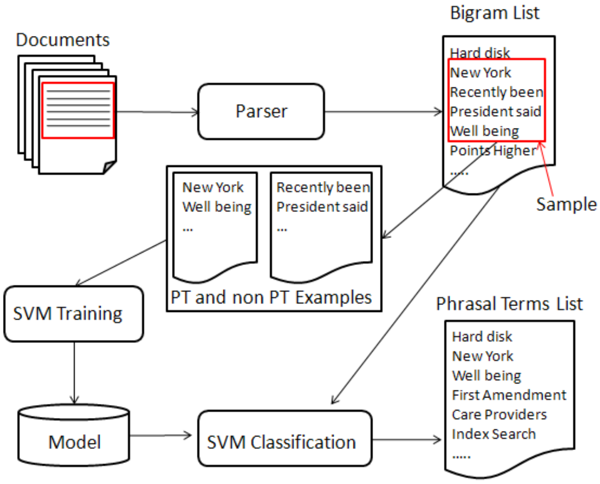

Fig. 1. The phrasal term (PT) detection process.

Our detection method, based on machine learning, uses statistical features of the word bigrams found in the documents to classify each of them as phrasal terms or not. The intuition behind this approach is that phrasal terms share similar statistical features that might be able to distinguish them from non-phrasal sequences. We believe that this happens because the co-occurrence of words in a language is governed by a fixed set of rules—the grammar. However, it is important to stress that the proposed approach is not based on the grammar itself, but in examples of phrasal terms and non-phrasal sequences.

In the experiments of this paper, we focus on the detection of two-word phrasal terms. It can, nevertheless, be easily expandable to n-word sequences. Also, we ignored n-grams that contained numerical characters and stop-words, in order to prune the number of n-grams considered. Preliminary experiments done without the exclusion of these bigrams showed that they introduce noise into the classification, with no distinguishable advantage over their exclusion.

The proposed method can be divided into three main stages, as illustrated in Figure 1. In the first stage, we extract all existing bigrams from the document collection and compute their features. To do so, the method parses all documents in the collection, collecting statistics about the individual terms and about the bigrams. This stage has a computational cost linear to the number of documents, and it is important to note that this stage can be trivially implemented during the indexing of the collection.

In the second stage, we manually label a small subset of bigrams as phrasal terms or non-phrasal sequences. This subset can then be used as training data for a Support Vector Machine (SVM) classifier.

Finally, in the third stage, we use the SVM classifier to process all bigrams and determine which are proper phrasal terms. It is important to stress that this classification is fairly inexpensive, especially because the method uses a reduced number of features.

3.1 Identifying Phrasal Term Candidates

As we mentioned, we have worked in this article with only phrasal terms composed of two individual words. At first glance, identifying all phrasal terms existing within all possible bigrams of text in a text collection can be seen as a very complex operation. In fact, if we consider all possible two-word sequences, the maximum number of candidates is $|V|^2$, where $|V|$ is the size of the vocabulary.

However, in real scenarios, the majority of possible sequences never actually occur, largely reducing this number. A simple algorithm, linear in the size of the collection, to find all bigrams present in a collection is:

(1) For each document in the collection.
    (a) Parse the document and extract every sequence of two words and its statistics.
    (b) Whenever main memory is full, dump all collected sequences to secondary storage.
(2) Join all runs of sequences produced into a single list of candidates.

As stated before, in this step the method ignores bigrams with stop-words and with numeric characters, which reduces the number of bigrams to be processed.

3.2 Statistical Features

We argue that only a basic set of word occurrence and co-occurrence statistics is needed to determine whether a bigram is a meaningful phrasal term or not. Even though no language-dependent grammar or syntactic information is used, we believe such statistics carry enough information to provide an accurate description of what is and what is not a phrasal term.

