# Enhanced Genre Classification through Linguistically Fine-Grained POS Tags∗

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.** We propose the use of fine-grained part-of-speech (POS) tags as discriminatory attributes for automatic genre classification and report empirical results from an experiment that indicate substantial accuracy gain by such features over the conventional bag-of-words approach through word unigrams. In particular, this paper reports our research to investigate the performance of a fine-grained tag set when tested with the British component of the International Corpus of English. Ten different genre classification tasks were identified and the performance of the tags was evaluated in terms of F-score. Our results show that the use of linguistically fine-grained POS tags produces superior accuracy when compared with word unigrams, particularly for a rich set of 32 different genres with Naïve Bayes Multinominal Classifier. Through a comparison with an impoverished tag set, our results further demonstrate that the superior performance is due to the rich linguistic information embodied in the 400-strong different POS tags.

**Keywords:** automatic genre classification, ICE-GB, fine-grained POS tag, linguistic granularity, AUTASYS.

## 1 Introduction

Text classification has been conventionally based on content matters and sentiment polarities. There are situations where genre classification is required for the identification of, for example, formal and informal sources of information. Genre classification of text is a process of classifying texts or documents according to the criterion of genre, such as style, form, or purpose, based on the assumption that "a document can be represented by the values of features that seem to express the attribute of a genre" (Lim *et al.* 2005:1264). Part-of-speech (POS) tags have been employed in automatic genre classification in that they do not "reflect the topic of the document, but rather the type of text used in the document" (Finn and Kushmerick, 2003) and that their distribution has been observed to vary across different genres (e.g. Nakamura, 1993; Rayson et al., 2002). Nevertheless, a majority of past studies have included POS tags with other features to form a combined feature set. For example, Karlgren and Cutting (1994) included 6 POS tags (i.e. *adverb, preposition, 2nd person pronoun, 1st person pronoun, noun* and *present verb*) in classifying genres of the Brown Corpus. They carried out the classification tasks in terms of 2, 4 and 15 genre classes according to Brown categories. The combined feature set achieved an accuracy of 96%, 73% and 52% in the three classification tasks respectively. Dewdney et al. (2001) included POS tags of content words (i.e. *noun, verb, adjective* and *adverb*), where verbs were further defined in past, present and future tenses. Again, with a

∗ The work described in this paper was supported partially by research grants (Nos 9610126, 7008002, 7002387 and 7002190) from City University of Hong Kong. The authors would like to thank the reviewers for their valuable comments and suggestions.

combined feature set, the performance of classifying 7 genre classes reached 92%. Eissen and Stein (2004) included 10 POS tags (i.e. *noun, verb, relative pronouns, relative preposition, adverb, article, pronoun, modals, adjective* and *alphanumeric words*) in classifying 8 genre classes. The performance of the combined feature set was 70%. Some other studies have not specified the POS tags, while they do report the performance using a combined feature set. For instance, Boese and Howe (2005) reported an accuracy of 79.6% when classifying 5 genre classes, and an accuracy of 74.8% for 7 genre classes. Lim *et al.* (2005) reported a much lower performance of about 38%. Still, some studies have treated POS tags as independent feature set for automatic genre classification. For example, Finn and Kushmerick (2003) used 36 POS features in subjectivity classification (3 genre classes) and review classification (2 genre classes), and achieved 84.7% and 61.3% accuracy respectively. More recently, Stein and Eissen (2008) used 10 POS tags to classify 8 genre classes and reported an accuracy of 74%. Santini (2004) further computed POS tags into unigram, bigram and trigram. When classifying 10 genre classes, POS trigram achieved the best performance with 82.6% accuracy, compared with 77.6% for bigram and 77.3% for unigram. The study also investigated 4 spoken and 6 written genre classes, and POS trigram again performed the best. To sum up, past studies have shown encouraging and suggestive results of using POS tags in genre classification, and yet there are some limitations. For example, it is difficult to evaluate whether POS tags are discriminatory features for a given classification task when they are included in a complex feature set. Limited studies have regarded POS tags as independent feature set. It is also noticeable that the number of genre classes is comparatively small.

