Investigating Genre and Method Variation in Translation Using Text Classification

Marcos Zampieri[1,2] and Ekaterina Lapshinova-Koltunski[1]

[1] Saarland University, Germany
[2] German Research Center for Artificial Intelligence (DFKI)

Abstract. In this paper, we propose the use of automatic text classification methods to analyse variation in English-German translations from both a quantitative and a qualitative perspective. The experiments described in this paper are carried out in two steps. We trained classifiers to 1) discriminate between different genres (fiction, political essays, etc.); and 2) identify the translation method (machine vs. human). Using semi-delexicalized models (excluding all nouns), we report results of up to 60.5% F-measure in distinguishing human and machine translations and 45.4% in discriminating between seven different genres. More than the classification performance itself, we argue that text classification methods can level out discriminative features of different variables (genres and translation methods) thus enabling researchers to investigate in more detail the properties of each of them.

Keywords: human and machine translation, text classification, genres

1 Introduction

Text classification is an important area of research in Natural Language Processing (NLP) and it has been applied in a wide range of tasks such as spam detection [1] and temporal text classification [2]. From a purely engineering perspective, researchers are interested in how well classification methods can distinguish between two or more classes and what kind of features and algorithms deliver the best performance in each task. In recent work [3, 4], however, state-of-the-art text classification methods were proposed to investigate language variation across corpora. These methods were successfully applied in the identification of languages, varieties and dialects, as well as genres.

The present study is an attempt to use the same techniques for the identification of translation varieties – translations which differ in genres, e.g. essays, fiction,or methods, i.e. human and machine. We train classifiers to distinguish translated texts according to either their genre or method of translation, using the VARTRA corpus [5], a collection of English to German translations. More than the classification results *per se*, we use (semi-)delexicalized representations aiming to reduce topical bias and, therefore, levelling out interesting linguistic features that can be further used in linguistic analysis and NLP applications.

2 Related Work and Theoretical Background

Genre-specific variation of translation is related to studies within register and genre theory, e.g. [6], [7], which analyse contextual variation of languages. In lexico-grammatical terms, this variation is reflected in the distribution of linguistic patterns, i.e. subject/objects, evaluative patterns, negation, modal verbs, discourse phenomena (e.g. coreference or discourse markers).

Multilingual genre analysis is concerned with the distribution of such lexico-grammatical patterns not only across genres but also across languages, comparing the settings specific for the languages under analysis, e.g. [7] on English, Nukulaelae Tuvaluan, Korean and Somali, [8] and [9] on English and German. Moreover, the latter two also consider genres in translations. Applying a quantitative approach, Neumann (2013) [9] analyses an extensive set of features and shows to what degree translations are adapted to the requirements of different genres. Other scholars [10–13], also integrate register analysis in translation studies. However, they either do not account for distributions of these features, or analyse individual texts only. De Sutter et al. (2012) [14] and Delaere & De Sutter (2013) [15] in their analysis of translated Dutch also pay attention to genre variation, but concentrate on lexical features only.

Whereas attention is paid to genre settings in human translation analysis, they have not yet been considered much in machine translation. There exist some studies in the area of SMT evaluation, e.g. errors in translation of new domains [16]. However, the error types concern the lexical level only, as the authors operate solely with the notion of domain and not genre. Domains represent only one of the genre parameters and reflect what a text is about, i.e. its topic, and further settings are thus ignored. Although some NLP studies, e.g. those employing web resources, do argue for the importance of genre conventions, see e.g. Santini et al. (2010) [17], genre remains out of the focus of machine translation. In the studies on adding in-domain bilingual data to the training material of SMT systems [18] or on application of in-domain comparable corpora [19], again, only the notion of domain is taken into consideration.

Studies involving translation methods mostly focus on translation error analysis, and human translation serves usually as a reference in MT evaluation tasks. Some of them do consider linguistic properties, or linguistically-motivated errors [20, 21]. The latter one includes style errors, which is partly related to genre.

To our knowledge, the only study investigating differences between human and machine translation is Volansky et al. (2011) [22]. The authors analyse human and machine translations, as well as comparable non-translated texts. They use a range of features based on the theory of *translationese* (see [23] or [24]) expecting that the features specific for human translations can also be used to identify machine translation. Some of the translationese features were investigated using NLP techniques [25–28] similar to the ones we propose in this paper. What is most important for our study, however, is the claim by Volansky et al. (2011) [22] that some features of human translations coincide with those of machine-translated texts, whereas other features are diversifying between these two translation methods.

