Typed Skipgrams Investigated

Bachelor's Thesis
Computer Science Program
Radboud University Nijmegen

Niklas Weber
Student Number 0841420

August 2012

Supervisors: Prof. Cornelis H.A. Koster
Eva D'hondt MSc MA
Second Examiner: Dr. Suzan Verberne

Abstract

Every year, the patent filing rates at the different patent offices around the world increase, and the patent examiners are struggling to catch up. As such, reliably categorizing patents is to aid the examination process of rising economic importance.

This paper investigates whether capturing multiword expressions (specifically: institutionalized phrases) is an important factor for improving automated patent pre-classification. To do so we describe a novel text representation based on filtering by combinations of Part-of-Speech tags: typed skipgrams. We then compare the performance of different text representations (unigrams, bigrams, skipgrams, typed skipgrams and unigrams in combination with any of the other) when classifying a subset of the CLEF-IP 2010 corpus. We examine if there is a link between classification accuracy and ability to capture multiword expressions. We furthermore carry out additional experiments and analyses to investigate the influence of specific combinations of Parts-of-Speech on the overall result.

We find that typed skipgrams in combination with unigrams perform significantly better (difference in F1 value: 0.7%) than the unigrams+bigrams baseline. We also find that typed skipgrams succeed in capturing multiword expressions and that typed skipgrams consisting of noun-noun and noun-adjective combinations are the most important factor for the overall success. Finally, we conclude that capturing multiword expressions is the crucial mechanism behind the improvement in classification scores.

This research provides a potential means to further the state-of-the-art in patent classification when combined with additional optimizations. They also give directions for future research, highlighting typed skipgrams and filtering for multiword expressions as viable paths. Finally, we expect our results to generalize to every kind of text in which multiword expressions play an important role. Examples are scientific abstracts and, more generally, technical texts.

Chapter 1

Introduction

The aim of the presented research is to gain insight into what kind of (textual) information is needed to improve the quality of automated text classification, especially for technical texts such as patent abstracts.

Automated text classification is a supervised learning task in which a set of documents has to be classified in predefined categories. One of the most distinct advantages of a classified corpus is that documents are easier to find. This is of paramount importance for everyone handling large amounts of texts, even more so if they have to do deliver to according to some content-dependent criteria. Examples are libraries, web search engines such as Google, encyclopedias such as Wikipedia or patent offices.

Focusing on that last example, it can be observed that patent filing rates increase on a yearly basis, bringing patent offices to the edge of their capacities (Benzineb and Guyot, 2010). The relevance of good automated text classification becomes apparent when considering the typical workflow of a patent office: On arrival, a patent is automatically pre-classified into two or three of the higher, more general categories of the International Patent Classification Hierarchy[1].

Let us take an invention of new type of birdcage as an example; The corresponding patent application might be pre-classified as belonging to "Section A — Human Necessities, Subsection Agriculture".[2] The patent is then forwarded to the departments specialized in those domains. After inspection by a member of this department, the patent might be sent back because the pre-classification has been erroneous. If the pre-classification has been fitting, the patent will be manually classified into one or more of the more specialized sub-classes of the hierarchy, say "A01K Animal Housing [..]; Care Of Birds [..]" and forwarded to experts on this narrower field. Depending on the size of the patent office, this procedure might be repeated here, until the patent arrives at an expert for the lowest level in the hierarchy: "A01K 31/00 Housing Birds". At this level, the specialist in charge of this document might still decide that another person is better suited to handle this application or that other specialists might also be interested and redirect or forward the application accordingly.

From this example it should become clear that this process might be improved significantly if the pre-classification can be more precise and if it can be carried out on a lower level of the hierarchy, leaving out intermediate manual classification steps. Improvements to this process can save time and money in all patent offices using such a classification hierarchy. Seeing that

[1] The International Patent Classification (IPC) hierarchy is a taxonomy structured into (from more general to more specific) sections, classes, subclasses groups and subgroups. In its latest edition, the IPC contains eight sections, ca. 120 classes, ca. 630 subclasses and approximately 69,000 groups.

[2] Pre-classification is mostly carried out on the "class" level. "Section" has been here for the sake of demonstration.

the IPC is part of the Strasbourg Agreement - which has 61 contracting parties including most European states, China, the United States of America and the Russian Federation - it is safe to assume that this holds for most of the world's patent offices.

In a broader context, improved classification will help people to find relevant texts, be they books in a large library or scientific texts in a search engine, more easily.

1.1 Document Representations

The standard text representation used in text classification is the bag-of-words representation, which, in its most inclusive form, consists of all of the words contained in the text. Prior work has focused on improving text classification by expanding this representation with additional information, often in the form of statistical or linguistical phrases (Koster and Beney, 2009).

On the standard test set used to evaluate new ideas for text classification, the Reuters-21578 set of newswire texts, phrases were found to be more representative but suffer from sparseness, which leads to little overall impact (Caropreso et al., 2001). The best performance for such texts is still achieved by using the bag-of-words representation (Bekkerman and Allan, 2003).

This changes when inspecting different kinds of texts: Ozgür and Güngör (2009) show significant improvements when using phrases as index terms for the classification of scientific abstracts. D'hondt et al. (2012) investigated different text representations for patent classification and compared the performance of a text classification system when using words and words combined with either bigrams (being an example of statistical phrases) or linguistic phrases as index terms for patent abstracts. Their unexpected result was that words and bigrams taken together give the best performance. A possible explanation is that bigrams best capture multiword expressions[3] such as *machine learning* or *liquid nitrogen*.

We expect these expressions to be very representative for the different classes due to the specific language use in patents, which is a very technical domain, using many domain-specific terms. Furthermore, patents are written to be as generic as possible to extent the scope of any claims made. To this end, an applicant will obfuscate the description of the invention and its parts. For example, a *hose* might instead be referred to as a *watering device*.

1.2 Hypothesis, Predictions and Method

The starting point of this paper is: Verifying that better capturing of multiword expressions actually is the reason behind the good performance when representing text as words with bigrams. Generalizing from this, we formulate our hypothesis as follows:

"Multiword expressions are the most informative phrasal features for text classification in the patent domain".

If we assume this hypothesis to be correct then a text representation also capturing multiword expressions standing one or more words apart should do even better. This is due to the fact that such expressions might be split by function words; consider for example *divide and conquer*.

To test this theory, we shall examine the performance of skipgrams (with a maximum of two skips and always consisting of two words) as terms for patent abstract classification. Thinking further, we expect the performance to increase even more due to heightened relevance of terms if we filter these skipgrams lexically so that only those most likely to actually capture a multiword

[3]While "multiword expression" may also refer to complete phrases, such as "biting the bullet" we use this term for a compound of two words, normally both found within one noun phrase, denoting one combined concept. Two examples are given above.

expression remain. Because the filtering is based on the linguistic types of the constituent words (as determined by a Part-of-Speech tagger) we call this variant typed skipgrams.

Should the above predictions prove correct, we will analyse the document profiles as well as the penetration levels and conduct a series of leave-one-out classification experiments[4], which will give us insight into which information captured by the different text representations is most valuable and how they contribute to the classification results.

[4]Leave-one-out in this case means: for every subtype of typed skipgrams (e. g. noun-noun or noun-verb) conduct a classification experiment in which this class is not present.

Chapter 2

Background

In this section we will summarize the work done on the central concepts of the work presented in this paper, namely: the following topics will be dealt with below: text representations in classification (Section 2.1), feature selection (Section 2.2), patent classification (Section 2.3) and, finally, multiword expressions (Section 2.4).

Due to the similarity in topic, we shall generally follow the outline of the background section as found in D'hondt et al. (2012).

2.1 Text Representations in Classification

2.1.1 Text Classification

Before inspecting text representations, an introduction to text classification in general is advisable. Text classification is a supervised learning problem. First, a classification algorithm (in this paper: Winnow) is given labelled examples. For patent application abstracts, this would mean: the abstract and the document's class labels. The new type of birdcage from the introduction would thus be labelled with "A01" when operating on the "class" level of the IPC hierarchy.

From this training data, the classifier builds a model. Forms such a model might take are decision trees or decision boundaries of some sorts. Winnow constructs a hyperplane to linearly discriminate the data in an n-dimensional feature space.

After training, the classifier then is given previously unseen and unlabelled examples, for which a label has to be predicted. When evaluating the classifier accuracy, these unseen examples typically come from the so-called test set. The test set is a part of the pre-labelled data, but not used for training. As the labels are known, predicted and actual labelling can be compared to measure the accuracy.

There are 121 classes in the CLEF-IP 2010 corpus. As such, we are dealing with *multiclass* classification. As an abstract typically has two to three labels, it is also an example of *multi-label* classification.

2.1.2 Text Representations

In designing a classification experiment, an important decision concerns the form in which the data is presented to the classification algorithm. When using text classification, the input has to be transformed into vectors in feature space: feature vectors. This is done by choosing a

text representation, implicitly defining the feature space (though it may be modified by feature selection, see Section 2.2 below).

We discriminate between two groups of features: unigrams and phrases[1] which can be linguistical (i.e. constructed using linguistically motivated choices) or statistical (i.e. constructed without any linguistic knowledge).

Phrases in general were found to suffer too much from data sparseness. Classification performance might actually deteriorate when using phrases instead of words (Lewis, 1992), confirmed by Apte et al. (1994). Again, we witness the trade-off between informativeness and sparseness.

Bekkerman and Allan (2003) on the other hand state that there is new research on this topic due to the increase in computational power and size of data sets. The former makes improved linguistic analysis possible, contribution to informativeness. The latter helps to alleviate issues arising from sparseness, as smaller data sets are more prone to suffer from such problems. Still, Bekkerman and Allan (2003) also report that positive results seldom are significant improvements.

Let us briefly inspect the different representations we are going to use one by one:

Unigrams have been established as the standard text representation, helped by the fact that they work very well for the standard text corpora used in testing, such as Reuters-21578. Using unigrams as the text representation means that every one word of the text becomes one feature for the classifier. One might note that any order among words and thus information conveyed through this order is lost. Variation is possible by choosing to include/exclude function words in a stop word list, or by filtering based on other criteria.

Bigrams are phrases constructed from two words from the text. The important characteristic is that those words have to appear sequentially. As such, bigrams are statistical phrases. Bigrams and unigrams generalize to n-grams: sequences of n words from the text.

Skipgrams are more verbosely specified as n-k-skipgrams, meaning a sequence of n words from the text (n-gram), but allowing up to a maximum of k skipped words in-between the chosen ones. skipgrams, by virtue of their definition, are statistical phrases. Higher values for n or k lead to greater problems resulting from sparseness - a great number of options for the construction of phrases leads to many of them being present only once or maybe a few times across the whole corpus (Guthrie et al., 2006). When later in the text we use "skipgrams" we typically refer to 2-2-skipgrams, deviations from this will be made explicit.

Typed skipgrams form a subset of skipgrams, filtered according to Part-of-Speech tags. As such, they are a form of feature selection and could have also been listed in Section 2.2 below. Information about the tags is not given as input to the classifier. As far as we know, this is a novel representation. Employing Part-of-Speech information for text classification however has been done before, see for example Feldman et al. (2009). Other usages involve the study of phraseology, which also considers multiword expressions, see Pinna and Brett (2009).

