Cross-Lingual Genre Classification for Closely Related Languages

Dirk Snyman[1], Gerhard B. van Huyssteen[1] & Walter Daelemans[2]
[1]Centre for Text Technology (CTexT®), North-West University, Potchefstroom, South-Africa
[2]CLiPS-CL, University of Antwerp, Antwerp, Belgium
[1]{Dirk.Snyman;Gerhard.Vanhuyssteen}@nwu.ac.za
[2]Walter.Daelemans@ua.ac.be

Abstract— **Resource-scarcity is a topic that is continually researched by the HLT community, especially for the South-African context. We explore the possibility of leveraging existing resources to help facilitate the development of new resources for under-resourced languages by using cross-lingual classification methods. We investigate the application of an Afrikaans genre classification system on Dutch texts and see encouraging results of 63.1% when classifying raw Dutch texts. We attempt to optimise the performance by employing a machine translation pre-processing step, boosting performance of the Afrikaans system on Dutch data to 67.2%. Further investigation is required as we conclude that the robustness of the Afrikaans genre classification system needs improvement.**

Keywords – cross-lingual; genre classification; resource scarce languages; closely related languages; Afrikaans; Dutch

I. INTRODUCTION

When working with the indigenous South-African languages, one is always faced with resource scarcity. In [4] we describe the automatic classification of genre in a resource scarce environment, where experiments were done for six of the indigenous South-African languages. We concluded that the sparseness of available training data, data due to the resource scarceness of the languages in question, causes erratic results (due to overfitting) when using machine learning techniques to classify the genre of a text and that techniques to alleviate these symptoms should be investigated [4]. Therefore, this article investigates the application of technology recycling for the use in genre classification systems.

By adapting existing technologies for closely related languages, the development of resources for resource-scarce languages can be fast-tracked. This process is known as technology recycling [1]. Given a technology, created for a well sourced language *L1*, which is needed in another language *L2* which is resource-scarce, it would be faster and cheaper to adapt the *L1* technology for *L2* than to redevelop the *L2* technology from the ground up [1].

We investigate the effect of the language differences on genre classification and investigate methods by which existing technologies for a well resourced language could be leveraged for a resource-scarce language. We evaluate a genre classification system when classifying a strange language and then implement approaches to enhance its performance. Dutch and Afrikaans have been used successfully in technology recycling experiments because these two languages are similar enough [1][5] and as a result thereof, Dutch and Afrikaans will be used as the languages in question for this article.

We first give an overview of related research pertaining to cross-lingual genre classification in Section 2, after which we describe the experimental setup in Section 3 and the results of the experiments are shown and discussed in Section 4. We conclude this article in Section 5 and we give view to future work.

II. RELATED WORK

Relatively little research is available for the evaluation of a genre classification system that is based on one language, on data that is written in another language. The first research on actual "Cross-Lingual Genre Classification" was made available by Petrenz [2] although cross-lingual methods have been explored for other text classification tasks (other than genre, that is) as will be described later on. Petrenz [2] states that a lot of research aims to develop language independent approaches to text classification rather than cross-lingual approaches, but are seldom able to give definitive empirical proof that these approaches are actually viable. As an example, Petrenz [2] recalls one of the few research reports on genre classification on more than one language (English and Russian) done by Sharoff [3],[3] which suggests that encoding part of speech (POS) data and combining that with variation of common words as feature sets, will be a viable language independent approach. According to Petrenz [2], the claim of this approach being a language independent one is false as the construction of these features are based on the language they are constructed for, although constructed in the same manner for each language. Language neutrality of said approaches can thus be described as a "holy grail"-type pursuit as language specific information will always be implicitly included when constructing these kinds of feature sets. The experiments are also conducted on a per language basis, i.e. the English genre classification system is only evaluated on unseen English data and the Russian system on Russian data so "real" language independent performance is not evaluated. Petrenz [2] chooses to call these features, "stable" features for cross-lingual experiments as they are easily extracted for any language without any prior language knowledge or expertise and do not rely on existing technologies like POS-taggers. How can cross lingual genre classification then be done, if capable language

independent approaches for direct cross-lingual classification do not exist?

