Cross-Lingual Genre Classification

Philipp Petrenz
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, UK
p.petrenz@sms.ed.ac.uk

Abstract

Classifying text genres across languages can bring the benefits of genre classification to the target language without the costs of manual annotation. This article introduces the first approach to this task, which exploits text features that can be considered *stable genre predictors* across languages. My experiments show this method to perform equally well or better than full text translation combined with monolingual classification, while requiring fewer resources.

1 Introduction

Automated text classification has become standard practice with applications in fields such as information retrieval and natural language processing. The most common basis for text classification is by topic (Joachims, 1998; Sebastiani, 2002), but other classification criteria have evolved, including sentiment (Pang et al., 2002), authorship (de Vel et al., 2001; Stamatatos et al., 2000a), and author personality (Oberlander and Nowson, 2006), as well as categories relevant to filter algorithms (e.g., spam or inappropriate contents for minors).

Genre is another text characteristic, often described as orthogonal to topic. It has been shown by Biber (1988) and others after him, that the genre of a text affects its formal properties. It is therefore possible to use cues (e.g., lexical, syntactic, structural) from a text as features to predict its genre, which can then feed into information retrieval applications (Karlgren and Cutting, 1994; Kessler et al., 1997; Finn and Kushmerick, 2006; Freund et al., 2006). This is because users may want documents that serve a particular communicative purpose, as well as being on a particular topic. For example, a web search on the topic "crocodiles" may return an encyclopedia entry, a biological fact sheet, a news report about attacks in Australia, a blog post about a safari experience, a fiction novel set in South Africa, or a poem about wildlife. A user may reject many of these, just because of their genre: Blog posts, poems, novels, or news reports may not contain the kind or quality of information she is seeking. Having classified indexed texts by genre would allow additional selection criteria to reflect this.

Genre classification can also benefit Language Technology indirectly, where differences in the cues that correlate with genre may impact system performance. For example, Petrenz and Webber (2011) found that within the New York Times corpus (Sandhaus, 2008), the word "states" has a higher likelihood of being a verb in letters (approx. 20%) than in editorials (approx. 2%). Part-of-Speech (PoS) taggers or statistical machine translation (MT) systems could benefit from knowing such genre-based domain variation. Kessler et al. (1997) mention that parsing and word-sense disambiguation can also benefit from genre classification. Webber (2009) found that different genres have a different distribution of discourse relations, and Goldstein et al. (2007) showed that knowing the genre of a text can also improve automated summarization algorithms, as genre conventions dictate the location and structure of important information within a document.

All the above work has been done within a single language. Here I describe a new approach to genre classification that is cross-lingual. Cross-lingual genre classification (CLGC) differs

Proceedings of the EACL 2012 Student Research Workshop, pages 11–21,
Avignon, France, 26 April 2012.

from both poly-lingual and language-independent genre classification. CLGC entails training a genre classification model on a set of labeled texts written in a source language L_S and using this model to predict the genres of texts written in the target language $L_T \neq L_S$. In poly-lingual classification, the training set is made up of texts from two or more languages $S = \{L_{S_1}, \ldots, L_{S_N}\}$ that include the target language $L_T \in S$. Language-independent classification approaches are mono-lingual methods that can be applied to any language. Unlike CLGC, both poly-lingual and language-independent genre classification require labeled training data in the target language.

Supervised text classification requires a large amount of labeled data. CLGC attempts to leverage the available annotated data in well-resourced languages like English in order to bring the aforementioned advantages to poorly-resourced languages. This reduces the need for manual annotation of text corpora in the target language. Manual annotation is an expensive and time-consuming task, which, where possible, should be avoided or kept to a minimum. Considering the difficulties researchers are encountering in compiling a genre reference corpus for even a single language (Sharoff et al., 2010), it is clear that it would be infeasible to attempt the same for thousands of other languages.

2 Prior work

Work on automated genre classification was first carried out by Karlgren and Cutting (1994). Like Kessler et al. (1997) and Argamon et al. (1998) after them, they exploit (partly) hand-crafted sets of features, which are specific to texts in English. These include counts of function words such as "we" or "therefore", selected PoS tag frequencies, punctuation cues, and other statistics derived from intuition or text analysis. Similarly language specific feature sets were later explored for mono-lingual genre classification experiments in German (Wolters and Kirsten, 1999) and Russian (Braslavski, 2004).

