Identification of Translationese: A Machine Learning Approach

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov

Research Institute in Information and Language Processing,
University of Wolverhampton, Wolverhampton, United Kingdom
iustina.ilisei@wlv.ac.uk
School of Information Technology and Engineering,
University of Ottawa, Ottawa, Canada
diana@site.uottawa.ca
Department of Translation and Interpreting,
University of Málaga, Málaga, Spain
gcorpas@uma.es
Research Institute in Information and Language Processing,
University of Wolverhampton, Wolverhampton, United Kingdom
r.mitkov@wlv.ac.uk

Abstract. This paper presents a machine learning approach to the study of translationese. The goal is to train a computer system to distinguish between translated and non-translated text, in order to determine the characteristic features that influence the classifiers. Several algorithms reach up to 97.62% success rate on a technical dataset. Moreover, the SVM classifier consistently reports a statistically significant improved accuracy when the learning system benefits from the addition of simplification features to the basic translational classifier system. Therefore, these findings may be considered an argument for the existence of the Simplification Universal.

1 Introduction

The characteristics exhibited by translated texts compared to non-translated texts have always been of great interest in Translation Studies. Translated language is believed to manifest certain universal features, as a consequence of the translation process. Translations exhibit their own specific lexico-grammatical and syntactic characteristics [1–3]. These"fingerprints"left by the translation process were first described by Gellerstam and named translationese [4]. Fairly recently, it has been stated that there are common characteristics which all translations share, regardless of the source and the target languages [5]. Toury proposed two laws of translation: the law of standardisation and the law of interference [6], and it was Baker who defined four possible translation universals [5, 7]. However, the notion of these universals is based on intuition and introspection. Laviosa continued this line of research by proposing features for simplification in a corpus-based study [8]. Despite some evidence of the existence of such a

phenomenon, there is still a remarkable challenge in defining the features which characterise the simplification universal.

The aim of this study is twofold: first, to model a language-independent learning system able to distinguish between translated and non-translated texts. The main advantages of such a data representation are obvious: the system has a wide applicability for other languages, and thus, the "universal"label of this hypothesis is easier to investigate. Second, the goal is to investigate the validation of the simplification hypothesis and to explore the characteristic features which most influence the translated language.

2 Related Work

The simplification universal is described as the tendency of translators to produce simpler and easier-to-follow texts [5]. 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 [8]. Laviosa investigates lexical patterns for English and the obtained results show a relatively low proportion of lexical words over function words in translated texts, and a high proportion of high-frequency words compared to the low-frequency words. Moreover, great repetition of the most frequent words and less variety in the most frequently used words has been emphasised [9].

Recently, a corpus-based approach which tests the statistical significance of features proposed to investigate the simplification universal has been exploited for Spanish [10, 11]. The experiments were on both the medical and technical domains, and the translated texts were produced by both professional and semi-professional translators. In [10] the simplification universal is confirmed only for lexical richness. The results for the following parameters appear to be against this universal: complex sentences, sentence length, depth of syntactical trees, information load, senses per word. The experiments in [11] revealed that translated texts exhibit lower lexical density and richness, seem to be more readable, have a smaller proportion of simple sentences and appear to be significantly shorter, and discourse markers were used significantly less often. Simplification fingerprints were found on the technical translation and seemed to show that texts written by non-professional translators do not have such simplification traits.

A different perspective over the same line of research is employed by Baroni and Bernardini [12], who exploit machine learning techniques for the task of classifying Italian texts as translated or non-translated texts. The results obtained show that the SVM classifier depends heavily on lexical cues, the distribution of n-grams of function words and morpho-syntactic categories in general, and on personal pronouns and adverbs in particular. Therefore, it is proved that shallow data representations can be sufficient to automatically distinguish professional translations from non-translated texts with an accuracy above the chance level, and hypothesise that this representation captures the distinguishing features of translationese. Moreover, human accuracy on the same task seems to be much lower compared to the success rate of the learning system. In this study, the

exploitation of n-grams indicators is avoided because of their language dependence.