To bridge the gap between languages, cross-lingual methods often rely on target language adaptation [2]. Target language adaptation can be achieved by making use of techniques like syntactic reordering [1], morphological adaptations [1], lexical transfer [1] and full- or partial translation [6] to name but a few. Translation is the method which is favoured by Bel *et al*. [6] and by being one of the earliest reports on cross-lingual text classification (for English and Spanish), has set the tone for subsequent research to follow and has had a great influence on the direction that cross-lingual text classification experiments have taken [2], i.e. using machine translation as a pre-processing step when classifying another language. Bel *et al*. [6] state that, when attempting to classify an *L2* text with a *L1* classifier, the discrepancy between the source and target language vocabularies causes incompatibility between the classifier model and the test cases, resulting in very low classifier performance. This discrepancy can be (at least partially) solved with translation by using one of the following translation strategies [6]:

- Terminology translation: Terminology lists are compiled in the classifier language on a per class basis and only the terms which are deemed relevant (by some or other measure, e.g. information gain) to the classification of the specific class are translated in the target language (*L1*).
- Profile based translation: Only the words that occur in the training data for the classifier are translated in the target language (*L1*).
- Full text translation: The entire text is translated in the target language.

Bel *et al*. [6] however criticise the full text translation approach due to the high financial costs and time consuming nature of translation and the questionable translations rendered by machine translation. Petrenz [2] however reports good results on full text evaluations with machine translation systems, as the target language only has to be adapted and does not need to be translated in its entirety. This also compares to the findings of Pilon *et al*. [1] where simple lexical conversion is used in the same manner for POS-tagging experiments with Afrikaans and Dutch, yielding good results. Machine translation should therefore be more than sufficient to bridge the gap in vocabularies for the purpose of this research.

A prerequisite for cross-lingual genre classification using machine translation is that there is a certain set of minimum resources that have to be available for both *L1* and *L2*:

- An *L1* text classification system (i.e. a classifier model trained with genre-specific information);
- A compatible L2 test corpus (i.e. a corpus that is annotated with the same genre specific information as the *L1* classifier model, or which has genre annotations which can be adapted to match *L1*); and
- A machine translation or similar system for target language adaptation.

The next section describes the experimental setup for testing the abovementioned combination of resources for cross-lingual genre classification.

III. Experimental Setup

A. Genre classification system

For the purposes of this article we will use the Afrikaans genre classification system based on the Multinomial Naive Bayes (MNB) algorithm as described in [4] to classify previously unseen Dutch texts according to their genre. The roles of the two languages for traditional technology recycling experiments are reversed in such a way that Afrikaans acts as the well resourced language and Dutch acts as the resource scarce counterpart. This is because a Dutch genre classification system that matches the scope of the Afrikaans classifier could not be found to be used experimentally. A genre classification system with competitive results is already readily available and because Dutch corpora are generally genre annotated in some way and it would be easier to map the genre annotations to the Afrikaans classifier. Petrenz [2] shows the results for cross-lingual genre classification experiments for Spanish and English. From the results reported for these two languages, it can be seen that the directionality of such experiments do not affect the outcome thereof as the reported results for both directions are quite similar.

WEKA [10] is a suite of machine learning algorithms offered as an experimental environment. It holds the benefit of providing access to pre-processing scripts for text to vector conversion with a range of feature extraction options. The Dutch data pre-processing will be done in WEKA as well as the evaluation of the Afrikaans genre classification system, classifying Dutch data.

B. Data

The Afrikaans genre classifier is based on texts that have been extracted from public domain government websites as described in [4]. The classes for the genre classification system mentioned in [4] have been compacted to three classes in order to deal with the sparseness of class representations due to resource scarcity discussed in [4]. Afrikaans showed a good coverage of all the previous classes but for compatibility with the other indigenous languages in planned future work, the shift to a three class genre classification scheme will be used with Afrikaans already.

Class name	**# Training instances**
Expressive	229
Appellative	439
Informative	536
Total	1204

Table 1. Genre classes and instances per class: Afrikaans

These three classes have been adopted from Wachsmuth and Bujna [8] which identify the three classes as follows:

- Personal (expressive). Text that aims to express the personal attitude of an individual towards a product of interest.
- Commercial (appellative). Text that follows commercial purposes with respect to a product of interest.
- Informational (informative). Text that reports on a product of interest in an objective and journalistic manner.