In subsequent research, automatically generated feature sets have become more popular. Most of these tend to be language-independent and might work in mono-lingual genre classification tasks in languages other than English. Examples are the word based approaches suggested by Stamatatos et al. (2000b) and Freund et al. (2006), the image features suggested by Kim and Ross (2008), the PoS histogram frequency approach by Feldman et al. (2009), and the character n-gram approaches proposed by Kanaris and Stamatatos (2007) and Sharoff et al. (2010). All of them were tested exclusively on English texts. While language-independence is a popular argument often claimed by authors, few have shown empirically that this is true of their approach. One of the few authors to carry out genre classification experiments in more than one language was Sharoff (2007). Using PoS 3-grams and a variation of common word 3-grams as feature sets, Sharoff classified English and Russian documents into genre categories. However, while the PoS 3-gram set yielded respectable prediction accuracy for English texts, in Russian documents, no improvement over the baseline of choosing the most frequent genre class was observed.

While there is virtually no prior work on CLGC, cross-lingual methods have been explored for other text classification tasks. The first to report such experiments were Bel et al. (2003), who predicted text topics in Spanish and English documents, using one language for training and the other for testing. Their approach involves training a classifier on language A, using a document representation containing only content words (nouns, adjectives, and verbs with a high corpus frequency). These words are then translated from language B to language A, so that texts in either language are mapped to a common representation.

Thereafter, cross-lingual text classification was typically regarded as a domain adaptation problem that researchers have tried to solve using large sets of unlabeled data and/or small sets of labeled data in the target language. For instance, Rigutini et al. (2005) present an EM algorithm in which labeled source language documents are translated into the target language and then a classifier is trained to predict labels on a large, unlabeled set in the target language. These instances are then used to iteratively retrain the classification model and the predictions are updated until convergence occurs. Using information gain scores at every iteration to only retain the most predictive words and thus reduce noise, Rigutini et al. (2005) achieve a considerable improvement over the baseline accuracy, which is a simple translation of the training instances and subsequent

mono-lingual classification. They, too, were classifying texts by topics and used a collection of English and Italian newsgroup messages. Similarly, researchers have used semi-supervised bootstrapping methods like co-training (Wan, 2009) and other domain adaptation methods like structural component learning (Prettenhofer and Stein, 2010) to carry out cross-lingual text classification.

All of the approaches described above rely on MT, even if some try to keep translation to a minimum. This has several disadvantages however, as applications become dependent on parallel corpora, which may not be available for poorly-resourced languages. It also introduces problems due to word ambiguity and morphology, especially where single words are translated out of context. A different method is proposed by Gliozzo and Strapparava (2006), who use latent semantic analysis on a combined collection of texts written in two languages. The rationale is that named entities such as "Microsoft" or "HIV" are identical in different languages with the same writing system. Using term correlation, the algorithm can identify semantically similar words in both languages. The authors exploit these mappings in cross-lingual topic classification, and their results are promising. However, using bilingual dictionaries as well yields a considerable improvement, as Gliozzo and Strapparava (2006) also report.

While all of the methods above could technically be used in any text classification task, the idiosyncrasies of genres pose additional challenges. Techniques relying on the automated translation of predictive terms (Bel et al., 2003; Prettenhofer and Stein, 2010) are workable in the contexts of topics and sentiment, as these typically rely on content words such as nouns, adjectives, and adverbs. For example, "hospital" may indicate a text from the medical domain, while "excellent" may indicate that a review is positive. Such terms are relatively easy to translate, even if not always without uncertainty. Genres, on the other hand, are often classified using function words (Karlgren and Cutting, 1994; Stamatatos et al., 2000b) like "of", "it", or "in". It is clear that translating these out of context is next to impossible. This is true in particular if there are differences in morphology, since function words in one language may be morphological affixes in another.

Although it is theoretically possible to use the bilingual low-dimension approach by Gliozzo and Strapparava (2006) for genre classification, it relies on certain words to be identical in two different languages. While this may be the case for topic-indicating named entities — a text containing the words "Obama" and "McCain" will almost certainly be about the U.S. elections in 2008, or at least about U.S. politics — there is little indication of what its genre might be: It could be a news report, an editorial, a letter, an interview, a biography, or a blog entry, just to name a few. Because topics and genres correlate, one would probably reject some genres like instruction manuals or fiction novels. However, uncertainty is still large, and Petrenz and Webber (2011) show that it can be dangerous to rely on such correlations. This is particularly true in the cross-lingual case, as it is not clear whether genres and topics correlate in similar ways in a different language.