The features we adopt here are mostly based on frequency counting, and have a straightforward calculation, adding little computational cost to the feature extraction. Given an *ordered pair of words* (t_1, t_2), we propose the following set of features:

(1) **Pair probability** $(P(t_1, t_2))$: The number of times pair (t_1, t_2) occurs in the collection, over the number of pairs in the collection. A large frequency may indicate that a bigram is a possible phrasal term. On the other hand, a small frequency is not necessarily an indication that the bigram is not a phrasal term.
(2) **Pair Probability given** t_1 $(P((t_1, t_2)|t_1))$: The number of times pair (t_1, t_2) occurs, over the number of times that t_1 occurred. This feature indicates how likely it is that the presence of t_1 is an indication that the next word is t_2. We argue that t_1 being often followed by t_2 is an indication that the pair is a phrasal term.
(3) **Pair Probability given** t_2 $(P((t_1, t_2)|t_2))$: The number of times pair (t_1, t_2) occurs, over the number of pairs that end with t_2. Similar to the Pair Probability given t_1 feature.
(4) **Document Probability of** (t_1, t_2) $(P_D((t_1, t_2)))$: The number of documents where pair (t_1, t_2) occurs, over the total number of documents. Notice that this is different from Pair Probability, since it relates to the number of documents where the pair appears, instead of its absolute number of occurrences.
(5) **Document Probability of** t_1 $(P_D(t_1))$: The number of documents where word t_1 occurs, over the total number of documents. This feature should reflect the importance of word t_1, in a role similar to that of the inverse document frequency (IDF), traditionally used in the Vector Space Model [Baeza-Yates and Ribeiro-Neto 1999].
(6) **Document Probability of** t_2 $(P_D(t_2))$: The number of documents where word t_2 occurs, over the total number of documents.
(7) **Raw Pair Probability given** t_1 $(P_R(t_1))$: The number of distinct pairs starting with t_1, over the number of distinct pairs.
(8) **Raw Pair Probability given** t_2 $(P_R(t_2))$: The number of distinct pairs ending with t_2, over the number distinct of pairs.

It is important to notice that, while the proposed set of features is fairly simple, the experiments presented in section 5 show that these features are good enough to allow for a satisfactory precision in the phrasal terms detection task. The importance of each feature is studied in section 5.2.1.

After all the bigrams and their statistics are extracted, examples of phrasal terms and non-phrasal sequences are manually extracted from a random sample of bigrams, and used to train an SVM classifier model. The remaining feature vectors for all the bigrams are applied to this classifier model, which will determine whether each bigram is a meaningful phrasal term or not. As a result of these steps, a list of detected phrasal terms is created.

We note that, while in this work we use an SVM classifier to select phrasal terms, other classification techniques could be used.

In the following section, we describe how these discovered phrasal terms can be used for enhancing search in text collections.

4. USING PHRASAL TERMS

After the detection of phrasal terms, it is important to study how this information can be used in Information Retrieval tasks. In this work, we study the particular case of text document searching. We chose to evaluate the impact of phrasal terms using the traditional Vector Space Model (VSM) with TF-IDF term weighting [Baeza-Yates and Ribeiro-Neto 1999] and propose two different approaches to apply phrasal terms in a search task:

(1) **Phrasing:** Phrasing consists in parsing the user query in order to find phrasal terms. Whenever one is found, it is treated by the search engine as a *phrasal query*. Phrasal queries are queries in which, for a document to be considered relevant, it is mandatory that it possesses the designated phrase. The phrases are usually represented by the use of quotes. For instance, if our method detected "Los Angeles" as a phrasal term, the query '*Los Angeles Hotel*' would be processed by the search engine as the query '*"Los Angeles" Hotel*', with the search engine considering "Los Angeles" as a phrase.

(2) **Expansion with Phrasal Terms:** In this case, all the detected phrasal terms in the collection are treated as single-word terms, and indexed by the search engine as such. All individual words are still indexed. In practice, we are expanding the documents in the collection to contain, besides all its individual words, also all its phrasal terms. All document norms are recomputed taking this into account. The same process is applied to the query. For instance, in a scenario where the proposed approach detected "Los Angeles" as being a phrasal term, the query '*Los Angeles Hotel*' would be expanded into a query with four distinct terms: "*Los*", "*Angeles*", "*Hotel*", and it would be expanded to contain the phrasal term "*Los Angeles*". The same would apply to all documents.

5. EXPERIMENTS

The experiments are divided in two main parts: phrasal term detection and search impact evaluation. Before presenting them, we detail the experimental setup.

5.1 Experimental Setup

In the experiments performed, we use four datasets: FBIS, Wt10G, L.A. Times and OHSUMED. We chose these four datasets since they contain documents of different natures, topics and sizes, which allowed us to examine the behavior of our method in distinct scenarios. Furthermore, all these datasets have an avaiable previously evaluated set of queries, which allowed us to use them in the search experiments.

