Robust Cross-Lingual Genre Classification through Comparable Corpora

Philipp Petrenz, Bonnie Webber

University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB, UK
p.petrenz@sms.ed.ac.uk, bonnie@inf.ed.ac.uk

Abstract

Classification of texts by genre can benefit applications in Natural Language Processing and Information Retrieval. However, a mono-lingual approach requires large amounts of labeled texts in the target language. Work reported here shows that the benefits of genre classification can be extended to other languages through cross-lingual methods. Comparable corpora – here taken to be collections of texts from the same set of genres but written in different languages – are exploited to train classification models on multi-lingual text collections. The resulting genre classifiers are shown to be robust and high-performing when compared to mono-lingual training sets. The work also shows that comparable corpora can be used to identify features that are indicative of genre in various languages. These features can be considered stable genre predictors across a set of languages. Our experiments show that selecting stable features yields significant accuracy gains over the full feature set, and that a small amount of features can suffice to reliably distinguish between different genres.

Keywords: Genre, Text Classification, Cross-Lingual, Comparable Corpora

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 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. Recent work by one of the current authors (Petrenz, 2012) demonstrated a new approach to genre classification that is cross-lingual (CLGC) in that it trains a genre classification model solely on labeled texts from one language L_S and then uses this model to predict the genres of texts written in another language L_T. As such, CLGC differs from both poly-lingual and language-independent genre classification in requiring *no labeled training data in the target language* (L_T). Instead, it 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.

What is new in the current work is that we show that there is even greater benefit to be gained from the use of a comparable corpus, comprising texts in several languages, in training a genre classifier for texts of the target language (L_T), different from any in the comparable corpus.

The paper is structured as follows: Section 2. describes prior work on genre classification, including our own. Section 3. describes our approach based on a comparable corpus, Section 4. describes the set of experiments we carried out and Section 5. discusses the results. Finally, Section 6. concludes with thoughts on taking this work forward.

2. Prior work

Work on automated genre classification was first carried out by Karlgren and Cutting (1994). Like Kessler et al. (1997) after them, they exploited hand-crafted sets of features, which are specific to texts in English. 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 include word based approaches (Argamon et al., 1998; Stamatatos et al., 2000b; Freund et al., 2006), PoS trigrams (Argamon et al., 1998) and PoS history frequencies (Feldman et al., 2009), image features (Kim and Ross, 2008), and character n-gram approaches (Kanaris and Stamatatos, 2007; Sharoff et al., 2010), all of which were tested exclusively on English texts. One of the few researchers to assess the language-independence of their approach 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, although in both cases his experiments were mono-lingual.

The only work on CLGC to date has been that of Petrenz (2012). This makes use of a set of hand-crafted stable features to bridge the language gap between English and Chinese, and then a bootstrapping technique to exploit unlabeled data in the target language. The approach performs equally well or better than a baseline in which texts are automatically translated and a mono-lingual genre classifier applied to the result. However, classifiers were only trained on a single language (English or Chinese), rather than exploiting the additional knowledge that might be available in comparable corpora. The notion of *stable features* used by Petrenz and Webber (2011) to specify features that are unaffected (i.e., stable) in the face of changing topics, could be applied here to specify features that are stable in the face of changing languages.

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 to some extent on statistical machine translation. This makes applications dependent on parallel corpora, which may not be available for poorly-resourced languages. It also suffers 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 comparable corpus 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, they also report considerable from using bilingual dictionaries.

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 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 ambiguity. Genres, on the other hand, are often classified using function words (Karlgren and Cutting, 1994; Stamatatos et al., 2000b) like "of", "it", or "in", which are next to impossible to translate out of context, especially when morphological differences between the languages can mean that function words in one language are 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 lexical identities in the two 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 — it is less indicative of genre: The text could be *inter alia* a news report, an editorial, a letter, an interview, a biography, or a blog entry, although correlations between topics and genres would probably rules out genres like instruction manuals or product reviews. However, uncertainty is still large, and Petrenz and Webber (2011) show that it can be dangerous to rely on such correlations.

3. Approach

The experiments described in Section 4. exploit features that are comparable across languages and a corpus of comparable texts across the same set of languages. We describe

both here before going into detail about the experiments.