3 Approach

The approach I propose here relies on two strategies I explain below in more detail: *Stable features* and *target language adaptation*. The first is based on the assumption that certain features are indicative of certain genres in more than one language, while the latter is a less restricted way to boost performance, once the language gap has been bridged. Figure 1 illustrates this approach, which is a challenging one, as very little prior knowledge is assumed by the system. On the other hand, in theory it allows any resulting application to be used for a wide range of languages.

3.1 Assumption of prior knowledge

Typically, the aim of cross-lingual techniques is to leverage the knowledge present in one language in order to help carry a task in another language, for which such knowledge is not available. In the case of genre classification, this knowledge comprises genre labels of the documents used to train the classification model. My approach requires no labeled data in the target language. This is important, as some domain adaptation algorithms rely on a small set of labeled texts in the target domain.

Cross-lingual methods also often rely on MT, but this effectively restricts them to languages for which MT is sufficiently developed. Apart from the fact that it would be desirable for a cross-lingual genre classifier to work for as many

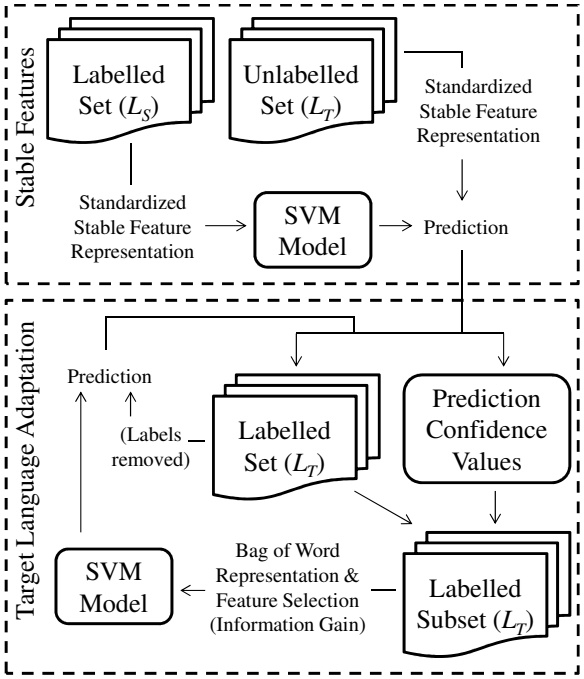


Figure 1: Outline of the proposed method for CLGC.

languages as possible, MT only allows classification in well-resourced languages. However, such languages are more likely to have genre-annotated corpora, and mono-lingual classification may yield better results. In order to bring the advantages of genre classification to poorly-resourced languages, the availability of MT techniques, at least for the time being, must not be assumed. I only use them to generate baseline results.

The same restriction is applied to other types of prior knowledge, and I do not assume supervised PoS taggers, syntactic parsers, or other tools are available. In future work however, I may explore unsupervised methods, such as the PoS induction methods of Clark (2003), Goldwater and Griffiths (2007), or Berg-Kirkpatrick et al. (2010), as they do not represent external knowledge.

There are a few assumptions that must be made in order to carry out any meaningful experiments. First, some way to detect sentence and paragraph boundaries is expected. This can be a simple rule-based algorithm, or unsupervised methods, such as the Punkt boundary detection system by Kiss and Strunk (2006). Also, punctuation symbols and numerals are assumed to be identifiable as such, although their exact semantic function is unknown. For example, a question mark will be identified as a punctuation symbol, but its function (question cue; end of a sentence) will not. Lastly, a sufficiently large, unlabeled set of texts in the target language is required.

3.2 Stable features

Many types of features have been used in genre classification. They all fall into one of three groups: *Language-specific features* are cues which can only be extracted from texts in one language. An example would be the frequency of a particular word, such as "yesterday". *Language-independent features* can be extracted in any language, but they are not necessarily directly comparable. Examples would be the frequencies of the ten most common words. While these can be extracted for any language (as long as words can be identified as such), the function of a word on a certain position in this ranking will likely differ from one language to another. *Comparable features*, on the other hand, represent the same function, or part of a function, in two or more languages. An example would be type/token ratios, which, in combination with the document length, represent the lexical richness of a text, independent of its language. If such features prove to be good genre predictors across languages, they may be considered *stable* across those languages. Once suitable features are found, CLGC may be considered a standard classification problem, as outlined in the upper part of Figure 1.

