

Phrase-based Document Categorization revisited

Cornelis H.A. Koster
Computing Science Institute
University of Nijmegen
The Netherlands
kees@cs.kun.nl

Jean G. Beney
Dept. Informatique
INSA de Lyon
France
jean.beney@insa-lyon.fr

ABSTRACT

This paper takes a fresh look at an old idea in Information Retrieval: the use of linguistically extracted phrases as terms in the automatic categorization (aka classification) of documents. Until now, there was found little or no evidence that document categorization benefits from the application of linguistics techniques. Classification algorithms using the most cleverly designed linguistic representations typically do no better than those using simply the bag-of-words representation. Shallow linguistic techniques are used routinely, but their positive effect on the accuracy is small at best.

We have investigated the use of dependency triples as terms in document categorization, which are derived according to a dependency model based on the notion of aboutness. The documents are syntactically analyzed by a parser and transduced to dependency graphs, which in turn are unnested into dependency triples following the aboutness-based model. In the process, various normalizing transformations are applied to enhance recall.

We describe a sequence of large-scale experiments with different document representations, test collections and even languages, presenting evidence that adding such triples to the words in a bag-of-terms document representation may lead to a significant increase in the accuracy of document categorization.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Linguistic Processing—*linguistic terms*; H.3.3 [Information Search and Retrieval]: Selection process—*document classification*

General Terms

Experimentation

Keywords

text categorization, linguistic terms, dependency triples, aboutness.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PaIR'09, November 6, 2009, Hong Kong, China.

Copyright 2009 ACM 978-1-60558-809-4/09/11 ...\$10.00.

1. INTRODUCTION

The goal of our work is to improve the accuracy of classification algorithms for patent documents by using, besides words, also dependency triples as part of the document representation.

The document representation most widely used today in Information Retrieval is still the bag-of-words model, both in traditional query-based retrieval and in document categorization. In this representation, the word order of the document plays no role at all, whereas our intuition tells us that important information about the category of a document must be available in certain groups of more than one word found in the document. It is as if we are trying to determine the function of a building (church? castle? monastery? dwelling? shop? railway station?) by dismantling it into bricks and then looking exclusively at the frequency distribution of the individual types of bricks, disregarding any coherent substructures but simply taking them apart into individual bricks (the bag-of-bricks model). Intuitively it seems obvious that larger units (combinations of words into meaningful phrases) could serve as better terms for the categorization, yet there is surprisingly little evidence for it.

1.1 Previous work

The use of linguistically derived phrases as indexing terms has always fascinated researchers in IR (for an overview see [13]). Many different kinds of phrases have been tried:

- statistical phrases, chunks or collocations: sequences of k non-stop words occurring consecutively [7], stemmed and ordered bigrams of words [6], even collocations of varying length taken from a specially prepared domain terminology [3];
- syntactical phrases identified by shallow parsing [1], template matching or finite state techniques [10];
- syntactical phrases (Head/Modifier pairs) obtained by shallow parsing [20] or deep parsing [9, 25].

It is a disappointing fact that over the years no classification experiment using any sort of linguistic phrases has shown a marked improvement over the use of single keywords, at least for English. In the case of statistical phrases, the experience is only a little bit brighter [6, 20]. In the last decade, there was a consensus that the value of Natural Language Processing to IR is dubious, even among people who tried hard to make linguistically-based IR work [20, 24]. The predominant feeling, as voiced in [13], is that only ‘shallow’

linguistic techniques like the use of stop lists and lemmatization are of any use to IR, the rest is a question of using the right statistical techniques.

For highly inflected languages this may be different (but we are not aware of any pertinent classification experiments described in literature). For Chinese, positive effects are reported using n-grams of characters [23, 21], but these are largely due to the problems of word segmentation particular to written Chinese.

Our previous experiments in Text Categorization with linguistically motivated phrases [3] and Head/Modifier pairs [15] have also failed to show a worthwhile improvement over keywords, in spite of a heavy investment in linguistical resources.