The Wt10G collection [Soboroff 2002] is comprised of 1,692,097 documents selected from a crawl of the World Wide Web. This dataset is useful to show how the methods behave in a free-publication environment, such as the Web. While Wt10G may be considered small when compared to true web search scenarios, previous studies have shown that this dataset, despite being only a sample, retains many of the properties of larger web crawls [Soboroff 2002].

The OHSUMED document collection [Hersh et al. 1994] is a subset of MEDLINE, containing 348,566 medical articles. For this work, we indexed only the title and abstract of the documents.

The FBIS dataset is part of the TREC document collection (disk 5) [Voorhees 1999]. It contains 130,471 randomly selected newspaper articles. This collection, while having fewer documents than OHSUMED, actually contains more text, since it contains full newspaper articles, regarding news from all around the world.

Finally, the L.A. Times collection is also a part of the TREC document collection (disk 5) [Voorhees 1999]. It contains about 131,896 randomly selected newspaper articles, published between 1989 and 1990 in the Los Angeles Times newspaper. Unlike FBIS, all articles are taken from a single newspaper, hence the collection has a more limited vocabulary and set of topics.

Table I shows some statistics about the databases. We can clearly see that, while containing more documents than L.A.Times and FBIS, OHSUMED has a smaller vocabulary and a smaller number of bigrams. It is also important to notice that the removal of bigrams with numbers and with stop-words had a huge impact in the number of bigrams, removing about 50% of the bigrams present in Wt10G.

	Voc. Size	Total Bigrams	Considered
OHSUMED	164407	5590058	2971596
FBIS	400073	7595417	3842587
L.A.Times	248660	9440633	5518509
Wt10G	5646913	91143120	47885424

Table I. Statistics about the text collections utilized in the experiments. Total Bigrams is the absolute number of bigrams in the database. Considered is the number of bigrams without stop-words and numeric characters.

5.1.1 *Classification Setup.* As described in Section 3, in this work we used an SVM classifier to discover sets of phrasal terms from a set of bigrams. The SVM implementation used was *libsvm* [Chang and Lin 2001]. The kernel adopted was the RBF kernel and the parameters were optimized using scripts included in libsvm.

Bigrams with frequency values smaller than 10 were discarded as phrasal term candidates, since they contain too little co-occurrence information. Also, since these bigrams are so infrequent, the possible gain yielded by detecting phrasal terms among them may be small, in comparison with the overhead necessary to classify them all.

For each database used in this paper, 500 phrasal terms and 500 non-phrasal sequences were randomly sampled from the bigrams found in the databases, to be used as training and testing sets. The training set was stratified [Witten and Frank 2000] in this manner because the set of bigrams that is not a phrasal term is much larger than the set of phrasal term bigrams, and thus a purely random sample would also have a disproportional number of non-phrasal terms, which could bias the classifier, and preliminary experiments done with an unstratified set of examples led to a classifier with overall much worse results. All tests were performed using 10-fold cross validation. The precision and recall values obtained represent the average of the 10 runs. While in the experiments presented we used 1000 examples in the construction of our classification model, in all databases when using as few as 100 examples to construct the model our method yielded similar precision and recall values as using all the 1000 examples.

In the case of OHSUMED, which is compromised mostly of medical terms, technical expertise was needed in order to identify the phrasal and non-phrasal terms. Thus, to do so we asked for a Pharmacy Msc to do this identification.

To evaluate the phrasal term detection, we used the traditional classification metrics of micro-averaged precision, macro-averaged precision, macro-averaged recall and macro-averaged $F1$ [Witten and Frank 2000]. We also present the total precision and recall for the phrasal terms class and non-phrasal sequences class. These measures are defined as follows:

Let C be the set of bigrams used for testing. We define $C_{phrasal} \subseteq C$ as the set of phrasal terms within C and $C_{not} \subseteq C$ as the set of non-phrasal sequences within C. Let $C'_{phrasal}$ be the set of bigrams classified as phrasal terms by the SVM classifier, and let C'_{not} be the set of bigrams classified as non-phrasal sequences by the SVM classifier.

