Use of Terms and Term-Related Units as Feature Sets for Automatic Text Classification

Alex Chengyu Fang[1] and Jing Cao[2]

Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong SAR
[1]acfang@cityu.edu.hk
[2]cjing3@student.cityu.edu.hk

Abstract. The current study investigates how terminologically-informed features would contribute to automatic text classification. In particular, we examine the use of terms and term-related units as feature sets in different classification tasks. A sub-corpus of 80 texts was created out of the British component of the International Corpus of English. Three classification tasks were determined according to subject domains, registers and text categories. The performance of the selected feature sets was evaluated in terms of F-score through machine learning techniques. Such performance was also compared with that of conventional lexical and grammatical feature sets. Although it is a comparatively small corpus, the empirical results show that while features determined according to the lexical criterion have a consistent performance, the use of terms produced superior classification performance when classifying texts according to subject domains.

Keywords: Automatic text classification, feature generation, terms, machine learning.

1 Introduction

Substantial research has been carried out in the generation and selection of discriminatory features for better automatic text classification (ATC) performance. Features have been generated according to lexical, grammatical and knowledge-based criteria. The conventional bag-of-words (BOW) approach has often employed as the baseline and has shown surprisingly good performance [1]. Part-of-speech (POS) based features have also been examined as a useful supplement to existing ATC technologies. For instance, [2] tested eight feature sets, including nouns, nouns + adjectives, nouns + adjectives + proper names, nouns + proper names, adjectives + proper names, verbs, nouns + verbs and nouns + verbs + adjectives. The results show that nouns tend to be most influential, while verbs the least. [3] showed that it was possible to use the particular sets of adjectives and adverbs to classify genres. In particular, speaker-oriented adverbs are found to be more effective than trait adjectives and adverbs. [4] examined adjective use and the results show that

adjectives can be an effective indicator for text categories and that different domains tend to have their own preference for the use of adjectives. In addition, knowledge-based features such as ontology have started to attract more and more attention. Encouraging results have also been found and yet past studies have been confined to limited genres or domains. For example, [5] investigated the news articles from Reuters RCV1 using ontology created from Wikipedia. [6] focused on the gene ontology to tackle the text categorisation in the field of biology. [7] proposed a bag-of-concepts model to classify news articles from Reuters-21578 and 20 Newsgroups, and medical abstracts from MEDLINE as well.

Motivated by previous studies, the work to be reported in this article aims at an evaluation of terminologically-informed features including terms and term-related units in ATC. In addition, classification tasks were designed to cover different registers, genres and text categories. A series of experiments were designed with the least interference of statistical techniques. Our purpose was not to seek the best classification performance but to see how the use of terminologically-informed features would contribute to ATC.

A sub-corpus of 80 texts was created out of the British component of the International Corpus of English (ICE-GB; [8]). Three different classification tasks were identified for the same set of texts: textual classification according to corpus categories, stylistic classification according to registers such as learned and popular, and topical classification according to subject domains. The performance of terms and term-related units was evaluated through machine learning techniques. Meanwhile BOW feature set was tested as the baseline and POS-based feature sets as a comparison. Although a relatively small number of texts were used in the experiments, the empirical results show that while features determined according to the lexical criterion have a consistent performance, the use of terminologies produced superior classification performance when classifying texts according to subject domains.

The rest of the paper is orgainsed as follows. Section 2 is a description of the methodology and explains the experimental setup, the corpus and machine learning tools. Section 3 describes the creation of a subcorpus arranged according to the three classification tasks and then explains the generation and extraction of the feature sets. Section 4 presents the results before Section 5, which draws some preliminary conclusions and suggests some future research.

2 Methodology

In this section we will first explain the experimental setup, describe the corpus and then briefly introduce the machine learning tools.

2.1 Experimental Setup

A goal of the series of experiments that we designed was to investigate the performance of terminological feature sets in three different classification tasks. The

TC tasks were identified as follows: 1) textual classification into categories, 2) stylistic classification of the texts into registers, and 3) topical classification according to subject domains. As mentioned earlier, the performance of the proposed feature sets was also compared with conventional lexical and grammatical feature sets. To be more specific, we were interested in the following feature sets: 1) terms and term-related units, 2) the bag of words (BOW) through word unigrams, and 3) POS-based features. All the performance results were evaluated in terms of precision, recall and F-score (F_1).