3 Methodology

The approach in this paper is based on supervised machine learning techniques which aim to distinguish between translated and non-translated, spontaneous texts. Therefore, a training dataset and a test dataset were constructed comprising random instances from both classes. By using Weka [1] [13, 14], the classifiers are trained including and excluding the features proposed for the simplification universal within the data representation, and afterwards the T-test evaluates the statistical significance between the accuracies obtained in both cases. Therefore, if the success rate of the learning system including the simplification indicators in the feature vector is high, then it may be stated that this is an argument for the existence of the simplification universal.

As is proposed, these universals can be studied by comparing translations with non-translations in the same language [15], thus strictly avoiding any foreign interference [16]. The resource exploited is the monolingual comparable corpora for Spanish language extensively described in [10], which comprise three pairs of translated and non-translated texts, as follows:

- 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).

The training set comprises 450 randomly selected instances and the overall test set has 150 randomly selected instances from all three pairs of comparable texts. The same proportion of texts is kept for both selected training and test datasets. In order to extract the feature vector for the learning process, all the texts of the corpora were parsed with the Connexor Machinese [17], which provides the dependency parser for the Spanish language model.

The learning system exploits twenty-one language-independent features. Some of these parameters are designed to capture the simplicity characteristic of texts, which is expected to improve the performance of the classifiers, on the assumption that the simplification universal is valid. Additionally, in order to prevent learning to classify according to the topic of a text, the current approach avoids the bag-of-words model.

The first set of features which grasp general characteristics of texts, considered to stand for the translationese effect, are the following:

- the proportion in texts of grammatical words, nouns, finite verbs, auxiliary verbs, adjectives, adverbs, numerals, pronouns, prepositions, determiners, conjunctions, and the proportion of grammatical words to lexical words.

[1] http://www.cs.waikato.ac.nz/ml/weka/

For the last parameter above, the following parts of speech are considered to belong to the class of grammatical words: determiners, prepositions, auxiliary verbs, pronouns, and interjections. Lexical words, also known as content words, are represented by nouns, verbs, adjectives, adverbs, and numerals.

The data representation for the learning system comprises all the above parameters and includes the proposed simplification features described below:

1. average sentence length,
2. sentence depth as the parse tree depth,
3. proportion of simple sentences, complex sentences and sentences without any finite verb,
4. ambiguity as the average of senses per word[2],
5. word length as the proportion of syllables per word,
6. lexical richness,
7. information load as the proportion of lexical words to tokens.

Most of the features employed (1-4, 6-7) in the data representation were originally proposed in [10] for the investigation of the simplification universal. The experiments in [11] deal with the universal in a slightly different manner (e.g. using readability measures), hence the results previously mentioned are slightly different from the ones reported in [10] but by and large compatible.

The next stage of the study consists of evaluation on separate datasets corresponding to each corpus domain, in order to determine the performance of the text classification for each type and genre. Therefore, the system is trained on the entire training dataset and it is tested on the following datasets: the technical domain written by professional translators dataset, and on the medical domain written by students dataset. As the medical domain written by professionals dataset has insufficient class instances, no separate dataset was considered.

The machine learning classifiers applied on the categorisation task are the following: Jrip, Decision Tree, Naïve Bayes, BayesNet, SVM, Simple Logistic and one meta-classification algorithm: the Vote meta-classifier with the Majority Voting combination rule, which considers the Decision Tree, Jrip and Simple Logistic classifiers output. To assess the statistical significance of the improvement of the machine learning system when including simplification features compared to the learning system without these features, the paired two-tailed t-test has been applied with 0.5 significance level.

4 Evaluation

The accuracy obtained with the data representation including the simplification features is compared to the accuracy obtained by the system without the simplification features. The assumption is the following: if the lack of simplification features causes a statistically significant difference, this can be considered as an argument for the existence of the simplification universal.

[2] Note that the ambiguity parameter is obtained exploiting the Spanish Wordnet synsets [18].