The precision of the SVM classifier for a given class $c \in \{phrasal, not\}$ is the percentage of bigrams correctly classified as being of class c, over all bigrams classified as being of class c:

$$Pr_c = \frac{|C'_c \cap C_c|}{|C'_c|} \tag{1}$$

Recall is defined as the percentage of bigrams correctly classified as being of class c, over all bigrams that actually belong to class c:

$$Rc_c = \frac{|C'_c \cap C_c|}{|C_c|} \tag{2}$$

Micro-averaged precision is defined as the percentage of correctly classified candidates within the set of all classified candidates:

$$\mu Pr = \frac{|C'_{phrasal} \cap C_{phrasal}| + |C'_{not} \cap C_{not}|}{|C|} \tag{3}$$

Macro-averaged precision is defined as the average precision for all classes:

$$Pr = \frac{Pr_{phrasal} + Pr_{not}}{2} \tag{4}$$

Macro-averaged recall is defined as the average recall for all classes:

$$Rc = \frac{Rc_{phrasal} + Rc_{not}}{2} \tag{5}$$

Macro-averaged $F1$ is the harmonic mean between macro-averaged recall and macro-averaged precision:

$$F1 = \frac{2PrRc}{Pr + Rc} \tag{6}$$

5.1.2 *Document Searching Setup.* Search results for both the original datasets and after the inclusion of the detected phrasal terms were evaluated in terms of Mean Average Precision (MAP) and precision at the top 10 results provided for each query (Pg@10) [Baeza-Yates and Ribeiro-Neto 1999]. Topics queries used in our experiments are: from 1 to 106 of OHSUMED, from 351 to 400 of FBIS, from 401 to 450 of L.A. Times, from 451 to 500 of Wt10G, where only the titles portion of the queries were considered.

5.2 Phrasal term detection

Table II shows the phrasal term detection results achieved by our method in the four databases. In general, results show that our method is a robust solution to the phrasal term detection problem. In terms of precision, in all collections, satisfactory results were obtained.

	μPr	$Pr_{phrasal}$	Pr_{Not}	Rc_{Comp}	Rc_{Not}	Mac_{Prec}	Mac_{Rec}	Mac_{F1}
OHSUMED	70,80%	70,72%	70,88%	71,00%	70,60%	70,80%	70,80%	70,80%
FBIS	76,10%	77,94%	74,48%	72,80%	79,40%	76,21%	76,10%	76,16%
L.A.Times	81,90%	85,05%	79,27%	77,40%	86,40%	82,16%	81,90%	82,03%
Wt10G	94,40%	95,87%	93,02%	92,80%	96,00%	94,45%	94,40%	94,42%

Table II. Results of phrasal term detection in the Wt10G, FBIS, L.A.Times, and OHSUMED datasets.

The lowest values obtained were for the OHSUMED collection. Results, in all the metrics, were around 71%. By analyzing the data, we can conclude that the size of the documents had an important role in this behavior. Documents in the OHSUMED database are fairly small, containing only titles and abstracts, which reduces co-occurrence information available for the bigrams. However, as can be seen on section 5.3, even with relatively small precision values for OHSUMED, gains were still achieved when using the detected phrasal terms to search documents.

In the remaining databases, we observed that precision for class “phrasal terms” is higher than for class “non-phrasal sequences”. The SVM classifier appears, therefore, to be conservative when choosing phrasal terms. This could, of course, be changed by further tuning the classifier. However, this is not the focus of our work. Precision values were around 71%, 78%, 85%, and 95% for the OHSUMED, FBIS, L.A. Times, and Wt10G collections, respectively.

These results, even with a difference of about 24% when comparing different databases, show that the method is a simple and good alternative to phrasal term detection. However, it is important to verify why the results in the different databases had that much difference. The F1 values obtained in WT10G are on par with the results shown in [Zhang et al. 2007], that were around 90% of F1. However, a direct comparison is not possible since they used a different dataset.

These results may indicate that the size of the database might have a strong influence in the overall quality of the classifier, since bigger databases had better results. In order to investigate whether the assumption that the size of the database affects the phrasal term detection holds true, we repeated the experiments using the Wt10G collection and splitting it into portions of different sizes: 1/2 of the Wt10G database, 1/5 of the Wt10G and 1/10 of the Wt10G database. The examples used in these experiments were mostly the same as in the previous experiment, with the exception of the examples that had a frequency smaller than 10 in that slice of the database, which were replaced by new randomly sampled examples. The results are presented in Figure 2.