The resulting Afrikaans training data is composed as shown in Table 1. The number of available training instances for each class differs, but the best results for the Afrikaans genre classifier are seen when using all of the available data, when compared to balancing the classes. The best results noted for the Afrikaans genre classifier, based on cross validation experiments, are a precision of 0.931, a recall of 0.930 and a resulting *f*-score of 0.929.

For the Dutch test corpus an excerpt from the original LASSY corpus [7] is used. An official extract from the corpus which is known as LASSY Small is a million word corpus, annotated with syntactic information, as well as POS-tags and lemmas. Genre annotations are also present, but are a little harder to come by. The genre annotations are not explicitly mentioned in the corpus or corpus meta data, but there is mention of the genres in LASSY in the project documentation[1]. The genre classes can be identified by matching the classes mentioned in the documentation to the file names of the corpus' .xml files. The corresponding files are then mapped to the abovementioned genre classes. The initial composition for the Dutch testing corpus is shown in Table 2. There are some of the LASSY corpus files for which a genre could not be identified from the corpus documentation and these files were therefore excluded when compiling the Dutch test instances.

Class name	**# Training instances**
Expressive	75
Appellative	546
Informative	107
Total	728

Table 2. *Genre classes and instances per class: Dutch*

The abovementioned datasets will be encoded in standard binary word occurrence vectors, also known as a bag of words approach (BOW). BOW is one of the stable features for cross-lingual genre classification as described by Petrenz [2].

C. Machine Translation System

For the machine translation component the "Dutch to Afrikaans Converter" (D2AC) by Van Huyssteen and Pilon [5] will be used. D2AC is a rule-based machine translation system based on the orthographic, morphosyntactic and lexical differences between Afrikaans and Dutch. D2AC is not a complete machine translation system as it only applies lexical transfer because it was developed with technology recycling as motivation. They report a precision of 71% for word-level evaluation and a BLEU score of 0.2519 for D2AC [5]. The experiments will be repeated with the Dutch-Afrikaans Google Translate (GT) as machine translation system to verify the results obtained for D2AC

D. Evaluation

The evaluation method used is *n*-fold cross validation (*n*=10), with 90% of the data used for training and 10% of the data used for testing. The standard information retrieval measures, Precision, Recall and F-measure are used to evaluate the effectiveness of classification for the system [9].

Class C_i		**Actual Class**	
		Yes	No
Classifier class	Yes	TP	FP
	No	FN	TN

Table 3. Standard information retrieval methods[9]

The formulas for Recall, Precision, and F-Measure of Ci (see Table 3) are shown in the following three equations (1)(2)(3), Where TP = True Positive, TN = True Negative, FN = False Negative and FP = False Positive classifications.

$$\textbf{R (Recall)} = \textbf{TPi / TPi + FNi}, \quad (1)$$

$$\textbf{P (Precision)} = \textbf{TPi / TPi + FPi}, \quad (2)$$

$$\boldsymbol{f_1}\ \textbf{(}\boldsymbol{f}\textbf{-Measure) = 2(R*P) / (R+P)} \quad (3)$$

E. Baseline System

As a baseline for the experiments a random class baseline (representing a one out of three chance of guessing the correct class) is used. This would result in a 33.33% chance of choosing the correct class. This does, however, not reflect the class distributions. When taking into account the difference in the training instances available to each class, the random baseline can be adjusted to 36.7%. A most frequent class baseline of 44.52% (obtained by dividing the number of instances for the most frequent class by the total number of instances. i.e. always selecting the "Informative" class) is also used.

In the next section the results for the following set of experiments will be discussed:

- Classifying unseen Dutch instances with an Afrikaans genre classifier;
- Translating the Dutch instances to Afrikaans with D2AC and GT and reclassifying the now Afrikaans(-like) instances; and
- Compare the results of these two experiments with each other and with the baselines set above.