3.1. 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 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, serve a similar role in two or more languages. An example would be type/token ratios, which, in combination with 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. If suitable features can be identified, CLGC may be considered a standard classification problem.

The approach we propose, like the one in (Petrenz, 2012), makes use of stable features that are mainly structural rather than lexical (cf. Section 4.2.) since the latter tend to vary by topic and are thus *unstable* with respect to genre (Petrenz and Webber, 2011). It does not assume the availability of machine translation, supervised PoS taggers, syntactic parsers, or other supervised tools. The only resources required are a way to detect sentence and paragraph boundaries in both source and target languages (e.g., a simple rule-based algorithm or an unsupervised method), and a sufficiently large, unlabeled set of target-language texts.

3.2. Hypotheses related to comparable corpora

The experiments described in Section 4. are designed to test two hypotheses: First, a comparable corpus of texts written in different languages but from the same distribution of genres can be used to train a classification model that is more robust for cross-lingual classification tasks than a model trained on a mono-lingual training set whose genre-related differences might not be the same as those in the target language. Adding more languages to the training set will result in a classification model which can separate genres in multiple languages. This makes it more likely to perform well on the target language.

The second hypothesis is that selecting features based on the cross-lingual performance within a separate comparable corpus can prevent a classifier from over fitting to the idiosyncrasies of the training language. Using a supervised feature selection technique on a set of several languages may yield features that have predictive power in more than one language. Cross-lingual genre classification can be regarded a special case of a domain adaptation problem, where feature selection techniques have been applied successfully before (Pan et al., 2010). Here, we apply a simple feature-ranking method, using information gain to determine the value of a feature to predict genres. Information gain is defined as

$$IG(Class, Feature) = H(Class) - H(Class|Feature)$$

where $H(X)$ is the entropy of variable X. A subset of features can then be obtained by choosing the top k features in this ranking of n features. While the availability of domain knowledge would allow this parameter to be set manually, here we determine it automatically, by finding the maximum cross-validation accuracy on the comparable corpus, where each fold corresponds to training on a single language and testing on all remaining languages. While this involves an exhaustive search over all possible values of k, using the information–gain ranking greatly reduces the possible numbers of feature subsets from $2^n - 1$ to n.

Note that, unlike the method in Gliozzo and Strapparava (2006), discussed in Section 2., the current approach does not require the comparable corpus to include texts from the target language.

4. Experiments

4.1. Data

Our experiments use three publicly available corpora, each of which included texts from a single genre written in several languages: the Reuters volume 1+2 corpus (Rose et al., 2002), the Europarl corpus (Koehn, 2005), and the JRC-ACQUIS corpus (Steinberger et al., 2006). All three corpora contain a large number of texts in Danish, English, French, German, Italian, Portuguese, Spanish, and Swedish. (Although all three also contain texts in Dutch, there are comparatively few Dutch texts in the Reuters corpus, so Dutch texts are not used in our experiments.) We reorganized the source corpora to obtain a comparable corpus that contains texts in eight languages and three genres: newswire texts, transcribed speech, and legal texts. Note that the corpus is *comparable* since it contains texts from a fixed set of genres, but not necessarily topics.

Since the source corpora are in different formats, some preprocessing was necessary. The XML markup was removed from the Reuters newswire texts, and only the contents of the tags `<headline>`, `<byline>`, `<dateline>`, and `<text>` were kept. Paragraph markers were kept in the text. The texts in the Europarl corpus were divided up by speaker: that is, we considered each speech to be a distinct document. We then removed the `<speaker>` tags, but kept the paragraph markers. We ignored missing speeches: The only requirement was that each text contains at least one token. The JRC-ACQUIS corpus comprises several sub-genres within the legal domain, including treaties, agreements and proposals. We therefore restricted ourselves to using documents from CELEX[1] sector 3 (legislation), as this is the largest group within the corpus. We extracted the text within the `<body>` tags, again keeping the paragraph structure intact.

All texts were segmented into sentences using the unsupervised *Punkt* algorithm (Kiss and Strunk, 2006) implemented in the NLTK (Bird et al., 2009) framework. Since

[1] CELEX (Communitatis Europeae Lex) is a database for European Union law documents. All texts in the JRC-ACQUIS corpus are classified by CELEX sector and document type.