These results confirm the conclusion that the size of the database indeed influences the phrasal term detection. It is also important to notice that as the database grows, the impacted caused by the addition of new pages diminishes, which is an indication that after a certain number of pages, using more pages for the phrasal term detection will have almost no effect in the precision of the detection procedure.

At first glance this does not hold true when considering the results for the FBIS collection and the L.A. Times collection, since they have about the same number of documents and similar sizes, yet different detection performance. However, the FBIS database is comprised of articles with a much broader spectrum of topics in comparison to L.A. Times, since FBIS is a translation of foreign texts regarding many different countries while L.A. Times is composed of news articles that are pertinent to the Los Angeles area citizens, having, thus, a much more tight domain. This broader spectrum of FBIS leads to a larger vocabulary size, as shown in table I.

Table III shows the number of bigrams found and considered in each dataset (i.e. bigrams with frequency higher than 10 and without stop words), and the number of bigrams classified as phrasal terms by our method. However it is important to notice that these are the bigrams that our method decided are phrasal terms, and not necessarily actual phrasal terms.

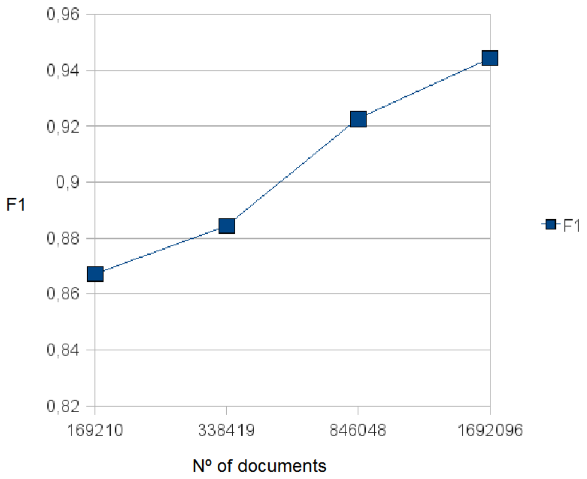

Fig. 2. F1 scores obtained when applying the phrasal term detection method to portions of different sizes of the Wt10G Collection.

	# of bigrams	# of Phrasal Terms
OHSUMED	141724	38687
FBIS	163143	62451
L.A.Times	160370	34013
Wt10G	3038832	69180

Table III. Total of bigrams considered (Frequency larger than 10 and no stop words) and total detected as phrasal terms by our method.

By examining these results, It is interesting to notice that, as expected for being the largest, Wt10G yielded the largest number of bigrams. However, this large number of bigrams did not translate into a much larger number of phrasal terms in comparison with the other databases, specially FBIS. Notice also that there is no obvious correlation between the number of bigrams and the number of phrasal terms detected.

5.2.1 *Feature Impact study.* In this section, we study the impact that each feature has on the phrasal term detection task. To do so, we first calculate the information gain of each feature, in order to verify which features are the best class discriminators in each database. Afterwards, we studied the impact of each feature in the final classification result, and analyzed the impact of removing several combinations of the less discriminative features from the classification in each database. For a better visualization, the names of the features in this section will be replaced by numbers, given by table IV.

Id	Feature	Id	Feature
1	$P(t_1, t_2)$	5	$P_D((t_1))$
2	$P((t_1, t_2)\|t_1)$	6	$P_D((t_2))$
3	$P((t_1, t_2)\|t_2)$	7	$P_R(t_1)$
4	$P_D((t_1, t_2))$	8	$P_R(t_2)$

Table IV. Identifiers for each feature in the following tables.

Table V shows the information gain of each feature for each database. A first observation is that feature 1(pair frequency) is among the best features in most cases, except for OHSUMED collection.

Wt10G		L.A.Times		FBIS		Ohsumed	
Id	InfGain	Id	InfGain	Id	InfGain	Id	InfGain
1	0.6945	1	0.4066	2	0.2138	2	0.0824
4	0.6372	4	0.3665	1	0.2012	5	0.0707
2	0.4864	2	0.2509	4	0.1636	7	0.0564
3	0.4609	3	0.2281	3	0.112	3	0.0461
6	0.0178	8	0.0691	7	0.0481	6	0.0439
8	0	6	0.041	8	0.0443	8	0.0393
5	0	7	0.0227	5	0.0439	4	0.0243
7	0	5	0	6	0	1	0.0147

Table V. Information gain of each feature in all databases. The features are identified according to the Ids in table IV

As it can be seen, in OHSUMED all features had small information gain values, which explains the worse classification results in this collection.