3 Methods

3.1 Data

For the purpose of our study we use VARTRA [5], a corpus of multiple translations from English into German. These translations were produced by: (1) human professionals (PT1), (2) human student translators (PT2), (3) a rule-based MT system (RBMT), (4) a statistical MT system trained with a large quantity of unknown data (SMT1) and (5) a statistical MT system trained with a small amount of data (SMT2). The genres available in VARTRA are: political essays (ESS), fictional texts (FIC), instruction manuals (INS), popular-scientific articles (POP), letters of share-holders (SHA), prepared political speeches (SPE), and touristic leaflets (TOU). Each subcorpus represents a translation variety, a translation setting which differs from all others in both method and genre (e.g. PT1-ESS or PT2-FIC, etc.).

Before classification was carried out, we split the corpus into sentences (of size between 12 and 24 tokens). This created 6,200 instances. The data was then split into a training (80%) and a test (20%) set.

The features used in different experiments include bag-of-words (bow), word bigrams, word trigrams and word 4-grams. The novelty of our approach is that we substitute all nouns with placeholders in some of the experiments. This results in what we call a semi-delexicalized text representation, which lies between fully delexicalized representations [3] and the classical bag-of-words or n-gram language models. Previous studies [4, 29] show that named entities significantly improve the result of text classification systems, so we decided to use this semi-delexicalized representation to minimize topic variation. The decision was motivated by both our goal of investigating translation variation influenced by both genre and method, and our aim to obtain a robust classification method that could perform well on different corpora.

3.2 Algorithms

We used two algorithms in our experiments. The first is a Naive Bayes (NB) classifier using bag-of-words as features. Naive Bayes classifiers work based on an independence assumption (the presence of a particular feature of a class is not related to the presence of any other feature), which is particularly useful for supervised learning and makes them extremely fast.

The second algorithm is based on a likelihood function calculated over n-gram language models as described by Zampieri and Gebre (2014) [30]. The language models can contain characters and words (e.g. bigrams and trigrams), linguistically motivated features such as parts-of-speech (POS) or morphological categories [4], or (semi-) delexicalized models such as the one we explore here.

4 Results

In this section, we present the results we obtained in different classification experiments. For the evaluation step we used standard NLP metrics such as

Precision, Recall and F-Measure. The linguistic analysis and discussion of the most important differences between both method and genre variation will be presented later in section 4.5.

4.1 Genres and Methods

The first experiment shows why it is important to use semi-delexicalized features in a dataset that represents both dimensions of variation in translation (genre and method). The question posed at this stage is simple: how different are the samples with respect to methods and genres? We use the aforementioned Naive Bayes classifier trained on (non-delexicalized) bag-of-words. In Table 1 we present the results as well as a baseline computed based on the random assignment of all documents to a particular class.

Type	**Classes**	**Precision**	**Recall**	**F-Measure**	**Baseline**
Genres	7	57.4%	57.8%	57.3%	14.2%
Methods	5	35.9%	36.2%	35.3%	20.0%

Table 1. Naive Bayes: Genres and Methods

The results of this preliminary experiment show that the classifier was able to distinguish between the seven translation genres with up to 57.3% F-Measure and between the five translation methods with up to 35.3% f-measure. The method is aided by named entities and content words that are domain specific and, therefore, influence the performance of the classifier. Therefore, we use placeholders to substitute nouns (both named entities and common nouns) to minimize topical bias in the following experiments. At the same time, the results of the present experment will allow us to compare classification performance of non-delexicalized vs. semi-delexicalized representations.

4.2 Translation Methods

In the next experiment, we take a closer look at the differences between five methods of translation (PT1, PT2, RBMT, SMT1 and SMT2) while minimizing topic influence, i.e. trying to distinguish between them excluding all nouns.

Classes	**Precision**	**Recall**	**F-Measure**	**Baseline**
5	35.1%	35.9%	34.9%	20.0%
4	43.2%	44.9%	43.1%	25.0%

Table 2. Naive Bayes: Translation Method

The data contains outputs of two different SMT systems and in this step, we decide to merge them into a unique class of SMT. This was mainly done to answer the question of whether this kind of distinction is meaningful in practical terms, and whether the outputs of SMT1 and SMT2 are significantly different.

The results (presented in Table 2) improved substantially after the grouping. In the five-class setting the f-measure obtained for class SMT1 was the lowest of all 26.4%, whereas the SMT class could obtain the best result in the four-class setting (58.5%). This indicates that the outputs of both systems contain similar features.

