Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach

Ekaterina Lapshinova-Koltunski and Mihaela Vela
Saarland University
A2.2 University Campus
D-66123 Saarbrücken
{e.lapshinova,m.vela}@mx.uni-saarland.de

Abstract

In this paper, we apply text classification techniques to prove how well translated texts obey linguistic conventions of the target language measured in terms of registers, which are characterised by particular distributions of lexico-grammatical features according to a given contextual configuration. The classifiers are trained on German original data and tested on comparable English-to-German translations. Our main goal is to see if both human and machine translations comply with the non-translated target originals. The results of the present analysis provide evidence for our assumption that the usage of parallel corpora in machine translation should be treated with caution, as human translations might be prone to errors.

1 Introduction: Motivation and Goals

In the present paper, we demonstrate that both manually and automatically translated texts differ from original texts in terms of *register*, i.e. language variation according to context (Halliday and Hasan, 1989; Quirk et al., 1985). Similar observations were made in other studies, such as those by Gellerstam (1986), Baker (1995) and Teich (2003), who show that translations tend to share a set of lexical, syntactic and/or textual features. Several studies, including (Ozdowska and Way, 2009; Baroni and Bernardini, 2006; Kurokawa et al., 2009) and (Lembersky et al., 2012), employ computational techniques to investigate these differences quantitatively, mainly applying text classification methods.

Our main aim is to show that human translations, which are extensively deployed as data for both training and evaluation of statistical machine translation (SMT), do not necessarily obey the conventions of the target language. We define these conventions as register profiles on the basis of comparable data in the form of original, non-translated texts in the target language. These register-specific profiles are based on quantitative distributions of features characterising certain registers derived from theories described in Section 2.1 below. The non-translated data set and the corresponding register-specific features are used to train classifiers, for which we apply two different classification methods (see Section 3.4). The resulting classes serve as approximation for the standards of the target language. For the test data, we use multiple translations of the same texts produced by both humans and machines. The results of this analysis provide evidence for our assumption that we should treat the application of human translations in multilingual technologies, especially SMT (for instance, its evaluation), with caution. Our results show that there is a need for new technologies which would allow a machine-translated text to be a closer approximation to the original text in terms of its register. However, we are not aiming to provide solutions for this problem in the paper, but rather to show the importance of registers for both human and machine translation.

2 Related Work

2.1 Main notions within register theory

Studies related to register theory, e.g. by Quirk et al. (1985), Halliday and Hasan (1989) or Biber (1995), are concerned with contextual variation of languages, and state that languages vary with respect to usage context within and across languages. For example, languages may vary according to the activity of the involved participants or the relationship between speaker and addressee(s). These parameters correspond to the variables of (1) *field*, (2) *tenor* and (3) *mode* de-

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 122–131,
Lisbon, Portugal, 17 September 2015.

fined in the framework of systemic functional linguistics (SFL), which describes language variation according to situational contexts; see, for instance, studies by Halliday and Hasan (1989) and Halliday (2004). These variables are associated with the corresponding lexico-grammatical features. Field of discourse is realised in term patterns or functional verb classes, such as activity (*approach, supply*, etc.), communication (*answer, inform, suggest*, etc.) and others. Tenor is realised in modality expressed by modal verbs (*can, may, must*, etc.) or stance expressions (used by speakers to convey personal attitude to the given information, e.g. adverbs like *actually, certainly, amazingly, importantly*). And mode is realised in information structure and textual cohesion, e.g. coreference via personal (*she, he, it*) and demonstrative (*this, that*) pronouns. Thus, differences between registers can be identified through the analysis of occurrence of lexico-grammatical features in these registers; see Biber's studies on linguistic variation (Biber, 1988; Biber, 1995; Biber et al., 1999). The field of discourse also includes *experiential domain* realised in the lexis. This corresponds to the notion of domain used in the machine translation community. However, it also includes colligation (morpho-syntactic preferences of words), in which grammatical categories are involved. Thus, domain is just one of the parameter features a register can have.