For our main experiment we have allowed phrases involving combinations of nouns, adjectives and verbs (the precise filtering rules and examples can be seen in Section 3.2.3). This leads to all phrases containing function words to be filtered out. As we are using linguistic knowledge (in the form of PoS tags), typed skipgrams can be considered linguistic phrases, although of a rather weak kind.

Typed skipgrams are related to dependency triples (Koster and Beney, 2009), which consist of two words, chosen intra-sententially, and their syntactical connection with each other, e.g. the one being the object to the other. They thus capture and encode head-modifier pairs - many of

[1] "Phrase" here means "an indexing term that corresponds to the presence of two or more single word indexing terms" (Lewis, 1992). As such, it does not necessarily correspond to the definition of phrase as used in the study of syntax. In the remainder of the text, "phrase" will refer normally refer the indexing-term usage. Exceptions, such as in "X is a noun phrase" should be clear by context.

which are also present as (typed) skipgrams. Additionally, many institutionalized phrases are noun-phrases and, thus, are captured by head-modifier pairs. But in contrast to dependency triples, the relation between the constituent words of typed skipgrams is not computed, and thus no explicit factor in constructing them. The linguistical information that actually is used is not explicitly integrated in into the (typed) skipgrams.

Other variants of filtered skipgrams have been tried. Consider for example Orthogonal Sparse bigrams, described in Siefkes et al. (2004). They have been designed to capture the same information as provided by sparse binary polynomial hashing (loc. cit. page five) but to be more space-efficient. By remembering how many words have been skipped and selecting OSBs cleverly, all of the original features can be reconstructed as linear combinations of the new features.

OSBs are, in contrast to typed skipgrams, purely statistical phrases. Siefkes et al. (2004) have applied them to spam filtering, also using the Winnow algorithm. While at first the application of a variant of filtered skipgrams to a text classification problem makes that project appear similar to the work described here, there are important differences:

1. OSBs are quite different in aim and characteristics to typed skipgrams.
2. The textual domains differ in nature: as Goldstein and Sabin (2006) have shown verbs are important when classifying email, whereas we are primarily interested in combinations of nouns with either nouns or adjectives.
3. Patent classification at the IPC "class" level is, at this moment, considerably harder than spam classification. A good indicator for this is the number of classes: two for spam filtering (spam versus no spam), 121 for patent classification. As such, simply by guessing one achieves, on average, 50% for the former and about 6% for the latter. Furthermore, while there are differences in data sets, goals etc. it may be noted that the baseline for patent classification is considerably lower than for spam filtering (74.79 versus 98.88), again hinting at the greater difficulty of our classification task.

2.2 Feature Selection

Feature selection is a kind of dimensionality reduction and thus an alternative to, e.g., feature extraction. It is used for purposes as removing irrelevant features, removing noisy features, dispelling the curse of dimensionality or speeding up tasks such as classification or regression analysis.

In addition to the feature selection present in typed skipgrams other ways of feature selection regarding phrases have been researched.

In what follows, results refer to the Reuters-21578 data set, unless otherwise mentioned. As such, there is an inherent difference between the use of language used in our work (patents, long and complicated sentences, obfuscation, domain-specific vocabulary) and most of the papers cited below (newspaper articles, aiming for clarity and understandability).

Caropreso et al. (2001) show that many bigram features are more important than unigrams features. Nevertheless, when the unigram/bigram ratio for a fixed number of features is changed in favor of bigrams, classification accuracy decreases.

Braga et al. (2009) use the Multinomial Naive Bayes classification algorithms in two different setups to combine unigrams and bigrams: a) two classifiers, one using unigrams and the other using bigrams. Their label rankings are then combined in a later step; b) one classifier using both unigrams and bigrams as a feature for the same classifier. They find that there is nearly no difference. The bigram-only classifier generally assigns the same labels with a lower confidence - when combined with the unigrams-only classifier they simply affirm each other. The authors

suggest to combine unigrams only with those bigrams that, in that domain, are more meaningful together than apart.

It is especially that last part that makes this research important to our work. It can be seen as a description of institutionalized phrases (and thus: multiword expressions, for specifics see Section 2.4 below). It must be noted, that the authors did not have institutionalized phrases explicitly in mind. Nevertheless, their usage of a different name for the same phenomenon does not diminish the relevance of their work.

Our research might be seen as implementing Braga et al. (2009)'s advice: combining unigrams with institutionalized phrases. The important differences are, that we are explicitly looking for such phrases and thus have adopted our terms to capture them. Such adaptions include, for example, the abandonment of bigrams in favour of skipgrams and typed skipgrams.

Tan et al. (2002) describe a procedure to select those highly representative and meaningful bigrams, based on Mutual Information scores. Scores of the words in a bigram compared to the unigrams class models were compared. The top two percent of those bigrams were selected. This resulted in an significant increase over the unigrams baseline. Bekkerman and Allan (2003) note, that this baseline was not state-of-the-art.

The filtering employed by Tan et al. (2002) can be said to again implicitly aim for institutionalized phrases. Again it has not been explicitly set into this context by the authors. Their results are promising, but seeing that they used the Reuters-21578 and the Yahoo Science corpora (the latter consisting of science-related web sites) generalization to the patent domain must be tested. Furthermore, they have not inspected the interaction between such special bigrams and unigrams in combined representations. As we shall see later, it is this interaction that is important to the performance of typed skipgrams. Finally, their terms, while sharing some intention, significantly differ from the ones employed here.

More research has been conducted using similar selection criteria: Crawford et al. (2004) classified emails using the filtering devised by Tan et al. (2002). The unigram baseline could not be improved.

As stated above, the reason might be sought in the different use of language in emails - capturing multiword expressions in general and using bigrams in particular is not as important as it is for patent pre-classification.

In contrast to the research described above, we shall thus employ a term selection technique specifically designed to choose multiword expressions, on a corpus for which they are important.

2.3 Patent Classification

Patent classification has to deal with some special complexities and requirements with respect to text classification in general. Great interest in improving patent classification for practical applications has spawned research specializing in this domain.

2.3.1 Complexities

Some of those complexities are:

First, multiple temporal variants of the same document may be present. This leads to a further imbalance of classes, possibly having a small negative effect on classification accuracy (Kolcz et al., 2003).

Secondly, patents feature a difficult use of language: non-standard terms abound, which sometimes are invented by the author Atkinson (2008) to make the invention seem more innovative. Acronyms, terminology and general terms are employed frequently. "General terms" here refers to expressions as *watering device*, used instead of *hose*, to increase coverage of the patent. As

there is no convention regarding these things acronyms etc. are chosen differently by different people, which leads to sparseness issues being more severe.

Thirdly, patents usually have more than one IPC code. As it is unspecified which part of a document leads to it being labelled with which code training gets more difficult. Furthermore, multi-class classification in itself often is problematic due to classification algorithms being only applicable to two-class problems. Earlier approaches to handle this include selecting only one class per document (Fall et al., 2003) or adjusting the algorithm to multi-class training (Koster et al., 2003). The work presented in this paper uses a third approach: the classification algorithm (here: Winnow) still is a mono-classifier. For every of the 121 classes one such mono-classifier is trained (see Section 3.5 for more detail).

Finally, patent classification has to deal with a very diverse domain, covering all possible technical areas.

2.3.2 Earlier Work in Patent Classification

Work on patent classification is plentiful. In what follows, we try to give a concise overview, focusing on research close to this project. A complete overview of the work done in patent classification up to 2002 can be found in Fall and Benzineb (2002), while Benzineb and Guyot (2010) offer a general introduction to this topic. The historical beginning is normally given with Larkey (1999) who was the first to develop a fully automated patent classification system but did not report results on overall accuracy.

Koster et al. (2003) used the EPO-alpha corpus, classifying with a combined representation of unigrams and head-modifier pairs. These pairs were derived from the EP4IR parser. No improvement on the unigrams baseline could be found, the reason behind this being phrase sparseness.

Returning to our overview it is noteworthy that since 2009 the CLEF-IP track is organized by the Information Retrieval Facility (IRF), providing very large real-life patent data sets. The three best results from CLEF-IP 2010 will be discussed below.

Guyot et al. (2010) were most successful using Balanced Winnow as the classifier and a combination of unigrams and collocations of variable length as the text representation.

Collocations are groups of words that appear more often together than would be expected by chance. This definition does resemble the one given above for institutionalized phrases. Given the accordance regarding data set and classification algorithm, these results seem to be promising for our own work.

The main difference lies in the focus of the work: we shall try to trace the influence of multiword expressions (being a superset of the institutionalized phrases), not aim for state-of-the-art classification scores. As such, we do not include documents other than the English ones (Guyot et al. (2010) also use German and French ones). Our representations, (typed) skipgrams, also differ, being constructed explicitly to capture MWEs and typed skipgrams also including linguistical information in contrast to the purely statistical collocations.

Verberne et al. (2010b) and Beney (2010) used a combined representation of unigrams and dependency triples, derived from an English and a French parser. They reported a slight increase in classification accuracy by adding those triples to the unigrams. As with the other research using dependency triples one may note that these triples are “more syntactical” than typed skipgrams and may thus lie at another point of the informativeness / sparseness trade-off.

To conclude this overview let us once again return to the research regarding dependency triples. As a follow-up to the above-mentioned study Koster et al. (2010) investigated the influence of different syntactic phrases. They report that attributive phrases, i.e. combinations of nouns with either adjectives or nouns, to be the most important for the patent domain. Small, but significant

improvements could be achieved by adding triples to unigrams.

The special relevance of this study lies in its insight into the nature of the most valuable phrases. When comparing with the description of institutionalized phrases given above, we see that they, too, are expected to be composed mainly of noun-noun or adjective-noun combinations. Typed skipgrams consisting of such words are thus expected to be the most influential. This expectation shall be tried in sections 5.1 and 5.2.

These similarities strengthen our claim that typed skipgrams and dependency triples capture the same meaningful word combinations. We expect typed skipgrams to outperform triples due to less issues related to sparseness and filtering that better captures multiword expressions, which we hold to be a decisive factor for classification performance. Performance of triples might also be impaired by consistent parser errors: terms such as "software engineering conference" and "software engineering tutorial" might both be analyzed as right-headed, ignoring the compound of "software engineering". This leads to inconsistent triples. Consistent errors for Part-of-Speech tagging, on the other hand, are less likely to be problematic. If certain words are consistently miss-classified among all classes, this is unlikely to severely impact the classifier.[2]

2.4 Multiword Expressions

What multiword expressions exactly are and what specific subgroups fall under them still is subject to ongoing debate. For our work, we will define them to be, very generally, groups of words which, when together, denote a special meaning. Following the taxonomy laid out by Sag et al. (2002)[3], "special meaning" can refer to two concepts, forming the two great branches of multiword expressions: lexicalized and institutionalized expressions.

Lexicalized expressions are characterised by having at least partially idiosyncratic meaning or syntax. The former refers to phrases in which words combine to a new meaning which normally does not follow from the composition of meanings of the individual words. Examples include "to kick the bucket" or "to spill the beans". On the other hand, expressions can be said to have idiosyncratic syntax when words are combined in otherwise unusual ways to form them. "Long time no see" or "every which way" constitute examples of this category.