4.3 Different Genres: Different Language?

In the next step, we try to automatically distinguish between different genres represented in the dataset. For this experiment, we also use the semi-delexicalized features.

Classes	Precision	Recall	F-Measure	Baseline
7	45.5%	46.1%	45.4%	14.2%

Table 3. Naive Bayes: Genres in Translation

As seen from Table 3, the automatic distinction between the seven genres achieved ca. 45% of both Precision and F-measure. This performance is substantially above the 14.2% baseline which indicates that the genres in VARTRA are essentially different. However, we are also interested in whether the classifier's performance for genre discrimination was consistent across all translation methods. For this step, we perform genre classification within each translation method (with both SMT outputs in one class) and the result can be seen in figure 1.

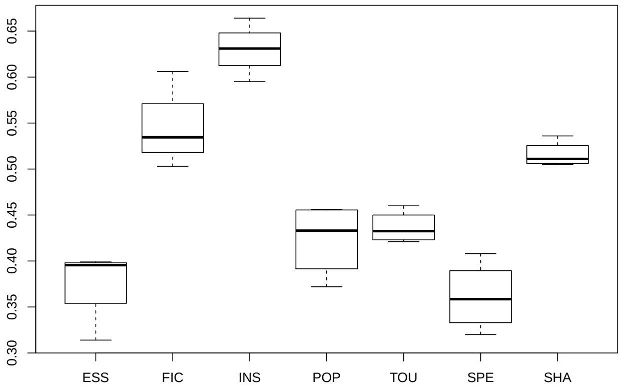

Fig. 1. Genre Distinction Across Method

The performance across all seven genres is constant regardless of the translation method applied. For example, instruction manuals (INS) followed by fiction (FIC) are the easiest genre to identify in all four translation methods, whereas speech (SPE) and essays (ESS) are consistently regarded as the most problematic ones. All the results are significantly higher than the expected baseline accuracy. The nominal values used to generate figure 1 are presented in table 4 on a scale from 0 to 1 with three decimal digits. The baseline we consider is once again the majority class, 14.2% f-measure.

Method	**ESS**	**FIC**	**INS**	**POP**	**TOU**	**SPE**	**SHA**
PT2	0.399	0.533	0.595	0.372	0.421	0.346	0.536
PT1	0.314	0.606	0.664	0.456	0.425	0.371	0.507
RBMT	0.397	0.536	0.632	0.411	0.440	0.320	0.515
SMT	0.394	0.503	0.630	0.455	0.460	0.408	0.505

Table 4. Genre Distinction Across Method

4.4 Human vs. Machine

In the last experiment, we investigate whether the differences between the four translation methods are weaker than between a less fine-grained classification into human and machine translation. For this step, we unify PT1 and PT2 into one class, and RBMT and SMT1 and SMT2 into the other. We also tested different sets of semi-delexicalized features, i.e. bigrams, trigrams and 4-grams to find out which allow the best classification results, see Table 5. In all three scenarios, the model performs above the expected baseline of 50.0% F-measure. The best performance, however, is obtained for the trigram model (60.5% f-measure and 61.1% precision).

Features	**Precision**	**Recall**	**F-Measure**	**Baseline**
bigrams	53.3%	53.3%	53.3%	50.0%
trigrams	61.1%	60.0%	60.5%	50.0%
4-grams	55.2%	54.2%	54.7%	50.0%

Table 5. N-grams: Human x Machine

As the amount of training data is not large, from 4-grams onwards the method seems to suffer from data sparsity and as can be expected, performance drops.

4.5 Feature Analysis

This section aims to identify the most informative features from the semi-delexicalized n-grams in our experiments. This step is manual and carried out by looking through the most informative features and thus discriminative for certain genres and methods in our translation data. We evaluated the trigrams, as the performance of trigram models achieved the best results in the classification task. The list of the features specific either to human or machine translation is shown in Table 6. Using the same strategy, we generate a list of features discriminating genre pairs (for the sake of space, we display political essays and fictional texts) in Table 7.

human	**machine**
full nominal phrase (with def./indef. modif.)	full nominal phrase (with def./indef./poss. modif.)
personal reference (1st pers. plural)	personal reference (1st pers. sg)
extended reference (demonst.)	extended reference (pers.)
prepositional phrase with local meaning	prepositional phrase with different meanings
discourse markers with additive meaning	discourse markers with adversative meaning