2.2 Register in translation

Whereas attention is paid to register settings in human translation as described by House (2014), Steiner (2004), Hansen-Schirra et al. (2012), Kruger and van Rooy (2012), De Sutter et al. (2012), Delaere and De Sutter (2013) and Neumann (2013), registers have not yet been considered much in machine translation. There are some studies in the area of SMT evaluation, e.g. those dealing with the errors in translation of new domains (Irvine et al., 2013). However, the error types concern the lexical level only, as the authors operate solely with the notion of domain (field of discourse) and not register (which includes more parameters, see Section 2.1 above). Domains reflect what a text is about, its topic. So, consideration of domain alone would classify news reporting on certain political topics together with political speeches discussing the same topics, although they belong to different registers. We expect that texts from the latter (political speeches) translated with a system trained on the former (news) would be lacking in persuasiveness, argumentation and other characteristics reflected in their lexico-grammatical features, for instance, imperative verbal constructions used to change the addressee's opinion, or interrogatives as a rhetorical means. The similarity in domains would cover only the lexical level, in most cases terminology, ignoring the lexico-grammatical patterns specific for the given register (see the discussion on domain vs. register in (Lapshinova-Koltunski and Pal, 2014)). More recently, Zampieri and Lapshinova-Koltunski (2015) and Lapshinova-Koltunski (inpress) have shown the dominance of register-specific features of translated texts over translation-method-specific ones. Although some NLP studies, for example, those employing web resources, do argue for the importance of register conventions, see (Santini et al., 2010) among others, register remain out of the focus of machine translation. One of the few works addressing the relevance of register features for machine translation is (Petrenz, 2014), in which the author uses text features to build cross-lingual register classifiers.

2.3 The impact of target and source texts in translation quality

If languages differ in their register settings (Hansen-Schirra et al., 2012; Neumann, 2013), the register profiles of the source and the target are also different. In his work on translation quality, Steiner (2004) applies 'the guiding norms' for evaluation derived from both the target language and the register properties of the source. In MT evaluation, various methods and metrics of evaluation commonly rely on reference translations, which means that the relation between machine-translated texts and human translations is considered. We believe that we cannot judge the quality of a translation by merely comparing a source and a (reference) translation. Quality assessment also requires consideration of the target language conventions, i.e. those derived from comparable texts (belonging to the same registers) in a target language.

Some recent corpus-based studies on translation (Baroni and Bernardini, 2006; Koppel and Ordan, 2011) have shown that it is possible to automatically predict whether a text is an original or a

translation. Furthermore, automatic classification of original vs. translated texts found application in machine translation, especially in studies showing the impact of the nature (original vs. translation) of the text in translation and language models used in SMT. Kurokawa et al. (2009) show that for an English-to-French MT system, a translation model trained on an English-to-French data performs better than one trained on French-to-English translations. However, the 'better performance' of an SMT system is measured by BLEU scores (Papineni et al., 2002), indicating to which extend an SMT output comply with a reference, which is a translation itself. Inspired by Kurokawa et al. (2009)'s work, Lembersky et al. (2012) show that the BLEU score can be improved if they apply language models compiled from translated texts and not non-translated ones. They also show that language models trained on translated texts fit better to reference translations in terms of perplexity. In fact, this confirms the claim that machine translations comply more with translated rather than with non-translated texts produced by humans. It results in the improvement of the BLEU score, but not necessary leading to a better quality of machine translation. Several studies have confirmed the fact that BLEU scores should be treated carefully, see (Callison-Burch et al., 2006; Vela et al., 2014a; Vela et al., 2014b).

3 Methodology and Resources

3.1 Research questions

Following the assumption that translated language should normalise the linguistic features (like those described in 2.1 above) in order to adapt them to target language conventions, we use a classification method (using German original data for training, and translations for testing) to prove if register settings in translations correspond to those of the comparable originals. It is not our intention to directly measure the differences between originals and translations in the same language. This has been a common practice in numerous corpus-based translation studies that concentrate mostly on features in isolation, not paying much attention to their correlation: see Section 2.3 above.

Instead, we want to investigate if the register-related differences modelled for non-translated texts also apply for translation, and if they are sensitive to the variation according to the translation method involved. In fact, we model register classes for German non-translated texts, and test them on German translations from English source texts which are comparable to German non-translated ones in terms of registers. We expect that for some types of translations (e.g. human vs. machine), registers are identified more easily than for the others. We measure the accuracy scores (*precision*, *recall* and *f-measure*) which are class-specific numbers obtained for various sets of data: see details in Section 3.4.

Our classification analysis is structured according to the following questions: (1) Do translations from English into German correspond to German originals in their register settings? (2) Which translation can be classified best in terms of register? (3) Is there any difference between human (PT1 and PT2) and machine translations (RBMT and SMT), if register settings are concerned?