In this paper, we are going to explore the use of dependency triples as terms in Information Retrieval and Document Classification. We use a “deep” parsing and transduction process to bring documents into this document representation, to which we are going to apply methods from IR which are statistical rather than logical in nature. This has two consequences:

- we are not striving for a representation of the *meaning* of the text, suitable for logical inference and reasoning, but rather for a representation of what is called in IR the aboutness of the text – see section 1.2;
- we expect the precision of triples as search terms to be higher than that of words, but it is easy to predict that their recall will be lower, due to the fact that the probability of the literal re-occurrence of a word in another text is higher than that of a triple (a pair of words).

These thoughts are elaborated further in the following sections, introducing the notion of aboutness, describing an indirect way to measure it and formulating the Aboutness Hypothesis.

In section two of this paper we will describe the use of dependency triples as terms for IR, introducing and motivating a dependency model based on the notion of aboutness and discussing the syntactic normalization of phrases, which is crucial for achieving recall.

In section three we describe the software and data resources used in our experiments: the test corpora, parsers and classification engines used. In section four we describe our experiments to test the classificatory value of triples and the plausibility of the aboutness hypothesis and present the experimental results.

1.2 The notion of Aboutness

The notion of *aboutness* is highly central to Information Retrieval: the user of a retrieval system expects the system, in response to a query, to supply a list of documents which are *about* that query. Practical retrieval systems using words as terms are based on a surprisingly simpleminded notion of aboutness[5]:

If the word x occurs in the document then the document is *about* x .

This notion is then made less naive by introducing a measure for the *similarity* between the query and the document, based on aboutness and other notions such as term frequency and document frequency.

A model-theoretic basis for the notion of aboutness was described in [4]:

An information carrier i will be said to be *about* information carrier j if the information borne by j holds in i

The rather abstract sounding notion of “information carrier” can denote a single term, but also a composition of terms into a structured query or a document.

For phrase-based retrieval, we need a notion of aboutness which is appropriate for phrases, which in its simplest form can be defined in the same way:

If the phrase x occurs in the document then the document is *about* x .

Although intuitively it seems obvious that phrases provide a more informative document representation than keywords, the above formulation is not helpful in deciding what phrases from a document to choose, which parts to eliminate and how to represent them.

1.3 Indirect measurement of aboutness

In fact we cannot measure aboutness directly for lack of a well-defined measure, but we can compare the accuracy (precision and recall) achieved in the categorization of a given set of documents using different document representations. The representation that gives the highest accuracy must have best captured the aboutness of the documents.

Text Categorization provides a good vehicle for the study of document representations, because many suitable test sets are available and accuracy can be accurately measured. Categorization does not suffer from the problem of measuring recall, unlike query-based search. Solid classification algorithms with a firm statistical basis allow objective and repeatable experiments, that can easily be replicated by others.

1.4 The Aboutness Hypothesis

According to Information Retrieval folklore, the best classification terms are the words from the open categories, in particular nouns and adjectives. The words from the closed categories are just stop words. In a classification experiment reported in [2], it was indeed found that nouns, verbs and to a lesser degree adjectives and adverbs are the only words that contribute measurably to the aboutness of a text.

These observations lead us to the following **Aboutness Hypothesis**:

- of all the words in a text, the words from the open categories (nouns, verbs, adjectives, modifiers) carry most of the aboutness;
- of all possible dependency triples (words from the text joined by a syntactic relator), the triples containing only words from the open categories will carry most of the aboutness;
- we expect a better classification result when using triples as terms in addition to words than when using words alone.

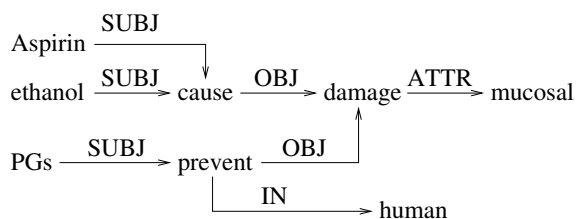
We are going to investigate the elements of this hypothesis in one the hardest document classification tasks: the classification of patent documents.

In order to capture also the aboutness of the words not included in a dependency triple and the non-words in the documents (non-standard names, formulae, various markers) we shall also study the effect of taking as terms all words in the document together with the triples extracted from it.

2. DEPENDENCY GRAPHS AND DEPENDENCY TRIPLES

By a *dependency graph* for a phrase we mean an acyclic graph (a tree with possibly additional confluent arcs) whose nodes are marked with words from that phrase and whose arcs are marked with directed relations. We are particularly interested in *syntactic relations*, those that can reliably be found by parsing.