2.2 Corpus

A newly built corpus [9], based on the British component of the International Corpus of English (ICE-GB; [8]), was used for the experiments. The corpus was originally used to explore the syntactic characteristics of terminological expressions across different text types and subject domains in contemporary English. In the current study, the annotated terms and term-related units in the corpus were employed as feature sets to test possible application in the field of text classification.

Table 1 presents the corpus composition. As can be seen in Table 1, the new corpus comprises 80 texts and each component text has about 2,000 word tokens. There are four parallel subject domains (humanities, social sciences, natural sciences and technology) according to two registers (learned and popular).

Table 1. The structure of corpus

Register	Subject Domain	# of Texts	# of Tokens
Learned	Humanities	10	21,467
	Social Sciences	10	21,527
	Natural Sciences	10	21,484
	Technology	10	21,282
Popular	Humanities	10	23,700
	Social Sciences	10	20,955
	Natural Sciences	10	20,803
	Technology	10	21,143
Total		80	172,361

A particularly useful feature of the corpus is that it is richly annotated at lexical, grammatical, syntactic, and terminological levels. Consider (a) as an example:

(a) *The fibres of group B are found in the autonomic nervous system.*

The same example is represented in the corpus according to a formalism exemplified in Figure 1.

```
PU CL(main,montr,pass,pres)
 SU NP(term)
  DT DTP()
   DTCE ART(def) {The}
  NPHD N(com,plu) {<t>fibres</t>}
  NPPO PP()
   P PREP(ge) {of}
   PC NP()
    NPHD N(com,sing) {group B}
 VB VP(montr,pres,pass)
  OP AUX(pass,pres) {are}
  MVB V(montr,edp) {found}
 A PP()
  P PREP(ge) {in}
  PC NP(term)
   DT DTP()
    DTCE ART(def) {the}
   NPPR AJP(attru)
    AJHD ADJ(ge) {<t>autonomic}
   NPPR AJP(attru)
    AJHD ADJ(ge) {nervous}
   NPHD N(com,sing) {system</t>}
     PUNC PUNC(per) {.}
```

Fig. 1. An example of multi-layer annotations of (a)

As illustrated above, the tree structure for (a) is annotated with grammatical, syntactic and terminological information. At the grammatical level, words are coded with part-of-speech (POS) tags that include a head tag (such as nouns, verbs, and adjectives) with a set of attributes indicating the sub-categorizations of the head tag. For instance, the verb `found` enclosed within a pair of curly brackets is tagged as `V(montr,edp)`, namely, a mono-transitive verb in past participial form. As another example, `{The}` is assigned a label `ART(def)`, meaning it is a definite article, and `{fibres}` is a common noun in its plural form. Syntactically, each node comprises two labels: one representing its syntactic category (such as noun phrase and adjective phrase) and the other the syntactic function. Take the node `SU NP()` as an example, which indicates that it is a noun phrase (`NP`) functioning as the subject (`SU`) of the clause. The same NP comprises a determiner (`DT`), the head (`NPHD`) and a post-modifier (`NPPO`). The definite article `The` constitutes the central determiner (`DTCE`), a daughter node of `DT`. At the terminological level, terms are marked with '`<t>`' at the beginning and with '`</t>`' at the end in the tree diagram, and the resulting NP is described by an additional attribute '`term`'. For example, there are two NPs that are marked as terms (i.e. `fibres` and `autonomic nervous system`). In the current study, terms primarily correspond to noun-phrase (NP) groups and consist of words that are single nouns or complex noun phrases. Following [10], we also consider terms in a pragmatic sense. Take text `w2a-031` for example. The text is about "blind shaft drilling" under the domain of *technology*. In addition to terms in technology and engineering, we may also mark up terminological entities from related domains such as *environment*. See [9] for a more detailed description of, especially, the annotation of terminologies.