3.2 Feature selection

The input for the classifiers represents a set of features derived from register studies described in Section 2.1 above. These features constitute lexico-grammatical patterns of more abstract concepts, i.e. textual cohesion expressed via pronominal coreference or other cohesive devices, evaluative patterns (e.g. *it is interesting/important that*) and others. Several studies (Biber et al., 1999; Neumann, 2013), successfully employed these features for cross-lingual register analysis, showing that they reflect intra-lingual linguistic variation. In our previous work, see (Lapshinova-Koltunski, inpress), we applied a similar set of features to analyse register variation in translation.

Register features should reflect linguistic characteristics of all texts under analysis, be content-independent (do not contain terminology or keywords), be easy to interpret yielding insights on the differences between variables under analysis. So, we use groupings of nominal and verbal phrases instead of part-of-speech n-grams, as they are easier to interpret as n-grams. The set of selected features for the present analysis is outlined in Table 1. The first column denotes the extracted and analysed patterns, the second represents the corresponding linguistic features, and the third denotes the three context parameters according to register theory as previously described in Section 2.1.

The number of nominal and verbal parts-of-speech, chunks and nominalisations (*ung-*

nominalisations) reflect participants and processes in the field parameter. The distribution of abstract or general nouns and their comparison to other nouns gives information on the vocabulary (parameter of field). Modal verbs grouped according to different meanings defined by Biber et al. (1999), and evaluation patterns express modality and evaluation, i.e. the parameter of tenor. Content words and their proportion to the total number of word in a text represent lexical density, which is an indicator of the parameter of mode. Conjunctions, for which we analyse distributions of logico-semantic relations, belong to the parameter of mode as they serve as discourse-structuring elements. Reference, expressed either in nominal phrases or in pronouns, reflects textual cohesion (mode). Overall, we define 21 features[1] representing subtypes of the categories given in Table 1.

3.3 Corpus resources

German non-translated texts (GO=German originals) used as training data for classifiers are extracted from CroCo (Hansen-Schirra et al., 2012), a corpus of both parallel and comparable texts in English and German. The dataset contains 108 texts which cover seven registers: political essays (ESSAY), fictional texts (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to share-holders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The decision to include this wide range of registers is justified by the need for heterogeneous data for our experiment. Therefore, the dataset contains both frequently machine-translated texts, e.g. SPEECH, ESSAY and INSTR, and those, which are commonly not translated with MT systems, such as FICTION or POPSCI. The number of texts per register in GO comprises approximately 36 thousand tokens.

The translation data set is smaller (50 texts) and contains multiple German translations (both human and machine) of the same English texts, see (Lapshinova-Koltunski, 2013). Translations vary in (1) translator expertise, which differentiate them into professional (PT1), and novice (PT2) translations; and in (2) translation tools, which include rule-based (RBMT) and statistical machine translation (SMT). PT1 was exported from the above mentioned corpus CroCo (Hansen-Schirra et al., 2012), which contains not only GO but also comparable German translations from English originals covering the same registers as in GO. PT2 was produced by trainee translators with at least BA degree, who have little experience in translation. All of them produced translations using different translation memories (available via OPUS[2]) with the help of Across[3], a computer-aided translation tool which can be integrated into the usual work environment of a translator. The rule-based machine translation variant was produced with SYSTRAN6[4] (Systran, 2001), whereas for statistical machine translation, a Moses-based system was used which was trained with EUROPARL, a parallel corpus containing texts from the proceedings of the European parliament (Koehn, 2005). Every translation subcorpus has the same number of texts, as the data represent multiple translations of the same texts.

To extract the occurrences of register features described in 3.2, we annotate all subcorpora with information on token, lemma, part-of-speech (pos), syntactic chunks and sentence boundaries using Tree Tagger (Schmid, 1994). The features are then defined as linguistic patterns in form of the Corpus Query Processor regular expressions (Evert and Hardie, 2011), available within the CWB tools (CWB, 2010). As the procedures to annotate and to extract features are fully automatic, we expect them to influence some of the results, e.g. lexical density, which is entirely based on the pos categories assigned by Tree Tagger. So, the erroneous output of the tagger could also effect the results on the features. However, a gold-standard corpus is needed to evaluate the performance of the feature extraction, which is beyond the goals of the present work.

3.4 Classification methods

For our classification task, we train two different models by using two different classifiers on German original data. The applied techniques include (1) *k-nearest-neighbors* (KNN), a non-parametric method, and (2) *support vector machines* (SVM) with a linear kernel, a supervised method, both commonly used in text classification.