As an example, the sentence *In humans, PGs prevent the mucosal damage caused by aspirin and ethanol* has the following dependency graph:



A dependency graph represents the main structure of a sentence in an abstract way, much more compactly than a constituent tree (parse tree), in terms of certain syntactic relations between words of the sentence, from which semantic relations can be derived relatively easily. In that sense it is close to a semantical representation.

By a *dependency triple* we mean a triple [word,relation,word], which forms part of a dependency graph, from which it can be obtained by *unnesting* the graph to the triples contained in it. Triples are a syntactic way of representing phrases, closely related to the Head/Modifier pairs used before in IR and to the Index Expressions of [4, 11].

2.1 A taxonomy of triples

The following table shows an example triple for each of the most important dependency relations:

subject relation	[device,SUBJ,comprise]
object relation	[cause,OBJ,lesion]
predicate relation	[aspirin,PRED,nsaid]
attribute relation	[breaker,ATTR,earth_crust]
prepos relation (noun)	[coating,PREPof,albumin]
prepos relation (verb)	[charge,PREPinto,container]
prepos relation (adj.)	[free,PREPto,operate]
modifier relation (verb)	[extending,MOD angularly]
modifier relation (adj)	[efficient,MOD,highly]

2.2 Syntactic normalization

In writing a text, people prefer to avoid literal repetition; they will go out of their way to choose another word, another turn of phrase, using anaphora, synonymy, hypernymy, periphrasis, metaphor and hyperbole in order to avoid the literal repetition of something they have said or written before. From a literary standpoint this is wonderful, but it gives complications in IR: we have to compensate for this human penchant for variety.

Rather than using literal phrases taken from the text as terms, we shall therefore in the transduction process reduce each phrase to a normal form which expresses only the bare bones of its aboutness. We will eliminate all dispensable ornaments and undo such morphological, syntactic and semantic variation as we can. In this way, we strive to regain recall while surrendering little precision.

The syntactical normalization is in our case expressed in the grammar itself, using a mechanism for describing compositional transduction: every construct in the grammar is described along with its transduction to an output format (dependency graphs).

- all elements which are deemed not to contribute to the aboutness of the text will be elided during transduction: articles, quantifiers, connectors - which is much like applying a stop list;
- many sentences contain embedded constructions like relative clauses and participial constructions, from which additional basic sentences are derived;
- one of the most effective normalizing transformations is *de-passivation*: transforming a passive sentence into an active sentence with the same aboutness. For this transformation, the sentence *this picture was painted by Ruysdael* is considered equivalent to *Ruysdael painted this picture*.

Morphological normalization by lemmatization is applied to the nouns and verb forms occurring in the resulting triples. As an example, the triples

[model,SUBJ,stand] [stand,PREP at,window]

can be obtained from the model stood at the window, a model standing at the window and from two models were standing at the window.

2.3 Dependency triples as terms

In principle it is not difficult to use triples as terms in classification (by considering the triple as a string), but there are severe practical problems to solve: There are many more two-word phrases than words in a language. The classification algorithm must be able to cope with very high numbers of features, demanding sophisticated Term Selection. But for the same reason, the feature space is very sparse. Triples may be high-precision terms, but they threaten to have a very low recall: a triple is never more frequent than the lowest frequency of the words from which it is composed. Low-frequency triples abound.

In classifying documents on e.g. gastric ulcers, it is therefore not immediately clear whether a triple like

[ulcer,ATTR,gastric]

is a more accurate search term than the cooccurrence of the words 'ulcer' and 'gastric'. It will probably have a higher precision but also certainly a lower recall. Compared to the string 'gastric ulcer' the triple will have the same precision but a slightly higher recall, since it catches also phrases like 'most of the ulcers were gastric'. Does the gain in precision compensate for the loss in recall?

We need experimentation in order to find out whether triples work.

3. RESOURCES USED

We are interested in comparing the accuracy achieved when using dependency triples as terms with that of words in the automatic classification of patent texts. In this section we describe the resources used in the experiments: the corpora used, the classification system and the parser. The next section describes the experiments made and their results.