Europarl and JRC-ACQUIS are parallel corpora, we ensured that no translation of the same text was used in any two sets in our experiments. For Europarl texts, we always used the language that the speech was made in, which is indicated in the meta-data. For JRC-ACQUIS, the choice was random, since the corresponding journal is published in all European languages simultaneously.

Splitting the legislation texts of the JRC-ACQUIS yielded 1,942 documents in each of the eight languages. To keep the genre distribution in our corpus balanced, we randomly sampled 1,942 documents from both the Reuters and the Europarl corpora. The resulting eight sets each contained 5,826 texts from a single language. A list with identifiers of the texts we used for our experiments can be found on our website[2], along with scripts to extract and clean texts from the source corpora mentioned before. There is, to the best of our knowledge, no publicly available corpus containing texts written in several languages from a common set of genres. Therefore, the method described above can be seen as a suggestion to facilitate research into cross-lingual genre classification and provide a common data set to compare approaches.

4.2. Features

We hypothesized that our experiments could produce a set of features that would serve as stable genre predictors across a range of languages, not just for a single one as in (Petrenz and Webber, 2011). To this end, we selected as candidate features, ones that would hold for texts in many languages. These included the frequencies of 32 common punctuation symbols, as well as simple text statistics (document length, sentence length mean and variance, paragraph length mean and variance, single-sentence-paragraph count and frequency over all sentences, single-sentence-paragraph distribution value, type/token ratio[3], and number/token ratio).

Single-sentence paragraphs are typically headlines, datelines, author names, or other structurally interesting parts. Their distribution value indicates how evenly they are distributed throughout a text, with high values indicating single-sentence paragraphs predominantly occurring at the beginning and/or end of a text. It is computed by averaging over the distance of all such paragraphs from the $(n/2)$th token in a text of length n.

To this set, we added features based on concepts from information retrieval. We used tf-idf weighting and marked the ten highest-weighted words in a text as relevant. We then treated the 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 us to compute an average precision (AP) value, which indicates the distribution of relevant words. A high AP score means that the top tf-idf weighted words are found predominantly in the beginning of a text. This follows the intuition that genre conventions may influence the location of important content words within a text. For example, Thomson et al. (2008) found that news articles in English, French, Japanese, and Indonesian are all structured according to the inverted pyramid principle (Pöttker, 2003), where important information appears in the beginning, followed by background information and other less important material. In addition, for each of the same ten words, we added its 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).

Finally, we also added the frequencies in the text of the 25 most common words in the respective language. Common word frequencies have been shown to have discriminative power in mono-lingual genre classification tasks (Stamatatos et al., 2000a). However, since the i^{th} most common word in language A differs semantically from the i^{th} most common word in language B, we expected these features to be of little value for a cross-lingual task and that they might have a negative impact on prediction accuracies. We included them in the feature set to find out whether this is the case and if so, whether they are filtered out in the feature selection process of our method.

The final set comprised 78 features, three of which were discarded, as they had zero values for all texts in one or more languages. After extracting the full set of features from the texts, their values were standardized. This was achieved by subtracting from each feature value the mean over all texts and dividing it by the standard deviation, so that each feature had zero mean and unit variance. Standardization was done separately for each language, to balance out differences between them. Because this step exploits only unlabeled data in order to make feature values more comparable (i.e., it does not require genre labels), standardization can be applied to the target language feature set, as long as enough target language texts are available.

4.3. Experimental Frameworks

To generate baselines, we evaluated classification models which were trained on one language and tested on another. To this end, we trained a separate Support Vector Machine (SVM) model for each of the eight mono-lingual sets, using all 75 features. Each model was then tested on the seven languages that were not used to train it. This performance is achievable without the use of a comparable corpus.

To exploit the genre labels in more than just one language, we then merged the representations of seven language sets into a single training set, holding one language back for testing. An example of this is illustrated in Figure 1. Naturally, the merged multi-lingual training set contained seven times as many texts as any mono-lingual baseline. Since supervised classification results tend to improve with larger training set sizes, we removed this bias by splitting the merged set into seven disjoint training sets, keeping the language and genre distributions intact. Thus, for each target language, the SVM model was trained seven times and evaluated by computing the average accuracy.

