Towards Simplification: A Supervised Learning Approach

Iustina Ilisei[1], Diana Inkpen[2], Gloria Corpas Pastor[3], and Ruslan Mitkov[4]

[1, 4] Research Institute in Information and Language Processing, University of Wolverhampton, UK
[2]School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada
[3]Department of Translation and Interpreting,University of Malaga, Malaga, Spain
[1]iustina.ilisei@gmail.com, [2]diana@site.uottawa.ca, [3]gcorpas@uma.es, [4]r.mitkov@wlv.ac.uk

Abstract

The aim of this study is to train a computer system to distinguish between translated and original text, in order to investigate the simplification phenomenon. The experiments are based on Spanish comparable corpora with two different genres: medical and technical texts. The classifiers achieve an overall accuracy of 87% on a test set, and the removal of the features related to simplification from the learning process leads to a decreased accuracy of the classifiers. Therefore, the obtained results may be interpreted as an argument for the existence of the simplification universal.

1 Introduction

In Translation Studies, the characteristics exhibited by translated texts compared to non-translated texts, have always been of great interest. The specific language of translations has certain universal features, as a consequence of the translation process. The translations exhibit their own peculiar lexico-grammatical and syntactic characteristics (Borin and Prütz, 2001; Teich, 2003). Fairly recently, it has been stated that there are common characteristics which all the translations share, regardless of the source and the target languages (Baker, 1993). Toury (1995) proposed two laws of translation: the law of standardisation and the law of interference, and Baker (1993, 1996) defined four possible translation universals. However, these explanations for the universals are based on intuition and introspection. Laviosa (2002) continued this line of research by proposing features for simplification in a corpus-based study. Despite some evidence of the existence of such a phenomenon, there is still a remarkable challenge in defining the features needed to investigate the simplification universal and its degree in translated texts.

The goal of the present study is to investigate the validation of the simplification hypothesis, and to use a language-independent feature vector in the process of training a system to distinguish between translated and non-translated texts. The main advantages of using only language-independent features are obvious: the system has a wide applicability for other languages, and more importantly, the universality characteristic of this hypothesis is easier to investigate.

2 Related Work

One of the translation universals defined by Baker (1993) is *simplification*, which is described as the tendency of translators to produce easier-to-follow and simpler texts. The follow-up research methodology in the investigation of translation universals is based on comparable corpora, and some empirical results sustaining the universal were provided (Laviosa, 2002).

More light was shed on this by proposing several features in support of the simplification universal and by testing their statistical significance on a Spanish comparable corpora (Corpas et al., 2008). The experiments are on medical and technical domain, and the translations are written by both professional and semi-professional translators. The study tackles the simplification and convergence universals in a different manner, using readability measures, but are largely compatible with the results outlined in (Corpas, 2008). Simplification universal has been contradicted for most of the features investigated, except for lexical richness (Corpas, 2008).

A different approach to this research topic is undertaken by Baroni and Bernardini (2006), reporting outstanding results using machine learning algorithms for the task of classifying Italian texts as translated or originals. They use a feature vector to represent a document, by changing both the size and the type of the units: unigrams, bigrams, trigrams, and word forms, lemmas, part of speech tags, and mixed, respectively. They show that the SVM classifier depends mainly on lexical cues, the distribution of n-grams of function words and the morpho-syntactic categories in general, and on personal pronouns and adverbs in particular. The results prove that shallow data representations can be sufficient to automatically distinguish professional translations from non-translated texts with an accuracy above the chance level, and thus hypothesise that this representation catches the distinguishing features of translationese.

3 Methodology

Our approach is based on supervised machine learning algorithms which aim to distinguish between translated and non-translated texts. We train classifiers by including in the data representation vector specific features which are proposed for the simplification universal. If the accuracy of the classifiers decreases when we remove the simplification features from the feature vectors, it can be stated that this is an argument towards the existence of the simplification universal.

For our experiments, we use three pairs of comparable corpora, described in Corpas (2008). They are Spanish comparable corpora of non-translated and translated texts. Two pairs are from the medical domain, written by translation students and professional translators, respectively. The third one is from the technical domain, written by professionals. The three paired corpora are the following:

- Corpus of Medical Translations by Professionals (MTP), which is comparable to the Corpus of Original Medical texts by Professionals (MTPC);
- Corpus of Medical Translations by Students (MTS), which is comparable to the Corpus of Original Medical texts by Students (MTSC);

- Corpus of Technical Translations by Professionals (TT), which is comparable to the Corpus of Original Technical texts by Professionals (TTC).