2.3 Machine Learning Tools

Weka [11], a general purpose machine learning software package, was employed to estimate classification performance in terms of weighted average precision, recall and F-score (F_1). All the primary results to be reported were obtained using Naïve Bayes (NB) classifier, and support vector machines (SVM) classifier (i.e. LibSVM in Weka) was also used for a complementary evaluation of the chosen feature sets. Considering data size, 10-fold cross validation was used to calculate the results.

3 Pre-processing and Feature Extraction

3.1 Text Organisation

Given the three classification tasks, 80 texts in the corpus were arranged accordingly into three settings for textual, stylistic and topical classification tasks respectively:

Task 1: Textual Classification. By textual classification, it is meant that the texts are to be classified according to the 8 original ICE categories. Table 2 describes the categories, together with number of texts and total word tokens in each category.

Table 2. The structure of the dataset for textual classification

Text Code	Explanation	# of Texts	# of Tokens
LHUM	humanities in learned writing	10	21,467
LSOC	social sciences in learned writing	10	21,527
LNAT	natural sciences in learned writing	10	21,484
LTEC	technology in learned writing	10	21,282
PHUM	humanities in popular writing	10	23,700
PSOC	social sciences in popular writing	10	20,955
PNAT	natural sciences in popular writing	10	20,803
PTEC	technology in popular writing	10	21,143
Total		80	172,361

Task 2: Stylistic Classification. In the current study we were also interested in examining whether selected features can classify texts according to registers, and therefore texts in the corpus were grouped into two registers for stylistic classification: learned vs. popular settings (see Table 3).

Table 3. The structure of the dataset for stylistic classification

Registers	# of Texts	# of Tokens
Learned	40	85,760
Popular	40	86,601
Total	80	172,361

Task 3: Topical Classification. For topical classification, we were concerned with the classification according to subject domains and hence the texts were regrouped into 4 subject domains regardless of their registers. See Table 4 for the basic statistics of this dataset.

Table 4. The structure of the dataset for topical classification

Topics	# of Texts	# of Tokens
Humanities	20	45,167
Social sciences	20	42,482
Natural sciences	20	42,287
Technology	20	42,425
Total	80	172,361

3.2 Feature Generation and Extraction

As mentioned early, this work investigates the use of terminologically-informed feature sets. To be more specific, we examined the use of noun phrases marked as terms. In addition, term adjectives (i.e. adjectives used in terminological expressions) were investigated as a term-related feature set incorporating both the knowledge-based and linguistically informed strategies. Meanwhile, for each TC task, we employed the conventional BOW approach as the baseline experiment, which is a list of word types (*bow*) filtered without functional items. We also examined POS features as a comparison, and the focus was on the open classes, namely, nouns, verbs, adjectives, and adverbs. In all, seven feature sets were generated and exploited in the experiments:

- *term* (terms)
- *term-adj* (adjectives occurring in terms)
- *bow* (bag-of-words without functional items)
- *n* (nouns)
- *v* (verbs)
- *adj* (adjectives)
- *adv* (adverbs)

The extraction of different feature sets is quite straightforward, thanks to the multi-layer annotations of the corpus. As can be seen in Figure 1, terms were extracted by identifying the '<t>' at the beginning and with '</t>' at the end for each term. With the help of the POS annotations, adjectives that occur in multi-word terms were extracted. For the bag of words, all the words in the curly brackets were extracted and functional items were filtered according to their POS tags. For POS features, words with POS tags of the four open classes were extracted accordingly.

4 Results

In this section we report the results of three TC tasks in our experimental study. All of the classification results were derived from the presence of the chosen features though feature frequency was also tested in our experiments, which generally showed a poorer performance, therefore omitted from the report.

The performance of the chosen feature sets were evaluated using F-score (F_1). As noted earlier on, all the primary results obtained with Naïve Bayes (NB) classifier are reported first (i.e. Section 4.1 – 4.3), followed by a complementary evaluation using the LibSVM classifier (i.e. Section 4.4).

4.1 Textual Classification

Figure 2 illustrates the learning curve of the 7 features sets in textual classification with the increase of training data size, from 10% to 100%. It is noticeable that discriminatory attributes are unevenly distributed across the document texts. Take the performance of *bow* for example: With 10% of the training data, the generated feature set achieved an accuracy of about 20%; with 50% of the training texts, the accuracy of the generated feature set reached 40%, and with all of the training data, the ultimate accuracy reached over 60%.