4.1 Classification results

The accuracies for the 10-fold cross-validation evaluation on the training data and the accuracy for the test dataset evaluation are reported in Table 1. The training dataset comprises 450 instances, with 156 for the translation class and 294 for non-translation class instances, and the test dataset comprises 148 instances, with 52 for the translation class and 96 for non-translation class.

An asterisk by the accuracy value indicates that a statistically significant improvement is registered when including the simplification features compared to the same classifier without the simplification features. There are no worse cases, therefore only improvement is marked.

Table 1. Classification Results: Accuracies for several classifiers

	Including Simplification Features		**Excluding Simplification Features**	
Classifier	*10-fold cross-validation*	*Test set*	*10-fold cross-validation*	*Test set*
Baseline	65.33%	64.86%	65.33%	64.86%
Naive Bayes	*76.67%	79.05%	69.33%	75.00%
BayesNet	78.67%	79.73%	75.11%	77.03%
Jrip	79.56%	83.11%	73.33%	77.03%
Decision Tree	78.22%	81.76%	78.22%	81.76%
Simple Logistic	*77.33%	83.11%	71.11%	80.41%
SVM	*79.11%	*81.76%	69.33%	73.65%
Meta-classifier	*80.00%	87.16%	73.33%	85.81%

The baseline classifier, ZeroR, considers the majority class from the dataset. As the majority class is the non-translated class, the baseline is 64.5%. The meta-classifier, which takes the majority vote between Decision Tree, Jrip and Simple Logistic classifiers, reaches 87.16% for the randomly selected test set and 80% for 10 fold cross-validation.

4.2 Experiments on separate test datasets

The experiments continue with the evaluation of the system on three subsets of the test set according to the three types of corpora: the test set pair 1 for MTP-MTPC, test set pair 2 for MTS-MTSC, and test set pair 3 for TT-TTC. The same proportion of class instances is kept as in the previous stage: test set pair 2 has 66 and 36 instances for non-translated and translated class, respectively; test set pair 3 has 28 non-translated class instances and 14 translated class instances. As pair 1 has only 5 instances for both classes, it is not relevant to test the classifiers on such a small dataset.

In Table 2, the accuracies for the classifiers tested on these three datasets are reported. As expected from the previous experiment, none of them report worse

results when adding the simplification features. Moreover, the SVM classifier shows a statistically significant improvement for the technical domain written by professionals, reaching the highest performance of 97.62% accuracy. Nevertheless, BayesNet, Simple Logistic, and the meta-classifier register similar values for the same pair (technical domain), not statistically significant according to the t-test.

Table 2. Classification accuracy results on the medical and technical test datasets.

	Including Simplification Features		**Excluding Simplification Features**	
Classifier	MTS-MTSC	TT-TTC	MTS-MTSC	TT-TTC
Baseline	64.71%	66.67%	64.71%	66.67%
Naive Bayes	71.57%	95.24%	71.57%	80.95%
BayesNet	73.53%	97.62%	71.57%	92.86%
Jrip	79.42%	95.24%	72.55%	92.86%
Decision Tree	77.45%	92.86%	75.49%	95.24%
Simple Logistic	77.45%	97.62%	79.41%	83.33%
SVM	75.49%	*97.62%	74.51%	69.05%
Meta-classifier	82.35%	97.62%	78.43%	92.86%

The learning system retrieves outstanding results for the technical domain, with all the classifiers having above 95% success rates.

Aiming to determine the most salient features which led to these results, the following subsection provides the feature analysis output from the learning system and the attribute evaluators selection.

4.3 Results analysis

A deeper result analysis is undertaken and the rules considered by the classifiers are described in figures 1 and 2. The Jrip and the Decision Tree classifiers are two algorithms which provide an intuitive output for analysis [19].

As can be noticed from the pruned tree output in Figure 1, the most informative feature is undoubtedly lexical richness, followed by sentence length and proportion of grammatical words by lexical words. Both lexical richness and sentence length are features considered to be indicative of the simplification hypothesis, widely discussed and studied in the past decade. Sentence length is a characteristic which posed a certain difficulty in its interpretation in the study undertaken by [10, 11]. Additionally, the proportion of grammatical words and lexical words makes a valuable contribution in the classification. This is a feature first proposed for this task, and considered to stand for the translationese phenomenon in general, rather than for any particular universal. On the third level is the proportion of pronouns and conjunctions in texts.