[1] http://www.let.rug.nl/~vannoord/Lassy/deliverable1-1.pdf

IV. RESULTS

A. Classifying Dutch instances with an Afrikaans genre classifier

When classifying the unseen Dutch test instances (where the genre annotations extracted from the LASSY project documentation) with the Afrikaans genre classification system, we see some rather disappointing results where the classification precision of 42.3% (Table 4) exceeds the random baseline of 36.7%, but doesn't exceed the most frequent class baseline of 44.52%. But, Bel *et al.* [6] states a precision of 10.75% when evaluating English and Spanish in a pure cross-lingual text classification situation, which puts the performance of pure cross lingual systems in some perspective. They attribute the overlap in the two languages causing the 10.75% precision, to proper nouns and acronyms which are shared between the training and test sets. One would, however, expect a much larger overlap between languages which are said to be closely related, and would therefore expect a somewhat higher score, taking into account we already see an improvement of 28.45% over the English-Spanish results. When translating the Dutch to Afrikaans (as in the next section) only a 3.1% increase in precision was noted. This however didn't hold to Bel *et al.*'s [6] findings of accuracies ranging from 53.8% to 84.5% for translated cross-lingual classifications. These discrepancies prompted a review of all the variables which have an impact on the results.

Language	Precision	Recall	*f*-Measure
Dutch	0.392	0.281	0.277

Table 4. Initial results for Afrikaans classifier and Dutch data

When taking a closer look at the Dutch texts, it was noted that some of the texts which were annotated with the extracted genre classes, weren't supposed to be annotated as such. It was noted that the classes were very noisy and it would need to be remapped to ensure the class representations were indeed representative of the said class. When the genre classes were extracted from the LASSY documentation, there was no indication of how the classes in LASSY were defined, seeing as the genre annotations for LASSY aren't an explicit part of the corpus, it wouldn't be needed to include this kind of descriptions. It is suspected that the interpretation of what a specific genre class constituted differed from what the class in LASSY actually constituted. The Dutch training set was therefore reclassified by hand, making sure the instances were attributed to the correct class. The reclassified test set is presented in Table 5

Class name	# Training instances
Expressive	321
Appellative	391
Informative	16
Total	728

Table 5. Genre classes and instances per class: Dutch reclassified

The cross-lingual Dutch-Afrikaans experiment was repeated, this time with encouraging results (see Table 6). We now see a precision of 63.1%, which exceeds both the random baseline of 36.7% as well as the most frequent class baseline of 44.52% and also satisfies Bel *et al.*'s [6] findings for translated cross-lingual classification, even without being translated yet. In the following section, the results for the translated cross-lingual classification are presented.

Language	Precision	Recall	*f*-Measure
Dutch	0.631	0.284	0.318

Table 6. *Results for Afrikaans classifier and reclassified Dutch data*

B. Classifying translated Dutch instances with an Afrikaans genre classifier

When translating the data with both D2AC and GT we see an increase in the performance, which is above the baselines that were set and even further approximates the highest result of 84.5% as reported by Bel *et al.*'s [6] for translated cross-lingual experiments. The results are shown in Table 7.

Language	Precision	Recall	*f*-Measure
D2AC: Dutch	0.660	0.385	0.438
GT: Dutch	0. 672	0. 429	0. 485

Table 7. Results for Afrikaans classifier and translated Dutch data

Table 8 shows the confusion matrix for the best results seen in Table **7**, i.e. the Dutch test set, translated with GT and classifier with the Afrikaans genre classification system. The classes seem to be confused across the board with the highest confusion noted between Expressive texts being classified as Appellative and Informative texts being classified as Expressive texts. This could be due to erroneous translations or the choice of words for a translation which could be non-prototypical of the class representation of the classifier, which could lead to a misclassification.

		Classified class		
		a	**b**	**c**
Actual class	**a = Appellative**	324	45	22
	b = Expressive	173	50	98
	c = Informative	4	8	4

Table 8.Confusion matrix for GT: Dutch and Afrikaans classifier

The gain in precision which is seen from translating the text is still substantially lower than the gain seen by Bel *et al.* [6]. The results obtained for D2AC and GT seem to be consistent, with only a small variation in the performance being noted. One possible explanation for this occurrence could once again be found in the differences and similarities of the vocabularies of the Afrikaans training data and the Dutch test instances. Using WEKA [10] the words were analysed to ascertain their contribution to classification or in other words, how informative each word is with respect to the classification task. This was done by ranking the words according to their information gain (IG). The top 10 Dutch

words with the highest IG are listed in *Table 9*. These words all have counterparts in Afrikaans which effectively means that these words do not have to be translated because they exist in both vocabularies. The translation of the text therefore only improves the vocabulary compatibility on words which do not contribute very much to the classification and because of that, the gain seen when translating the text before classification is minimal.