I propose an approach that makes use of such stable features, which include mostly structural, rather than lexical cues (cf. Section 4). Stable features lend themselves to the classification of genres in particular. As already mentioned, genres differ in communicative purpose, rather than in topic. Therefore, features involving content words are only useful to an extent. While topical classification is hard to imagine without translation or parallel/comparable corpora, genre classification can be done without such resources. Stable features provide a way to bridge the language gap even to poorly-resourced languages.

This does not necessarily mean that the values of these attributes are in the same range across languages. For example, the type/token ratio will typically be higher in morphologically-rich languages. However, it might still be true that novels have a richer vocabulary than scientific articles, whether they are written in English or Finnish. In

order to exploit such features cross-linguistically, their values have to be mapped from one language to another. This can be done in an unsupervised fashion, as long as enough data is present in both source and target language (cf. Section 3.1). An easy and intuitive way is to standardize values so that each feature in both sets has a mean value of zero mean and variance of one. This is achieved by subtracting from each feature value the mean over all documents and dividing it by the standard deviation.

Note that the training and test sets have to be standardized separately in order for both sets to have the same mean and variance and thus be comparable. This is different from classification tasks where training and test set are assumed to be sampled from the same distribution. Although standardization (or another type of scaling) is often performed in such tasks as well, the scaling factor from the training set would be used to scale the test set (Hsu et al., 2000).

3.3 Target language adaptation

Cross-lingual text classification has often been considered a special case of domain adaptation. Semi-supervised methods, such as the expectation-maximization (EM) algorithm (Dempster et al., 1977), have been employed to make use of both labeled data in the source language and unlabeled data in the target language. However, adapting to a different language poses a greater challenge than adapting to different genres, topics, or sources. As the vocabularies have little (if any) overlap, it is not trivial to initially bridge the gap between the domains. Typically, MT would be used to tackle this problem.

Instead, my use of stable features shifts the focus of subsequent domain adaptation to exploiting unlabeled data in the target language to improve prediction accuracy. I refer to this as *target language adaptation* (TLA). The advantage of making this separation is that a different set of features can be used to adapt to the target language. There is no reason to keep the restrictions required for stable features once the language gap has been bridged. In fact, any language-independent feature may be used for this task. The assumption is that the method described in Section 3.2 provides a good but enhanceable result, that is significantly below mono-lingual performance. The resulting decent, though imperfect, labeling of target language texts may be exploited to improve accuracy.

A wide range of possible features lend themselves to TLA. Language-independent features have often been proposed in prior work on genre classification. These include bag-of-words, character n-grams, and PoS frequencies or PoS n-grams, although the latter two would have to be based on the output of unsupervised PoS induction algorithms in this scenario. Alternatively, PoS tags could be approximated by considering the most frequent words as their own tag, as suggested by Sharoff (2007). With appropriate feature sets, iterative algorithms can be used to improve the labeling of the set in the target domain.

The lower part of Figure 1 illustrates the TLA process proposed for CLGC. In each iteration, confidence values obtained from the previous classification model are used to select a subset of labeled texts in the target language. Intuitively, only texts which can be confidently assigned to a certain genre should be used to train a new model. This is particularly true in the first iteration, after the stable feature prediction, as error rates are expected to be high. The size of this subset is increased at each iteration in the process until it comprises all the texts in the test set. A multi-class Support Vector Machine (SVM) in a k genre problem is a combination of $\frac{k\times(k-1)}{2}$ binary classifiers with voting to determine the overall prediction. To compute a confidence value for this prediction, I use the geometric mean $G = (\prod_{i=1}^{n} a_i)^{1/n}$ of the distances from the decision boundary a_i for all the n binary classifiers, which include the winning genre (i.e., $n = k - 1$). The geometric mean heavily penalizes low values, that is small distances to the hyperplane separating two genres. This corresponds to the intuition that there should be a high certainty in any pairwise genre comparison for a high-confidence prediction. Negative distances from the boundary are counted as zero, which reduces the overall confidence to zero. The acquired subset is then transformed to a bag of words representation. Inspired by the approach of Rigutini et al. (2005), the information gain for each feature is computed, and only the highest ranked features are used. A new classification model is trained and used to re-label the target language texts. This process continues until convergence (i.e., labels in two subsequent iterations are identical) or until a pre-defined iteration limit is reached.

