Label Propagation for Fine-Grained Cross-Lingual Genre Classification

Philipp Petrenz
School of Informatics
University of Edinburgh
P.Petrenz@sms.ed.ac.uk

Bonnie Webber
School of Informatics
University of Edinburgh
bonnie@inf.ed.ac.uk

Abstract

Cross-lingual methods can bring the benefits of genre classification to languages which lack genre-annotated training data. However, prior work in this field has been evaluated on coarse genres only. To predict fine-grained genres across languages, we propose a label propagation method, which combines separate sets of features. The results are promising, as the approach outperforms most baselines in our experiments.

1 Introduction

Automated genre classification exploits features in a text to make inference about its communicative purpose. This can benefit users of information retrieval systems, who would like to filter search results based on such criteria [1–3], and text summarization applications, since the structure and position of important information within a text correlates with its genre [4, 5]. Since genres also impact other linguistic properties [6, 7], reliable identification promises to improve further NLP applications.

Work on automated genre classification has been carried out for almost two decades. However, most of the development and evaluation has involved only English texts, due in part to the lack of genre-annotated corpora in other languages. Since genre classification is a supervised machine learning task, any mono-lingual method is restricted to languages for which these resources exist in sufficient quality and quantity. Cross-lingual techniques, on the other hand, can bring the benefits of genre classification to languages, for which no suitable training data exists.

We have reported our prior work on cross-lingual genre classification (CLGC) in [8] and [9]. The method in [8] relies on a set of easily extractable surface features, such as MEAN SENTENCE LENGTH and TYPE/TOKEN RATIO, to bridge the language gap. It then exploits a set of unlabeled target language texts by iteratively re-labeling them, using Support Vector Machines (SVM) and a Bag of Words (BoW) feature set. This approach requires little knowledge about the target language — no Machine Translation (MT), no parser, and no Part of Speech (PoS) tagger — but performs as well as or better than a combination of full-text MT and a mono-lingual classifier for coarse-grained sets of genres (2-way and 4-way classification). In [9], comparable corpora are exploited to improve a 3-way classifier using multi-lingual training sets and feature selection.

While such results are promising, the methods are limited, unless they can also be applied to a more fine-grained set of genres. The approach presented here aims to achieve this by exploiting several disparate sets of text features. This idea was used before in mono-lingual genre classification [10], where features often fall into different categories such as structural, lexical, or presentational [2, 11].

The featurs and algorithm we propose are described in Section 2 of this paper. Section 3 covers the data, evaluation metrics, and baselines used in our experiments, while Section 4 discusses the results and Section 5 briefly summarizes our findings.

2 Method

2.1 Features

Our experiments use three separate feature sets: Cross-lingual features, PoS histogram features, and BoW features. The cross-lingual set is based on the features used in [8], but extended by the frequencies of the 12 universal PoS tags introduced in [12]. Mapping PoS tags from different languages to this set makes them comparable across languages. The set comprises nouns, verbs, adjectives, adverbs, pronouns, determiners, pre/post-positions, numerals, conjunctions, particles, punctuation marks, and a tag for other categories. We also added features for the mean and standard deviation of word lengths, for a total of 33 features. As in [8], they are standardized to zero mean and unit variance for each language separately, to remove the language impact on absolute values.

A PoS histogram feature set has been used before in mono-lingual genre classification [13]. While the authors only evaluated their method on English texts (using the Penn PoS tag set), it can be applied to other languages and PoS tag sets. Here, we adopt their histogram statistics with a sliding window of 5 PoS tags, but skip the PCA step, since the feature set is already small and further reduction did not improve results in preliminary tests.

The BoW features are binary, indicating whether or not the corresponding word occurs in a text. Similar sets have been used in [14] and [3] to classify genres. To reduce the dimensionality, we rank the features by their variance in the collection of source language texts and keep only the top 500 features. This corresponds to eliminating words which occur very frequently or very infrequently.

Note that, unlike the features used in [8], this approach assumes the availability of a PoS tagger in both the source and target languages. It also requires a set of unlabeled texts in the target language.

2.2 Graph-based learning

In order to combine separate heterogeneous sets of features and exploit unlabeled target data, we formulate genre classification as a graph-based label propagation problem, where genres are propagated through the network, using an algorithm based on [15].

To this end, we define a weighted directed graph $G = (V, E, W)$, where each node $v_i \in V$ represents a text from the union of source and target document sets $D = D_S \cup D_T$. Each pair of nodes is connected by two weighted directed edges e^f_{ij} and e^f_{ji} with weights w^f_{ij} and w^f_{ji} exist between them for each feature space f (Section 2.1). Associated with each node is a vector of genre probabilities, with each source language text having a probability of 1 for its (known) genre class, and 0 for all others. Genres for target texts are initialized to 0.