Dutch	IG
nog	0.302910
is	0.295780
maar	0.291490
dit	0.283370
die	0.280110
van	0.244970
al	0.239600
dat	0.239070
was	0.235000
wat	0.228150

Table 9. Information gain of Dutch words

V. Conclusion and Future Work

In this article we investigated the application of an Afrikaans genre classification system on Dutch data. We reported on a precision of 63.1% on the aforementioned. We then experimented with machine translations of the Dutch data as a pre-processing step, by using a Dutch to Afrikaans lexical convertor (D2AC) and the Dutch-Afrikaans Google translate, obtaining accuracies of 66% and 67.2% respectively. This kind of technology recycling could be used to help in bootstrapping training data for an under-resourced language, but to be used as a core technology in real world systems, further development is needed to improve the performance. Machine translation systems for some of the indigenous languages have already been developed in the Autsumato project [11] and further development is taking place to further the development for more of the indigenous languages. As these resources become available, the approach described in this research could be tested for these languages. We note however that there are a range of problems that arose from applying cross-lingual genre classification between Afrikaans and Dutch. The compatibility of the training set in the well resourced language and the test set in the underrepresented language is of cardinal importance. By ensuring compatibility for the Afrikaans and Dutch data sets, we noted an increase in performance of 26.8%. The Dutch data was classified by hand which translates to a time consuming, as well as costly process. We therefore propose further research in the compatibility of genre classified corpora, with special regard to automatic methods.

We noted only a small improvement of the performance when using machine translation as a pre-processing step which seems to be contrary to the findings of Bel *et al.* [6] and Petrenz [2]. The reason why only a small increase in performance was seen was noted to be due to an overlap in the vocabularies of Dutch and Afrikaans. This however should intuitively mean a better compatibility but seems to hamper the possibility for growth, rather than improve it. This brings the robustness of the genre classification system into question. Most of the words that overlap are function words that do not necessarily contribute to knowledge about a specific class and is falsely deemed informative. We would like to investigate the use of stop word lists (i.e. lists of words to exclude from training data) and other approaches in an attempt to improve the robustness of the system and eliminated the system's reliance on falsely informative features. Experiments with other machine learning approaches (like support vector machines) could also be performed to determine the suitability of MNB for this task. Initial experiments could also be performed for indigenous language pairs, implementing human translations (where machine translation is not yet available) and one of the less intensive translation strategies as mentioned in Section II.

VI. Acknowledgements

All fallacies remain our own.

References

[1] Pilon, S., Van Huyssteen, G.B., Augustinus, L., "Converting Afrikaans to Dutch for Technology Recycling," in *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, pp. 219–224 , 2010.

[2] Petrenz, P., "Cross-Lingual Genre Classification," in *Proceedings of the EACL 2012 Student Research Workshop*, Avignon, France, pp. 11-21, 2012.

[3] Sharoff, S., "Classifying web corpora into domain and genre using automatic feature identification," in *Proceedings of Web as Corpus Workshop*, Louvain-la-Neuve, 2007.

[4] Snyman, D.P., Van Huyssteen G.B., Daelemans, W. "Automatic Genre Classification for Resource Scarce Languages". in *Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa*, Vanderbijlpark, South Africa, pp. 132-137, 2011.

[5] Van Huyssteen, G.B. & Pilon, S., "Rule-based conversion of closely-related languages: A Dutch to Afrikaans convertor", in *Proceedings of the 20th Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, pp. 23-28, 2009.

[6] Bel, N., Koster, C., Villegas, M., "Cross-lingual text categorization," Research and Advanced Technology for Digital Libraries, 2769, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp. 126–139, 2003.

[7] Van Noord, G., "Huge Parsed Corpora in LASSY," in *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, Utrecht, The Netherlands, pp. 115–126, 2009.

[8] Wachsmuth, B., Bunja, K., "Back to the roots of Genres: Text Classification by Language Function," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, pp. 632–640, 2011.

[9] Yi-Hsing, C., Hsiu-Yi, H., "An Automatic Document Classifier System based on Naïve Bayes Classifier and Ontology," *in Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*. Kunming, 2008.

[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., "The WEKA Data Mining Software: An Update," SIGKDD Explorations, 11,1, 2009.

[11] Groenewald, H.J., Du Plooy, L., "Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project," AfLaT, pp. 27-30, 2010.