3.1 The EPO corpus

The European Patent Office (EPO) in Rijswijk, the Netherlands, has been investigating for some time the use of Natural Language Processing (NLP) and automatic classification techniques for the pre-classification of its incoming patent applications. To that end, it has organized during the year 2000 an open competition in order to assess the state-of-the-art in supervised learning for document classification (see [19]). We have reused two corpora which were provided for the second test in that competition.

The EPO2 corpus consists of 68418 pre-classified documents, each available either as the full text (EPO2F) or abstract only (EPO2A). Some statistics:

	documents	lines	words	characters
full-text	68418	44.7M	401.3M	2607.3M
abstracts	68418	1.0M	9.0M	59.5M

All documents have been carefully labeled by EPO with one or more *directorate* numbers, where a directorate is an organizational unit within EPO, responsible for all patent applications in a certain domain. This manual labeling is of very high quality, representing the state-of-the-art in patent classification.

There are 44 different directorates (numbered from 1 to 44). For each of the 68418 documents an average of 1.43 classes is given. The maximum number of classes per document is 7. The total number of <document, class> pairs is 97996. EPO2A contains only the abstracts of the documents, whereas EPO2F comprises the full text of the Patent Applications.

The full-text documents (EPO2F) contain about 6000 words each and the number of different words (types) in the corpus is 557 790.

For extracting the dependency triples from the documents, we have used the EP4IR (English Phrases for IR) parser [17] which is available in the public domain [18]. It is a rule-based best-only dependency parser for English, extended with a large lexicon of technical terms.

The analysis of the whole EPO2F corpus resulted in a total of about 650 million triples (9500 per document), 49.1 million different triples. Obviously, there are many more triples than words.

3.2 The WIPO corpus

We also make use of the corpus which was made available to us by WIPO for an experiment in 2003. It is a collection of 555396 abstracts in French, labeled according to the International Patent Classification (IPC) at class level, subclass level and main group level. We used the categorization into 119 classes, with 1 to 10 (avg 1.29) classes per document and 46 to 48026 (avg 4200) documents per class, 60 to 5046 (avg 600) words per document.

The texts were parsed with the (experimental) French parser FR4IR developed at the INSA de Lyon. The ab-

stracts contained a total of 435 986 different word forms. In some experiments we added to the abstracts the names and addresses of the applicants, giving 932 739 different terms. The number of different triples derived from the abstracts using EP4IR was 13 057 658.

3.3 The LCS classification engine

The Linguistic Classification System (LCS) used in these experiments [16] was developed at the University of Nijmegen for the experimentation with different document representations. The Balanced Winnow algorithm that it uses [22, 8] is highly robust against large but sparse sparse feature sets.

In the experiments we have compared the Balanced Winnow classification algorithm of LCS with the Support Vector Machine algorithm (SVM light [12]) in a large multi-classification task, using words and triples as terms.

Based on previous experiments with patent documents [14], the following Winnow parameters were chosen (which differ quite widely from the values given in [8]):

- for words: promotion factor 1.01, demotion factor 0.99
for triples: promotion factor 1.02, demotion factor 0.98;
- thresholds $\theta^+ = 2.0$, $\theta^- = 0.5$;
- number of iterations = 10.

4. EXPERIMENTAL RESULTS

In the following experiments we are going to investigate the Aboutness Hypothesis, using document classification:

- what is the relative contribution of the open classes?
- which relations contribute most to the aboutness?
- is classification on triples significantly more accurate than classification on words?

We start with the last (and most important) question. We shall also compare the performance of Winnow with that of SVM, using words and triples, because we suspect that the use of triples will have a different effect for different classification algorithms.

As the measure of accuracy, we used the Micro-averaged F1, assuming a loss function attaching equal weight to precision and recall, which appears to be representative for most applications in patent classification.

All experiments were performed on the Large Data Collider (LDC) of the Information Retrieval Facility (www.ir-facility.org).

4.1 EPO2F words and triples

Our main experiment uses the EPO2F corpus and the Balanced Winnow algorithm from LCS. Each experiment on this corpus was performed 10 times (10-fold cross-validation). The documents were split at random into a train set (80 %) and a test set (20 %). The average value of the accuracy obtained and its standard deviation were measured.

We report the accuracy achieved with three different document representations: words only, triples only and words plus triples. The numbers following \pm represent the standard deviation in the preceding numbers.