The current study introduces a new set of linguistically fine-grained POS tags generated by AUTASYS (Fang, 1996 and 2007) for automatic genre classification. We will report in this paper an experiment designed to investigate the impact of the proposed feature set when compared and contrasted with word unigrams as a bag of words (BOW) and an impoverished POS tag set. Machine learning tools were used to evaluate the classification performance in terms of F-score. The British component of the International Corpus of English (ICE-GB; Greenbaum, 1996) was employed as a resource of different text genres. Ten different genre classification tasks were identified based on the existing ICE-GB categories, which are grouped according to different granularities. As our results will show, the use of linguistically rich POS tags as discriminative features produces superior accuracy when compared with BOW for fine-grained genre classification. Our results will further demonstrate that the superior performance is due to the rich linguistic information since an impoverished tag set yielded worse classification results.

The rest of the paper is organised as follows. Section 2 is a description of the methodology, covering the experimental setup, the genre resource, and machine learning tools. Section 3 explains the feature sets including the proposed linguistically fined-grained POS tags, bag of words and impoverished POS tags. Section 4 presents and discusses the experiment results from ten different genre classification tasks. Finally, section 5 draws some preliminary conclusions and suggests some future research.

## 2 Methodology

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

### 2.1 Experimental Setup

A goal of the experiment that we designed was to investigate the performance of a set of linguistically fine-grained POS tags for various levels of genre classification tasks. Currently, we are more interested in verifying the contribution of such a feature set in the classification task than ascertaining the comparative performance of different feature selection methods. The bag-of-words (BOW) approach were used to generate the baseline statistics, which has been commonly used in past studies (e.g. Scott and Matwin, 1999; Diederich *et al.* 2003; Koster and

Seutter, 2003; Gupta and Ratinov, 2008; Li *et al.* 2009). Besides, an impoverished POS tag set was also examined for indication of effect of linguistic granularity on classification performance. All the performance results were evaluated according to F-score, which is defined as:

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

A series of genre classification tasks were identified based on the division of corpus in terms of different genre granularities, and also on the division of each granularity into speech vs. writing.

## 2.2 Corpus

Given the purpose of investigating genre attribute performance, the British component of the International Corpus of English (ICE-GB; Greenbaum, 1996) was employed as the genre resource. See Table 1 for the composition of the ICE-GB, where the numbers indicate the number of texts of about 2,000 word tokens each. Altogether, there are 500 component texts, with 300 for speech and 200 for writing.

**Table 1:** The composition of ICE-GB

| **Speech** | | | | **Writing** | | | |
|---|---|---|---|---|---|---|---|
| **Dialogue** | *Private* | | | **Non-Printed** | *Student Writing* | | |
| | S1A1 | Direct conversations | 90 | | W1A1 | Untimed essays | 10 |
| | S1A2 | Distanced conversations | 10 | | W1A2 | Timed essays | 10 |
| | *Public* | | | | *Correspondence* | | |
| | S1B1 | Class lessons | 20 | | W1B1 | Social letters | 15 |
| | S1B2 | Broadcast discussions | 20 | | W1B2 | Business letters | 15 |
| | S1B3 | Broadcast interviews | 10 | **Printed** | *Informational* | | |
| | S1B4 | Parliamentary debates | 10 | | W2A1 | Learned: humanities | 10 |
| | S1B5 | Legal cross-examinations | 10 | | W2A2 | Learned: social sciences | 10 |
| | S1B6 | Business transactions | 10 | | W2A3 | Learned: natural sciences | 10 |
| **Monologue** | *Unscripted* | | | | W2A4 | Learned: technology | 10 |
| | S2A1 | Spontaneous commentaries | 20 | | W2B1 | Popular: humanities | 10 |
| | S2A2 | Unscripted speeches | 30 | | W2B2 | Popular: social sciences | 10 |
| | S2A3 | Demonstrations | 10 | | W2B3 | Popular: natural sciences | 10 |
| | S2A4 | Legal presentations | 10 | | W2B4 | Popular: technology | 10 |
| | *Mixed* | | | | W2C1 | Press news reports | 20 |
| | S2B1 | Broadcast news | 20 | | *Instructional* | | |
| | *Scripted* | | | | W2D1 | Administrative writing | 10 |
| | S2B2 | Broadcast talks | 20 | | W2D2 | Skills and hobbies | 10 |
| | S2B3 | Non-broadcast talks | 10 | | *Persuasive* | | |
| | | | | | W2E1 | Press editorials | 10 |
| | | | | | *Creative* | | |
| | | | | | W2F1 | Fiction | 20 |

Based on the ICE-GB categories, four genre levels were identified according to granularity, namely, super, macro, micro and sub-micro. See David (2001) and Boese and Howe (2005) for a similar division of genre granularity. Table 2 is a summery of the four-level granularity of ICE-GB. The numbers within brackets indicate the number of genre classes at each level.