To evaluate whether a comparable corpus can be used to identify stable features from a set of candidates, even if the

[2] http://homepages.inf.ed.ac.uk/s0895822/BUCC2012/

[3] As the type/token ratio is known to correlate with document size, we recorded the ratio for a sliding window of 300 tokens. For shorter texts, this was estimated by computing a percentage of the average type/token ratio at the end of the text and multiplying this with the average value for 300 tokens.

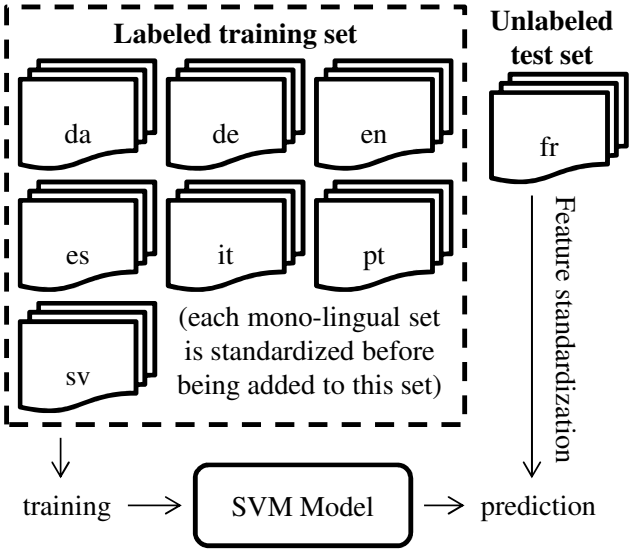

Figure 1: First experimental framework, example with French test set. Set of seven languages is used to train classification model. The full set of features is used.

set does not include texts written in the source or target languages, we conducted a second experiment. Here, we ranked features using a set of six languages. (Features were ranked by their information gain, as explained in Section 3.) Then, 6-fold cross-validation was used to determine the threshold parameter k. The feature sets of the seventh and eighth languages were reduced to the resulting subset, and then used for training and testing respectively. An example of this is illustrated in Figure 2. When compared with the baseline, the results will indicate to what extent feature selection on a separate comparable corpus can benefit cross-lingual genre classification applications.

5. Results and Discussion

Table 1 shows the classification accuracies for the 56 single language training experiments (i.e. baseline performances), as well as the accuracies yielded by the combined multi-lingual training set. The last row corresponds to the experimental framework illustrated in Figure 1. *For all eight target languages, accuracy based on the multi-lingual training set exceeded accuracy based on any of the seven mono-lingual baselines.* This significant (sign test; $p < 0.01$) improvement indicates that the knowledge represented by genre labels in different languages can be exploited to build robust cross-lingual genre classification models.

In the second experiment, we performed feature selection using the six languages that remained after choosing one language for training and a second for testing (cf. Figure 2). Table 2 shows the gains and losses in prediction accuracy when using only the top k features, as compared to the full feature set. For the 56 tasks, k ranged between 13 and 23, with the majority between 13 and 15. Most classification models benefited from this feature selection step. Although in some cases accuracy deteriorated, performance based on the reduced feature set was significantly better ($p < 1e^{-8}$), according to the sign test. Since these subsets were identified using a supervised ranking technique, the results in