EPO2F	seen	unseen	gain
words	81.58±0.20	66.86±0.23	baseline
triples alone	96.59±0.03	66.15±0.28	-0.71%
triples+words	96.09±0.04	70.03±0.21	+3.17%

The terms 'seen' and 'unseen' refer to testing the accuracy of the classifier on the train and the held-out test sets, respectively.

The results on unseen corpus are representative for the intended application (pre-classification of patent applications). The results on seen corpus are not very important for the application, but they show that as the quality of terms increases it is easier to achieve high accuracy on seen corpus; but this may also be interpreted as a sign of overtraining (the accuracy on the train set approaches 100%, but not on the test set).

The baseline (words) may look disappointing, but it is quite hard to improve upon it. Attempts using lemmatization, stop lists etc (not reported here) gave no significant improvement.

The difference of 3.17 percent points between the base line and the use of triples + words however is our main result: it is highly significant (six times the combined standard deviations).

Note that an accuracy close to the base line is achieved when using only triples as terms. In our previous experiments [15] with Head/Modifier pairs it was much lower; maybe the quality of the parser has in the mean time improved.

In this experiment each document was assigned to 1 up to 44 categories; the heuristic of demanding at least one category also leads to a slightly better result than this "pure" multi-classification in which the lower limit is 0 categories.

4.2 WIPO words and triples

Again we trained on 80% of the documents and tested on 20%, with ten-fold cross validation.

We also experimented with the use of the names and addresses of the applicants as additional terms.

The results on unseen documents were as follows.

WIPO	seen	unseen	gain
words only	65.72±0.04	54.51±0.09	base line
words+addresses	76.72±0.03	63.34±0.07	+8.8%
words+triples	86.12±0.02	68.26±0.07	+13.8%
words+triples+addr	88.49±0.02	72.94±0.10	+18.4%

In spite of the rather unfinished state of the French parser, the improvement when adding triples (13.8 %) is striking. But simply adding the names and addresses of the applicants already gains 8.8 %, a reminder that classification accuracy is not the only issue. The combination of these shows a staggering improvement in accuracy of 18.4 % over the base line.

4.3 The aboutness of word categories

In [2], the contribution to the accuracy by one category of words was measured by classifying a given corpus using only words from that category as terms. The category of words was determined using the Brill tagger. Of course this tagging was not very precise, because in English many words belong to more than one category, and it is well known that all nouns can be verbed.

We have repeated this experiment, using the parser as tagger: during the parsing, the parser has to disambiguate the Part-Of-Speech of every ambiguous word in the sentence. Both the Brill tagger and the parser can (and will) make errors, but in our approach we can be sure that each occurrence of a word obtains a part-of-speech which is appropriate in the context. The corpus used was EPO2F, without cross-evaluation.

The Parts-of-Speech A, N, V and X stand for the open word categories: adjective, noun, verb and adverb.

EPO2F	unseen	nmb of terms
words	66.72 %	485703
words + PoS	67.21%	492572
A	52.78 %	76870
N	65.77 %	370408
V	49.72 %	36194
X	19.49 %	8859
A+N+V+X	67.20 %	492331

The number of terms indicated is the number of different terms in the train set before term selection. Note that just marking each word with its Part-of-Speech has a small positive effect on the accuracy, probably because it resolves some of the word ambiguities.

As was to be expected, nouns appear to have the highest aboutness, but they are closely followed by adjectives and verbs. Of course as a measure for comparing the aboutness the classification accuracy provides merely a ranking. It does not take into account the relative frequencies of the different types of terms, and does certainly not yield a probability. The notion of aboutness and the measures for its quantification merit further research.

We also performed a leave-one-out experiment (again without cross-evaluation) to determine the effect of leaving out all words of a certain category.

EPO2F	unseen
words	66.72%
ANVX	67.20%
-A	68.33%
-N	68.11%
-V	68.00%
-X	67.92%

Curiously, there is in all cases a small improvement of the unseen accuracy – a phenomenon for which we have as yet no explanation.

4.4 The aboutness of relations

Here we mean by a relation the set of all triples having the associated relator. We want to measure for each relation what is its contribution to the accuracy, and hence to aboutness.