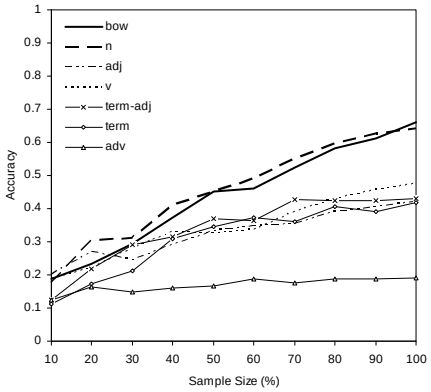

Fig. 2. Learning curve in textual classification

Table 5. Weighted average accuracy for textual classification

Feature Set	Precision	Recall	F_1
bow	0.775	0.750	0.750
n	0.739	0.713	0.712
adj	0.662	0.575	0.583
v	0.544	0.525	0.521
term-adj	0.692	0.525	0.501
term	0.447	0.525	0.437
adv	0.239	0.225	0.215

Table 5 summarises the performance of the 7 feature sets in textual classification task in terms of weighted average precision, recall and F-score (F_1). Feature sets are arranged according to F-score in descending order. As is shown, *bow* performed the best, followed by nouns and adjectives. Term-adjectives outperformed terms, although the difference is not significant. Adverbs turn out to be the weakest feature set. All differences in F-scores are statistically significant ($p < 0.005$), with the exception of the difference between three pairs –adjectives vs. verbs, term-adjectives vs. verbs, and terms vs. term-adjectives.

4.2 Stylistic Classification

Figure 3 illustrates the learning curve of the same 7 features sets in stylistic classification with the increased training data size, from 10% to 100%. Take verbs for example. With the first 10% of the training texts, the accuracy was about 50%; with 50% of the training texts, the accuracy reached over 70%; and with all of the training data, the ultimate accuracy reached over 80%. As a whole, the classification accuracies tend to be comparatively higher in general than those in the textual classification task.

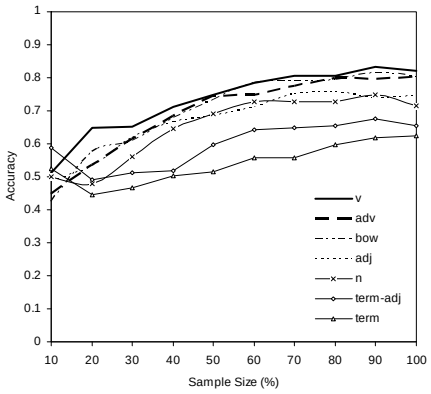

Fig. 3. Learning curve in stylistic classification

Table 6. Weighted average accuracy for stylistic classification

Feature Set	Precision	Recall	F_1
v	0.826	0.825	0.825
adv	0.825	0.825	0.825
bow	0.789	0.788	0.787
adj	0.776	0.775	0.775
n	0.730	0.725	0.723
term-adj	0.652	0.650	0.649
term	0.625	0.625	0.625

With regard to stylistic classification, the average performance of the 7 feature sets is presented in Table 6 according to F-score in descending order. All differences in F-scores are statistically significant ($p < 0.005$), with the exception of the difference between nouns vs. adjectives and terms vs. term-adjectives. It is noticeable that verbs and adverbs obtain the same F-score (0.825) and their performance is surprisingly higher than other feature sets, suggesting that these two word classes are perhaps more indicative of stylistic differences between texts. It is also observable that terms and term adjectives turned out to be the weakest features in this task, although their performance is better than that in textual classification task.

4.3 Topical Classification

Figure 4 illustrates the learning curve of the 7 features sets in topical classification with the increased training data size. Compared with the previous two tasks, the overall performance in this task seems to cover a wider range of accuracy, ranging from 10% to 90%, which indicates that the selected feature sets tend to have more distinctive properties in terms of subject domains or topics. It is also can be observed that terms seem to have the best performance. In particular, with all of the training data, the ultimate accuracy of terms reached over 90%, which can be considered satisfactory results, higher than 86.1% accuracy in [5] or 84.09% in [12].