The other great branch is formed by institutionalized phrases. In contrast to the lexicalized phrases, they are syntactically and semantically compositional but are statistically idiosyncratic. This means: they are groups of words which could be combined differently in principle but happen to appear in a specific combination in an unusually high frequency. Examples are "traffic light" or "turning signal" or "middle management" While "traffic director" or "corner light" might appear in principle, they are nowhere near as frequent as "traffic light".

Inside these two main branches the taxonomy continues to refine, reflecting things as variation in syntactic flexibility. For the work presented here, these distinctions mostly are too fine-grained, the reader is referred to the paper mentioned above for the details. The complete taxonomy of multiword expressions can also be found in the appendix, see Figure B.1.

For our research, the focus lies on the second of the main branches: the institutionalized phrases. The reason for this lies in the language predominant in patent applications, which is very technical and contains a large amount of domain-specific vocabulary. Specific fields (more or less sharply represented by the different IPC classes) have their own special terms, examples

[2] One might discriminate between two cases: a) the PoS tag is wrong, but the same features are generated because a word that should (not) be filtered still will (not) be filtered. b) the PoS tag is wrong and additional features are generated or features are lacking because the filtering now also goes wrong. Case b) will be more problematic but still be consistent among all classes, thus diminishing this error's impact on classification results.

[3] A diagram of the complete taxonomy can be found in Figure B.1, on Page 34.

including *machine learning*, *liquid nitrogen* or *divide and conquer*. All of these can be seen as institutionalized phrases.

Chapter 3

Experimental Setup

In this section we describe the data selection (Section 3.1), the preprocessing applied to this data (Section 1), including the generation of the various representations. A short description of the Part-of-Speech tagger is given in Section 3.3. Following this we give short summaries of the corpus statistics after preprocessing (Section 3.4) and, finally, of the classification framework and settings (Section 3.5).

In general, we follow the approach described by D'hondt et al. (2012), so as to maintain comparability and prevent potential new sources of error, such as faulty preprocessing or a wrong configuration of the classification system, from arising.

3.1 Data Selection

Our experiments were conducted on a subset of the CLEF-IP 2010[1] corpus, which is a subset of the MAREC patent collection. It contains 2.6 million patent documents, which pertain to a total of about 1.3 million individual patents (each patent possibly having multiple patent documents). The patents included in the corpus have been published between 1985 and 2001.

The documents are encoded in a customized XML format and contain text in English, French and German and consist of the following patent sections: title, abstract, claims and description. Furthermore, they also include meta-information, such as inventor, date of application, assignee, among others. Of all this data, we only use the abstracts for our experiments. The rationale behind this is that we are trying to investigate into the effect certain text representations (viz. unigrams, bigrams, skipgrams and their typed variant). Choosing only one (highly informative) part of the document will not lead to the best classification accuracy compared to other research[2], but, as we are more interested in comparing the different text representations, we focus only on the relative gains between the representations. As such, this selection will not change our findings but reduce the amount of data to a more manageable level.

Classification is carried out on the class level in the IPC8 hierarchy. Consequently, only documents having at least one IPC class in the <classification-ipcr> field have been used. The selection is further narrowed down by only choosing documents containing an English abstract. The IPC class has been extracted on the document level. This results in documents being left out which do not have both an English abstract and an IPC class, although the patent as a whole (to which this document belongs) may have both.

[1] Available through the IRF at `http://www.ir-facility.org/collection`

[2] Verberne and D'hondt (2011) show that classification accuracy is higher when using description and abstract instead of only the latter

Filtering based on these criteria leaves us with 532,264 abstracts, divided into 121 classes. The majority of these documents have one to three category labels, with an average of 2.12 labels per document. For classification, these documents have been split in a train set of 425,811 (80% of the corpus) and a test set of 106,453 (20%) documents, respectively.[3]

3.2 Data Preprocessing

General preprocessing consisted of the following steps: cleaning up character conversion errors, removing image references, removing claim references and, finally, splitting up the text into sentences, employing a list of abbreviations and acronyms that occur frequently in technical texts. Furthermore, the creation of each of the different text representations also includes decapitalization, lemmatisation and the removal of all punctuation except for "-". The latter steps are not part of the general preprocessing because the Part-Of-Speech Tagger uses this information as context features for tagging.

Lemmatisation for phrases is carried out by splitting them into their two constituent parts, lemmatising those and then recombining them. The impact of lemmatisation is outlined in Section 3.4 and considered in more detail in D'hondt et al. (2012), Section 3.2.4.

The special punctuation rule for "-" is present because the hyphen frequently connects two words which, together, form one unit of sense (e.g. *data-driven*, see also the examples below). As such it is useful to treat the resulting complex as one word. To clarify consider Example 1 below. A sentence after general preprocessing might look like this:

(1) Performance of data-driven processing can be increased.

3.2.1 Unigrams and Bigrams

To construct unigrams, sentences were split on whitespaces. The resulting words were lemmatized using the AEGIR lexicon. Bigrams were fashioned likewise. One may note that only intra-sentential bigrams were created. For our sample sentence the output is given in Table 3.1 below:

Type of Terms	Resulting Terms after Preprocessing
Unigrams	performance; of; data-driven; processing; can; be; increase
Bigrams	performance_of; of_data-driven; data-driven_processing; processing_can; can_be; be_increase

Table 3.1: unigrams and bigrams, both lemmatized, constructed for the example sentence

60 million unigrams and 58 million bigram tokens have been constructed for the whole corpus. For a more detailed overview incorporating all of the representations please refer to Table 3.4 on Page 14 below.

[3]No cross-validation has been carried out, based on the results of Verberne et al. (2010b), who demonstrated that for this corpus there is little variance between different train/test splits (with a standard deviation of less than 0.3%).

3.2.2 Skipgrams

Skipgrams, too, were constructed only intra-sententially.[4] As for the unigrams, sentences were split on whitespaces. After removing any punctuation, 2-skip-2-grams are created. For every possible value (here: 2-1-0) for the number of skips, one pass over the data has been carried out, constructing skipgrams with exactly this number of skips. In other words, the constructed skipgrams consist of two words (2-grams), with a varying number of gaps between those words (2-1-0). One may see k-skip-2-grams thus as a form of generalized bigram.

For sentences shorter than four or three words it is impossible to generate skipgrams with two skips or one skip respectively. The remaining skipgrams with less gaps were still constructed. After construction, the skipgrams underwent the same lemmatisation as the bigrams. Furthermore, no information about what words have been skipped or how many of them have been skipped is encoded in the resulting skipgrams.

Let us inspect our example. It should also clarify why the hyphen has not been filtered out: "data-driven" would have been split despite it being one connected word. The generated skipgrams are[5]:

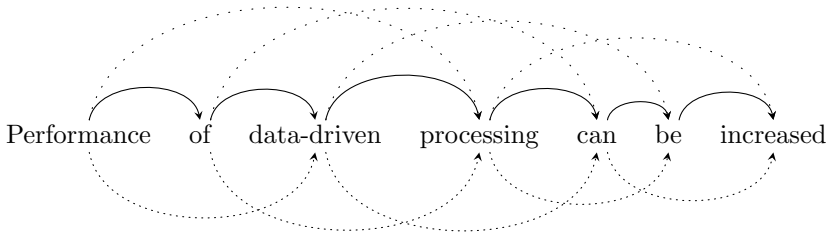

Figure 3.1: Creation of skipgrams.

The loosely dotted arrows above the text in Figure 3.1 depict the construction of the skipgrams containing two gaps. During the second iteration skipgrams with one skip are added, shown as dotted arrows below the text. Finally, bigrams, being skipgrams with zero skips, are constructed as indicated by the smaller solid arrows above the text. See Table 3.2 below for the resulting preprocessed data:

# Skips	Resulting Terms after Preprocessing
2 skips	performance_processing; of_can; data-driven_be; processing_increase
1 skip	performance_data-driven; of_processing; data-driven_can; processing_be; can_increase
0 skips	same as bigrams

Table 3.2: Lemmatized skipgrams created for the example sentence

About 168 million tokens have been created for the whole corpus. Again, more details can be

[4]The reasons for this are a) that we wish to maintain compatibility to bigrams and b) that our subject of interest, the multiword expression, are found inside the confinements of one sentence.

[5]Please note that the dot, ".", is not included for reasons of convenience and ease of display. In the actual algorithm it is considered as an individual element and filtered out.

found in Table 3.4 below.

3.2.3 Typed Skipgrams

The typed skipgrams were created as follows: We first ran an in-house Part-of- Speech tagger, developed at the Linguistics Department (see Halteren (2000); also see Section 3.3 immediately below), on the preprocessed sentences. The tagger was trained on the annotated subset of the British National Corpus and uses the CLAWS-6 tag set[6]. From these tagged sentences we then created typed skipgrams, using the algorithm described above.

To ease later filtering, the detailed CLAW6 tag set has been mapped to a more basic set of PoS tags (see Section B.1 on page 33). Most important for this work is the fact that all noun-related tags ($N*$) were mapped to N, all verb-related tags ($V*$) to V and adjectives (JJ) to A. For our main experiment, type information was purely used as a filter criterion and not given as input to anything else. Consider the example sentence, annotated with the mapped tagger output, found in Example 2 below. The example follows the syntax: *Tag1*:Word1 *Tag2*:Word2 etc.

(2) N:Performance *PREP*:of A:data-driven N:processing V:can V:be V:increased *UNK*:.

As can be seen in Example 2, every word (and token of punctuation except for "-") has been assigned a tag. After removing tagged punctuation the same (decapitalized) skipgrams as above are generated, but only for pairs of words that are assigned tags matching a pre-defined filter.

Assume that, for example, we only allowed the following type combination: type1=N, type2=N. Only one skipgram would be created, viz. *performance_processing*. Other combinations are textually too far apart or do not pass the type filter.

Note that the tag filter is directed: AV is not the same thing as VA. The actual filter we have used is partly derived from our hypothesis: as we are looking for multiword expressions and, more specifically, for institutionalized phrases, we allow type combinations (NN, NA, AN) which typically indicate such phrases.

For the sake of avoiding confirmation bias, we have also allowed combinations involving the V type. As a consequence, typed skipgrams now form the subset of skipgrams in which no phrases involving function words are contained. We shall show in Section 5.2 that this extension is not needed to significantly outperform the unigrams+bigrams baseline, although it does increase classification scores.

Please see Table 3.3 below for a list of allowed combinations, illustrated by example. For those combinations intended to capture multiword expressions, the respective example has been set in boldface. Please note that the examples are lemmatized, which explains the otherwise unexpected forms of the verbs. After filtering according to these criteria, 45 million tokens remained.

3.3 Part-of-Speech Tagger

For Part-of-Speech tagging we have used an in-house developed tagger, see Halteren (2000). To determine the PoS tag for a word it uses both knowledge about word frequencies and machine learning techniques. Furthermore, it is highly customizable, which has been the primary reason to select it for this task.

For this experiment, we have adapted it to use word frequency information about the patent domain, taken from the AEGIR lexicon. For our task this is rather valuable, as patents feature a very distinct use of language.