**Table 2:** Four levels of genre classes

| Super (2) | Macro (4) | Micro (11) | Sub-micro (32) |
|---|---|---|---|
| Speech | Dialogue | Private | direct conversation, distanced conversation |
| | | Public | class lessons, broadcast discussions, broadcast interviews, parliamentary debates, legal cross-examinations, business transaction |
| | Monologue | Unscripted | spontaneous commentaries, unscripted speeches, demonstrations, legal presentations |
| | | Mixed | broadcast news |
| | | Scripted | broadcast talks, non-broadcast talks |
| Writing | Non-printed | Student Writing | untimed essays, timed essays |
| | | Correspondence | social letters, business letters |
| | Printed | Informational | learned humanities, learned social sciences, learned natural sciences, learned technology, popular humanities, popular social sciences, popular nature sciences, popular technology, press news reports |
| | | Instructional | administrative writing, skills and hobbies |
| | | Persuasive | press editorials |
| | | Creative | fiction |

As can be seen in Table 2, the genre system of ICE-GB can be seen as a systemic hierarchy, with each level commanding a number of sub-divisions. For example, the super genre *Speech* has 2 macro genres (*Dialogue* and *Monologue*), which in turn command 5 micro genres (such as *Private* and *Public*) to be divided into 15 sub-micro classes such as *direct conversation* and *class lessons*.

## 2.3 Machine Learning Tools

Weka (Witten and Frank, 2005), a general purpose machine learning software package, was employed to estimate classification performance in terms of average weighted F-score. Naïve Bayes Classifier (NB) was used to evaluate the present or absent property of features, while Naïve Bayes Multinominal Classifier (NB-MN) was used to evaluate the frequency of features. Considering data size, 10-fold cross validation was used to calculate the results.

# 3 Feature Sets

## 3.1 Fine-Grained POS Tags (F-POS)

We propose the use of linguistically fine-grained part-of-speech tags (F-POS) as a feature set for automatic genre classification. The proposed F-POS tags are produced by a probabilistic tagger named AUTASYS (Fang, 1996 and 2007) according to a tag-feature hierarchy that comprises a head tag indicating general classes such as nouns and verbs augmented with a subcategorisation feature such as common nouns and monotransitive verbs. Often the tag also includes an additional feature indicating the grammatical status, such as singular common nouns and present-tense monotransitive verbs. Consider (a) as an example:

(a) *The workshop was held to collect current data on the related laboratory investigations.*

When tagged by AUTASYS, (a) is represented as:

```
The <tag ART(def)> workshop <tag N(com,sing)> was <tag AUX(pass,past)>
held <tag V(montr,edp)> to <tag PRTCL(to)> collect <tag
V(montr,infin)> current <tag ADJ(ge)> data <tag N(com,sing)> on <tag
PREP(ge)> the <tag ART(def)> related <tag ADJ(edp)> laboratory <tag
N(com,plu):1/2> investigations <tag N(com,plu):2/2> . <tag PUNC(per)>
<#000000>
```

As illustrated above, the tag-feature hierarchy for different part-of-speech in (a) can be analyzed as:

| Word | Head Tag | Subcategory | Additional feature | Meaning |
|---|---|---|---|---|
| the | ART | def | n.a. | article, definite |
| workshop | N | com | sing | noun, common, singular |
| was | AUX | pass | past | auxiliary, passive, past tense |
| held | V | montr | edp | verb, monotransitive, *-ed* participle |
| to | PRTCL | to | n.a. | particle *to* |
| related | ADJ | edp | n.a. | adjective, *-ed* participle |
| on | PREP | ge | n.a. | preposition, general |

As a result, the pre-processing of the grammatical annotation extracted 487 different types of POS tags for the whole corpus, with 449 for spoken genres and 319 for written genres.

### 3.2 BOW

A bag of words (BOW) through word unigrams were tested as the baseline experiment. In the current study, the BOW has been filtered with a stoplist of functional items, and the orthographical word forms are retained without lemmatization. A total of 35,758 word types were found for the whole corpus and subsequently used as BOW attributes, with 21,198 for spoken genres and 27,305 for written genres.