Table 6. Features discriminating between human and machine translations

ESS	**FIC**
passive constructions	active verbs
modal verbs with the meaning of volition and obligation	
to-infinitives	
prepositional phrase	predicative adjectives
demonstrative reference	personal reference
discourse markers with additive meaning	discourse markers with adversative meaning

Table 7. Features discriminating between political essays and fictional texts

Semi-delexicalized trigrams consist of a sequence of words and placeholders, e.g. *können PLH PLH*, *zu erfüllen hat*, *das PLH*, *aber*, etc. Intuitively, we try to recognize more abstract categories, i.e. modal verbs with the meaning of possibility, infinitive clauses, discourse markers with adversative function for the given trigrams. As seen from the lists, both translation methods have similar discriminating features, i.e. full nominal phrases, coreferring expressions, prepositional phrases and discourse markers. However, the differences between them can be identified on a more fine-grained level: if we take into account morphological preferences and the scope of referring expressions, the meaning of prepositional phrases and discourse markers. All these phenomena seem to be related to participants and structures of textual discourse.

The features that turn out to be specific for genres include verbs and verbal constructions, further types of phrases, and also different types of coreferring expressions and discourse markers. Genre-discriminating features are also, as in case of methods, on a more fine-grained level. However, the level of description is not on morphological, but rather on syntactic level (active vs. passive, prepositional vs. adjectival phrases). Moreover, they describe rather processes than participants of discourse. The last features coincide in both tables (additive vs. adversative construction), which means that they are informative in both genre and method classification.

Our preliminary observations on features coincide with the results of empirical analyses on genres, e.g. those obtained by Neumann (2013) [9]. For instance, the author point to personal pronouns, predicative adjectives, mental and verbal processes as indicators of narration and casual style which are specific for fictional texts. Polticial essays, which are characterized as expository texts with rather neutral style, contain relational processes, verbs of declarative mood, frequent nominalsations and almost no personal pronouns.

We believe that we need a more detailed analysis of the resulting features to have firm basis to build upon in our final conclusions on the features. For instance, the definition of mood and tense of verbal phrases, as well as their membership in a certain semantic verb class would contribute to a better specification of genres. Moreover, this step can be automatized with the help of existimg morphological tools, taggers and wordnets, which is however, beyond the scope of the present paper.

The resulting lists of features can be beneficial for not only genre classification task but also for machine translation task, as they can help to automatically differentiate between human and machine translation.

5 Conclusion and Outlook

This paper is, to our knowledge, the first attempt to use text classification techniques to discriminate methods and genres in translations and to identify their specific features and relevant systemic differences in a single study. We report results of up to 60.5% F-measure in distinguishing human and machine translations and 45.4% in discriminating between seven different genres.

We used different algorithms and sets of features to study variation in English-German translation data. For that we used not only the classical bag-of-words and n-gram language models but also the use of (semi-)delexicalized representations along with classical bag-of-words and n-gram language models, which helps us to decrease the thematic bias in classification. The aim was both the discrimination of methods and genres *per se*, and also the identification of relevant systemic differences across genres and methods of translation.

The results of our analysis can find application in both human and machine translation. In the first case, they deliver valuable knowledge on the translation product, which is influenced by the methods used in the process and the context of text production expressed by the genre. In case of machine translation, the results will provide a method to automatically identify genres in translation data thus helping to separate out-of-genre data from a training corpus.

The aforementioned practical applications of the results are part of our future work, which will also include tests with other classification algorithms such as the popular support vector machines [31] used in Petrenz and Webber (2012) for a similar task [32]. We also plan to automate the generation of more abstract categories for the informative features as well as to experiments other kinds of de-lexicalized representations such as the one used by Quiniou et al. (2012) [33]. Finally, we would like to carry out further and more detailed linguistic analysis.