[6] `http://ucrel.lancs.ac.uk/claws6tags.html`

Tag 1	Tag 2	Example Expression
N	N	**machine learning**
N	V	catheter comprise
N	A	**mother superior**
V	N	use medicament
A	N	**transverse wave**
V	A	work fast
A	V	liquid spill

Table 3.3: Allowed combinations of tags and examples of corresponding expressions

There are, however, two caveats. First, we have not retrained it on a patent corpus. This means that it is still trained on the label distributions that are typical of the original training texts, the British National Corpus. Second, due to constraints of time, we have not tested its accuracy on this new corpus. As errors in tagging propagate to typed skipgram filtering, they might influence classification scores and, probably, change the semantics of captured phrases. If a noun is not classified as a noun, we might not be able to correctly find multiword expressions. We have tried to account for this in the discussion by manual inspection of the results, see Section 5.

3.4 After-Preprocessing Corpus Statistics

A summary of the statistics for the different representations after preprocessing is given in Table 3.4 below. The following values are given per representation:

1. The number of tokens which is the number of instances in the corpus. When considering an example corpus only consisting of the tokens *A, B* and, again, *B*, the token count would amount to three.

2. The number of types, which is: the number of unique terms in the corpus. For the *A, B, B* example this would amount to two.

3. The token/type ratio, which gives an indication of the spread of the data. If it is high it means that there are many duplicate tokens. In general, this is desirable because types being instantiated by a very small amount of tokens tend to be insignificant to the classifier. As such, a low token/type ratio indicates a potential data sparseness problem.

4. The number of hapaxes, i.e. unique tokens (or, put differently: types instantiated only by one token). A high number of hapaxes is generally undesirable for the same reasons a low token/type ratio is (see above). Such terms will also most likely be removed by the classifier, as something which occurs only once in the corpus is of no use when trying to identify classes of documents.

Unless mentioned otherwise, the numbers given include lemmatisation. To display its impact, we have included the data for non-lemmatised skipgrams. As one can see, lemmatisation decreases the number of types for skipgrams ca. 3,000,000, thereby increasing the token/type ratio. Furthermore, a reduction of the number of hapaxes by approximately 23% can be noted. A more detailed account for unigrams and bigrams can be found in D'hondt et al. (2012), Section 3.2.4.

The impact of filtering by types can be seen when comparing skipgrams and typed skipgrams (both lemmatised). The amount of data is reduced by nearly 75 %. This has an unexpected effect:

Representation	#Tokens	#Types	#Tokens/#Types	#Hapaxes
Uni	60,065,858	419,171	143.30	215,448
Bi	57,499,818	4,226,210	13.61	2,124,847
Skip (non-lem)	168,845,555	14,829,363	11.39	7,541,208
Skip	168,808,226	11,794,377	14.31	5,842,074
Tskip	45,913,342	6,959,461	6.60	3,467,690

Table 3.4: Corpus statistics for various representations.

the token/type ratio is reduced to less than half of its previous value. This is caused by filtering out phrases containing function words. As function words appear frequently, types containing them tend to be instantiated by many tokens.

This decrease in token/type ratio illustrates the fundamental informativeness/sparseness trade-off. As we select more informative phrases through linguistic filtering they are at risk of having less impact due to resulting sparseness. The impact on classification accuracy requires an analysis in greater detail, see Section 4.2 below.

3.5 Classification Experiments

Classification was done using the Linguistic Classification System (LCS, cf. Koster et al. (2003)). Within this framework one may select a classifier form the following set: Naive Bayes, Balanced Winnow. Earlier work (Verberne et al., 2010b) has shown SVM Light and Balanced Winnow to lie on an equal level, both outperforming Naive Bayes. Of those two, Balanced Winnow offers the higher speed. Again following D'hondt et al. (2012) we therefore choose to use Balanced Winnow.

We also use the same LCS configuration, viz.:

1. Global Term Selection: minimal document frequency = 2, minimal term frequency = 3
2. Local Term Selection: Simple Chi Square (Galavotti et al., 2000), selecting the 10,000 most representative term per class.
3. After local term selection all of the remaining terms are combined into one vocabulary which is then used as a starting point for training the individual classes, i.e. aggregation of term vocabularies
4. Term Strength Calculation: LTC algorithm
5. Training Method: Ensemble learning based one-versus-rest binary classifiers. This means that there is not one classifier assigning all the class labels, but every class has its own binary classifier. Each of these classifiers independently assigns a score to every given document, representing the confidence that this document belongs to that class. To each document is assigned at least one and at most four of these class labels (if the classifier confidence score is greater than the threshold of 1.0).
6. Winnow Configuration: $\alpha = 1.02, \beta = 0.98, \theta+ = 2.0\theta- = 0.5$, with a maximum of 10 training iterations. This setting has also been used in Koster et al. (2010).

Chapter 4

Results and Initial Analysis

The results of our classification experiments are depicted in Table 4.1 (micro-averaged). In addition to the measure value, the confidence ranges at a 95% confidence level have been given. The best results per representation and measure have been set in boldface. The macro-averaged results fall within too large confidence intervals to draw any significant conclusions. For the sake of completeness they can be found in the appendix, Section A.1.

Terms	% Precision	% Recall	% F1
Unigrams	76.82 ± 0.17	66.51 ± 0.19	71.29 ± 0.19
Bigrams	79.31 ± 0.17	67.54 ± 0.19	72.95 ± 0.18
Uni+Bi	79.37 ± 0.17	70.72 ± 0.19	74.79 ± 0.18
Skipgrams	79.34 ± 0.17	69.06 ± 0.19	73.84 ± 0.18
Uni+Skip	79.30 ± 0.17	71.14 ± 0.19	75.00 ± 0.18
Typed Skipgrams	79.69 ± 0.17	67.03 ± 0.19	72.81 ± 0.18
Uni+TSkip	**80.17 ± 0.16**	**71.33 ± 0.19**	**75.49 ± 0.18**

Table 4.1: Classification scores, micro-averaged

4.1 Measures and Averaging Methods

Let us first briefly introduce the measures used. Using the terminology from the pattern matching literature, all of the measures can be defined in terms of true positives (instances of a class which have been correctly labelled as such), false positives (an instance which has been wrongly labelled as belonging to class x) and, *mutatis mutandis*, true and false negatives.

Precision is defined as:

$$\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

If the classifier picks x instances as belonging to class y this is the fraction of instances which actually do belong to this class. In information retrieval terminology: what fraction of returned documents is relevant?

Recall is defined as:

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

If there are x instances of a class, what fraction of these did the classifier classify correctly? Or, again, rephrased: what fraction of relevant documents has been returned?

The F1 value is the harmonic mean between precision and recall. For the variables $x_1, ..., x_n$ it is defined as:

$$\frac{n}{\frac{1}{x_1} + ... + \frac{1}{x_n}}$$

The harmonic mean tends towards the lowest value given. Large outliers have less effect than for the geometric mean, but many small outliers tend to aggravate.

Precision, Recall and F1 have been averaged in two ways. The *micro-average* is an average over every class instance in the test set (of which there are 226071 (one for every class label / document combination)). As such, classes having many documents will have greater influence on this average.

Macro-average, on the other hand, is an average over every class, of which there are 121. This lends an equal weight to every class, independent of its size.

As a direct consequence of its definition, scores resulting from the latter averaging method have a much wider confidence interval. When looking at Table A.1 in Section A.1 one can see that these intervals actually are so wide that none of the results is significantly different from any other (which would, here, require a difference of ca. 18%, no combination of values lies this far apart). As such, we will only consider micro-averaged results in the discussion below.

4.2 Initial Analysis

First, we were able to reproduce the results of D'hondt et al. (2012) which showed that combined representations outperform unigrams. When bigrams, skipgrams or typed skipgrams are combined with unigrams, classification scores always increase with respect to the solitary unigrams classification.

Secondly, it may be observed that all (isolated) phrases yield better scores than unigrams. This is noteworthy insofar as it contrasts earlier findings in the literature, see for example Lewis (1992). Also see Koster and Beney (2009), who state that "It is a disappointing fact that over the years no classification experiment using any sort of linguistical phrases has shown a marked improvement over the use of single keywords, at least for English.". Here one may note that only the typed skipgrams can be considered to be weakly linguistical phrases. The reason for this deviation from earlier results might be found in the different language use in patents. Technical terms and domain-specific vocabulary are more prevalent than in, for example, newswire texts and they are better captured by phrases than by words alone. Domain-specific vocabulary can be seen to, for most parts, fall under the institutionalized phrases subset of multiword expressions. As such, success of phrasal representations provides first support for our hypothesis, which tried to explain the good performance of words + bigrams by their ability to capture such MWEs.

Thirdly, while unigrams and typed skipgrams perform better than unigrams combined with skipgrams, skipgrams on their own yield better recall and F1 scores than pure typed skipgrams. It may also be observed, that typed skipgrams still provide a higher precision. Several reasons might explain this behaviour. Typed skipgrams have a much lower token/type ratio (14.31 versus 6.60) and a vastly lower number of tokens in general (168 million vs 45 million, confer Table 3.4 on Page 13) and thus suffer from data sparseness problems, causing a lower recall. On the other hand, they also are highly informative, meaning that if a document fits their model, the labelling is likely to be correct.

Fourthly, even when ignoring the reasons behind the low recall and high precision of typed skipgrams, we can see that they combine with unigrams very well. The information that the two

different representations provide seems to complement each other, yielding the highest scores across all representations.

Fifthly, finally and most importantly (and implicitly mentioned before), unigrams + typed skipgrams significantly outperform the unigrams + bigrams baseline[1]. Unigrams + typed skipgrams also significantly outperform unigrams + skipgrams. The latter performs better than unigrams + bigrams, but not significantly.

We thus see that these results are as predicted by our hypothesis: The combined results of unigrams + x increase with an increasing ability of x to capture multiword expressions.

One might raise the objection that unigrams+skipgrams only work (slightly) better due to an increased number in tokens. That this is not the case is made clear by unigrams+typed skipgrams outperforming both other combined representations significantly. As has been remarked before, the typed skipgrams form a small subset (ca. 45 million tokens vs ca. 168 million) of the skipgrams. If thus better results for unigrams+skipgram were to be explained by a higher token count alone then lowering this count below the level of bigrams (ca. 57 million) should not result in an increased performance.

[1]This baseline is slightly higher than in D'hondt et al. (2012) due to improvements made to the preprocessing step.

Chapter 5

Discussion

In the last section we have found evidence supporting our initial hypothesis[1] Yet, what we have actually found out is the following: typed skipgrams and words taken together are good features. As of now, we have assumed this to be the case because the typed skipgrams capture multiword expressions. Other observations (see above) make this assumption appear plausible, but fail to completely remove reasonable doubt.

In what follows, we aim to give a further analysis of the results and thereby find further evidence that our hypothesis is true. The central question we pose is this: Can the improvement in performance of typed skipgrams + words be traced back to the influence of multiword expressions? If we find this to be the case, we will inquire further into the nature of the multiword expressions which proved useful to the classification task. To this end, we shall proceed as follows:

1. We shall conduct a series of leave-one-out experiments, inspecting the influence of specific subtypes of typed skipgrams, see Section 5.1.
2. We will perform an additional run, combining the two most informative subtypes of typed skipgrams with unigrams, in Section 5.2.
3. The ten most influential terms for one class will be compared among different representations in Section 5.3.
4. In Section 5.4, the 100 most influential typed skipgrams for that same class will be manually annotated.
5. Penetration values among different representations and ranks will be compared in Section 5.5.