We extract a training dataset of 450 randomly selected instances and a test set of 150 randomly selected instances from all the three pairs of comparable texts. We keep the same proportion of texts of each kind in the selected training and test sets. The set of language-independent features proposed for the training of our system are as follows: the first twelve are general parameters, while the next nine are designed to catch the simplicity characteristic of texts. On the assumption that the simplification universal is valid, the latter features are expected to improve the performance of the classifiers.

The first twelve features are the proportion in each text of the following: grammatical words, nouns, finite verbs, auxiliary verbs, adjectives, adverbs, numerals, pronouns, prepositions, determinants, conjunctions, and the ratio of grammatical words per lexical words.

The simplification features considered, most of them originally proposed in (Corpas, 2008; Laviosa, 2002), are the following: the average sentence length, the parse tree depth, the proportion of simple sentences, complex sentences and sentences without any finite verb, the ambiguity level of sentences as the average of the proportion of senses for each word of the sentence, the word length as the proportion of syllables per word, the ratio of lemmas divided by the number of tokens, and the ratio of lexical words by total number of tokens.

To exploit all of these features, the corpora was parsed with the Connexor Machinese' dependency parser for the Spanish language model described in Tapanainen and Jarvinen (1997). Also, the Spanish Wordnet has been exploited to compute the ambiguity parameter ratio (Verdejo, 1999).

The algorithms used for the classification are the following (Witten and Frank, 2005): Jrip, Decision Tree (J48), Naïve Bayes, BayesNet, SVM, Simple Logistic and one meta-classification algorithm: using the results from three algorithms: J48, Jrip and Simple Logistic.

To assess the statistical significance of the improvement of the machine learning system when including simplification features comparing to the learning system without these features, we apply the paired two-tailed t-test, with 0.5 significance level. T-tests have been applied for the evaluation measurements computed: the accuracy, the precision, the recall and the f-measure of the classifiers.

4 Experiments

In order to investigate the simplification universal, we compare the accuracy of the classification task considering the entire features vector to the accuracy of the learning system trained with all the parameters except the simplification features ones. Our assumption is as follows: if the lack of simplification features causes a statistically-significant difference, this may be seen as an argument towards the existence of such an universal.

1.1 Classification results

In Table 1, we present the main results of the classification when all the three corpora are used as one larger corpus. We report accuracy results for 10-fold cross-validation on the training data to see how well the classifiers were able to learn, and for the test data to confirm that what was learnt on the training data is valid when the classifiers are applied to unseen test data.

Throughout all the table cells, a star near the value of the result for a classifier indicates that the result is better in a statistically significant manner, when including the simplification features, than the same classifier without the simplification features. We only added stars on the side of the classifier that included all twenty-one features, in case the improvement brought by the simplification features is statistically significant.

Accuracy (%)	**Including** Simplification Features		**Excluding** Simplification Features	
	10-fold cross-validation	Test set	10-fold cross-validation	Test set
Baseline (ZeroR)	65	65	65	65
Naive Bayes	*77	79	69	75
BayesNet	79	80	75	77
Jrip	80	83	73	77
Decision Tree	78	82	78	82
Simple Logistic	*77	83	71	80
SVM	*79	*81	69	73
Meta-classifier	*80	**87**	73	86

Table 11 - 1: Classification Results: Accuracies for several classifiers.

The baseline in our experiments is the ZeroR classifier from Weka which takes into account the majority class from the data set, in our case being the non-translated class. Therefore, the baseline is 65% in general, as we followed the same proportion of instances for both the training dataset and test dataset.

The meta-classifier, which takes the majority vote between J48, Jrip and Simple Logistic, reaches 87% for the randomly selected test set and 80% for 10 fold cross-validation.

In Table 2, the test set results for the positive class (translated class) reach up to 0.83 precision, and 0.63 recall, with a statistically-significant improvement in f-measure of 0.69 for the SVM classifier, when the simplification features are included in the data representation. BayesNet is a classifier which exhibits a constant significant improvement for all three evaluation measurements in the case of including the simplification features; excluding them would reduce the results up to 0.08 f-measure.

	Including Simplification Features			**Excluding** Simplification Features		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Naive Bayes	0.77	*0.64	0.68	0.8	0.52	0.61
BayesNet	*0.55	*0.43	*0.41	0.07	0.09	0.08
Jrip	0.61	0.55	0.56	0.67	0.68	0.65
Decision Tree	0.64	0.58	0.59	0.75	0.61	0.65
Simple Logistic	0.77	0.66	0.7	0.69	0.54	0.58
SVM	0.83	*0.63	*0.69	0.73	0.47	0.54
Meta-classifier	0.76	0.63	0.66	0.73	0.65	0.66

Table 11 - 2: Classification results for the test set (precision, recall, and f-measure).