[1] Note that we select 18 only for the final classification, see details in Section 3.4.

[2] http://opus.lingfil.uu.se/

[3] http://www.across.net/

[4] Note that SYSTRAN6 is a rule-based system. With the release of SYSTRAN7 in 2010, SYSTRAN implemented a hybrid (rule-based/statistical) machine translation technology which is not involved in this analysis.

pattern	feature	parameter
nominal and verbal chunks	participants and processes	field
ung-nominalisations and general nouns	vocabulary and style	
modals with the meanings of permission, obligation, volition	modality	tenor
evaluative patterns	evaluation	
content vs. functional words	lexical density	mode
additive, adversative, causal, temporal, modal conjunctive relations	logico-semantic relations	
3rd person personal and demonstrative pronouns	cohesion via reference	

Table 1: Features under analysis

When using KNN, the input consists of the K closest training examples in the feature space, and the output is a class membership. This method is instance-based, where each instance is compared with existing ones using a distance metric, and the distance-weighted average of the closest neighbours is used to assign a class to the new instance (Witten et al., 2011).

For our experiments we have to determine the final number for K and the most appropriate number of features used in the classification, for which the Monte Carlo cross-validation method is used (as this method provides a less variable, but more biased estimate). Having the most significant features in the set, we calculate the distribution of errors by cross-validating 10 pairs of training-validation sets and choosing K[5] and the tuple (*numberOfFeatures=17, K=11*) is selected for our classification analysis. The classification is then performed on the translation (test) data, using the *knn* package (Ripley, 1996; Venables and Ripley, 2002).

Because the features that we select for classification have different measurement scales in our data, both the training and the test data are standardised using Formula 1 below.

$$x_s = \frac{x - Min}{Max - Min} \quad (1)$$

Applied to our corpus, the classification algorithm is supposed to store all available cases in GO (108 data points) and classify new cases in translation data (50 data points) based on a distance function measure, for which Euclidean distance is used.

[5]with in an interval between 3 and 19

When using SVM models (Vapnik and Chervonenkis, 1974), the learning algorithm tries to find the optimal boundary between classes by maximising the distance to the nearest training data of each class. Given labelled training data, the algorithm outputs an optimal hyperplane which categorises new instances. One of the reasons why SVM are used often is their robustness towards overfitting as well as their ability to map to a high-dimensional space.

We apply SVM on the same data set as for KNN, meaning that the same standardised training (108 data points) and test (50 data points) sets, as well as the same features were selected. We also apply the same procedures, training the SVM classifier on the German originals and testing the resulting model on the German translations.

First, both classifiers are tested in the 10-fold cross-validation step (Section 4.1). Judging the performance scores in terms of *precision*, *recall* and *f-measure*, we decide on classes (registers) used to answer the research questions formulated in Section 3.1. As already mentioned above, these scores are class-specific and indicate the results of automatic assignment of register labels to certain non-translated texts. In case of precision, we measure the class agreement of the data with the positive labels given by the classifier. For example, there are ten German fictional texts in our data. If the classifier assigns FICTION labels to ten texts only, and all of them really belong to FICTION, then we will achieve the precision of 100%. With recall, we measure, if all translations of a certain register were assigned to the register class they should belong to. So, if we have ten fictional texts, we would have the highest recall if all of them are assigned with the FICTION label. F-measure combines both precision and recall, and is under-

stood as the harmonic mean of both. For the tests on translation data, we select registers for which we could achieve at least 60% of f-measure.

Next, we apply the classifiers on the translation data, which is split into different variables according to the posed research questions in Section 3.1, i.e. all translation variants or human vs. machine. As in the previous step, we also analyse the scores for precision, recall and f-measure, as our assumption is that these values would indicate if German translated texts correspond with their register settings to the non-translated German. Hence, the higher the values, the better a translation correspond to comparable originals.

4 Classification analysis

4.1 Classifier performance

In the first step, we validate the performance of our classifiers trained on German originals with the selected set of features. As we don't have comparable data in German at hand to test the classifier, we perform 10-fold cross-validation for both KNN and SVM classifiers. The results of the cross-validation are presented in Table 2.