### 3.3 Impoverished Tags (I-POS)

The third feature set was generated from F-POS but contains only the head tags without the subcategorisation features and hence linguistically impoverished. Again take the seven words in (a) for example.

| Word | Head Tag |
|---|---|
| the | ART |
| workshop | N |
| was | AUX |
| held | V |
| to | PRTCL |
| related | ADJ |
| on | PREP |

I-POS was used in the experiment in order to ascertain the effect of grammatical granularity on classification performance. As a result, there were altogether 36 I-POS attributes for the total corpus, 36 for spoken genres, and 27 for written genres.

## 4 Experiment Results

In this section we report the results of a series of genre classification tasks in our experimental study. As noted earlier on, all results were obtained from two Naïve Bayes Classifiers (i.e. NB and NB-MN) in Weka and presented as average weighted F-scores. The first sub-section will be devoted to the classification results based on the presence of the selected features. The second part of this section will present the results obtained according to feature frequency, followed by the discussion section.

## 4.1 Results Obtained from NB Classifier

As mentioned earlier, Naïve Bayes Classifier was used to evaluate the three feature sets according to presence or absence of genre attributes. Table 3 summarises the performance of the 3 feature sets in genre classification in terms of average weighted F-score. The first column lists the four levels of genres. The second column shows 10 genre classification tasks, where *S* stands for speech, *W* stands for writing, and the number indicates the number of genre classes in a given classification task.

Several interesting patterns can be observed in Table 3. First of all, there tends to be a continual drop in accuracy with the increase in number of classes in general. Take F-POS for example. The F-score of F-POS in *SW* classification tasks starts from 0.998 in *SW-2* and then decreases to 0.842 in *SW-4*, 0.747 in *SW-11* and finally drops to 0.582 in *SW-32*. Secondly, genre classification tasks regarding spoken texts generally receive better results than those of written texts. This is perhaps due to those F-POS tags that are specific to speech only. One example is `REACT` for 'reaction signal' such as *um*, *yeah* and *wow*, which practically occur exclusively in transcribed speech. Thirdly, F-POS achieves better performance than BOW in 6 classification tasks, and yields a competing performance in 2 tasks (i.e. *SW-4* and *W-17*) where the difference is not statistically significant. Finally, F-POS performs better than I-POS in almost all of the 10 classification tasks, indicating that fine-grained POS tags with rich linguistic information can better represent text genres than simple POS tags.

**Table 3:** Average weighted F-score (NB)

| Genre Granularity | Code | BOW | F-POS | I-POS |
|---|---|---|---|---|
| Super Genre | SW-2 | 0.871 | 0.998 | 0.998 |
| Macro Genre | S-2 | 0.885 | 0.917 | 0.858 |
| | W-2 | 0.886 | 0.742 | 0.704 |
| | SW-4 | 0.855 | 0.842 | 0.798 |
| Micro Genre | S-5 | 0.802 | 0.749 | 0.566 |
| | W-6 | 0.709 | 0.769 | 0.513 |
| | SW-11 | 0.746 | 0.747 | 0.549 |
| Sub-micro Genre | S-15 | 0.561 | 0.606 | 0.341 |
| | W-17 | 0.586 | 0.550 | 0.216 |
| | SW-32 | 0.551 | 0.582 | 0.288 |

In addition to the proposed new feature set, the current study also extended the genre classes up to 32 categories. Next we take a closer look at the three classification tasks (i.e. *SW-32*, *S-15* and *W-17*) at the sub-micro level. Figures 1, 2 and 3 illustrate the learning curves of the three feature sets with the increased training data set (from 10% to 100%) in the three tasks respectively.

Three interesting patterns emerge in the learning curves. Firstly, the accuracy of performance increases when more training texts are added. Take F-POS in *SW-32* for example. With 10% of the training data, F-POS achieves an accuracy of about 0.20 in terms of F-score; with 50% of the training texts, the F-score reaches to 0.40, and with all of the training data, the ultimate F-score reaches over 0.50. Secondly, F-POS performs better than BOW in both *SW-32* and *S-15*, while BOW outperforms F-POS in *W-17*. Finally, F-POS outperforms I-POS in all the three tasks, indicating that fine-grained POS tags with rich linguistic information can better represent the type of texts.