5.1 Leave-One-Out Experiment

To examine the influence of various subtypes of typed skipgrams we have conducted a series of leave-one-out experiments. For each iteration, we have chosen a pair of tags originally allowed (cf. Table 3.3 on Page 13), filtered it out (symmetrically, so if we filter a N-A combination we also filter A-N) and re-run the classification. Please note, that we did not so in a cumulative manner. Rather, every iteration of this experiment stands on its own, meaning, that each time we filtered out only one combination and did not also filter out the ones from previous iterations. The results are shown in Table 5.1 (micro-averaged).

[1] Viz.: "Multiword expressions are the most informative phrasal features for text classification in the patent domain".

Again, as with the results of the main experiment described in Section 4.2, the macro-averaged results prove not to be significantly different from one another. Consequently, no conclusions will be drawn from their differences. They are listed in Section A.2.

The tables are formatted in the following way: the first row gives our typed skipgrams baseline. The following rows give, per representation and measure, the difference, compared to the baseline, followed by the actual result in parentheses. Furthermore, the confidence interval range at 95% confidence is given. Per representation and measure the value deviating the most from our baseline is set in boldface.

Terms	% Precision	% Recall	% F1
Tskipgrams baseline	79.69 ± 0.17	67.03 ± 0.19	72.81 ± 0.18
tskip noNN	**-1.93 (77.76) ± 0.17**	**-4.12 (62.91) ± 0.20**	**-3.26 (69.55) ± 0.19**
tskip noNA	-1.12 (78.57) ± 0.17	-2.75 (64.28) ± 0.20	-2.10 (70.71) ± 0.19
tskip noNV	-0.75 (78.94) ± 0.17	-2.14 (64.89) ± 0.20	-1.58 (71.23) ± 0.19
tskip noVA	-0.07 (79.62) ± 0.17	-0.23 (66.80) ± 0.19	-0.16 (72.65) ± 0.18

Table 5.1: Micro-averaged classification scores for the leave-one-out experiment for typed skipgrams

The biggest drops in classification accuracy, compared to the typed skipgram baseline found in Table 4.1 on Page 15 above, can be witnessed when leaving out noun noun-tagged (NN) typed skipgrams, noun-adjective (NA), noun-verb (NV) and verb-adjective (VA), respectively.

For now, let us call the group of typed skipgrams indicated by the (unordered) tag combination of V A the VA subtype of typed skipgrams (and the same, *mutatis mutandis*, for other subtypes). When considering F1 scores, the VA subtype is significantly less important than any of the other ones, whereas the NN subtype is significantly more important than the rest. The difference between the NA and NV subtypes is not significant.

We explain these findings by the NN subtype of typed skipgrams holding the most valuable information, followed by NA. NV also has descriptive value, its impact seems to be a consequence of the high token count for this subtype. VA simply does not appear often enough to have much influence. To provide support for this reasoning it is necessary to also consider the frequency of the different subtypes, as depicted in Table 5.2.

Tag	#Occurrences	% of the whole corpus
VA/AV	4,692,536	2.78
NA/AN	12,794,063	7.58
NV/VN	14,111,043	8.36
NN	14,315,700	8.48

Table 5.2: Distribution of subtypes of typed skipgrams, absolute and relative to the whole corpus

As is evident, the VA subtype comprises a relatively few features, so that their removal does not impact classification greatly.

NN does have the highest token count but is closely (difference: ca 200.000 tokens) followed by NV. As such, the great difference in impact can not be explained by the difference in number. The same seems to be true when comparing NN and NA, but the case is less clear here.

Another point of interest is the significant difference between NA and NV. Whereas NV has a higher token count, NA has the higher impact. We thus conclude that NA terms seem to be more informative than NV terms, as they yield a greater change of scores while having less tokens.

This conclusion is supported by our findings of the manual inspection of the ten most important terms per representation, described in Section 5.3. We see, that many of the influential NV terms are formed by combining high-weight unigrams with auxiliary verbs, such as "be" or "have".

How does this relate to our hypothesis? Let us revisit what we are looking for: we are hoping to see that the subtypes of typed skipgrams with the highest influence are also the subtypes most likely to contain multiword expressions. As we are looking for terms akin to "green IT" (NA) or "software engineering" (NN) we expect the NN and NA subtypes to have the greatest impact. Since this is exactly what we observe this lends further support to our hypothesis.

5.2 Leave-Two-In Experiment

By conduction the leave-one-out experiment, see Section 5.1 above, we have found typed skipgrams being constructed from pairs of words tagged noun-noun (NN subtype) or noun-adjective (or adjective-noun) (NA class) to have the greatest impact on typed skipgram performance.

In this experiment we only allow those two subtypes of typed skipgrams. We compare the performance of this subset of typed skipgrams, on its own and in combination with unigrams, to earlier results. We hope to find that this new variant still significantly outperforms the unigrams+bigrams baseline.

The micro-averaged performance of this subtype of typed skipgrams has been depicted in Table A.3. The new representation has been labelled with "onlyNNNA", some previous results have been included for ease of comparison. As before, the macro-averaged results can be found in the appendix, in this case in Section A.3.

Terms	% Precision	% Recall	% F1
UniBi	79.37 ± 0.17	70.72 ± 0.19	74.79 ± 0.18
UniSkip	79.30 ± 0.17	71.14 ± 0.19	75.00 ± 0.18
UniTskip	**80.17 ± 0.16**	**71.33 ± 0.19**	**75.49 ± 0.18**
uni+tskip onlyNNNA	79.88 ± 0.17	71.06 ± 0.19	75.21 ± 0.18
Skip	79.34 ± 0.17	69.06 ± 0.19	73.84 ± 0.18
Tskip	79.69 ± 0.17	67.03 ± 0.19	72.81 ± 0.18
tskip onlyNNNA	78.86 ± 0.17	64.82 ± 0.20	71.16 ± 0.19

Table 5.3: Classification scores for the leave-two-in experiment, micro-averaged

When considering typed skipgrams restricted to the NN/NA subtypes (we shall call them minimal typed skipgrams, or m-t-skipgrams, for now) we can see that they perform worse than the less restricted typed skipgrams we have used before. While scores for all measures decline significantly it is noteworthy that the loss in recall (2.21) is much higher than the loss in precision (0.81). We thus see, that the NN/NA subtypes are the main factor behind the overall precision of typed skipgrams. The greater loss of recall might be caused by NV and VA being important subtypes for this, the loss of ca 19 million (out of ca 46 million) tokens or, most likely, a combination thereof.

If we assume NN/NA to capture multiword expressions (and this is supported by our later manual analysis in sections 5.3 and 5.4) we may draw the conclusion that multiword expressions have high discriminatory power and lead to high precision scores when using them as features.

If we turn our attention to m-t-skipgrams combined with unigrams we see that they perform better than skipgrams+unigrams, but worse than regular typed skipgrams+unigrams. The

differences between the classification scores of the unigrams+m-t-skipgrams and the two other representations are not significant.

The loss in recall (0.27) and precision (0.29), compared to regular typed skipgrams+unigrams, is much smaller than when contrasting our two versions of typed skipgrams without unigrams. As we have seen before in Section 4.2 typed skipgrams seem to combine very well with unigrams, complementing each other. This also seems to hold for m-t-skipgrams. Especially the loss in recall is made less severe by combining with unigrams.

When relating these findings to our hypothesis the main observation is that the unigrams + bigrams baseline is significantly outperformed by unigrams + m-t-skipgrams. This indicates that NN/NA are the important subtypes of typed skipgrams. As these subtypes are also the ones capturing multiword expressions we again find support for our initial hypothesis: multiword expressions being captured by bigrams was the mechanism underlying the success of unigrams + bigrams.

We have furthermore shown that a possible alternative interpretation of our results, viz.: "unigrams + typed skipgrams work better than unigrams + skipgrams only due to functions words being filtered out", can not be upheld. Were this the case then leaving out noun verb or verb adjective combinations should have a more severe impact. As this is not the case we conclude that the improvement in scores actually is in the phrases better capturing multiword expressions (although this does happen by filtering out skipgrams containing function words).

Finally, we see that while the initial number of tokens for m-t-skipgrams (26 million) is far lower than the one for bigrams (50 m.) the results are better nonetheless. Furthermore, comparison to typed skipgrams (45 m. initial features) yields comparable performance. This leads to the conclusion that the filtering was strict, but no significant amount of useful features has been discarded by accident.

5.3 Manual Analysis of Most Influential Terms Across Representations

In this section per representation the ten terms given most positive weight by the classifier[2] for the A61 class, the largest class in the corpus, are shown, see Table 5.4. Our aim is to illustrate the differences in terms chosen by the classifier and to start a first investigation into why these terms are chosen, how much overlap there is and whether multiword expressions play a role. This is not a large-scale, automatized analysis and findings thus cannot hold up against the same standard of validity as in the previous sections. Nevertheless, there is value in an illustrative example to understand what happens in the classifier.

Before the observations of this experiment are summarized, a short reiteration might be helpful. We are especially interested in the institutionalized phrases subclass of multiword expressions (examples: traffic light, greedy search, ...; cf. Section 2.4). By merit of their definition whether a specific combination of two words is such a phrase is relative to the chosen corpus. This is the basis on which we have decided whether to label a phrase as multiword expression in the analysis below.

One could argue that, as we are looking for phrases characteristic for individual and largely disjoint classes each such class constitutes a frame of reference against which institutionalized-ness

[2] "Weight" here means "weight assigned by the Balanced Winnow classifier". Balanced Winnow gives two weights to a term, thus keeping two hyperplanes in the feature space, separating the target class from everything else. One of these hyperplanes works as in Perceptron or Support Vector Machine algorithms, the other one is a negative boundary: words are given more negative weight if they are more important for not being in the target class. When ordering terms by weight for our further experiments and analyses we have only taken into account the positive weights of the terms.

Uni	Bi	Skip	TSkip
dental	a_catheter	an_implantable	catheter_be
orthopedic	an_implantable	a_catheter	dental_comprise
implantable	a_dental	a_surgical	catheter_have
catheter	a_surgical	an_orthopedic	absorbent_article
endoscope	the_catheter	a_dental	prosthesis_be
prosthesis	an_endoscope	an_endoscope	cosmetic_composition
prosthetic	for_dental	the_catheter	surgical_instrument
denture	an_orthopedic	absorbent_article	dental_material
surgical	absorbent_article	for_dental	prosthesis_have
suture	a_wheelchair	a_wheelchair	prosthesis_comprise

Table 5.4: Top ten highest weighted terms, according to Winnow weight, of the A61 class per representation.

must be measured.

Regarding unigrams we can see that all terms are content words, divided equally between nouns and adjectives.

Regarding bigrams it can be stated that they are generally constructed from content words (here: adjectives and nouns, let us call this the semantically heavy part), prefixed by functions words (mainly articles, we shall refer to it as the semantically light part). The only exception in this top en is "absorbent_article" which is a meaningful multiword expression.

