Source and Translation Classification using Most Frequent Words

Zahurul Islam
AG Texttechnology
Institut für Informatik
Goethe-Universität Frankfurt
zahurul@em.uni-frankfurt.de

Armin Hoenen
AG Texttechnology
Institut für Informatik
Goethe-Universität Frankfurt
hoenen@em.uni-frankfurt.de

Abstract

Recently, translation scholars have made some general claims about translation properties. Some of these are source language independent while others are not. Koppel and Ordan (2011) performed empirical studies to validate both types of properties using English source texts and other texts translated into English. Obviously, corpora of this sort, which focus on a single language, are not adequate for claiming universality of translation properties. In this paper, we are validating both types of translation properties using original and translated texts from six European languages.

1 Introduction

Even though it is content words that are semantically rich, function words also play an important role in a text. Function words are more frequent and predictable than content words. Generally, function words carry grammatical information about content words. High frequency function words are relatively shorter than mid/low frequency function words (Bell et al., 2008). Due to their high frequency in texts and their grammatical role, function words also indicate authorial style (Argamon and Levitan, 2005). These words could play an important role in translated text and in the translation process.

Source and translation classification is useful for some Natural Language Processing (NLP) applications. Lembersky et al. (2011) have shown that a language model from translated text improves the performace of a Machine Translation (MT) system. A source and translation classifier can be used to identify translated text. This application also can be used to detect plagiarism where the plagiarised text is translated from another language.

From the early stage of translation studies research, translation scholars proposed different kinds of properties of source text and translated text. Recently, scholars in this area identified several properties of the translation process with the aid of corpora (Baker, 1993; Baker, 1996; Olohan, 2001; Laviosa, 2002; Hansen, 2003; Pym, 2005). These properties are subsumed under four keywords: *explicitation*, *simplification*, *normalization* and *levelling out*. They focus on the general effects of the translation process.

Toury (1995) has a different theory from these. He stated that some *interference* effects will be observable in the translated text. That is, a translated text will carry some fingerprints of its source language. Specific properties of the English language are visible in user manuals that have been translated to other languages from English (for instance, word order) (Lzwaini, 2003). Recently, Pastor et al. (2008) and Ilisei et al. (2009; 2010) have provided empirical evidence of simplification translation properties using a comparable corpus of Spanish.

Koppel and Ordan (2011) perform empirical studies to validate both theories, using a subcorpus extracted from the *Europarl* (Koehn, 2005) and IHT corpora (Koppel and Ordan, 2011). They used a comparable corpus of original English and English translated from five other European languages. In addition, original English and English translated from Greek and Korean was also used in their experiment. They have found that a translated text contains both source language dependent and independent features.

International Joint Conference on Natural Language Processing, pages 1299–1305,
Nagoya, Japan, 14-18 October 2013.

Obviously, corpora of this sort, which focus on a single language (e.g., English), are not adequate for claiming the universal validity of translation properties. Different languages (and language families) have different linguistic properties. A corpus that contains original and translated texts from different source languages will be ideal for this kind of study. In this paper, we are validating both types of translation properties using original and translated texts from six European languages. As features, we used frequencies of the 100 most frequent words of each target language.

The paper is organized as follows: Section 2 discusses related work, followed by an introduction of our corpus in Section 3. The experiment and evaluation in Section 4 are followed by a discussion in Section 5. Finally, we present conclusions and future work in Section 6.

2 Related Work

Corpus-based translation studies is a recent field of research with a growing interest within the field of computational linguistics. Baroni and Bernardini (2006) started corpus-based translation studies empirically, where they work on a corpus of geo-political journal articles. A Support Vector Machine (SVM) was used to distinguish original and translated Italian text using n-gram based features. According to their results, word bigrams play an important role in the classification task.

Van Halteren (2008) uses the *Europarl* corpus for the first time to identify the source language of text for which the source language marker was missing. Support vector regression was the best performing method.

Pastor et al. (2008) and Ilisei et al. (2009; 2010) perform classification of Spanish original and translated text. The focus of their works is to investigate the *simplification* relation that was proposed by (Baker, 1996). In total, 21 quantitative features (e.g. a number of different POS, average sentence length, the parse-tree depth etc.) were used where, nine (9) of them are able to grasp the simplification translation property.

Koppel and Ordan (2011) have built a classifier that can identify the correct source of the translated text (given different possible source languages). They have built another classifier which can identify source text and translated text. Furthermore, they have shown that the degree of difference between two translated texts, translated from two different languages into the same target language reflects, the degree of difference of the source languages. They have gained impressive results for both of the tasks. However, the limitation of this study is that they only used a corpus of English original text and English text translated from various European languages. A list of 300 function words (Pennebaker et al., 2001) was used as feature vector for these classifications.