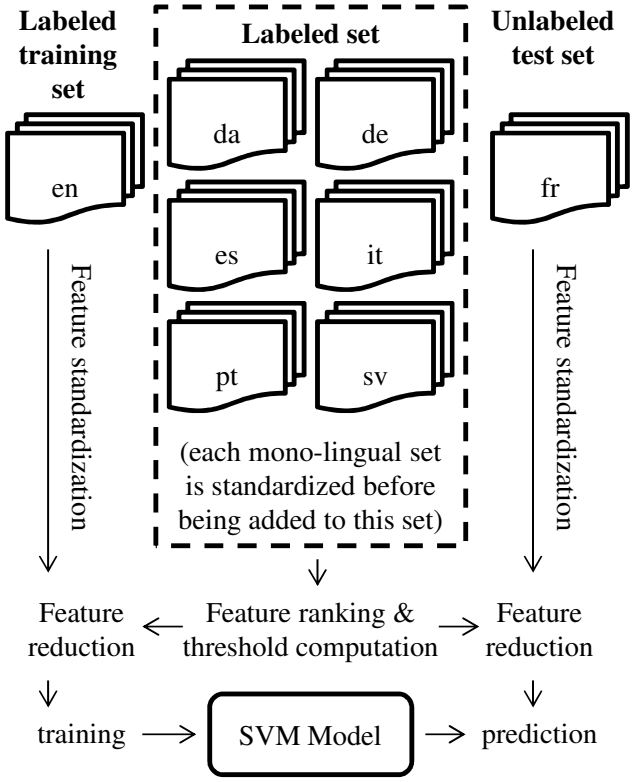

Figure 2: Second experimental framework, example with English training set and French test set. Set of six languages is used to rank features and determine the threshold k. The languages used for training and testing are not represented in this set.

Table 2 suggest that comparable corpora can also be used to identify features with strong discriminative powers for cross-lingual genre classification tasks. They also show that this is possible even if neither the source nor the target language is included in the comparable corpus.

An important question is whether the algorithm can find a good value for the threshold k. Using the results in Table 2, we picked the combination that gained the most from the feature reduction (training on Spanish texts, testing on German texts: es→de) and the one that suffered the most (training on Portuguese texts, testing on English texts: pt→en). We also picked the combination that used the largest number of features (training on Danish texts, testing on Italian texts: da→it). For these three combinations, we recorded the performance when removing features from the set one by one, starting at the performance of the full set shown in Table 1. Figure 3 illustrates the prediction accuracies as functions of the number of features used. The arrows indicate the threshold chosen by the algorithm. The es→de classifier performs clearly better when selecting between 12 and 22 features from the ranking. The threshold (14) happens to be a very good choice and yields significant[4]

[4]We assume that the number of misclassifications is approximately normally distributed with mean $\mu = e * n$ and standard deviation $\sigma = \sqrt{\mu * (1 - e)}$, where e is the percentage of misclassified instances and n is the size of the test set. The 95% confidence interval is then $\mu \pm 1.96 * \sigma$.

	da	de	en	es	fr	it	pt	sv	μ
Danish (da)	—	.959	.951	.961	.930	.965	.937	.971	.953
German (de)	.943	—	.925	.934	.897	.957	.933	.954	.935
English (en)	.948	.942	—	.961	.934	.962	.942	.972	.952
Spanish (es)	.960	.920	.952	—	.946	.963	.927	.973	.949
French (fr)	.961	.952	.965	.974	—	.973	.940	.967	.962
Italian (it)	.959	.963	.955	.962	.948	—	.949	.953	.956
Portuguese (pt)	.955	.948	.945	.954	.928	.954	—	.961	.949
Swedish (sv)	.965	.949	.948	.963	.911	.947	.928	—	.944
Multi-lingual	**.979**	**.968**	**.973**	**.979**	**.967**	**.980**	**.971**	**.986**	**.975**

Table 1: Prediction accuracies for the cross-lingual genre classification tasks. Rows 2-9 denote the training language, Columns 2-9 denote the testing language. The accuracies in row 10 were achieved by training the model on the seven languages which it was not tested on. Column 10 contains the average of each row. The best accuracy for each column is highlighted.

	da	de	en	es	fr	it	pt	sv
Danish (da)	—	+.005	+.013	+.009	**+.033**	+.011	+.013	−.009
German (de)	+.015	—	+.016	**+.031**	**+.035**	+.009	−.002	−.001
English (en)	**+.021**	+.018	—	**+.022**	**+.040**	+.010	+.005	+.010
Spanish (es)	+.005	**+.062**	**+.021**	—	**+.024**	+.017	**+.035**	+.004
French (fr)	+.015	+.016	+.011	+.017	—	+.000	+.018	+.010
Italian (it)	−.003	+.017	+.011	**+.025**	+.019	—	+.010	+.017
Portuguese (pt)	**+.024**	−.001	**−.026**	**+.025**	+.011	**+.022**	—	+.011
Swedish (sv)	+.009	+.011	**+.025**	+.019	**+.061**	**+.030**	+.017	—

Table 2: Difference in prediction accuracy after feature selection when compared to the corresponding results in Table 1. As in Table 1, rows 2-9 denote the training language, columns 2-9 denote the testing language. Differences of more than .02 are highlighted.