As has also been observed in D'hondt et al. (2012), the heavy parts of the individual bigrams mainly stem from the high-ranking unigrams: when excluding "absorbent_article" eight out of nine bigrams have one of the important unigrams as heavy part. When looking at the inverse relation we see that six out of ten unigrams appear as heavy part of one of the most influential bigrams.

One might think that this relation (bigrams stemming from high-ranking unigrams versus high-ranking unigrams appearing in bigrams) should be symmetric because every unigram should only appear once in the unigram list and once in the bigram list. That this is not the case is explained by the fact that one heavy part of the bigrams (possibly being one of those unigrams) may be connected to different light parts. Only this combination as a whole has to be unique.

A possible explanation for those characteristics of bigrams is that their importance is mainly derived from important unigrams. The high weighing of combinations of unigrams and articles is due to high frequency thereof: articles in general appear very often, even more so when inspecting the direct vicinity of adjectives and nouns.

The fact that one multiword expression is among those ten phrases is consistent with the observation that unigrams+bigrams perform significantly better than unigrams alone. If this sample is indicative for the rest of the terms it would fit well into our hypothesis.

Regarding skipgrams it is the case their sample of phrases is a permutation of the bigrams. This indicates that skipgrams with zero skips (which are identical to the bigrams) feature higher in the class profiles and are more readily selected by the classifier. As classification results differ for skipgrams and bigrams it is likely that more divergence will be found when increasing the sample size.

Regarding typed skipgrams one can observe that six out of ten are constructed from a noun or adjective (again: the heavy part) and an auxiliary verb (again: the light part). "Auxiliary" in a dual sense of the word: these are auxiliary verbs but they also play the same role as the articles for bigrams and skipgrams. They carry nearly no additional meaning.

Subtype	#Occurences	Fraction	#MWEs	Fraction	Capture Ratio
NN	25	0.25	16	0.41	0.64
NA	30	0.30	23	0.59	0.77
NV	34	0.34	0	0.00	0.00
VA	11	0.11	0	0.00	0.00

Table 5.5: Distribution of tag subtypes and multiword expressions among the best 100 typed skipgrams

As before, we explain the presence of such heavy/light combinations by their distribution: auxiliary verbs are frequent and close to every word (at least when allowing skips). As such, they also are close to the important unigrams. Because function words have been filtered out, they are the best next candidate.

As far as meaningfulness goes, auxiliary verbs are close to functions words. When filtering all typed skipgrams involving verbs and combining with unigrams the unigrams+bigrams baseline is still outperformed significantly, as can be expected from these observations. This is shown in Section 5.2.

Filtering out only auxiliary verbs (as opposed to all verbs) and seeing, whether scores for unigrams + typed skipgrams increase might be an interesting experiment. Unfortunately, we could not include it anymore.

Returning to a more general inspection of the typed skipgrams it can be seen that there are four phrases which are meaningful combinations of words; three of those seem to be multiword expressions (they appeared as units in text relating to the medical field, frequency analyses have not been carried out).

Due to the small sample size, only tentative conclusions can be drawn. It does seem, however, that the linguistic filtering applied to typed skipgrams does help to promote multiword expressions. Less 'light' parts can be found among the sample.

5.4 Manual Annotation of the 100 Most Influential typed skipgrams

This section describes our findings of a manual annotation of the 100 typed skipgrams given the most weight by the classifier for the A61 class. We inspected from what combination of Part-of-Speech tags they were created and, furthermore, which of these combinations best capture multiword expressions. The results of our annotation are depicted in Table 5.5.

"Subtype" is here used as in Section 5.1 before, meaning: the "NA" subtype consists of the typed skipgrams constructed from pairs of words tagged either "N" and "A" or "A" and "N". This scheme holds,*Mutatis mutandis*, for all subtypes. To recap Section 3.2.3, the tags N, A and V try to capture nouns, adjectives and verbs, respectively.

"Capture Ratio" is the number of captured multiword expressions divided by the number of typed skipgrams for this subtype.

When inspecting the distribution of the different subtypes we see that (in this order) NV, NA and NN have the highest number of occurrences and VA is left far behind. This does not reflect the typed skipgram distribution described in Table 7. Although the NN subtype has the most tokens when considering the whole corpus, and the greatest impact in the leave-one-out experiments, relatively few NN features are selected as high-impact features by the classifier.

Before extracting a conclusion, let us first have a look at the distribution of multiword expressions amongst those subtypes. Two annotators have found 39 and 24 such expressions respectively[3] There are 39 such expressions in total. To determine whether a term is a multi word expressions we have conducted an Internet search for that term and checked, whether it appeared as an institutionalized term in the first results. This, admittedly, is a very vague and non-deterministic method, but is more certain than only using intuition. A thorough analysis would require statistical comparison of this term to the rest of the class or inspection by an expert of the relevant domain; both of those could, due do the limited amount of time available, not be carried out. We see, that only NN and NA capture these multiword expressions.

Again, there is a disagreement between these numbers and the leave-one-out experiment: here the NA subtype captures more multiword expressions than NN. This disagreement turns out to be unproblematic as the lower impact in the leave-one-out experiment can also be explained by the lower token count. One might also take into account the low sample size considered here.

On a greater scale, one can see that NN and NA are still the most important subtypes regarding multiword expressions (or rather: the only important ones)[4]. This fits in with earlier observations: when constructing typed skipgrams using only these two subtypes, the unigrams+bigrams baseline is still outperformed (cf. Section 5.2. Furthermore, these subtypes were shown to have the biggest impact on typed skipgrams performance (cf. Section 5.1. As such, these observations again support our hypothesis that capturing of multiword expressions was the crucial mechanism behind the success of unigrams+bigrams.

5.5 Penetration of Phrases

Penetration, introduced by Caropreso et al. (2001), measures the relative number of phrases among the top k terms for a class in the class profile. In general, this value tells us about the tendency of a classifier to select phrases, instead of words, as features that can discriminate a certain class against the rest of the corpus.

On a purely technical level this means that, for higher values of penetration, the Winnow algorithm has given more weight to phrasal terms , compared to unigrams. Winnow is a mistake-driven classification algorithm: The winnow weights of the features are only updated during training when a test document is assigned a wrong label or no label. In the latter case the winnow weights of all the terms that the document shares with positives examples of its class in the training set are then multiplied by the promotion factor (α), conversely, the weights of all the terms that the document shares with negative training material are multiplied by the demotion factor (β). Consequently, high-ranking terms in the class profile are not necessarily the most representative terms for that specific class but rather are the terms that capture such specific information to distinguish this class from all others in the corpus.

Higher penetration levels means the classifier has found that more specific, i.e. phrasal terms distinguish better from other classes then unigrams. The penetration level itself however is a contestable measure as it is also dependant on the number of types (features) that are offered to the classifier, i.e. the feature space in which the classifier operates.

Penetration values at different ks for the three biggest classes per representation can be found

[3]The inter-annotator agreement has been measured using Cohen's Kappa, yielding 0.746. The scale for Cohen's Kappa has two important points: 0, which means that the inter-annotator agreement is as high as one would expect if all annotators simply guessed. 1, which means that all annotators are in perfect agreement. There is no established standard on what constitutes a good value. Following the rules of thumb established in Landis and Koch (1977) and Fleiss et al. (2003, p. 218), 0.746 appears to indicate a high agreement.

[4]For this specific sample. That this finding generalizes to the whole corpus, especially when focusing on institutionalized phrases, is indicated by the previous leave-one-out and leave-two-in experiments.

Class	Representation	Rank: 20	50	100	1000	10000	100000
A61	Uni+bigrams	0.00	0.04	0.11	0.41	0.70	0.88
	Uni+skipgrams	0.00	0.06	0.19	0.54	0.82	0.93
	Uni+typed skipgrams	0.00	0.04	0.08	0.32	0.65	0.89
C07	Uni+bigrams	0.05	0.22	0.30	0.53	0.73	0.85
	Uni+skipgrams	0.10	0.26	0.35	0.68	0.84	0.93
	Uni+typed skipgrams	0.00	0.16	0.21	0.45	0.70	0.86
H04	Uni+bigrams	0.10	0.16	0.24	0.56	0.81	0.90
	Uni+skipgrams	0.10	0.24	0.31	0.72	0.89	0.95
	Uni+typed skipgrams	0.10	0.12	0.21	0.48	0.80	0.92

Table 5.6: Penetration of phrases for different ranks, phrases and classes

in Table 5.6. A subset of this data, restricted to the A61 class, is visualized in Figure 5.1. It can be seen that, independent of the class, unigrams+skipgrams has the highest penetration value (starting from rank 50). It is furthermore the case that unigrams+bigrams show a slightly higher penetration than unigrams+typed skipgrams.

At first glance, it is surprising that the highest-scoring text representation (unigrams + typed skipgrams) has the lowest penetration levels. If the typed skipgrams capture the most specific information of the classes, we would expect them to feature quite high in the class profiles and it would be evident that they play a large role in the classification process.

The reason for their low penetration levels, compared to skipgrams, might be twofold: On the one hand skipgrams have a larger feature space with much more phrases compared to unigrams (28 skipgram terms to each unigram term; and 16 typed skipgram terms to each unigram term, respectively). Hence, the statistical possibility of selecting a phrasal term as a distinguishable feature is much higher for skipgrams.

Furthermore, the more specific typed skipgrams are much sparser and, consequently, less likely to feature a lot in either positive and/or negative training material. Consider, the Token-Type ratio of typed skipgrams in Table 3.4. On average, there are only 6 occurrences per typed skipgram term in the entire corpus. From this we can conclude that typed skipgrams will have a limited impact during training. skipgrams, on the other, hand contain many function words, creating general terms that can appear in many documents and in many classes. These will therefore have more impact during training both as positive and negative training material. Analysis in Section 5.3 showed high-ranking skipgrams are mainly class-specific unigrams combined with function words, which seems to support the previous claim. While the typed skipgrams may not feature heavily in the top ranks, penetration at lower levels show that the classifier still selected a lot of typed skipgrams with smaller winnow weight. It seems that for combined unigram+typed skipgram classification, most of the work is left to unigrams but the complementary typed skipgrams fill up the gaps.

This implies that the quality of the phrases must be higher. As such, linguistic filtering from skipgrams to typed skipgrams appears to be successful. As this filtering happened under the assumption that phrases should come closer to multiword expressions, we again find support for our hypothesis.