References

1. Medlock, B.: Investigating classification for natural language processing tasks. Technical report, University of Cambridge - Computer Laboratory (2008)
2. Niculae, V., Zampieri, M., Dinu, L.P., Ciobanu, A.M.: Temporal text ranking and automatic dating of texts. In: 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014). (2014)
3. Diwersy, S., Evert, S., Neumann, S.: A semi-supervised multivariate approach to the study of language variation. Linguistic Variation in Text and Speech, within and across Languages. (2014)
4. Zampieri, M., Gebre, B.G., Diwersy, S.: N-gram language models and POS distribution for the identification of Spanish varieties. In: Proceedings of TALN2013, Sable d'Olonne, France (2013) 580–587
5. Lapshinova-Koltunski, E.: VARTRA: A comparable corpus for analysis of translation variation. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, ACL (2013) 77–86
6. Halliday, M., Hasan, R.: Language, context and text: Aspects of language in a social-semiotic perspective. Oxford University Press, Oxford (1989)
7. Biber, D.: Dimensions of Register Variation. A Cross Linguistic Comparison. Cambridge University Press, Cambridge (1995)
8. Hansen-Schirra, S., Neumann, S., Steiner, E.: Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German. de Gruyter, Berlin, New York (2012)
9. Neumann, S.: Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German. De Gruyter Mouton, Berlin, Boston (2013)
10. House, J.: Translation Quality Assessment. A Model Revisited. Günther Narr, Tübingen (1997)
11. Steiner, E.: An extended register analysis as a form of text analysis for translation. In Wotjak, G., Schmidt, H., eds.: Modelle der Translation – Models of Translation. Leipziger Schriften zur Kultur-, Literatur-, Sprach- und Übersetzungswissenschaft, Leipzig (1996) 235–256
12. Steiner, E.: A register-based translation evaluation. TARGET, International Journal of Translation Studies **10 (2)** (1997) 291–318
13. Steiner, E.: Translated Texts. Properties, Variants, Evaluations. Peter Lang Verlag, Frankfurt/M. (2004)
14. De Sutter, G., Delaere, I., Plevoets, K.: Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. In: Quantitative Methods in Corpus-based Translation Studies: a Practical Guide to Descriptive Translation Research. Volume 51. John Benjamins Publishing Company, Amsterdam, The Netherlands (2012) 325–345

15. Delaere, I., De Sutter, G.: Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. Belgian Journal of Linguistics **27** (2013) 43–60
16. Irvine, A., Morgan, J., Carpuat, M., III, H.D., Munteanu, D.S.: Measuring machine translation errors in new domains. TACL **1** (2013) 429–440
17. Santini, M., Mehler, A., Sharoff, S.: Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., Santini, M., eds.: Genres on the Web: Computational Models and Empirical Studies. Springer (2010) 3–30
18. Wu, H., Wang, H., Zong, C.: Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: Proceedings of COLING-2008, Manchester, UK (2008) 993–1000
19. Irvine, A., Callison-Burch, C.: Using comparable corpora to adapt MT models to new domains. In: Proceedings of the ACL Workshop on Statistical Machine Translation (WMT). (2014)
20. Popovic, M., Ney, H.: Towards automatic error analysis of machine translation output. Computational Linguistics **37**(4) (2011) 657–688
21. Fishel, M., Sennrich, R., Popovic, M., Bojar, O.: Terrorcat: a translation error categorization-based mt quality metric. In: 7th Workshop on Statistical Machine Translation. (2012)
22. Volansky, V., Ordan, N., Wintner, S.: More human or more translated? original texts vs. human and machine translations. In: Proceedings of the 11th Bar-Ilan Symposium on the Foundations of AI With ISCOL. (2011)
23. Gellerstam, M.: Translationese in Swedish novels translated from English. In: Translation Studies in Scandinavia. (1986) 88–95
24. Baker, M., et al.: Corpus linguistics and translation studies: Implications and applications. Text and technology: In honour of John Sinclair **233** (1993) 250
25. 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
26. Ilisei, I., Inkpen, D., Pastor, G.C., Mitkov, R.: Identification of translationese: A machine learning approach. In: Computational Linguistics and Intelligent Text Processing. Springer (2010) 503–511
27. Volansky, V., Ordan, N., Wintner, S.: On the features of translationese. Literary and Linguistic Computing (2013)
28. Ciobanu, A.M., Dinu, L.P.: A quantitative insight into the impact of translation on readability. Proceedings of the 3rd PITR workshop (2014) 104–113
29. Gebre, B.G., Zampieri, M., Wittenburg, P., Heskens, T.: Improving native language identification with tf-idf weighting. In: Proceedings of the BEA, Atlanta, USA (2013)
30. Zampieri, M., Gebre, B.G.: Varclass: An open source language identification tool for language varieties. In: Language Resources and Evaluation (LREC). (2014)
31. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning (ECML), Springer (1998)
32. Petrenz, P., Webber, B.: Robust cross-lingual genre classification through comparable corpora. In: The 5th Workshop on Building and Using Comparable Corpora. (2012)
33. Quiniou, S., Cellier, P., Charnois, T., Legallois, D.: What about sequential data mining techniques to identify linguistic patterns for stylistics? In: Computational Linguistics and Intelligent Text Processing. Springer (2012) 166–177