Popescu (2011) uses *string kernels* (Lodhi et al., 2002) to study translation properties. A classifier was built to classify English original texts and English translated texts from French and German books that were written in the nineteenth century. The *p-spectrum* normalized kernel was used for the experiment. The system works on a character level rather than on a word level. The system performs poorly when the source language of the training corpus is different from the one of the test corpus.

We can not compare our findings directly with Koppel and Ordan (2011) even though we use text from the same corpus and similar techniques. The English language is not considered for this study due to unavailability of English translations for some languages included in this work. Furthermore, instead of the list of 300 function words used by Koppel and Ordan (2011), we used the 100 most frequent words for each candidate language.

3 Data

The field of translation studies lacks a multilingual corpus that can be used to validate translation properties proposed by translation scholars. There are many multilingual corpora available used for different NLP applications. A customized version of the Europarl corpus (Islam and Mehler, 2012) is freely available for corpus-based translation studies. However, this corpus is not suitable for the experiment we are performing here. We extract a suitable corpus from the Europarl corpus in a way similar to Lembersky et al. (2011) and Koppel and Ordan (2011). Our target is to extract texts that are translated from and to the languages considered here. We trust the source language marker that has been put by the respective translator, as did Lembersky et al.(2011) and Koppel and Ordan (2011).

To experiment with stylistic differences in translated text, a list of function words and their

	German	Dutch	French	Spanish	Polish	Czech
German	-	2,574,110	4,757,076	2,035,736	584,114	215,212
Dutch	4,881,949	-	4,386,270	2,682,935	446,702	149,235
French	5,241,411	659,001	-	2,724,897	659,001	226,435
Spanish	4,020,898	1,925,157	3,696,393	-	662,718	247,219
Polish	451,357	112,274	695,360	194,724	-	82,312
Czech	378,300	105,058	684,061	187,236	214,959	-

Table 1: The customized corpus for source language identification (number of words per language)

	German	Dutch	French	Spanish	Polish	Czech
German	-	197	197	198	201	197
Dutch	197	-	197	198	198	191
French	148	147	-	148	149	157
Spanish	148	147	148	-	148	148
Polish	151	141	149	148	-	129
Czech	140	164	149	148	151	-

Table 2: Source language identification corpus (chunks)

respective native frequencies is necessary. Since for many languages such a list does not exist, we pursue an alternative strategy. A list of the 100 most frequent words is available for many languages and since at the same time the majority of these first 100 most frequent words of any language are function words, we use these lists. The 100 most frequent German words are taken from the Deutscher Wortschatz.[1] The most frequent Czech word list is taken from the freely available Czech national corpus.[2] The 100 most frequent Spanish words are taken from the book *A Frequency Dictionary of Spanish: Core Vocabulary for Learners* (Davies, 2006). The French most frequent words are taken from the *A Frequency Dictionary of French: Core Vocabulary for Learners* (Lonsdale and Bras, 2009). The 100 most frequent Dutch words are taken from snowball.[3] The most frequent Polish word list are collected from the Polish scientific publisher PWN.[4]

4 Experiment

In order to validate two different kinds of translation properties mentioned in Section 1, two different experiments will be performed. For the first experiment, our hypothesis is that texts translated into the same language from different source languages have different properties, a trained classifier will be able to classify texts based on different sources. Our second hypothesis is that translated texts are distinguishable from source texts; a classifier can be trained to identify translated and original texts. Note that we use the Naive Bayes multinomial classifier (Mccallum and Nigam, 1998) in WEKA (Hall et al., 2009) for classification. To overcome the data over-fitting problem, we randomly generate training and test set N times and calculate the weighted average of *F-Score* and *Accuracy*. In this experiment the value of N is 100. The randomly generated training sets contain 80% of the data while the remaining data is used as a test set. To evaluate the classification results, we use standard *F-Score* and *Accuracy* measures.

4.1 Source Language Identification

In this experiment, our goal is to validate the translation properties postulated by Toury (1995). He stated that a translated text inherits some fingerprints from the source language. The experimental result of Koppel and Ordan (2011) shows that text translated into English holds this property. If this characteristic also holds for text translated into other languages, then it will corroborate the claim by Toury (1995). If it does not hold for a single language then it might be claimed that this translation property is not universal. In order to train a classifier, we use texts translated into the same language from different source languages. Table 1 shows the statistics of the corpus used for source language identification experiments. Later, each corpus is divided into a number of chunks (see Table 2). Each chunk contains at least seven sentences. Our hypothesis is again similar to Koppel and Ordan (2011), that is, if the classifier's accuracy is close to 20%, then we cannot say that there is an *interference* effect in translated text. If the classifier's accuracy is close to 100% then our conclusion will be that *interference* effects exist in translated text. Table 3 and Table 4 show the evaluation results. Table 3 shows the *F-Scores* for translated text from different source languages. Rows represent translated texts and columns represent source languages.