Table V also shows that, in Wt10G, L.A. Times and FBIS, features 5, 6, 7 and 8 (Document probability of t_1 and t_2 and Raw probability of t_1 and t_2, respectively) yielded the smaller information gain values. This is an interesting and somewhat expected result, since these features regard information from single word occurrences (document probability of the words and number of bigrams the word is on) and not from the bigrams. Nonetheless, we believe that, even though these features are not good individual discriminators, their use in addition to other features may indeed lead to a positive impact in the results, especially in smaller databases where the bigram informations have a smaller information gain value.

To assert whether this belief is accurate, we performed the same experiments of section 5.2, but removing all the combinations of the features 5,6,7 and 8. Table VI shows the results for some of these combinations in terms of F1 when removing the feature sets 5,6,7,8, 5,6 and 7,8. Other combinations were omitted due to space restrictions and because they did not change the conclusions. As it can be seen, while in L.A. Times and FBIS databases the use of these features indeed translated into a higher F1 value, this does not hold true for Wt10G, where removing these features had little impact on the final F1 result. This may be an indication that, since Wt10G is much larger than the other databases, the bigram statistics are enough to allow for a precise classification in Wt10G.

	- 5,6,7,8	- 5,6	- 7,8
OHSUMED	0.6947	0.7096	0.6953
FBIS	0.7536	0.7506	0.7576
L.A.Times	0.7982	0.8231	0.8202
Wt10G	0.9452	0.9462	0.9451

Table VI. F1 values for the classification task without feature sets {5,6,7,8}; {5,6}; {7,8}.

5.3 Application to Search

In the above experiments, we have presented results that show our approach can achieve high precision levels when detecting phrasal terms, specially in larger collections. However, it is also important to show that these phrasal terms can indeed have a positive influence in the quality of information retrieval systems. In this section we present results obtained by the application of phrasal terms in document searching.

We start by applying our method to each complete document collection. 1000 example bigrams were used to train the SVM classifier. Table III shows the number of bigrams in each dataset (bigrams with frequency higher than 10 and without stop words), and the number of bigrams classified as phrasal terms by our method.

Following, we look for the detected phrasal terms in the queries used for testing, and submit these queries with the added phrasal terms to a Vector Space Model Search Engine. We also evaluate the results obtained when manually picking all phrasal terms from the queries. This last experiment is useful to verify the impact the method would have if all phrasal terms present in the queries were detected, and to give a better insight about how the proposed method impacts search in comparison with the best-case scenario, that is human classification.

Table VII presents the total number of queries and the number of queries that have phrasal terms, either detected by the method or manually detected, for each database. It can be seen that the OHSUMED collection has the largest number of queries with phrasal terms in the queries (75). This is due to the fact that the OHSUMED queries are substantially longer and have many medical terms. It is also interesting to notice that when manually picking phrasal terms, less queries were found (73). This is due to the fact that some of the phrasal terms found by our method were not actual phrasal terms . Conversely, only 11 of the Wt10G queries contained phrasal terms. Even though this small number may have a small impact in the overall results, the gain obtained when considering only the modified queries is expressive, as reported in the following.

	number of queries		
	total	automatic	manual
Ohsumed	106	75	73
FBIS	50	14	24
L.A.Times	50	9	21
Wt10G	50	11	12

Table VII. Number of queries modified for each database

	MAP					
	Baseline	Manual	Expansion	Gain	Phrase	Gain
Ohsumed	0,1853	0.1864	0,1934	4%	0,1872	1%
FBIS	0,1019	0.1082	0,1084	**6%**	0,1017	0%
L.A.Times	0,1424	0.1676	0,1555	**9%**	0,1429	0%
Wt10G	0,1043	0,1144	0,1116	**11%**	0,1025	-0,02%

Table VIII. MAP Results obtained when considering all queries.