The rules observed by the JRip classifier, according to which the classifier takes its decisions, is presented in Figure 2.

```
lexicalRichness <= 0.16
|   sentenceLength <= 16.81: non-translation (30.0)
|   sentenceLength > 16.81
|   |   ratioProns <= 0.05
|   |   |   lexicalRichness <= 0.11: translation (46.0/1.0)
|   |   |   lexicalRichness > 0.11
|   |   |   |  ...................................
|   |   |   ratioNumerals <= 0.03: non-translation (15.0/1.0)
|   |   |   ratioNumerals > 0.03
|   |   |   |  ...................................
lexicalRichness > 0.16
|   grammsPerLexics <= 0
|   |   ratioConjs <= 0.03
|   |   |   ratioAdjectives <= 0.09: translation (9.0/1.0)
|   |   |   ratioAdjectives > 0.09: non-translation (2.0)
|   |   ratioConjs > 0.03
|   |   |   ratioNumerals <= 0.06
|   |   |   |  ...................................
|   |   |   ratioNumerals > 0.06
|   |   |   |  ...................................
|   grammsPerLexics > 0
|   |   lexicalRichness <= 0.31: non-translation (88.0/2.0)
|   |   lexicalRichness > 0.31
|   |   |   ratioConjs <= 0.04: non-translation (11.0)
|   |   |   ratioConjs > 0.04
|   |   |   |   ratioNumerals <= 0.04
|   |   |   |   |  ................................
|   |   |   |   ratioNumerals > 0.04: translation (4.0)
```

Fig. 1. Pruned tree output from the Decision Tree classifier.

```
Rule 1: (lexicalRichness <= 0.16) and (ratioFiniteVerbs <= 0.08)
 => class=translation (86.0/15.0)
Rule 2: (simpleSentences >= 0.3) and (wordLength <= 2.46) and
 (sentenceLength >= 20.7) and (ratioNouns >= 0.33)
 => class=translation (24.0/3.0)
Rule 3: (ratioFiniteVerbs <= 0.09) and (ratioPreps <= 0.13)
 => class=translation (17.0/6.0)
Rule 4: => class=non-translation (323.0/53.0)
```

Fig. 2. JRip classifier rules output.

The first rule considers lexical richness and proportion of finite verbs, whilst sentence length, word length, proportion of nouns and prepositions appear in the second and third rule output from this classifier.

Furthermore, the feature selection evaluators output is exploited in order to see the ranking of the attributes, regardless of any classifier. The Information Gain and Chi-square algorithms provide the information from Figure 3. The notation of the sentences without any finite verb is marked in the program as the zeroSentences attribute.

As can be observed, the two feature selection algorithms acquire approximately the same knowledge, particularly for the top seven attributes. The slight variation in the ranking is minimal, and the most intriguing part is that ambiguity is listed as one of the less informative features in the classification system.

Table 3. Attributes Ranking Filters.

Information Gain	**Chi squared**
0.1 lexicalRichness	61.79 lexicalRichness
0.08 grammsPerLexics	43.55 grammsPerLexics
0.07 ratioFiniteVerbs	39.28 ratioFiniteVerbs
0.05 ratioNumerals	33.12 ratioNumerals
0.05 ratioAdjectives	23.89 ratioAdjectives
0.04 sentenceLength	23.55 sentenceLength
0.04 ratioProns	22.64 ratioProns
0.03 simpleSentences	21.07 wordLength
0.03 wordLength	19.74 simpleSentences
0.03 grammaticalWords	15.37 zeroSentences
0.03 zeroSentences	13.79 ratioNouns
0.02 ratioNouns	11.46 lexicalWords
.....	.....

Moreover, this attribute is consistently disregarded by both decision tree and Jrip classifiers. Therefore, the assumption that the more ambiguous a text is, the more probable that it is a non-translation is rejected by the employed learning system, in line with one of the findings by [10].