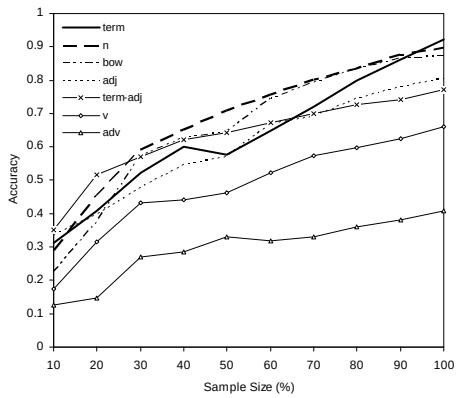

Fig. 4. Learning curve in topical classification

Table 7. Weighted average accuracy for topical classification

Feature Set	Precision	Recall	F_1
term	0.941	0.938	0.938
n	0.924	0.925	0.924
bow	0.865	0.863	0.861
adj	0.858	0.838	0.837
term-adj	0.811	0.800	0.803
v	0.722	0.713	0.714
adv	0.431	0.413	0.416

For topical classification, the average performance of each feature set is summarised in Table 7. Three interesting observations emerge. First, terms, which have been conventionally regarded as conceptual units, performed the best in the classification of texts according to subject domains, suggesting that conceptual units such as terms are perhaps the most suitable discriminatory feature for the purposes that concern us here. This observation also suggests that, with texts that have already been classified according to domains, it is feasible to extract a list of terms that can be held indicative of a particular domain, an interesting research thread that we plan to pursue in our ongoing research to automatically build terminological ontologies. Secondly, nouns are also highly indicative of subject domains, achieving a higher accuracy than *bow*. However, their difference is not statistically significant while the difference between terms and nouns is, suggesting that, compared with features selected according to lexical and grammatical strategies, terms have a unique and superior contribution in the classification of texts according to subject domains. Finally, adjectives and term adjectives (with $F_1 > 0.80$) have turned out to be POS-based and terminologically-informed features that are highly correlated with the content matter while verbs and adverbs are comparatively less so. All differences in F-scores are statistically significant ($p < 0.005$), with the exception of the difference between nouns and *bow*.

4.4 Results from SVM Classifier

A complementary evaluation of the chosen feature sets were performed by using support vector machines (SVM) classifier (i.e. LibSVM in Weka). Tables 8, 9 and 10 summarise the weighted average precision, recall and F-score (F_1) for textual, stylistic and topical classification respectively. The 7 feature sets are arranged according to F-score in descending order.

Table 8. Weighted average accuracy for textual classification

Feature Set	Precision	Recall	F_1
bow	0.762	0.675	0.653
v	0.607	0.538	0.514
n	0.677	0.538	0.489
adj	0.591	0.450	0.429
term	0.524	0.450	0.385
term-adj	0.487	0.388	0.306
adv	0.339	0.288	0.292

Table 9. Weighted average accuracy for stylistic classification

Feature Set	Precision	Recall	F_1
adj	0.753	0.750	0.749
adv	0.750	0.738	0.734
v	0.790	0.738	0.725
bow	0.753	0.638	0.591
term-adj	0.708	0.575	0.494
term	0.696	0.563	0.473
n	0.660	0.538	0.428

Table 10. Weighted average accuracy for topical classification

Feature Set	Precision	Recall	F_1
term-adj	0.818	0.813	0.813
term	0.843	0.763	0.756
bow	0.844	0.750	0.702
v	0.751	0.738	0.733
adj	0.806	0.750	0.722
n	0.815	0.713	0.640
adv	0.390	0.438	0.409

As can be noted in the above tables, the weighted average F-scores obtained from LibSVM classifier turned out to be comparatively lower than those from NB classifier. It is also worth noticing that nouns performed surprisingly worse than their performance in NB classifier. Nevertheless, when compared with the results obtained from NB classifier, most of the chosen feature sets achieved similar performance pattern in all the three classification tasks except nouns. In the task of textual classification (see Table 8), *bow* performed the best, followed by verbs and nouns, and adverbs are again the weakest feature. As for the stylistic classification (see Table 9), adverbs and verbs achieved a quite high accuracy, higher than *bow*, again suggesting that these two word classes are perhaps more indicative of stylistic differences between texts. Although adjectives are observed to have a slightly higher accuracy than adverbs, the difference is not statistically significant. In the last classification task, knowledge-based feature sets (i.e. *term-adj* and *term*) performed the best and adverbs were the weakest feature.