4 Experiments

4.1 Baselines

To verify the proposed approach, I carried out experiments using two publicly available corpora in English and in Chinese. As there is no prior work on CLGC, I chose as baseline an SVM model trained on the source language set using a bag of words representation as features. This had previously been used for this task by Freund et al. (2006) and Sharoff et al. (2010).[1] The texts in the test set were then translated from the target into the source language using *Google translate*[2] and the SVM model was used to predict their genres. I also tested a variant in which the training set was translated into the target language before the feature extraction step, with the test set remaining untranslated. Note that these are somewhat artificial baselines, as MT in reasonable quality is only available for a few selected languages. They are therefore not workable solutions to classify genres in poorly-resourced languages. Thus, even a cross-lingual performance close to these baselines can be considered a success, as long as no MT is used. For reference, I also report the performances of a random guess approach and a classifier labeling each text as the dominant genre class.

With all experiments, results are reported for the test set in the target language. I infer confidence intervals by assuming that the number of misclassifications is approximately normally distributed with mean $\mu = e \times n$ and standard deviation $\sigma = \sqrt{\mu \times (1 - e)}$, where e is the percentage of misclassified instances and n is the size of the test set. I take two classification results to differ significantly only if their 95% confidence intervals (i.e., $\mu \pm 1.96 \times \sigma$) do not overlap.

4.2 Data

In line with some of the prior mono-lingual work on genre classification, I used the Brown corpus for my experiments. As illustrated in Table 1, the 500 texts in the corpus are sampled from 15 genres, which can be categorized more broadly into four broad genre categories, and even more broadly into informative and imaginative texts. The second corpus I used was the Lancaster Corpus of Mandarin Chinese (LCMC). In creating the LCMC, the Brown sampling frame was followed very closely and genres within these two corpora are comparable, with the exception of Western Fiction, which was replaced by Martial Arts Fiction in the LCMC. Texts in both corpora are tokenized by word, sentence, and paragraph, and no further pre-processing steps were necessary.

Informative	Press (88 texts)	Press: Reportage
		Press: Editorials
		Press: Reviews
	Misc. (176 texts)	Religion
		Skills, Trades & Hobbies
		Popular Lore
		Biographies & Essays
	Non-Fiction (110 texts)	Reports & Official Documents
		Academic Prose
Imaginative	Fiction (126 texts)	General Fiction
		Mystery & Detective Fiction
		Science Fiction
		Adventure & Western Fiction
		Romantic Fiction
		Humor

Table 1: Genres in the Brown corpus. Categories are identical in the LCMC, except Western Fiction is replaced by Martial Arts Fiction.

Following Karlgren and Cutting (1994), I tested my approach on all three levels of granularity. However, as the 15-genre task yields relatively poor CLGC results (both for my approach and the baselines), I report and discuss only the results of the two and four-genre task here. Improving performance on more fine-grained genres will be subject of future work (cf. Section 6).

4.3 Features and Parameters

The stable features used to bridge the language gap are listed in Table 2. Most are simply extractable cues that have been used in mono-lingual genre classification experiments before: Average sentence/paragraph lengths and standard deviations, type/token ratio and numeral/token ratio. To these, I added a ratio of single lines in a text — that is, paragraphs containing no more than one sentence, divided by the sentence count. These are typically headlines, datelines, author names, or other structurally interesting parts. A distribution value indicates how evenly single lines are distributed throughout a text, with high values indicating single lines predominantly occurring at the beginning and/or end of a text.

[1] Other document representations, including character n-grams, were tested, but found to perform worse in this task.

[2] http://translate.google.com

Features	F	N	P	M	Features	F	N	P	M
Average Sentence	−0.5	0.6	0.1	0.0	Type/Token	0.0	−0.9	0.6	0.3
Length	−1.0	0.5	0.0	0.3	Ratio	0.0	−0.9	0.9	0.1
Sentence Length	−0.3	0.5	−0.1	0.0	Numeral/Token	−0.3	0.6	−0.1	−0.1
Standard Deviation	−0.5	0.4	0.0	0.1	Ratio	−0.7	0.7	0.4	−0.1
Average Paragraph	−0.4	0.3	−0.1	0.1	Single Lines/	0.3	0.1	−0.1	−0.2
Length	−0.4	0.4	−0.6	0.4	Sentence Ratio	0.0	−0.3	1.1	−0.4
Paragraph Length	−0.4	0.4	−0.2	0.1	Single Line	−0.3	0.2	0.0	0.1
Standard Deviation	−0.1	0.4	−0.6	0.1	Distribution	0.1	−0.1	0.1	0.0
Relative tf-idf values of	0.2	0.1	−0.1	0.0	Topic Average	−0.4	0.8	−0.3	0.0
top 10 weighted words*	0.4	−0.2	−0.5	0.1	Precision	−0.4	0.8	−0.2	−0.1