Unlike in the algorithm described in [15], edge weights are not assigned relative to the distance between nodes. Instead, Euclidean distances in a feature space f are used to compute a ranking (in ascending order) for each node. Incoming edge weights are based on position in this ranking, with a Gaussian function determining the value. The weight w^f_{ij} of the directed edge from node j to node i in feature space f is

$$w^f_{ij} = \exp(-\frac{(r^f_{ij} - 1)^2}{2\sigma^2})$$

where r^f_{ij} is the position of node j in the distance ranking of node i in feature space f. σ is a defined as $\frac{|D_T|}{|D_S|} \cdot \theta \cdot m$ where $|D_T|$ and $|D_S|$ are the number of texts in the target language and source language sets respectively, m is the number of texts in the smallest genre class within D_S, and θ is a parameter which we set to 0.1. Note that w^f_{ij} will typically not be the same as w^f_{ji}, since distance ranks are not used to compute outgoing edge weights, that is r^f_{ij} does not affect w^f_{ji}. This means that each node has the same total incoming weight, while the sums of outgoing weights differ. Therefore outliers, which are far from all other nodes, have only little impact.

For edges between target language nodes, weight assignment is repeated for each feature space. The resulting graph has three layers which share nodes, but not edge weights. The rank-based weights ensure that each feature space has the same impact, regardless of the absolute distance values. For the edges between source and target language nodes, only the cross-lingual feature set is used. For

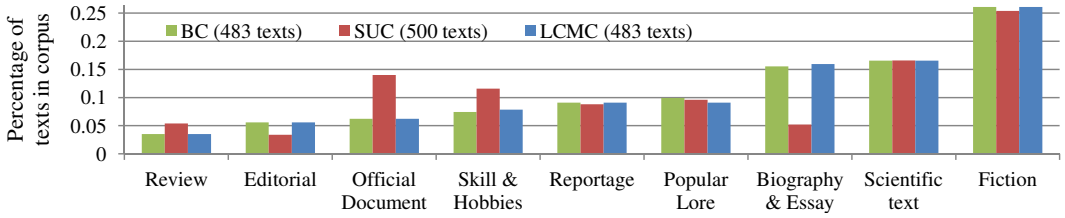

Figure 1: Percentage of texts per genre in the BC (English), SUC (Swedish), and LCMC (Chinese).

edges from source to target nodes, σ is a defined as $\theta \cdot m$. Edges in the opposite direction have 0 weight, as do all source-target edges for the other two feature spaces (i.e. PoS histogram and BoW).

The iterative label propagation process is similar to [15], with two modifications. Firstly, in addition to the input from other nodes, each node receives its own labels from the previous iteration. The weight of this input is equal to the combined weight of all incoming edges of a single feature space. Secondly, each node propagates the squares of its label probabilities. This is to reduce the overall impact of nodes that have their probability mass spread across several genre classes, since this indicates low confidence in their labels. Note that in the first iteration, only labels propagated through the source-target edges have any effect, since the initial target node label proabilities are 0.

As suggested in [15], the target node label probabilities are scaled after each iteration, so that the sums over all target node probabilities match the relative genre distribution of the source set (or one provided by an oracle). They are also scaled to add up to 1 for each node. These scaling steps are repeated until convergence. After the final iteration and scaling, each node is assigned whichever genre label has the highest probability in its vector.

3 Experiments

3.1 Data & Evaluation

For our experiments, we used the Brown Corpus (BC) of English texts, the Stockholm-Umeå Corpus (SUC) of Swedish texts, and the Lancaster Corpus of Mandarin Chinese (LCMC), since they are publicly available and comparable in their sampling methodology. However, their categorization scheme is not identical, with the SUC oriented more functionally. We removed all *religion* texts from the BC and LCMC, since it is not a main category in the SUC. We also assigned all fiction texts to a single category. Figure 1 illustrates the proportional numbers of texts that fall in each of the nine resulting genre categories. Note that the distribution of genres is both skewed and language-specific. This is challenging for a machine learning algorithm, but realistic for a cross-lingual task.

All three corpora are tokenized on the word, sentence, and paragraph level, and we used this tokenization without further processing. We assgined PoS tags using the Stanford log-linear tagger [16] trained on appropriate corpora (en: Wall Street Journal; sv: CoNLL 2006 shared task; zh: Chinese Treebank). For the cross-lingual feature set, PoS tags were transformed to the 12-class universal set (Section 2.1) after tagging.

We carried out separate experiments for English, Swedish, and Chinese as the target language. As our previous results suggest [9], multi-lingual training sets can prevent a learning algorithm from over-fitting to one specific language. Therefore, the texts from the two source languages are combined into one multi-lingual source (training) set for all algorithms including the baselines. In addition to prediction accuracy, we evaluated all algorithms with the mean F1-Score over all genre classes. Since the genre distributions are skewed, this measure indicates how balanced predictions are across genres.