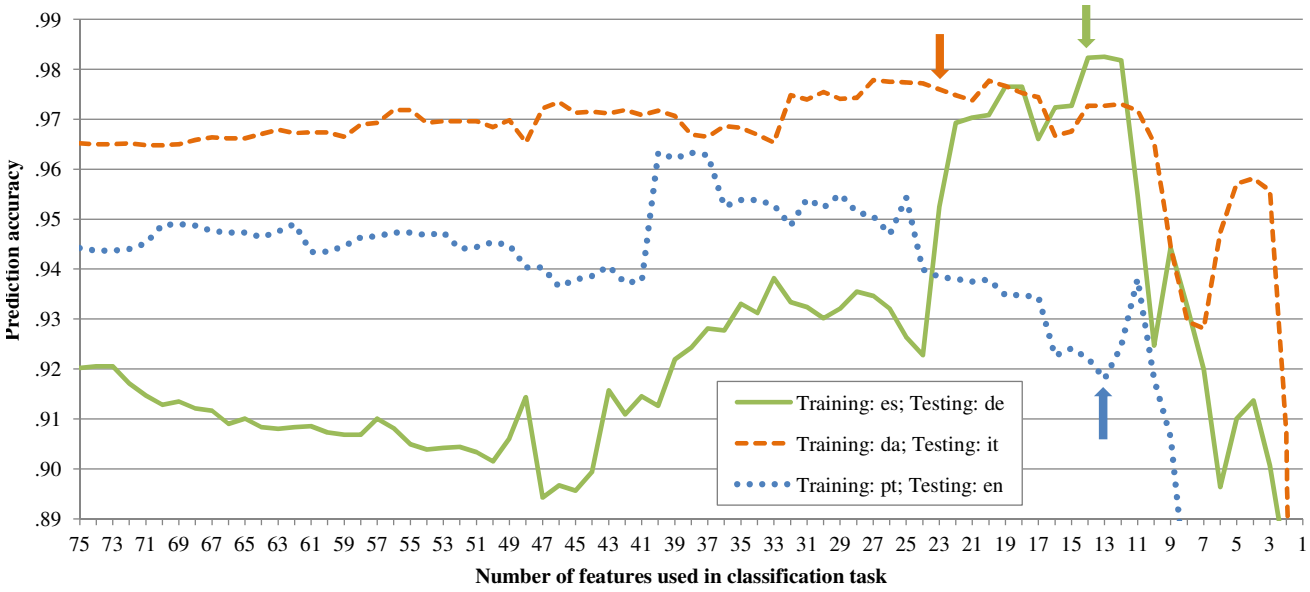

Figure 3: Prediction accuracies for the classifier trained on Spanish texts and tested on German texts (green, continuous line), the classifier trained on Danish texts and tested on Italian texts (orange, dashed line), and the classifier trained on Portuguese texts and tested on English texts (blue, dotted line). For all three classifiers, the accuracy achieved is given as a function of the number of top rank-ordered features used. The arrows denote the automatically determined number of features for these tasks (14, 23, and 13 respectively).

improvement over the baseline. The performance of the pt→en classifier stays mostly within the confidence interval of the baseline, although it clearly outperforms it for feature set sizes 37-40. Accuracy drops and falls below baseline level for fewer than 20 features. Here, the chosen threshold (13) is too low, since this classifier would benefit from additional features. The da→it classifier benefits slightly but significantly from a reduced feature set until accuracy drops sharply for less than 11 features. The threshold (23) is a good choice, although the exact value is less crucial than for the es→de and pt→en classifiers, in that small variations would have little effect on the result.

The majority of positive results in Table 2 suggests that the chosen threshold k is usually suitable to improve the prediction accuracy. In line with that, Figure 3 shows that the algorithm picks a near-optimal value for k for some training/testing combinations. However, the example of the pt→en classifier shows that this is not necessarily the case. On the other hand, it also illustrates that even where feature reduction leads to deteriorating performances, this could be due to a sub-optimal threshold choice. This is clearly the case for the pt→en classifier, where a set of 37-40 features would have improved baseline performance significantly. Optimizing the computation of this threshold, possibly by exploiting the unlabeled data in the target language, would be an interesting problem for future work.

In order to get an idea of the types of features which are typically selected, we ranked them by their information gain using a combined set that included all eight languages. The top 15 features are listed below. Note that the information gain of a certain feature varies depending on the exact set of languages used. However, the ranking in our experiments was fairly stable and the top 15 features rarely differed from the ones below.