	P10					
	Baseline	Manual	Expansion	Gain	Phrase	Gain
Ohsumed	0,2367	0.2333	0,2393	1%	0,2401	1%
FBIS	0,1364	0.1434	0,1491	**9%**	0,1365	0%
L.A.Times	0,1473	0.1727	0,1618	**10%**	0,1455	-1%
Wt10G	0,1327	1364	0,1418	7%	0,1291	0,03%

Table IX. P10 Results obtained when considering all queries.

Tables VIII and IX shows the results for all test queries. Column “Baseline” refers to the results obtained using VSM. Columns “Phrase” and “Expansion” refer to the use of the detected phrasal terms as phrases in the queries and to the use of phrasal terms to expand documents, respectively. Column “Manual” refer to the results obtained when manually picking bigrams.

It is noticeable that using the phrasal terms as phrases in the queries yielded little to no gain (even small losses in some cases). This is due to the small number of relevant documents in each collection, many of which did not contain the phrase being queried. Conversely, using the detected phrasal terms to expand the documents led to gains in all the databases, both in terms of MAP and in P@10. An important remind here is that the expansion methods results in less costly query results computation when compared to the option of phrases.

It is interesting to notice that, as the quality of the phrasal term detection increases, the gain obtained by the use of phrasal terms also increases. While in OHSUMED the gain was only of 4% in terms of MAP and 1% in terms of P@10, in Wt10G the gains were more expressive (around 10%).

It is also interesting to notice that, in comparison to manually detecting phrasal terms, the results obtained by the automatic classification were similar, and even greater as in FBIS and OHSUMED. This is a good indication that the impact of the proposed method was near the optimal possible impact yielded by finding phrasal terms in search queries.

	MAP					P10				
	Baseline	Expan.	Gain	Phrase	Gain	Baseline	Expan.	Gain	Phrase	Gain
Ohsumed	0,1916	0,2052	**7%**	0,1942	1%	0,2558	0,2642	3%	0,2606	2%
FBIS	0,1457	0,1636	**12%**	0,1450	-1%	0,2013	0,2273	**13%**	0,2013	0%
L.A.Times	0,1560	0,2119	**36%**	0,1585	2%	0,1616	0,2424	**50%**	0,1515	-6%
Wt10G	0,0383	0,0616	**23%**	0,0364	0%	0,0606	0,1212	**17%**	0,0404	0%

Table X. Results obtained when considering only queries with phrasal terms.

When considering only queries that contain phrasal terms, the results are quite impressive. We present in Table X values obtained when considering only queries that contain phrasal terms. As can be seen, the gains obtained were significant, ranging from 7% to 36% in terms of MAP. In terms of P@10, although there was only a small gain in OHSUMED, in the remaining databases the gain ranged from 13% to 50%. Once again, the use of phrasing yielded poor results when regarding only queries with phrasal terms. On the other hand, it is clear that the use of phrasal terms to expand documents in search tasks can lead to expressive gains. The results obtained in the Wt10G database may seen oddly small (ranging from 0.03 to 0.06), but this is due to the fact that a few of the queries that had detected phrasal terms did not yield any relevant result in the 100 first results, having thus MAP and P@10 values of 0.

It is also important to notice that, when expanding a document with the addition of phrasal terms, there is a small change, in VSM, in the document's norm, since the phrasal terms are added as extra terms in the representation of documents. The documents norm is increased, reflecting the addition of the phrasal term information to the documents. Thus, it is important to measure how this change affected the results. To do so, in table XI we show the results obtained by queries that do not contain phrasal terms. We do so because the results in these queries are only affected by the change in the norm of the documents. The results of using phrasal terms for phrasing are not shown because phrasing does not change the quantity of information present in the documents.

	MAP			P@10		
	Orig.	Exp.	Gain	Orig.	Exp.	Gain
Ohsumed	0,1703	0,1649	**-3%**	0,1906	0,1789	-6%
L.A.Times	0,1394	0,1431	**3%**	0,1441	0,1441	**0%**
FBIS	0,0898	0,0921	**3%**	0,1176	0,1257	**7%**
Wt10G	0,1502	0,1504	**0%**	0,1921	0,1921	**0%**

Table XI. Results obtained when considering only the queries that do not contain phrasal terms.