3.2 Baselines

We compare the performance of our method to three baselines. Firstly, the SVM wrapper algorithm used in [8], using the cross-lingual feature set described in Section 2.1 to bridge the language gap. This extends the features used in [8] and adapts it to the task of fine-grained genre classification. The

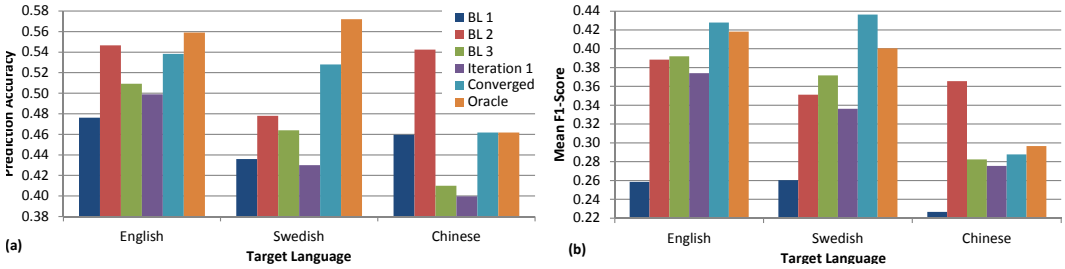

Figure 2: (a) Accuracy and (b) mean F1-Score for the 3 baselines and the proposed approach after the first iteration, after convergence, and with the correct genre distribution provided by an oracle.

BoW representation is used to iteratively re-label the target language texts, reduced in each iteration to the 500 features with the highest information gain score.

The second baseline combines full-text MT with a mono-lingual classifier, using *Google Translate* to translate each Swedish and Chinese text into English — regardless of which language is used as the target language. An SVM model is then trained using the reduced 500-dimension BoW representation described in Section 2.1. This baseline requires computational overhead and additional resources to automatically translate texts. It is therefore restricted to applications where such overhead is acceptable and appropriate resources are available.

The third baseline uses the same graph-based algorithm as our proposed method. However, instead of keeping separate feature sets (and thus several edges between two nodes in the graph), all features (cross-lingual, PoS histograms, and BoW) are combined in one set, except for source-target edges.

4 Results & Discussion

Figures 2a and 2b illustrate prediction accuracies and mean F1-Scores respectively for the three baselines and three versions of our proposed method: Only one iteration, full convergence, and full convergence with the correct target genre distribution provided by an oracle, rather than estimated from the set of source texts. The SVM wrapper algorithm (BL 1) performs relatively poorly, in particular when assessed by F1-Scores. Predictions are highly skewed towards the dominant genres, which makes this approach unsuitable for a fine-grained CLGC task. The MT based method (BL 2) performs well for English and Chinese as target languages, but less so for Swedish. This might be due to the fact that the genre distribution in the SUC differs significantly from the other two corpora. The graph-based method with the combined feature set (BL 3) does not achieve high accuracy, but outperforms BL1 when comparing F1-Scores.

When comparing the results of our approach after convergence to those after the first iteration, it becomes obvious that the algorithm is able to exploit the unlabeled target texts to improve both accuracy and F1-Scores. It outperforms BL3 for all three target languages, which indicates that a combination of separate feature sets benefits genre classification methods. It performs similarly to BL2 for English, better for Swedish, but worse for Chinese as target language. This may be due to the fact that both the PoS histogram and the BoW feature sets were designed for English genre classification and work less well on a language as different as Chinese. For English and Swedish, scaling to the oracle-provided genre distribution boosts prediction accuracy. This effect is larger for Swedish, where genre distributions in source and target sets differ more.

5 Conclusion

A graph-based algorithm with label propagation can be used to combine different feature sets in a classifier designed to predict fine-grained genres across languages. These features include a cross-lingual mapping of PoS tags, which had not been explored for this task before. For two of three tested target languages, our results show comparable or better results than a method using full text machine translation, while requiring fewer resources. Since MT exploits different features, we plan to experiment with a combination of the two, to see if results can be improved when MT is available.

References

[1] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[2] B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[3] L. Freund, C. L. A. Clarke, and E. G. Toms. Towards genre classification for IR in the workplace. In *Proceedings of the 1st international conference on Information interaction in context*, pages 30–36, New York, NY, USA, 2006. ACM.

[4] J. Goldstein, G. M. Ciany, and J. G. Carbonell. Genre identification and goal-focused summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 889–892, New York, NY, USA, 2007. ACM.

[5] V. A. Yatsko, M. S. Starikov, and A. V. Butakov. Automatic genre recognition and adaptive text summarization. *Autom. Doc. Math. Linguist.*, 44:111–120, June 2010.

[6] B. Webber. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, 2009.

[7] P. Petrenz and B. Webber. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393, 2011.

[8] P. Petrenz. Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.

[9] P. Petrenz and B. Webber. Robust cross-lingual genre classification through comparable corpora. In *The 5th Workshop on Building and Using Comparable Corpora*, page 1, 2012.

[10] J. Chaker and O. Habib. Genre categorization of web pages. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 455–464, Washington, DC, USA, 2007. IEEE Computer Society.

[11] C. Lim, K. Lee, and G. Kim. Multiple sets of features for automatic genre classification of web documents. *Information processing & management*, 41(5):1263–1276, 2005.

[12] S. Petrov, D. Das, and R. McDonald. A universal part-of-speech tagset. *Arxiv preprint ArXiv:1104.2086*, 2011.

[13] S. Feldman, M. Marin, J. Medero, and M. Ostendorf. Classifying factored genres with part-of-speech histograms. In *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. Association for Computational Linguistics, 2009.

[14] E. Stamatatos, G. Kokkinakis, and N. Fakotakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495, 2000.

[15] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[16] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.