In addition, analysing the confusion matrix, a high proportion of misclassified instances are due to the labelling of the translated text as non-translations. Thus, it can be asserted that this behaviour is expected, as the main purpose of a translation is to be easily confused with a spontaneous, non-translated text.

5 Conclusions and Further Research

This paper presents a new study on the investigation of universals of translations in Spanish. A supervised learning approach is employed to identify the most informative features that characterise translations compared to non-translated texts. The learning system is trained on two domains, medical and technical, and the novelty consists of its language-independent data representation. The outstanding accuracy provided by several classifiers is evidence that translations can indeed be identified.

On the categorisation task, the algorithms achieve an accuracy of 87.16% on a test set, and reach up to 97.62% for separate test datasets from the technical domain. However, the removal of the features related to simplification from the machine learning process leads to decreased accuracy of the classifiers. Therefore, the retrieved results may be considered as an argument for the existence of the simplification universal. A performance analysis of our classifiers' output reveals that the learning system relies heavily on the following features: lexical richness, proportion of grammatical words to lexical words, sentence length, word length and some morphological attributes like nouns, pronouns, finite verbs, conjunctions and prepositions.

The main research direction to be tackled in the future is the investigation of the other translation universals. An additional subject of investigation will be a deeper analysis of the indicative features which influence translated language.

References

1. Borin, L., Prütz, K.: Thorough a dark glass: part of speech distribution in original and translated text. In: Computational Linguistics in the Netherlands. Amsterdam: Rodopi (2001) 30–44
2. Hansen, S.: The Nature of Translated Text. Saarbrücken: Saarland University (2003)
3. Teich, E.: Cross-linguistic Variation in System and Text. Berlin:Mouton de Gruyter (2003)
4. Gellerstam, M.: Translationese in Swedish novels translated from English. Translation Studies in Scandinavia. Lund: CWK Gleerup (1986)
5. Baker, M.: Corpus Linguistics and Translation Studies – Implications and Applications. In: Text and Technology: In Honour of John Sinclair. Amsterdam & Philadelphia: John Benjamins (1993) 233–250
6. Toury, G.: Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins (1995)
7. Baker, M.: Corpus-based Translation Studies: The Challenges that Lie Ahead. In: Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. Amsterdam & Philadelphia: John Benjamins (1996) 175–186
8. Laviosa, S.: Corpus-based Translation Studies. Theory, Findings, Applications. Amsterdam & New York: Rodopi (2002)
9. Laviosa, S.: Core patterns of lexical use in a comparable corpus of English narrative prose. In: The Corpus-Based Approach. Volume Special Issue of Meta. Montréal: Les Presses de L'Université de Montréal (1998) 557–570
10. Corpas, G.: Investigar con corpus en traducción: los retos de un nuevo paradigma. Frankfurt am Main, Berlin & New York: Peter Lang (2008)
11. Corpas, G., Mitkov, R., Afzal, N., Pekar, V.: Translation universals: Do they exist? a corpus-based nlp study of convergence and simplification. In: Proceedings of the AMTA, Waikiki, Hawaii (2008)
12. Baroni, M., Bernardini, S.: A new approach to the study of translationese: Machine-learning the difference between original and translated text. Literary and Linguistic Computing **21, 3** (2006) 259–274
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. SIGKDD Explorations **11(1)** (2009) 10–18
14. Witten, I.H., Frank, E.: Data Mining : Practical Machine Learning Tools and Techniques. Second edition edn. Morgan Kaufman (2005)
15. Olohan, M.: Introducing Corpora in Translation Studies. Routledge (2004)
16. Pym, A.: On Toury's laws of how translators translate. In: Beyond Descriptive Translation Studies. Benjamins (2008) 311–328
17. Tapanainen, P., Jarvinen, T.: A non-projective dependency parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA. (1997) 64–71
18. Verdejo, F.M.: The spanish wordnet. Technical report, Universitat Politenica de Catalunya, Madrid, Spain (1999)
19. Quinlan, J.R.: Induction of decision trees. Machine Learning **1** (1986) 81–106