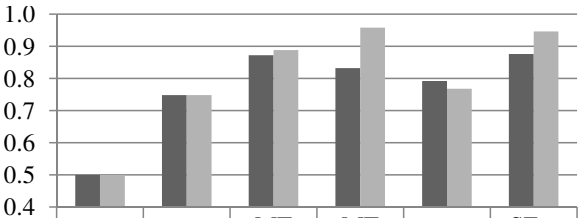
Table 2: Set of 19 stable features used to bridge the language gap. The numbers denote the mean values after standardization for each broad genre in the LCMC (upper values) and Brown corpus (lower values): **F**iction, **N**on-Fiction, **P**ress, and **M**iscellaneous. Negative/Positive numbers denote lower/higher average feature values for this genre when compared to the rest of the corpus. *Relative tf-idf values are ten separate features. The numbers given are for the highest ranked word only.

The remaining features (cf. last row of Table 2) are based on ideas from information retrieval. I used tf-idf weighting and marked the ten highest weighted words in a text as relevant. I then treated this text as a ranked list of relevant and non-relevant words, where the position of a word in the text determined its rank. This allowed me to compute an average precision (AP) value. The intuition behind this value is that genre conventions dictate the location of important content words within a text. A high AP score means that the top tf-idf weighted words are found predominantly in the beginning of a text. In addition, for the same ten words, I added the tf-idf value to the feature set, divided by the sum of all ten. These values indicate whether a text is very focused (a sharp drop between higher and lower ranked words) or more spread out across topics (relatively flat distribution).

For each of these features, Table 2 shows the mean values for the four broad genre classes in the LCMC and Brown corpus, after the sets have been standardized to zero mean and unit variance. This is the same preprocessing process used for training and testing the SVM model, although the statistics in Table 2 are not available to the classifier, since they require genre labels. Each row gives an idea of how suitable a feature might be to distinguish between these genres in Chinese (upper row) and English (lower row). Both rows together indicate how stable a feature is across languages for this task. Some features, such as the topic AP value, seems to be both a good predictor for genre and stable across languages. In both Chinese and English, for example, the topical words seem to be concentrated around the beginning of the text in Non-Fiction, but much less so in Fiction. These patterns can be seen in other features as well. The type/token ratio is, on average, highest in Press texts, followed by Miscellaneous texts, Fiction texts, and Non-Fiction texts in both corpora. While this does not hold for all the features, many such patterns can be observed in Table 2.

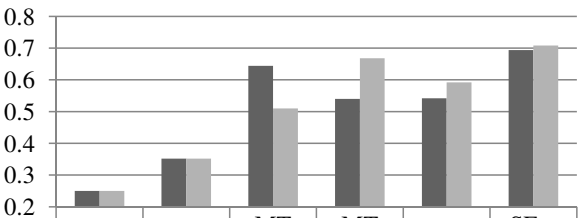
Since uncertainty after the initial prediction is very high, the subset used to re-train the SVM model was chosen to be small. In the first iteration, I used up to 60% of texts with the highest confidence value within each genre. To avoid an imbalanced class distribution, texts were chosen so that the genre distribution in the new training set matched the one in the source language. To illustrate this, consider an example with two genre classes A and B, represented by 80% and 20% of texts respectively in the source language. Assuming that after the initial prediction both classes are assigned to 100 texts in a test set of size 200, the 60 texts with the highest confidence values would be chosen for class A. To keep the genre distribution of the source language, only the top 15 texts would be chosen for class B.

In the second iteration, I simply used the top 90% of texts overall. This number was increased by 5% in each subsequent iteration, so that the full set was used from the fourth iteration. No changes were made to the genre distribution from the second iteration. To train the classification model, I used the 500 features with the highest informa-



	Rand.	Prior	MT Source	MT Target	SF	SF + TLA
t: zh	50.0%	74.8%	87.2%	83.2%	79.2%	87.6%
t: en	50.0%	74.8%	88.8%	95.8%	76.8%	94.6%

Figure 2: Prediction accuracies for the Brown / LCMC two genre classification task. Dark bars denote English as source language and Chinese as target language (en→zh), light bars denote the reverse (zh→en). Rand.: Random classifier. Prior: Classifier always predicting the most dominant class. The baselines MT Source and MT target use MT to translate texts into the source and target language, respectively. SF: Stable Features. TLA: Target Language Adaptation.