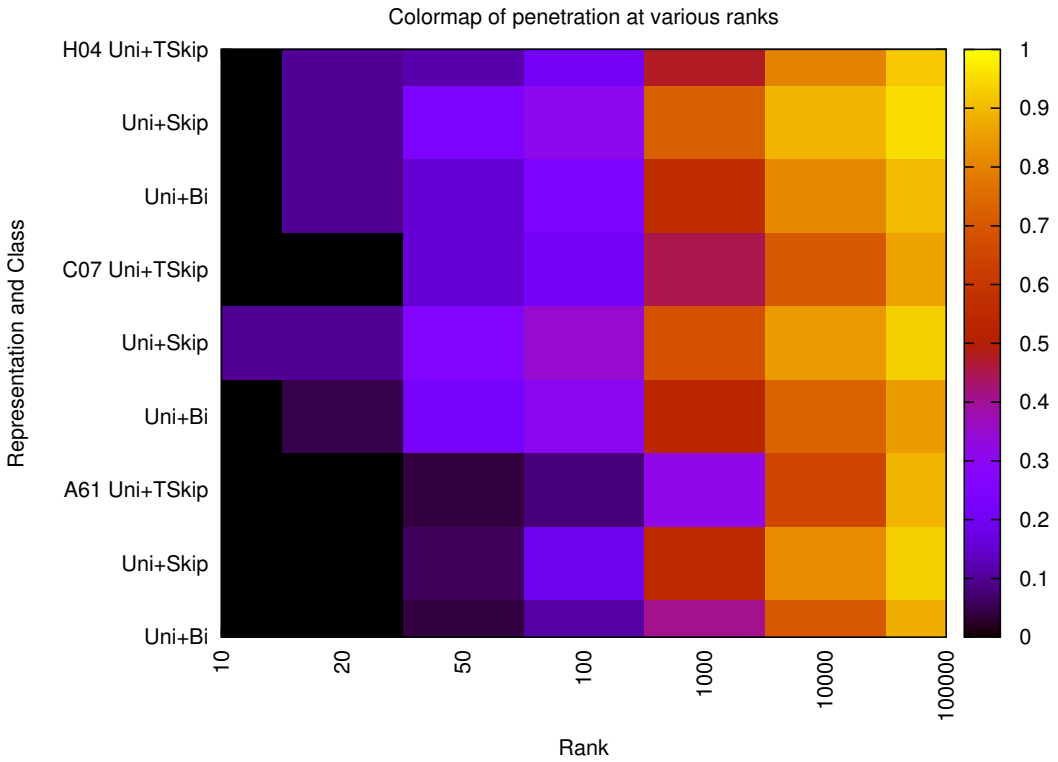

Figure 5.1: Penetration of phrases among all classes, ranks and representations. X-Axis: rank. Y-Axis: Representation and Class. Z-Axis (color): penetration.

Chapter 6

Conclusions

In this paper we have investigated whether the mechanism behind the success of the unigrams+bigrams text representation in D'hondt et al. (2012) is that representation's capability to capture multiword expressions, especially the subset called institutionalized phrases. To verify this hypothesis we have conducted a series of experiments:

Firstly, we have conducted multiple classification experiments on ca. 532.000 patent abstracts, originating from the CLEF-IP 2010 corpus. These experiments have been carried out on the "class" level of the IPC hierarchy, consisting of 121 different classes. Each document on average has 2.12 class labels. We thus solving a multi-class, multi-label supervised learning problem. For each of the following text representations a classification experiment has been carried out: unigrams, bigrams, skipgrams, typed skipgrams and unigrams in combination with any of the phrase-based text representations. For typed skipgrams, combinations of words being tagged as noun, adjective or verb by a Part-of-Speech tagger have been allowed. The rationale behind this was that the resulting terms would not contain function words, but would still capture multiword expressions (which we predicted to comprise noun/adjective combinations). Combinations with verbs have been allowed to evade confirmation bias and to prevent filtering from being too strict. In further experiments (see below), the influence of the different types of typed skipgrams has been tested. For all experiments recall, precision and F1-values have been measured an analyzed. In contrast to earlier results in the literature all phrase-based representations have been observed to outperform unigrams. We have furthermore found unigrams+typed skipgrams to perform best, followed by unigrams+skipgrams, both significantly surpassing the unigrams+bigrams baseline. They are ranked in accordance with their capability to capture multiword expressions, as predicted by our hypothesis.

Secondly, we have conducted leave-one-out experiment with typed skipgrams, each time filtering out phrases containing a particular Part-of-Speech combination (e.g. noun-noun or noun-adjective). By means of this we have gauged the impact of these different subsets of typed skipgrams. We found, that those constructed from noun-noun and noun-adjective combinations contribute the most, followed by noun-verb (all of these having significant impact) and, finally, noun-verb (having nearly no impact at all). The difference in impact could again be traced to the ability to capture multiword expressions, although the number of tokens also played a significant role.

Thirdly, we have compared the top ten terms given most weight by the classifier for different text representations in the class profiles of A61, the largest category in the corpus. It can be observed, that most of those important unigrams also appear in the important phrases. Overlap between different phrases was also found to be high: bigrams and skipgrams were a permutation

of each other. In general, high-ranking unigrams are combined with high-frequency words to construct high-impact phrases. Often this leads to "meaningless" combinations of unigrams with articles or (when impossible due to the filtering applied to typed skipgrams) with auxiliary verbs. Meaningful combinations of two unigrams among the top ten terms can only be found in greater numbers for typed skipgrams, proving the filtering to be effective.

Finally, we have manually annotated the 100 most influential typed skipgrams. A total of 39 multiword expressions has been found among them, providing evidence for a) the representation's capacity of capturing them and b) their influence on the results. All of those multiword expressions were captured either by noun-noun or adjective-noun combinations. Using only these combinations for typed skipgrams has been tested. In combination with unigrams results are not significantly worse than when combining unigrams with more inclusive typed skipgrams, although the unigrams+bigrams baseline was still outperformed significantly.

In conclusion, our findings are that there is considerable support for our hypothesis which explains the performance of unigrams+bigrams by their ability to capture multiword expressions. This hypothesis predicted the success of unigrams+typed skipgrams, which could be verified. It furthermore predicted the importance of noun-noun and noun-adjective typed skipgrams, which also could be verified. We expect these results to be generalizable for every sort of text relying heavily on institutionalized phrases. This means that unigrams + typed skipgrams should be effective features for such texts. Therefore, this representation might also prove helpful in classifying texts such as scientific abstracts. When combined with insights regarding parameter tuning (not covered in this paper) and when used in a more realistic classification setting (where, for example, also title and the beginning of the description section are being used) this new representation might help to increase the quality for automated text classification. Finding further ways of exploiting multiword expressions might also form a viable prospective for research. One can, for example, think of typed skipgrams consisting of three words, capturing things as "self-documenting extensible editor". The crucial aspect will again be to balance informativeness and sparseness, as Guthrie et al. (2006) have shown this to become a problem for unfiltered skipgrams involving four words.

There also are important remarks to the work presented here. First, the Part-of-Speech tagger has not been benchmarked. As decisions made by this component influence which typed skipgram are created, errors can potentially have a significant effect. Reversely, improving performance of the tagger might increase performance of the typed skipgrams. Secondly, an automated analysis of all classprofiles should be conducted. Conclusions about the distribution of multiword expressions could be verified and the influence of differences in class size could be analyzed. Finally, this experiment could be re-conducted for different classification algorithms. This would show whether the success of typed skipgrams depends on characteristics of the Winnow algorithm. Should this prove to be false, it would allow for a wider application of this new representation. Should it prove to be true, new insights on the behaviour of this classifier regarding feature selection could be gained.

Bibliography

Apte, C., Damerau, F., Weiss, S. M., Apte, C., Damerau, F., and Weiss, S. M. (1994). Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 12:233 – 251.

Arampatzis, A., Tsoris, T., Koster, C., and van der Weide, T. (1998). Phase-based information retrieval. *Information Processing & Management*, 34(6):693–707.

Atkinson, K. H. (2008). Toward a more rational patent search paradigm. In *Proceeding of the 1st ACM workshop on Patent information retrieval - PaIR '08*, page 37, New York, New York, USA. ACM Press.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.

Bekkerman, R. and Allan, J. (2003). Using bigrams in text categorization. Technical report, Center of Intelligent Information Retrieval, UMass Amherst.

Beney, J. (2010). LCI-INSA linguistic experiment for CLEF-IP classification track.

Beney, J. and Koster, C. H. A. (2003). SVM Paradoxes. In *Proceedings PSI 2003*, pages 545–554.

Benzineb, K. and Guyot, J. (2010). *Automated patent classification*. Springer-Verlag.

Braga, I. A., Monard, M. C., and Matsubara, E. T. (2009). Combining unigrams and bigrams in semi-supervised text classification. In Lopes, L. S., Lau, N., Mariano, P., and Rocha, L., editors, *Proceedings of Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, pages 489–500, Aveiro, Portugal.

Caropreso, M. F., Matwin, S., and Sebastiani, F. (2001). A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization.

Cheng, W., Greaves, C., and Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4):411–433.

Crawford, E., Koprinska, I., and Patrick, J. (2004). Phrases and Feature Selection in E-Mail Classification. In *Proceedings of the 9th Australasian Document Computing Symposium*, volume 2004, pages 59 – 62.

D'hondt, E., Verberne, S., Koster, K., and Boves, L. (2012). Text Representations for Patent Classification.

Dumais, S. (1998). Using SVMs for Text Categorization.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the seventh international conference on Information and knowledge management - CIKM '98*, pages 148–155.

Fall, C. and Benzineb, K. (2002). Literature survey: Issues to be considered in the automatic classification of patents. *World Intellectual Property Organization, Oct*, pages 1–64.

Fall, C. J., Törcsvári, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the international patent classification. *ACM SIGIR Forum*, 37(1):10–25.

Feldman, S., Marin, M., Medero, J., and Ostendorf, M. (2009). Classifying factored genres with part-of-speech histograms. In *NAACL-Short '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 173–176.

Fleiss, J. L., Levin, B., and Cho Paik, M. (2003). *Statistical Methods For Rates And Proportions*. Wiley, 3 edition.

Galavotti, L., Sebastiani, F., and Simi, M. (2000). Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization. In *ECDL '00 Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 59–68.

Goldstein, J. and Sabin, R. (2006). Using Speech Acts to Categorize Email and Identify Email Genres. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, pages 50b–50b. IEEE.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4.

Guyot, J., Benzineb, K., and Falquet, G. (2010). myclass: A mature tool for patent classification. In *CLEF (Notebook Papers/LABs/Workshops)*.

Halteren, H. V. (2000). The detection of inconsistency in manually tagged text. In *Proceedings of LINC-00. Luxembourg*.

Halteren, H. V., Zavrel, J., and Daelemans, W. (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics*, 27(2):199–229.

Held, P., Schellner, I., and Ota, R. (2011). Understanding the world's major patent classification schemes. Presented at the PIUG 2011 Annual Conference Workshop, Vienna.

Hotho, A., Sure, Y., and Getoor, L. (2004). Boosting for Text Classification with Semantic Features. In *International Workshop on Mining for and from the Semantic Web*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142.

Kolcz, A., Chowdhury, A., and Alspector, J. (2003). Data duplication: an imbalance problem? In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*.

Koster, C. (2004a). Head/Modifier Frames for Information Retrieval. In *Proceedings CICLing-2004*, pages 420–432.

Koster, C. (2004b). Transducing text to multiword units. In *Workshop on MultiWord Units MEMURA at the fourth International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal*.

Koster, C. H. and Beney, J. G. (2009). Phrase-based document categorization revisited. In *Proceedings of the PAIR 2009 workshop at CIKM 2009*, pages 49–55.

Koster, C. H. A. (2012). The Linguistic Classification System LCS User Manual.

Koster, C. H. A. and Bene, J. G. (2007). On the importance of parameter tuning in text categorization. In *Proceedings PSI 2006*, pages 269–280.

Koster, C. H. A., Beney, J., Verberne, S., and Vogel, M. (2010). *Phrase-based Document Categorization*. Springer-Verlag.