We experimented on the EPO2A and EPO2F corpora, representing each document by a bag of dependency triples, but leaving out all triples not belonging to a certain category. The results for the EPO2A dataset (short abstracts) are as follows:

EPO2A	accuracy unseen	percentage of triples
all triples	60.98	
only ATTR	45.38	33.61
only OBJ	23.09	15.57
only SUBJ	25.09	15.34
only Prep	19.84	26.15
only of	14.92	06.60
only MOD	8.46	5.58

For each relation, we show the accuracy that is achieved when omitting all others, and the percentage of that relation among all triples. We lacked the time for a leave-one-out experiment with cross-validation.

Looking only at the percentages, the ATTR relation, which is used in compound noun phrases, is the most important. Then follow the OBJ and SUBJ relations, relating a noun to a verb. After that the PREP relation (a catch all for all preposition relations). The PREP relation can relate a noun to either another noun, adjective or verb standing as the head of the triple. The of-relation concerns the preposition most frequent between two nouns; the MOD relation gives the smallest contribution to accuracy, but it is also the least frequent.

Obviously, the contribution of a relation to the classification depends also on its frequency in the corpus; that's why, in the last column we show the percentage of this relation among all triples. The MOD-relation has the smallest percentage.

We now turn to the results for EPO2F (large full-text documents).

EPO2F	accuracy unseen	percentage of triples
all triples	66.15	100%
only ATTR	67.26	30.62%
only OBJ	58.52	14.23%
only SUBJ	56.65	15.43%
only Prep	58.98	24.40%
only of	56.22	07.39%
only MOD	41.56	07.66%

This presents a rather different picture, and the accuracy achieved with each relation is higher than for abstracts. In particular the PREP relation turns out to be surprisingly important.

Further analysis is needed to get a better understanding of these issues, but the experiments make clear that, besides the traditional ATTR relation building compound terms, the relations involving verbs (SUBJ and OBJ) also contribute strongly to the aboutness.

4.5 Comparing Winnow and SVM

We have also compared the accuracy achieved by Winnow with that of SVM, for different corpora and representations, in one experiment without crossvalidation.

	seen corpus	unseen corpus
EPO2A Winnow words	70.48	56.89
words + triples	87.09	59.27
EPO2A SVM words	82.20	56.15
words + triples	98.67	55.92
EPO2F Winnow words	78.99	65.38
words + triples	95.56	70.49
EPO2F SVM words	93.61	65.68
words + triples	99.81	66.20

The results on words of SVM and Winnow are similar, but SVM does not improve much when using triples whereas Winnow shows a marked improvement. SVM does not appear to benefit from using triples. Our tentative explanation is as follows: SVM classification is based only on the support vectors (near the margin) while Winnow during training takes into account all documents that are temporarily misclassified (not so near to the final margin). For document classification, with so many terms, it is important to consider more documents.

4.6 Discussion

In classifying two very different corpora of patent material, the increase in accuracy obtained by adding triples as terms was statistically highly significant. There is no question that dependency triples are valuable terms for the classification of patent texts, in spite of the earlier failures reported. It is interesting to note that both parsers used in the experiment are still under development. It appears likely that when the parsers reach more accuracy on the very complicated documents from the patent world, the effect of using triples may increase further.

The results on the aboutness of word categories and relations require further analysis, but it is definitely plausible that the aboutness hypothesis is well founded. Certainly not busted, we would claim.

5. CONCLUSIONS

We have proposed the use of dependency triples as terms in document classification, using an aboutness-based dependency model rather than the more detailed dependency models preferred in linguistics. We made an experimental investigation of the classification accuracy reached when using these dependency triples as terms in the (pre)classification of patent applications. Our main result is that the addition of dependency triples as terms can, with a suitable classification algorithm, lead to a highly significant increase in precision and recall, well exceeding the state-of-the-art (as exemplified by the SVM classification algorithm).

We also formulated the Aboutness Hypothesis, stating essentially that the open word categories are the carriers of aboutness, thus motivating our choice of dependency model. Experiments corroborated the nearly exclusive value of these categories for classification.

Using a classification setting, we attempted to quantify the aboutness of different categories of words and dependency triples. Although this yielded rich evidence for the Aboutness Hypothesis, it still has to be analyzed and explained more deeply. It shows that the triples involving only the open categories, including the verbs, are indeed very good classification terms. For the same reasons we conjecture that they will be very good search terms in query-based search. The use of dependency triples as terms may lead to an important improvement in the effectiveness of search and classification in the patent world.