	Rand.	Prior	MT Source	MT Target	SF	SF + TLA
t: zh	25.0%	35.2%	64.4%	54.0%	54.2%	69.4%
t: en	25.0%	35.2%	51.0%	66.8%	59.2%	70.8%

Figure 3: Prediction accuracies for the Brown / LCMC four genre classification task. Labels as in Figure 2.

tion gain score for the selected training set in each iteration. As convergence is not guaranteed theoretically, I used a maximum limit of 15 iterations. In my experiments however, the algorithm always converged.

5 Results and Discussion

Figure 2 shows the accuracies for the two genre task (informative texts vs. imaginative texts) in both directions: English as a source language with Chinese being the target language (en→zh) and vice versa (zh→en). As the class distribution is skewed (374 vs. 126 texts), always predicting the most dominant class yields acceptable performance. However, this is simplistic and might fail in practice, where the most dominant class will typically be unknown.

Full text translation combined with monolingual classification performs well. Stable features alone yield a respectable prediction accuracy, but perform significantly worse than MT Source in both tasks and MT Target in the zh→en task. However, subsequent TLA significantly improves the accuracy on both tasks, eliminating any significant difference from baseline performance.

Figure 3 shows results for the four genre classification task (Fiction vs. Non-Fiction vs. Press vs. Misc.). Again, MT Source and MT Target perform well. However, translating from Chinese into English yields better results than the reverse. This might be due to the easier identification of words in English and thus a more accurate bag of words representation. TLA manages to significantly improve the stable feature results. My approach outperforms both baselines in this experiment, although the differences are only significant if texts are translated from English to Chinese.

These results are encouraging, as they show that in CLGC tasks, equal or better performance can be achieved with fewer resources, when compared the baseline of full text translation. The reason why TLA works well in this case can be understood by comparing the confusion matrices before the first iteration and after convergence (Table 3). While it is obvious that the stable feature approach works better on some classes than on others, the distributions of predicted and actual genres are fairly similar. For Fiction, Non-Fiction, and Press, precision is above 50%, with correct predictions outweighing incorrect ones, which is an important basis for subsequent iterative learning. However, too many texts are predicted to belong to the Miscellaneous category, which reduces recall on the other genres. By using a different feature set and concentrating on the documents with high confidence values, TLA manages to remedy this problem to an extent. While misclassifications are still present, recalls for the Fiction and Non-Fiction genres are increased significantly, which explains the higher overall accuracy.

6 Conclusion and future work

I have presented the first work on cross-lingual genre classification (CLGC). I have shown that some text features can be considered stable genre predictors across languages and that it is possible to achieve good results in CLGC tasks without

	Fict.	Non-Fict.	Press	Misc.
Fiction	65	2	8	51
Non-Fiction	4	59	2	45
Press	5	8	31	44
Miscellaneous	18	28	14	116
Precision	0.71	0.61	0.56	0.45
Recall	0.52	0.54	0.35	0.66

	Fict.	Non-Fict.	Press	Misc.
Fiction	102	0	2	22
Non-Fiction	0	83	0	27
Press	2	8	27	51
Miscellaneous	29	9	3	135
Precision	0.77	0.83	0.84	0.57
Recall	0.81	0.75	0.31	0.77

Table 3: Confusion Matrices for the four genre en→zh task. Left: After stable feature prediction, but before TLA. Right: After TLA convergence. Rows 2–5 denote actual numbers of texts, columns denote predictions.

resource-intensive MT techniques. My approach exploits stable features to bridge the language gap and subsequently applies iterative target language adaptation (TLA) in order to improve accuracy. The approach performed equally well or better than full text translation combined with monolingual classification. Considering that English and Chinese are very dissimilar linguistically, I expect the approach to work at least equally well for more closely related language pairs.

This work is still in progress. While my results are encouraging, more work is needed to make the CLGC approach more robust. At the moment, classification accuracy is low for problems with many classes. I plan to remedy this by implementing a hierarchical classification framework, where a text is assigned a broad genre label first and then classified further within this category.