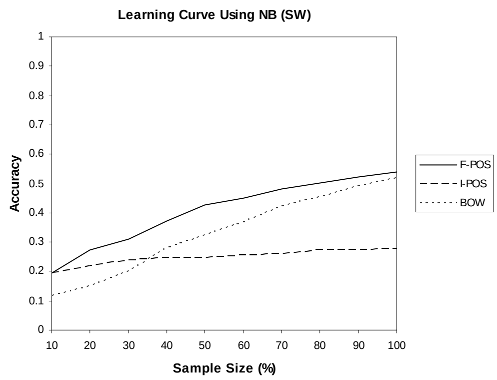



**Figure 1:** Learning curve for *SW-32*

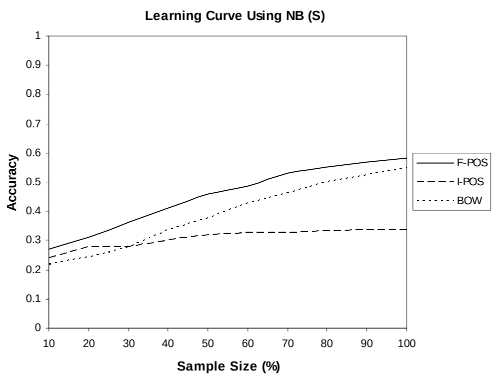



**Figure 2:** Learning curve for *S-15*

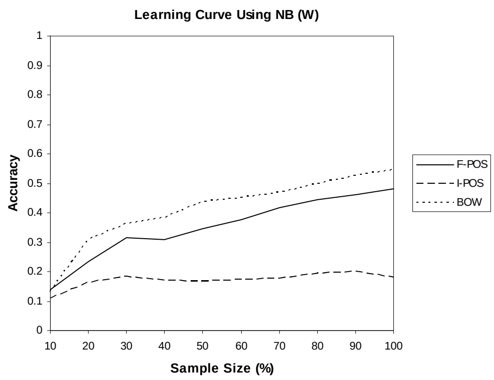



**Figure 3:** Learning curve for *W-17*

## 4.2 Results Obtained from NB-MN Classifier

Naïve Bayes Multinominal Classifier (NB-MN) was used to evaluate the three feature sets according to frequency of genre attributes. Table 4 summarises the performance of the 3 feature sets in genre classification in terms of average weighted F-score. Again, the first column lists the four levels of genres and the second column shows the 10 genre classification tasks. As can be seen in Table 4, the results are generally in line with the previous findings obtained from NB Classifier. First of all, a continual drop in accuracy gain can be observed in most cases with the increase in number of classes. Secondly, genre classification tasks regarding spoken texts generally receive better results than those of written texts. Thirdly, F-POS outperforms BOW with deeper genre classes. It is also worth noticing that BOW achieves better results than frequency-based features when the number of classes is small. Finally, F-POS performs better than I-POS in 9 out of 10 classification tasks.

Next, with regard to the classification at the sub-micro level, the learning curves of the 3 feature sets with the increased training data set (from 10% to 100%) are illustrated in Figures 4, 5, and 6. Again interesting patterns can be observed in the learning curves. Firstly, the accuracy of performance increases when more training texts are added. Secondly, F-POS demonstrates superior classification accuracy when compared with a bag of words and linguistically impoverished tags in all the three tasks.

**Table 4:** Average weighted F-score (NB-MN)

| Genre Granularity | Code | BOW | F-POS | I-POS |
|---|---|---|---|---|
| Super Genre | SW-2 | 0.988 | 0.984 | 0.998 |
| Macro Genre | S-2 | 0.904 | 0.898 | 0.898 |
| | W-2 | 0.892 | 0.778 | 0.728 |
| | SW-4 | 0.895 | 0.850 | 0.833 |
| Micro Genre | S-5 | 0.773 | 0.816 | 0.775 |
| | W-6 | 0.720 | 0.686 | 0.551 |
| | SW-11 | 0.703 | 0.781 | 0.688 |
| Sub-micro Genre | S-15 | 0.499 | 0.785 | 0.647 |
| | W-17 | 0.572 | 0.631 | 0.459 |
| | SW-32 | 0.438 | 0.726 | 0.588 |