The results in table XI show that the addition of phrasal terms had a small impact in all databases in terms of MAP, with the gain (or loss) being at most 3%, while in terms of P@10 the impact was slightly higher in some databases. It is interesting to notice that while most databases had a slight increase or no change in terms of MAP and P@10, in OHSUMED, the addition of phrasal terms to the documents yielded losses in both metrics. This is probably due to the nature of OHSUMED: it is composed of only the abstract and title of medical papers, which we verified that resulted in small, phrasal terms heavy documents, in which every phrasal term added to a document have a large influence in its norm. Further, the individual terms present on the phrasal terms from this collection are already quite specific and present low ambiguity when compared to the other two collections. This could cause the loss in precision presented in table XI. Nevertheless, these results indicate that, while the additional information had some impact on the results, the gain obtained by the methods was mostly due to the phrasal term existing in some of the queries, and not due to the information added to the documents.

These results are a strong indication that the method can lead to gains in search tasks. Further, the comparison with manually identifying phrasal terms show that these gains are near the best-case scenario for classifying phrasal terms.

6. CONCLUSIONS AND FUTURE WORK

In this work, we presented a method to detect phrasal terms in document collections. Experiments show that the proposed method yields good results, without a large computational overhead. The classification of phrasal terms achieved values of F1 ranging from 70,80% up to 94%.

We also presented the impact that the use of the phrasal term information has on document searching. When compared to the traditional Vector Space Model, the gain obtained by using phrasal terms was up to 11% in MAP, when considering all the queries, and up to 36% when considering only queries that do have phrasal terms. Thus, the use of phrasal terms is a strong solution to enhance the quality of search tasks.

In future work, we intend to further study the problem, including the possibility to use the proposed method to detect larger phrasal terms, verify how the proposed method behaves in datasets in different languages study the use of the phrasal terms as an independent source of information, test different features, such as HTML markup, capitalization, and placement on the page. Finally, we intend to measure the impact of the detection of phrasal terms in other text information retrieval tasks, such as classification and clustering.

REFERENCES

Baeza-Yates, R., Calderon-Benavides, L., and Gonzalez-Caro, C. The intention behind web queries. In *Proceedings of the International Conference on String Processing and Information Retrieval.* Glasgow, UK, pp. 98–109, 2006.

Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval.* Addison Wesley, 1999.

Chang, C.-C. and Lin, C.-J. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Ekkerman, R. B. and Allan, J. Using bigrams in text categorization. In *Tech Report.* pp. 1–10, 2003.

Hersh, W., Buckley, C., Leone, T. J., and Hickam, D. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the International ACM SIGIR Conference on Research & Development of Information Retrieval.* Dublin, Ireland, pp. 192–201, 1994.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. Wordnet: An on-line lexical database. *International Journal of Lexicography* vol. 3, pp. 235–244, 1990.

Mladenic, D. and Grobelnik, M. Word sequences as features in text-learning. In *Proceedings of the Electrotechnical and Computer Science Conference.* Portoroz, Slovenia, pp. 145–148, 1998.

Soboroff, I. Do trec web collections look like the web? *SIGIR Forum* 36 (2): 23–31, 2002.

Tan, C.-M., Wang, Y.-F., and Lee, C.-D. The use of bigrams to enhance text categorization. *Information Processing and Management: an International Journal* 38 (4): 529–546, 2002.

tau Yih, W., Goodman, J., and Carvalho, V. R. Finding advertising keywords on web pages. In *Proceedings of the International World Wide Web Conferences.* Edinburgh, Scotland, pp. 213–222, 2006.

Tesar, R., Strnad, V., Jezek, K., and Poesio, M. Extending the single words-based document model: a comparison of bigrams and 2-itemsets. In *Proceedings of the ACM Symposium on Document Engineering*. Amsterdam, The Netherlands, pp. 138–146, 2006.

Voorhees, E. M. Overview of the eighth text retrieval conference. In *Proceedings of the Text REtrieval Conference*. Gaithersburg, Maryland, pp. 1–24, 1999.

Witten, I. H. and Frank, E. *Data mining: Practical machine learning tools and techniques with Java implementations.* Morgan Kaufman, 2000.

Zhang, W., Liu, S., Yu, C., Sun, C., Liu, F., and Meng, W. Recognition and classification of noun phrases in queries for effective retrieval. In *Proceedings of the International Conference on Information and Knowledge Engineering*. Lisbon, Portugal, pp. 711–720, 2007.