Since TLA can only work on a sufficiently good initial labeling of target language texts, stable feature classification results have to be improved as well. To this end, I propose to focus initially on features involving punctuation. This could include analyses of the different punctuation symbols used in comparison with the rest of the document set, their frequencies and deviations between sentences, punctuation n-gram patterns, as well as the analyses of the positions of punctuation symbols within sentences or whole texts. Punctuation has frequently been used in genre classification tasks and it is expected that some of the features based on such symbols are valuable in a cross-lingual setting as well. As vocabulary richness seems to be a useful predictor of genres, experiments will also be extended beyond the simple inclusion of type/token ratios in the feature set. For example, *hapax legomena* statistics could be used, as well as the conformance to text laws, such as Zipf, Benford, and Heaps.

After this, I will examine text structure a predictor. While single line statistics and topic AP scores already reflect text structure, more sophisticated pre-processing methods, such as text segmentation and unsupervised PoS induction, might yield better results. The experiments using the tf-idf values of terms will be extended. Resulting features may include the positions of highly weighted words in a text, the amount of topics covered, or identification of summaries.

TLA techniques can also be refined. An obvious choice is to consider different types of features, as mentioned in Section 3.3. Different representations may even be combined to capture the notion of different communicative purpose, similar to the multi-dimensional approach by Biber (1995). An interesting idea to combine different sets of features was suggested by Chaker and Habib (2007). Assigning a document to all genres with different probabilities and repeating this for different sets of features may yield a very flexible classifier. The impact of the feature sets on the final prediction could be weighted according to different criteria, such as prediction certainty or overlap with other feature sets. Improvements may also be achieved by choosing a more reliable method for finding the most confident genre predictions as a function of the distance to the SVM decision boundary. Cross-validation techniques will be explored to estimate confidence values.

Finally, I will have to test the approach on a larger set of data with texts from more languages. To this end, I am working to compile a reference corpus for CLGC by combining publicly available sources. This would be useful to compare methods and will hopefully encourage further research.

Acknowledgments

I thank Bonnie Webber, Benjamin Rosman, and three anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Shlomo Argamon, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *Proceedings of First International Workshop on Innovative Information Systems*.

Nuria Bel, Cornelis Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In Traugott Koch and Ingeborg Slvberg, editors, *Research and Advanced Technology for Digital Libraries*, volume 2769 of *Lecture Notes in Computer Science*, pages 126–139. Springer Berlin / Heidelberg.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 582–590, Stroudsburg, PA, USA. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Douglas Biber. 1995. *Dimensions of Register Variation*. Cambridge University Press, New York.

Pavel Braslavski. 2004. Document style recognition using shallow statistical analysis. In *Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, pages 1–9.

Jebari Chaker and Ounelli Habib. 2007. 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. IEEE Computer Society.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

S. Feldman, M. A. Marin, M. Ostendorf, and M. R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784, Washington, DC, USA. IEEE Computer Society.

Aidan Finn and Nicholas Kushmerick. 2006. Learning to classify documents according to genre. *J. Am. Soc. Inf. Sci. Technol.*, 57(11):1506–1518.

Luanne Freund, Charles L. A. Clarke, and Elaine G. Toms. 2006. 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. ACM.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 553–560, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jade Goldstein, Gary M. Ciany, and Jaime G. Carbonell. 2007. 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. ACM.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic, June. Association for Computational Linguistics.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2000. A Practical Guide to Support Vector Classification.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, London, UK. Springer-Verlag.

Ioannis Kanaris and Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of the 19th IEEE International Conference on Tools with AI*, pages 3–10, Washington, DC.

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

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, Morristown, NJ, USA. Association for Computational Linguistics.

Yunhyong Kim and Seamus Ross. 2008. Examining variations of prominent features in genre classification. In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, pages 132–, Washington, DC, USA. IEEE Computer Society.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32:485–525, December.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 627–634, Morristown, NJ, USA. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Comput. Linguist.*, 37:385–393.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1118–1127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An em based training algorithm for cross-language text categorization. In *Proceedings of the Web Intelligence Conference*, pages 529–535.

Evan Sandhaus. 2008. New York Times corpus: Corpus overview. LDC catalogue entry LDC2008T19.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, March.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: Evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 3063–3070, Valletta, Malta, may. European Language Resources Association (ELRA).

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2000a. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814, Morristown, NJ, USA. Association for Computational Linguistics.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000b. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. 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: Volume 1 - Volume 1*, ACL '09, pages 235–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bonnie Webber. 2009. 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.

Maria Wolters and Mathias Kirsten. 1999. Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 142–149, Stroudsburg, PA, USA. Association for Computational Linguistics.