6. REFERENCES

- [1] M. Alonso, J. Vilares, and V. Darriba. On the usefulness of extracting syntactic dependencies for text indexing. *Lecture notes in computer science*, pages 3–11, 2002.
- [2] A. Arampatzis, T. Van der Weide, C. Koster, and P. Van Bommel. An Evaluation of

- Linguistically-motivated Indexing Schemes. In *Proceedings of BCS-IRSG 2000 Colloquium on IR Research, 5th-7th April 2000*, Sidney Sussex College, Cambridge, England, 2000.
- [3] N. Bel, C. H. Koster, and M. Villegas. Cross-Lingual Text Categorization. In *Proceedings ECDL 2003, Springer LNCS 2769*, pages 126–139, 2003.
- [4] P. Bruza and T. Huibers. Investigating Aboutness Axioms using Information Fields. In *Proceedings SIGIR 94*, pages 112–121, 1994.
- [5] P. Bruza and T. Huibers. A study of aboutness in information retrieval. *Artificial Intelligence Review*, 10(5):381–407, 1996.
- [6] M. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text Databases and Document Management: Theory and Practice*, pages 78–102, 2000.
- [7] W. Cohen and Y. Singer. Contextsensitive learning methods for text categorization. In *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval*, pages 307–315, 1996.
- [8] I. Dagan, Y. Karov, and D. Roth. Mistake-Driven Learning in Text Categorization. In *Proceedings of the Second Conference on Empirical Methods in NLP*, pages 55–63, 1997.
- [9] J. Fagan. *Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods*. PhD thesis, Cornell University, 1988.
- [10] G. Grefenstette. Light parsing as finite state filtering. In *Workshop on Extended finite state models of language, Budapest, Hungary, August 1996. ECAI'96.*, 1996.
- [11] F. Grootjen and T. Van der Weide. Effectiveness of index expressions. *Lecture notes in computer science*, pages 171–181, 2004.
- [12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML-98)*, pages 137–142. Springer Verlag, 1998.
- [13] K. S. Jones. The role of NLP in Text Retrieval. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 1–24. 1999.
- [14] C. Koster and J. Beney. On the Importance of Parameter Tuning in Text Categorization. In *Proceedings PSI 2006, SLNC 4378*, pages 269–280, 2006.
- [15] C. Koster and M. Seutter. Taming Wild Phrases. In *Proceedings 25th European Conference on IR Research (ECIR 2003)*, pages 161–176. Springer LNCS 2633, 2003.
- [16] C. Koster, M. Seutter, and J. Beney. Multi-classification of Patent Applications with Winnow. In *Proceedings PSI 2003, Springer LNCS 2890*, pages 545–554, 2003.
- [17] C. Koster, M. Seutter, and O. Seibert. Parsing the Medline Corpus. In *Proceedings RANLP 2007*, pages 325–329, 2007.
- [18] C. Koster and E. Verbruggen. The AGFL Grammar Work Lab. In *Proceedings of the FREENIX/Usenix conference 2002, Monterey, Ca.*, pages 13–18, 2002.
- [19] M. Krier and F. Zaccà. Automatic Categorization Applications at the European Patent Office. *World Patent Information*, (24):187–196, 2002.
- [20] D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings ACM SIGIR'92*, 1992.
- [21] Z. Li, P. Li, W. Wei, H. Liu, J. He, T. Liu, and X. Du. AutoPCS: A Phrase-Based Text Categorization System for Similar Texts. *Proceedings of the Joint International Conferences on Advances in Data and Web Management*, pages 369–380, 2009.
- [22] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, (2):285–318, 1988.
- [23] C.-L. Liu and C.-D. Hsieh. Exploring Phrase-Based Classification of Judicial Documents for Criminal Charges in Chinese. *Lecture Notes in Computer Science*, pages 681–690, 2006.
- [24] A. Smeaton. Using NLP and NLP resources for Information Retrieval Tasks. In T. Strzalkowski, editor, *Natural Language Information Retrieval*. 1997.
- [25] T. Strzalkowski. Natural Language Information Retrieval. *Information Processing and Management*, 31(3):397–417, 1995.