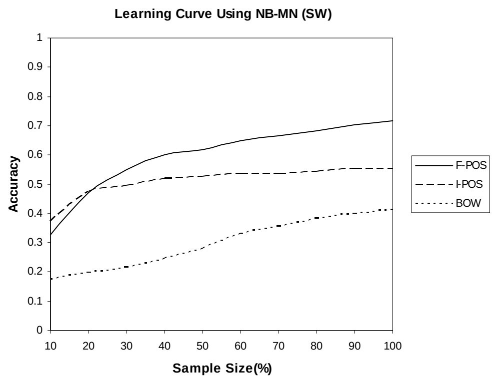
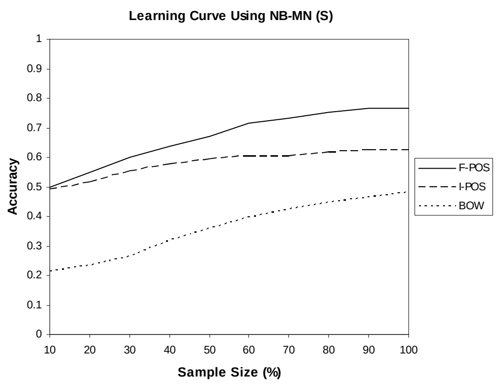



**Figure 4:** Learning curve for *SW-32*



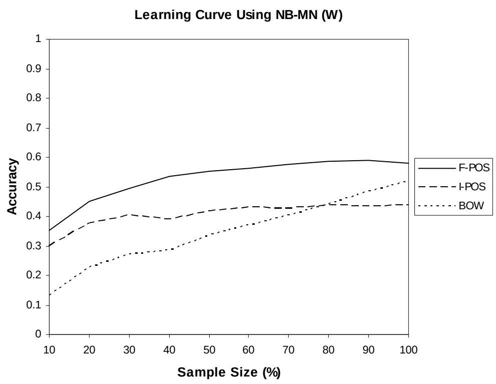




**Figure 5:** Learning curve for *S-15*

**Figure 6:** Learning curve for *W-17*

## 4.3 Discussion

Our investigation suggests that F-POS tag set is shown to provide better generalization than the BOW and that it also has a tremendous advantage over BOW in feature size. The investigation also indicates that the contribution of the proposed F-POS tags to genre classification is achieved through detailed linguistic information provided by the descriptive features. This is evident through the fact that performance dropped with the use of head tags without the features indicating the subcategorisation and grammatical status.

Table 5 presents an overview of results achieved through the use of POS tags as an independent feature set, including those obtained from three previous studies as well as from all the *SW* tasks in the current study.

**Table 5:** An overview of POS tag performance

| Past Studies | | | | Current Study | | |
|---|---|---|---|---|---|---|
| | # of Genre | Accuracy | | # of Genre | Accuracy (NB) | Accuracy (NB-MN) |
| Finn and Kushmerick (2003) | 2 | 61.3% | | 2 | 99.8% | 98.4% |
| | 3 | 84.7% | | / | / | / |
| / | / | / | | 4 | 84.2% | 85.0% |
| Stein and Eissen (2008) | 8 | 74.0% | | / | / | / |
| Santini (2004) | 10 | 77.3% | | / | / | / |
| / | / | / | | 11 | 74.7% | 78.1% |
| / | / | / | | 32 | 58.2% | 72.6% |

Although it is hard to compare the accuracy directly due to factors such as difference in genre class, corpus size, or evaluation model, it is safe to say that the proposed F-POS tags achieve satisfactory accuracy and that they obtain more consistent performance when feature frequency is considered.

## 5 Conclusion

This paper reported an experiment designed to investigate the performance of a linguistically fine-grained POS tag set in automatic genre classification when compared with word unigrams and a linguistically impoverished tag set. The British component of the International Corpus of English (ICE-GB) was employed as a resource of text genres. Ten different genre classification tasks were identified, with a maximum of 500 sample texts. Naïve Bayes and Naïve Bayes Multinominal Classifiers were used to evaluate the performance of the proposed feature set in terms of F-score.