5 Conclusion

In this paper, we described a series of experiments designed to investigate the use of terms and term-related units as feature sets in different ATC tasks. In particular, seven feature sets were generated according to terminological, lexical, and grammatical strategies. To cover different registers, genres and text categories, three different classification tasks were identified, including textual, stylistic and topical classifications. 80 texts were selected from a corpus richly annotated at lexical, grammatical, syntactic, and terminological levels and arranged variously to represent classes according to subject domains, registers and text categories. Naïve Bayes and LibSVM classifiers in Weka were chosen to estimate classification accuracy of the feature sets.

Results show that features selected according to the lexical criterion (i.e. *bow*) have a generally consistent performance, which is in line with past studies such as [1]. Our results also show that the performance of features selected according to different strategies varied depending on classification tasks. While BOW feature sets had a consistently good performance, terminologically and linguistically informed features yielded competing or better performance given different classification tasks. Verbs and adverbs form feature sets that demonstrate competing performance against *bow* when given the stylistic classification task. More importantly, the use of terms produced satisfactory classification performance when classifying texts according to subject domains.

Due to the need to classify the same group of texts in three different ways for the experiments, only a relatively small number of texts were used to produce the results. Our immediate future work will focus on the verification of the findings through the use of texts from other sources and of a larger number.

Acknowledgments
Research described in this article was supported in part by grants received from City University of Hong Kong (Project Nos 9610126, 7008002, 7002387 and 7002190). The authors also acknowledge the generous support received from the members of the Dialogue Systems Group at the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

References

1. Forman, G.: Feature Selection for Text Classification. In: Liu, H., Motoda, H. (eds) Computational Methods of Feature Selection, pp. 257--276. CRC Press/Taylor and Francis Group (2008)
2. Silva, C., Viera, R., Osorio, F. S., Quaresma, P.: Mining Linguistically Interpreted Texts. In: Proceedings of the Fifth International Workshop on Linguistically Interpreted Corpora. Geneva, Switzerland (2004)
3. Rittman, R.: Automatic Discrimination of Genres: The Role of Adjectives and Adverbs as Suggested by Linguistics and Psychology. VDM Verlag (2008)

4. Fang, A. C., Cao, J.: Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A Study Based on the British National Corpus. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation. Hong Kong (2009)
5. Janik, M., Kochut, K. J.: Wikipedia in Action: Ontological Konwledge in Text Categorization. In: Proceedings of the IEEE International Conference on Semantic Computing, pp. 268--275 (2008)
6. Seki, K., Mostafa, J.: Gene Ontology Annotation as Text Categorization: An Empirical Study. Information Processing and Management, 44(5), pp. 1754--1770 (2008)
7. Wang, X., Bai, R.: Applying RDF Ontologies to Improve Text Classification. International Conference on Computational Intelligence and Natural Computing, vol. 2, pp.118--121 (2009)
8. Greenbaum, S.: Comparing English World Wide: The International Corpus of English. Oxford: Oxford University Press (1996)
9. Fang, A. C., Cao, J., Song, Y.: A New Corpus Resource for Studies in the Syntactic Characteristics of Terminologies in Contemporary English. In: Proceedings of the 8th International Conference on Terminology and Artificial Intelligence. Toulouse, France (2009)
10. Eumeridou, E., Nkwenti-Azeh, B., McNaught, J.: An Analysis of Verb Subcategorization Frames in Three Special Language Corpora with View towards Automatic Term Recognition. Computers and the Humanities, 38, p. 37--60 (2004)
11. Witten, I. H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kauf-mann, San Francisco (2005)
12. Gu, H., Zhou, K.: Text Classification Based on Domain Ontology. Journal of Communication and Computer, 3 (5), 29--32 (2006)