A first minor observation can be made, in that the consistency of the results increases when analyzing them with respect to the concept of *language family*. The term *language family* is broadly used in linguistics as a denomination of groups of languages that have descended fom a common

[1] http://wortschatz.informatik.uni-leipzig.de/

[2] http://ucnk.ff.cuni.cz/syn2005.php

[3] http://snowball.tartarus.org/algorithms/dutch/stop.txt

[4] http://korpus.pwn.pl/stslow_en

	German	Dutch	French	Spanish	Polish	Czech
German	-	0.97	0.95	0.95	0.80	0.72
Dutch	0.90	-	0.90	0.89	0.62	0.67
French	0.96	0.96	-	0.95	0.78	0.71
Spanish	0.95	0.96	0.87	-	0.74	0.69
Polish	0.53	0.41	0.61	0.48	-	0.49
Czech	0.47	0.36	0.54	0.39	0.67	-

Table 3: Source language identification evaluation (F-Score)

Translated Text	Accuracy
German	88.2%
Dutch	81.1%
French	87.4%
Spanish	84.7%
Polish	51.3%
Czech	50.5%

Table 4: Source language identification evaluation (Accuracy)

ancestor. In the vast majority of cases, members of the same language family share a considerable number of words and grammatical structures. In the experiment, we consider three language families: Romance languages (French and Spanish), Germanic languages (German and Dutch), and Slavic languages (Polish and Czech).

With a Romance target language,[5] the identification of other Romance and of Germanic languages as translation sources performs high, with an F-Score of between 0.86 and 0.95. However, a noticeable drop in performance concerns the identification of the Slavic languages.

When we take a look at the confusion matrices for the respective classifications, we find that, for instance, most misclassifications in the French target language data are between the sources of Polish and Czech. For Germanic target languages, the pattern repeats: when translated into German or Dutch, Polish and Czech texts are hardest to identify as the correct source.

The Slavic target languages show a different pattern. Even in another Slavic target language, a Slavic source language cannot reliably be identified in our setting. In addition to this, translations into Slavic are harder to distinguish from each other. Misclassifications in this case show language family specific patterns: German is, for instance, most often misclassified as Dutch in both the Czech and the Polish data.

4.2 Source Translation Classification

Translated texts have distinctive features that make them different from original or non translated text. According to Baker (1993; 1996), Olohan (2001), Lavisoa (2002), Hansen (2003), and Pym (2005) there are some general properties of translations that are responsible for the difference between these two text types. Some of these properties are source and target language independent. According to their findings, a translated text will be similar to another translated text but will be different from a source text. In the past, researchers have used comparable corpora to validate these translation properties (Baroni and Bernardini, 2006; Pastor et al., 2008; Ilisei et al., 2009; Ilisei et al., 2010; Koppel and Ordan, 2011). Most of them used comparable corpora for two-class classification, distinguishing translated texts from the original texts. Only Koppel and Ordan (Koppel and Ordan, 2011) used English texts translated from multiple source languages. We perform similar experiments only for six European languages as shown in Table 1. In this experiment, the translated text in our training and test set will be a combination of all languages other than the target language. For example: when the original class contains original texts (source) in German, then the translation class contains texts that are translated German texts, translated from French, Dutch, Spanish, Polish, and Czech texts. Each class contains 200 chunks of texts, where as the translated class has 40 chunks from each of the source languages. The source language texts are extracted for the corresponding languages in a similar way from the Europarl corpus. Koppel and Ordan (2011) received the highest accuracy (96.7%) among all works noted above. The training and test data are generated in similar ways as in our previous experiment. That is, 80% of the data is randomly extracted for training and the rest of the data is used for testing. Expected F-Scores are calculated from 100 samples. Table 5 shows the evaluation results. Even though the classifier for German achieves around 99% accuracy, we cannot compare the result with Koppel and Ordan (Koppel and Ordan, 2011) as the amount of chunks for the classes are different. The classifiers for other languages also display very high accuracy.

The result of Table 5 shows that general translation properties exist for all languages used in this experiment.