As a result of the experiment, the linguistically rich POS set demonstrated superior classification accuracy when compared with a bag of words and linguistically impoverished tags. The finding highlights the importance of grammatical properties represented in the form of POS tags for the separation of texts according to a predefined hierarchy of genres. In addition, our results also indicate that good classification performance is derived predominantly from the rich linguistic information conveyed through subcategorisation features. This indication is evidenced by the fact that when removed of detailed, subcategorisation features the head tags produced inferior performance.

Future work will include the use of a much larger collection of texts to verify the actual performance of the fine-grained POS entity tags. Tag bigrams and trigrams will also be investigated to verify if additional accuracy gain can be achieved.

## References

Boese E. and A. Howe. 2005. Effects of Web Document Evolution on Genre Classification. *Proceedings of the CIKM' 05*. ACM Press.

David, L. 2001. Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, Vol.5, No.3, pp. 37-72.

Dewdney, N., C. VanEss-Dykema and R. MacMillan. 2001. The Form is the Substance: Classification of Genres in Text. *Proceedings of ACL Workshop on Human Language Technology and Knowledge Management*.

Diederich, J., J. L. Kindermann, E. Leopold and G. Paaß. 2003.Authorship attribution with support vector machines. *Applied Intelligence*, 19(1/2):109-123.

Eissen S. M. and B, Stein. 2004. Genre Classification of Web Pages: User Study and Feasibility Analysis. In S. Biundo, T. Fruhwirth and G. Palm, eds., *KI 2004: Advances in Artificial Intelligence*, pp. 256-269. Springer. Berlin-Heidelberg-New York.

Fang, A.C. 1996. AUTASYS: Automatic Tagging and Cross-Tagset Mapping. In S. Greenbaum, ed., *Comparing English World Wide: The International Corpus of English*, pp. 110-124. Oxford: Oxford University Press.

Fang, A.C. 2007. *English Corpora and Automated Grammatical Analysis*. Beijing: The Commercial Press.

Finn, A. and N. Kushmerick. 2003. Learning to Classify Documents According to Genre. *Proceedings of IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*.

Greenbaum, S. 1996. *Comparing English World Wide: The International Corpus of English*. Oxford: Oxford University Press.

Gupta, R. and L. Ratinov. 2008. Text Categorization with Knowledge Transfer from Heterogeneous Data Sources. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 842-847. Chicago.

Karlgren J. and D. Cutting. 1994. Recognizing Text Genre with Simple Metrics Using Discriminant Analysis. *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994).* Kyoto (Japan).

Koster, C. H. and M. Seutter. Taming wild phrases. 2003. In F. Sebastiani, editor, *Proceedings of ECIR'03, 25$^{th}$ European Conference on Information Retrieval*, pp. 161-176. Pisa, IT, Springer Verlag.

Li , Z., P. Li, W. Wei, H. Liu, J. He, T. Liu and X. Du. 2009. AutoPCS: A Phrase-Based Text Categorization System for Similar Texts. *Proceedings of the Joint International Conferences on Advances in Data and Web Management*, pp. 369-380. Suzhou, China.

Lim, C., K. Lee and G. Kim. 2005. Automatic Genre Detection of Web Documents. *Proceedings of Natural Language Processing (IJCNLP' 04)*, pp. 310-319. Springer.

Nakamura, J. 1993. Statistical Methods and Large Corpora - A New Tool for Describing Text Type. In Baker M., Francis G. and Tognini-Bonelli E. eds., *Text and Technology: In Honor of John Sinclair*, pp. 291-312. J. Benjamins, Philadelphia- Amsterdam.

Rayson, P., A. Wilson and G. Leech. 2002. Grammatical word class variation within the British National Corpus Sampler. In P. Peters, P. Collins and A. Smith, eds., *New Frontiers of Corpus Research*, pp. 295-306. Rodopi. Amsterdam – New York.

Santini M. 2004. A Shallow Approach to Syntactic Feature Extraction for Genre Classification". *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 2004)*. Birmingham (UK).

Scott, S. and S. Matwin. 1999. Feature engineering for text classification. In I. Bratko and S. Dzeroski, eds., *Proceedings of ICML'99, 16th International Conference on Machine Learning*, pp. 379-388. Morgan Kaufmann Publishers, San Francisco, US.

Stein, B. and S. M. Eissen. 2008. Retrieval Models for Genre Classification. *Scandinavian Journal of Information Systems*, Vol. 20: Iss. 1, Article 3.

Witten, I. H. and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Morgan Kauf-mann, San Francisco.