1. Single sentence paragraph count
2. Single sentence paragraph/sentence ratio
3. Paragraph length mean
4. Closing parenthesis frequency
5. Opening parenthesis frequency
6. Number frequency
7. Forward slash frequency
8. Single sentence distribution value
9. Colon frequency
10. Sentence length mean
11. Top 10 tf-idf average precision
12. Type/token ratio
13. Document length
14. Paragraph length standard deviation
15. Hyphen frequency

As expected, none of the 25 common-word frequency features was ranked among the top 15. This finding reinforces our intuition that common-word frequencies are useful in mono-lingual genre classification tasks, but harmful to cross-lingual models. While feature 11 above seems to have discriminative power, none of the other tf-idf based features is in the above list. This is likely due to the fact that these features have informative value only in combination with each other. However, information gain ranking evaluates only single features, not sets. A subset based selection approach might be more suitable to identify their strengths (cf. Guyon and Elisseeff (2003)).

Another observation is that features based on paragraph length dominate the ranking. This is likely due to the way texts of the three different genres are structured. Legal texts tend to have very short paragraphs, sometimes consisting of a single token (Example 1 below). Newswire paragraphs are mostly only one or two sentences long, but typically contain more than one token each (Example 2). In transcribed speech (Example 3), paragraphs tend to be longer.

1. **Legal text:**

```
<p>Commission Regulation (EC) No
1135/2006</p>
<p>of 25 July 2006</p>
<p>amending the import duties in the
cereals sector applicable from 26 July
2006</p>
<p>THE COMMISSION OF THE EUROPEAN
COMMUNITIES,</p>
<p>Having regard to the Treaty establishing
the European Community,</p>
```

2. **Newswire text:**

```
<p>The KFX top-20 index lost 0.20 point to
close at 126.29 in overall bourse turnover
of 1.944 billion crowns.  The KFX December
future rose 0.65 point to 126.40 with
10 contracts each worth 100,000 crowns
traded.</p>
<p>Novo Nordisk attracted a good deal of
attention following its announcement of
400 million crown rationalisation cuts for
1997 and 1998, finishing the day a solid 21
crowns up at 954.</p>
```

3. **Transcribed speech text:**

```
<p>Naturally I understand the honourable
Member's concern.  As far as the Commission
is concerned, we have never supported
financially the production or distribution
of school textbooks nor the preparation
of school curricula.  Assistance to the
educational system is focused mainly on
infrastructure, equipment for schools and
direct assistance for school expenses, for
example, salaries.  No request has ever
been made by the Palestinian Authority to
the Commission to finance school curricula
and textbooks.</p>
```

6. Conclusion

Our experiments with eight European languages show that cross-lingual genre classification (at least within these languages) is possible with a minimum of knowledge about the target language. Some features, which are easily extracted from plain texts, can be considered stable predictors of genre across languages. Applications exploiting such features may reduce the need for resources such as parallel corpora or supervised parsers in the target language. We

demonstrate that comparable corpora can be used to automatically identify stable features from a set of candidates. These can help to improve prediction accuracy, even when used in tasks with separate training and target languages.

We also show that using more than one language in the training set can prevent a cross-lingual genre classification model from over fitting the differences between genres in one language and thus improve its robustness. Exploiting a comparable corpus by either identifying stable features or using multi-lingual training sets significantly beats the baseline performances in our experiments.

Finally, we propose a method to construct a comparable corpus including legal texts, newswire texts, and transcribed speeches in eight European languages by remodeling three publicly available corpora. This can be used by researchers to compare cross-lingual genre classification methods.

7. 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.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

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.

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.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

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.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

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.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 751–760, New York, NY, USA. ACM.

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 clas-

sification of text genres. *Computational Linguistics*, 37(2):385–393.

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

Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid — when and why did it appear? *Journalism Studies*, 4(4):501–511.

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.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In Jude W. Shavlik, editor, *Proceedings of the Third Internaional Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria.

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.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. September.

Elizabeth A. Thomson, Peter R. White, and Philip Kitley. 2008. objectivity and hard news reporting across cultures. *Journalism Studies*, 9(2):212–228.

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.