Koster, C. H. A. and Seutter, M. (2003). Taming Wild Phrases. *Proceedings of 25th European Conference on IR Research (ECIR 25)*, pages 161 – 176.

Koster, C. H. A., Seutter, M., and Beney, J. G. (2003). Multi-classification of patent applications with Winnow. In *Proceedings PSI 2003*, pages 545–554.

Krenn, B., Pecina, P., and Richter, F. (2008). An Evaluation of Methods for the Extraction of Multiword Expressions. In *Towards a Shared Task for Multiword Expressions (MWE 2008)*, number june.

Lam, W., Ruiz, M., and Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *Knowledge and Data Engineering, IEEE Transactions on*, 11(6):865–879.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74.

Larkey, L. S. (1999). A patent search and classification system. In *Proceedings of the fourth ACM conference on Digital libraries - DL '99*, pages 179–187, New York, New York, USA. ACM Press.

Lewis, D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the 15th annual international ACM*.

Lewis, D. D. (1990). Representation Quality in Text Classification: An Introduction and Experiment. In *Proceedings of Workshop on Speech and Natural Language. Hidden*, pages 288–295.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318.

Liu, N., Zhang, B., Yan, J., Chen, Z., Liu, W., Bai, F., and Chien, L. (2005). Text Representation: From Vector to Tensor. *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 725–728.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Christiani, N., and Watkins, C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research 2*, 2:419–444.

Ozgür, L. and Güngör, T. (2009). Analysis of stemming alternatives and dependency pattern support in text classification. In *Tenth Internat. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing 2009), Mexico City*, pages 195–206.

Pinna, A. and Brett, D. (2009). Fixedness and Variability: Using PoS-grams to Study Phraseology in Newspaper Articles.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*.

Schweighofer, E., Rauber, A., and Dittenbach, M. (2001). Automatic text representation, classification and labeling in European law. *Proceedings of the 8th international conference on Artificial intelligence and law - ICAIL '01*, pages 78–87.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Siefkes, C., Assis, F., Chhabra, S., and Yerazunis, W. (2004). Combining winnow and orthogonal sparse bigrams for incremental spam filtering. *Knowledge Discovery in Databases: PKDD 2004*, pages 410–421.

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

Toutanova, K., Klein, D., and Manning, C. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of HLT-NAACL 2003*, pages 252–259.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 63–70.

Verberne, S. and D'hondt, E. (2011). Patent Classification Experiments with the Linguistic Classification System LCS in CLEF-IP 2011. In *CLEF (Notebook Papers/Labs/Workshop)*.

Verberne, S., D'hondt, E., Oostdijk, N., and Koster, C. H. A. (2010a). Quantifying the challenges in parsing patent claims. *International Workshop on Advances in Patent Information Retrieval (AsPIRe 2010)*, pages 14–21.

Verberne, S., Vogel, M., and D'hondt, E. (2010b). Patent Classification Experiments with the Linguistic Classification System LCS. In *CLEF (NotebookPapers/Labs/Workshop)*.

Zhang, W., Yoshida, T., and Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21(8):879–886.

Appendix

Appendix A

Additional Results

A.1 Main Experiment Macro-Averaged Results

Terms	% Precision	% Recall	% F1
Unigrams	74.05 ± 7.81	50.52 ± 8.91	58.47 ± 8.78
Bigrams	79.28 ± 7.22	49.93 ± 8.91	59.52 ± 8.75
Uni+Bi	78.26 ± 7.35	55.40 ± 8.86	63.51 ± 8.58
Skipgrams	79.45 ± 7.20	52.38 ± 8.90	61.54 ± 8.67
Uni+Skip	78.53 ± 7.32	55.79 ± 8.85	63.84 ± 8.56
Typed Skipgrams	**79.97 ± 7.13**	48.12 ± 8.90	57.98 ± 8.79
Uni+TSkip	78.30 ± 7.34	**56.30 ± 8.84**	**64.27 ± 8.54**

Table A.1: Classification scores, macro-averaged

A.2 Leave-One-Out Experiment Macro-Averaged Results

Terms	% Precision	% Recall	% F1
Tskip baseline	79.97 ± 7.13	48.12 ± 8.90	57.98 ± 8.79
tskip noNN	-1.13 (78.84) ± 7.28	**-5.21 (42.91) ± 8.82**	**-4.81 (53.17) ± 8.89**
tskip noNA	**-1.30 (78.67) ± 7.30**	-3.29 (44.83) ± 8.86	-2.96 (55.02) ± 8.86
tskip noNV	0.26 (80.23) ± 7.10	-2.85 (45.27) ± 8.87	-2.49 (55.49) ± 8.86
tskip noVA	-0.94 (79.03) ± 7.25	-0.19 (47.93) ± 8.90	-0.22 (57.76) ± 8.80

Table A.2: Macro-averaged classification scores for the leave-one-out experiment for typed skipgrams

A.3 Leave-Two-In Experiment Macro-Averaged Results

Terms	% Precision	% Recall	% F1
UniBi	79.37 ± 7.21	70.72 ± 8.11	74.79 ± 7.74
UniSkip	79.30 ± 7.22	71.14 ± 8.07	75.00 ± 7.72
UniTskip	**80.17 ± 7.10**	**71.33 ± 8.06**	**75.49 ± 7.66**
UniTskip onlyNNNA	79.88 ± 7.14	71.06 ± 8.08	75.21 ± 7.69
TSkip	79.69 ± 7.17	67.03 ± 8.38	72.81 ± 7.93
TSkip onlyNNNA	78.86 ± 7.28	64.82 ± 8.51	71.16 ± 8.07

Table A.3: Classification scores for the leave-two-in experiment, macro-averaged

Appendix B

Supplemental Information

B.1 Mapping Table for Types

Tag	Map	Tag	Map	Tag	Map	Tag	Map	Tag	Map
APPGE	D	II	PREP	NP1	N	RGR	P	VHG	V
AT	D	IO	PREP	NP2	N	RGT	P	VHI	V
AT1	D	IW	PREP	NPD1	N	RL	X	VHN	V
BCL	X	JJ	A	NPD2	N	RP	X	VHZ	V
CC	X	JJR	A	NPDM1	N	RPK	X	VM	V
CCB	X	JJT	A	NPDM2	N	RR	X	VMK	V
CS	X	JK	A	PN	P	RRQ	X	VV0	V
CSA	X	MC	Q	PN1	P	RRQV	X	VVD	V
CSN	X	MC1	Q	PNQO	P	RRR	X	VVG	V
CST	X	MC2	Q	PNQS	P	RRT	X	VVGK	V
CSW	X	MCGE	Q	PNQV	P	RT	X	VVI	V
DA	D	MCMC	Q	PNX1	P	TO	X	VVN	V
DA1	D	MD	A	PPGE	P	UH	X	VVNK	V
DA2	D	MF	Q	PPH1	P	VB0	V	VVZ	V
DAR	D	ND1	N	PPHO1	P	VBDR	V	XX	UNK
DAT	D	NN	N	PPHO2	P	VBDZ	V	YBL	UNK
DB	D	NN1	N	PPHS1	P	VBG	V	YBR	UNK
DB2	D	NN2	N	PPHS2	P	VBI	V	YCOL	UNK
DD	D	NNA	N	PPIO1	P	VBM	V	YCOM	UNK
DD1	D	NNB	N	PPIO2	P	VBN	V	YDSH	UNK
DD2	D	NNL1	N	PPIS1	P	VBR	V	YEX	UNK
DDQ	D	NNL2	N	PPIS2	P	VBZ	V	YLIP	UNK
DDQGE	D	NNO	N	PPX1	P	VD0	V	YQUE	UNK
DDQV	D	NNO2	N	PPX2	P	VDD	V	YQUO	UNK
EX	X	NNT1	N	PPY	P	VDG	V	YSCOL	UNK
FO	X	NNT2	N	RA	X	VDI	V	YSTP	UNK
FU	X	NNU	N	REX	X	VDN	V	ZZ1	UNK
FW	X	NNU1	N	RG	X	VDZ	V	ZZ2	UNK
GE	X	NNU2	N	RGQ	X	VH0	V	*all other*	PREP
IF	PREP	NP	N	RGQV	X	VHD	V		

Table B.1: Mapping table for types. 5x2 columns. As indicated in the last field, every tag not found in this table has been mapped to PREP

B.2 Multiword Expression Taxonomy

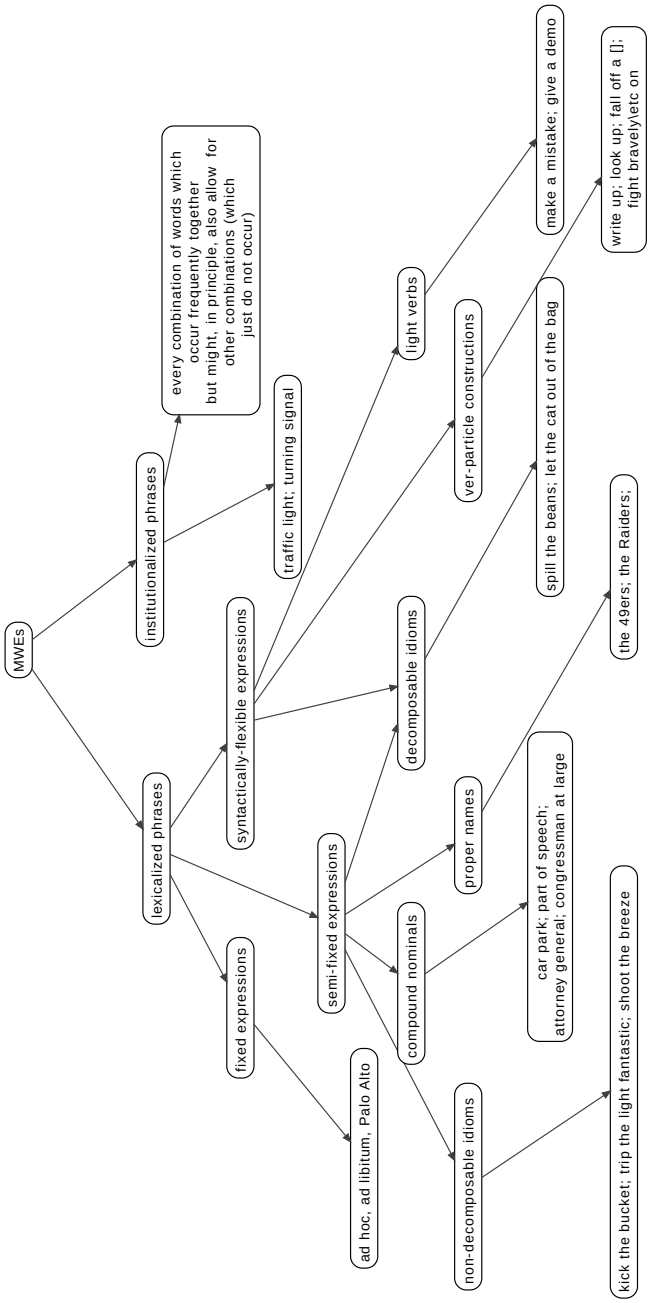

Figure B.1: A taxonomy of multiword expressions, extracted from Sag et al. (2002)