[5] Target language refers to text translated into the language

Language	Accuracy	F-Score
German	99.9%	0.99
Dutch	95.1%	0.95
French	81.9%	0.81
Spanish	94.4%	0.94
Polish	93.3%	0.93
Czech	81.1%	0.81

Table 5: Source translation classification

5 Discussion

The results show that training a classifier based on the 100 most frequent words of a language is sufficient to obtain interpretable results. We find our results to be compatible with Koppel and Ordan (2011) who used 300 function words. A list of the 100 most frequent words is easily obtainable for a vast number of languages, while lists consisting strictly of function words are rare and cannot be produced without considerable additional effort.

While the 100 most frequent words of a language are sufficient to train a classifier for Germanic or Romance languages, it fails to perform equally well for Slavic languages. Koppel and Ordan (2011) claim that Toury's (1995) findings of *interference* of a translation hold true; we find the assumption to be too simplistic, since for Slavic text either as a source or target language this statement cannot supported.

Although function words do exist in all the languages we examined, the language families differ in the degree to which it is necessary to use them. For instance, French lacks a case system (Dryer and Haspelmath, 2011), and makes instead use of prepositions. On the other hand, Polish and Czech most extensively use (inflectional) affixes (Kulikov et al., 2006). Regarding the distribution of word frequencies, for both Polish and Czech, the use of affixes causes a flatter Zipf curve. Kwapien et al. (2010) put it so :"...typical Polish texts have smaller α [as exponent of the formula $f(r) \sim r^{-\alpha}$] than typical English texts (on average: 0.95 vs. 1.05)." This means that on average a more frequent word does not differ as much in its frequency from a word 10 ranks further down in Polish as it does in English. Consequently, there will be fewer instances of the 100 most frequent words in the same portion of text. This is an obvious reason why a classifier's training must remain weaker in comparison to languages with a steeper Zipf curve. There is a positive correlation to language family when considering the probability of finding the same strategy (e.g. prepositions vs. affixes). In summary, the fact that Slavic uses more affixes, or is more inflectional in linguistic terms, explains to some extent why the classifier performs worst for Slavic target text.

However, for Slavic source texts, the classification results are equally unsatisfactory, which has to be explained differently. One phenomenon contributing here could be that Romance and Germanic have a recent history of mutual loans and calques, which increases the probability of finding synonyms where one has a Romance origin and one a Germanic origin. In the case of a translation, the translator, when confronted with such a synonym, might choose the item similar to the source language within the target language, as this minimizes the translation effort, complies thus to an economy principle and has virtually no effect on the translation.[6] Making this choice, the translator unintentionally distorts the native frequency patterns for the target language. This could be one of the processes generating an imprint of translated text in the frequency spectrum, since function words are also subject to loaning and synonymy.

If the translator has a choice for translating a preposition/affix and neither of the possibilities is similar to the source language, nor a loanword or structurally similar, he/she will go for the predominant word or structure of the target language (since he/she is a native of the target language by translation industry standard), making the translation less different from native text. The data can be influenced by many additional variables such as differing translation paradigms influencing the choice of structures (free translation vs. faithful translation), different industry standards, the size of the chunks,[7] the quality of the translation source marking, the native tongue of the translator(s), the time pressure for delivery, the payment, the membership of all sample languages to the European subbranch of Indo-European languages, the qualities of the lists of the most frequent 100 words, the genre of the Europarl corpus, and possibly many more.

This said, we believe the best hypothesis for the

[6] For French to English translations an example would be the translators choice of "intelligent" as a translation for French "intelligent" in a place where "smart" would have been slightly more natural.

[7] Since a short text contains fewer anaphora and thus personal pronouns.

interpretation of the data is that a good classification result is reached firstly for languages with a more isolating structure, since they make less use of affixes and therefore more of function words, and should display steeper Zipf curves. Secondly, the classification result should be better, the more instances the text contains, where the translator for one token (or for one structure) of the source language has the choice between at least two words or structures in the target language with one of those being similar to the source language, the other being different. The number of such instances most probably correlates positively with the degree and quality of language relationship and language contact since the number of cognates, loans and calques does. However, this number can also be "accidentally high" for two unrelated languages when they overlap in grammatical structure. As has been postulated for instance by Croft (2003), languages undergo a cyclic development from structurally more isolating towards agglutinative to inflectional and then back to isolating. When a language is in a state of transition, which practically all languages are, they offer two structural encoding possibilities for one specific grammatical property, e.g., a genitive (for instance, inflectional (an affix) as in *Peter's house* and isolating (a preposition) as in *the house of Peter*). All languages should share structural properties, since there are only three types and each language has practically at least two.