1.2 Preliminary results analysis

The decision tree classifier and Jrip algorithm offer an output for analysis more intuitive to humans (Quinlan, 1986). The decision tree which the system was able to learn has the following features: on the first level is the proportion of lemmas and tokens – a feature considered to be indicative for simplification (Corpas et al., 2008).

On the second level of the decision tree is the sentence length and the proportion of the grammatical words and lexical words. Sentence length is a characteristic widely discussed in similar studies and presented some difficulty in the interpretation of the results described in Corpas et al. (2008). The proportion of the grammatical words and lexical words is an original feature proposed in this paper, considered to stand for the translationese phenomenon rather than to be an indicator strictly for the simplification universal.

On the third level of the tree is the proportion of pronouns and conjunctions. Personal pronouns in particular have been considered before in similar studies, while in these experiments we take all the pronouns in general, regardless of their type. As conjunctions have not been studied as a feature in simplification, these results point to a new direction in the investigation of translation studies.

Thus, the top features taken into account by the decision tree algorithm are: the proportion in texts of lemmas by tokens, the proportion of grammatical words and lexical words, the sentence length, followed by the ratio of pronouns and conjunctions.

The Jrip classifier gives a readable format output of the rules employed in the classification task, pointing out that the most important features in the learning process: the first rule considers the proportion of lemmas by tokens and the proportion of finite verbs; the second one takes into account the sentence length, the proportion of nouns and the proportion of syllables per word.

5 Conclusions and further work

This study describes a supervised learning approach in the process of identification of the features that characterise translated texts vs. non-translated texts. The experiments are based on Spanish comparable corpora of medical and technical genres with translations written by both professional and semi-professional translators.

The novelty of our study consists in the learning model trained with language- and domain-independent features, including the parameters proposed for the simplification universal, which performs better than the system trained without the simplification features. On the categorisation task, our system has an accuracy of up to 87% on a test set, and the removal of the features related to simplification from the learning process causes a decreased performance of the classifiers exploited. This may be considered an argument towards the existence of the simplification universal.

In future work, a similar approach will be employed for the investigation of the other universals, such as explicitation hypothesis. Another line of research consists of a deeper analysis of the features which can be employed in the detection task of the explicitation universal.

References

Baker, M. (1993). 'Corpus Linguistics and Translation Studies – Implications and Applications'. In: M. Baker, M.G. Francis & E. Tognini-Bonelli (eds.). Text and Technology: In Honour of John Sinclair. Amsterdam & Philadelphia: John Benjamins. 233-250.

Baker, M. (1996). 'Corpus-based Translation Studies: The Challenges that Lie Ahead'. In: H. Somers (ed.). 1996. Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager. Amsterdam & Philadelphia: John Benjamins. 175-186.

Baroni, Marco and Silvia Bernardini. (2006). 'A new approach to the study of translationese: Machine-learning the difference between original and translated text'. Literary and Linguistic Computing. 21, 3: 259-274.

Borin, L. and Prütz, K. (2001). Thorough a dark glass: part of speech distribution in original and translated text. In Daelemans, W., Sima'an, K., Veenstra, J. and Zavrel, J. (eds), Computational Linguistics in the Netherlands 2000. Amsterdam: Rodopi, pp. 30–44.

Corpas Pastor, G. (2008). Investigar con corpus en traducción: los retos de un nuevo paradigma. Frankfurt am Main, Berlin & New York: Peter Lang.

Corpas Pastor, G., Mitkov R., Afzal N., Pekar V. (2008). Translation Universals: Do they exist? A corpus-based NLP study of convergence and simplification. In Proceedings of the AMTA (2008). Waikiki, Hawaii.

Frawley, W. (1984). 'Prolegomenon to a theory of translation'. In Frawley, W. (ed.), Translation: Literary, Linguistic and Philosophical Perspectives. Newark: University of Delaware Press, pp. 159–75.

Laviosa, S. (2002). Corpus-based Translation Studies. Theory, Findings, Applications. Amsterdam & New York: Rodopi.

Quinlan, J.R. (1986). 'Induction of Decision Trees'. Machine Learning, 1:81–106.

Tapanainen, P., Jarvinen, T. (1997). A non-projective dependency parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA. 64–71

Teich, E. (2003). Cross-linguistic Variation in System and Text. Berlin: Mouton de Gruyter.

Toury, G. (1995). 'Descriptive Translation Studies and Beyond'. Amsterdam: John Benjamins.

Verdejo, F.M. (1999) The spanish wordnet. Technical report, Universitat Politenica de Catalunya, Madrid, Spain

Witten, I. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Second Edition. Morgan Kaufmann.