Corroborating this rather complex hypothesis, we examine data on Bulgarian and Romanian. We take Bulgarian as the target language. The data showed that the classifier classifies Czech text no worse than Dutch or German and only slightly worse than French. When we replace Czech with Bulgarian and Spanish with Romanian in the German target language, the language family dependent pattern gets blurred and the identification of Polish performs quite well, that of Romanian relatively poor, while French is identified reliably. This together with the observation that Romanian is misclassified either as Polish or Bulgarian and Bulgarian is mostly misclassified as Romanian seems to be a strong hint towards the impact of the language specific usage of function words, linguistic structure, and the importance of language contact. Bulgarian and Romanian constitute the core of the most prominent linguistic contact zone or *sprachbund* ever written on the Balkans. This suggests that Romanian and Bulgarian translators may, due to grammatical convergence of their languages make, given two equivalent structures in any target language, the same (structurally motivated) choices and hence leave a very similar imprint. That is, sprachbund membership as well as language family could be decisive factors for a classifiers performance.

6 Conclusion

We have shown that *interference* as originally proposed by Toury (1995) is not supported by the data without making further assumptions. Language family and language contact should be considered separately for each language pair as sources for possible weak results of a classifier even when operating with function words as should be general structural similarity. As for the properties of translated text being universal, we found support for this in our data in a real n-ary validation setting. We have also shown that the much more easily obtainable lists of the 100 most frequent words work almost as well for classification as do longer lists that contain only function words.

7 Acknowledgments

We would like to thank Prof. Moshe Koppel and Dr. Noam Ordan for providing their data, and Prof. Dr. Alexander Mehler for suggestions and comments. We also thank Dr. Timothy Price for checking English and three anonymous reviewers. This work is funded by the LOEWE Digital-Humanities project at the Goethe-Universität Frankfurt.

References

Shlomo Argamon and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *The Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*.

Mona Baker. 1993. Corpus linguistics and translation studies - implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology. In Honour of John Sinclair*, pages 233–354. John Benjamins.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pages 175–186. Amsterdam & Philadelphia: John Benjamins.

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

Alan Bell, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2008. Predictability effects on durations of content and function words in conversational english. *Elsevier Journal of Memory and Language*, 60:92–111.

William Croft. 2003. *Typology and Universals*. Cambridge textbooks in linguistics. Cambridge University Press.

Mark Davies. 2006. *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. Taylor & Francis.

Matthew S. Dryer and Martin Haspelmath. 2011. The world atlas of language structures online.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations*, 11(1):10–18.

Silvia Hansen. 2003. *The Nature of Translated Text: An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Ph.D. thesis, University of Saarland.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2009. Towards simplification: A supervised learning approach. In *Proceedings of Machine Translation 25 Years On, London, United Kingdom, November 21-22*.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov, 2010. *Identification of translationese: A machine learning approach*, pages 503–511. Springer.

Zahurul Islam and Alexander Mehler. 2012. Customization of the europarl corpus for translation studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Leonid Kulikov, Andrej Malchukov, and Peter de Swart, editors. 2006. *Case, Valency and Transitivity*, volume 77 of *Studies in Language Companion Series*.

Jaroslaw Kwapien, Stanislaw Drozdz, and Adam Orczyk. 2010. Linguistic complexity: English vs. polish, text vs. corpus. *CoRR*, abs/1007.0936.

Sara Laviosa. 2002. *Corpus-based translation studies. Theory, findings, applications*. Amsterdam/New York: Rodopi.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.

Deryle Lonsdale and Yvon Le Bras. 2009. *A Frequency Dictionary of French: Core Vocabulary for Learners*. Routledge.

Sattar Lzwaini. 2003. Building specialised corpora for translation studies. In *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, Corpus Linguistics*.

Andrew Mccallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on 'Learning for Text Categorization'*.

Maeve Olohan. 2001. Spelling out the optionals in translation:a corpus study. In *Corpus Linguistics 2001 conference. UCREL Technical Paper number 13. Special issue.*

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? a corpus-based NLP study of convergence and simplification. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)*.

Jams W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Erlbaum Publishers.

Marius Popescu. 2011. Studying translationese at the character level. In *Recent Advances in Natural Language Processing*.

Anthony Pym. 2005. Explaining explicitation. In *New Trends in Translation Studies. In Honour of Kinga Klaudy*, pages 29–34. Akadmia Kiad.

Gideon Toury. 1995. *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam/Philadelphia.

Hans van Halteren. 2008. Source language markers in europarl translations. In *International Conference inComputational Linguistics(COLING)*, pages 937–944.