Overall, we achieve up to 80% of precision for the classification of GO with the register features. However, the performance of the classifier is dependent on the nature of the registers involved. Some of them seem to be more difficult to model than others: e.g. compare the results for fictional texts with those for SHARE or SPEECH.

	precision		**recall**		**f-measure**	
	KNN	SVM	KNN	SVM	KNN	SVM
ESSAY	0.43	0.64	0.70	0.61	0.53	0.62
FICTION	1.00	1.00	1.00	1.00	1.00	1.00
INSTR	1.00	1.00	0.64	0.79	0.78	0.88
POPSCI	0.75	0.89	0.90	0.80	0.82	0.84
SHARE	0.67	0.71	0.36	0.46	0.47	0.56
SPEECH	0.54	0.89	0.39	0.44	0.45	0.59
TOU	0.76	0.53	0.73	0.96	0.74	0.68
AVERAGE	0.74	0.81	0.67	0.72	0.69	0.74

Table 2: Classification results for GO per register

The best results are shown for fictional texts, popular-scientific texts and instruction manuals, for which the resulting f-measure amounts between 80-100%. SPEECH and SHARE reveal the lowest scores, and thus, are excluded from further analysis.

4.2 Question 1: Translations and register

Table 3 provides an overview of the f-measure values representing basically the diagonal of the confusion matrix of all classes (registers) under analysis, for the four different translation methods and two different classifiers. The table reveals that our classification algorithms perform differently depending on the register.

The best results are achieved for FICTION with both classification methods (lower performance is observed for PT2 with KNN and RBMT with SVM), where we observe f-measures up to 100%. This means that translations of English fictional texts best match the standards of German fiction. The worst results are observed for translations of political essays and popular-scientific texts, where missing correspondence with originals is observed for machine-translated texts in terms of SVM. The KNN values, although better, achieve the maximum of 53% for RBMT-POPSCI.

Misclassification results are observed for every class, varying in the translation method involved.

The classification results with both classifiers do not demonstrate the same results, e.g. SVM performs better for FICTION and INSTR, whereas KNN's best performance is observed for ESSAY, POPSCI and TOU. Therefore, we cannot claim that certain registers are generally more difficult to be identified in translated data than others, as the performance of the classifiers vary depending not only on the register but also the translation method involved.

4.3 Question 2: The best performance

To answer the second question, we compare the average values (for all classes) for precision, recall and f-measure for each translation variant in our data, as shown in Table 4.

	precision		**recall**		**f-measure**	
	KNN	SVM	KNN	SVM	KNN	SVM
PT1	0.56	0.49	0.71	0.72	0.61	0.51
PT2	0.53	0.68	0.67	0.58	0.55	0.44
RBMT	0.43	0.24	0.61	0.56	0.50	0.32
SMT	0.50	0.32	0.61	0.53	0.54	0.34

Table 4: Average values for the classification per translation variant

Ranking translations according to the calculated values, we observe the best performance of translations by humans with both classifiers. The differences between the KNN and SVM results are caused by the differences in the approach to learning: for KNN, all K neighbours influence the classification, whereas the SVM classifier draws a line

	ESSAY		FICTION		INSTR		POPSCI		TOU	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
PT1	0.45	0.13	0.86	0.86	0.62	0.52	0.55	0.60	0.43	0.44
PT2	0.52	0.27	0.75	1.00	0.35	0.35	0.67	0.29	0.40	0.30
RBMT	0.36	0.00	0.86	0.75	0.17	0.55	0.53	0.00	0.50	0.32
SMT	0.48	0.00	0.86	0.80	0.33	0.60	0.46	0.00	0.46	0.29
AVERAGE	0.45	0.10	0.83	0.85	0.37	0.51	0.55	0.22	0.45	0.34

Table 3: F-measure scores for classification per translation variant and register

to separate the data points. Significance analysis[6] confirms that the KNN results are similar for all translation varieties, as no significant difference can be observed (p-value of 0.99). This means that all translation variants correspond to comparable originals in a similar way. By contrast, the SVM values reveal variation, as the calculated p-value equals 0.03 (which is below the significance level of 0.05). Thus, we see that PT2 comply more with the register settings of the target language.

4.4 Question 3: Human vs. machine

In the following step, we compare the values for human and machine translations, analysing them per class (register). The results (see Table 5) show that both human and machine translations perform similarly, although both classifiers perform better on human translations (with the average f-measures of 0.58 vs. 0.48 for KNN and 0.52 vs. 0.33 for SVM). Our significance tests show that the results for HU vs. MT differ in terms of SVM (p-value of 1.59e-11), and is similar in terms of KNN (p-value of 0.08).

A more detailed analysis of the calculated values (presented in Figure 1) reveals much variation across registers in the results. Human translation performs better for certain registers only, i.e. ESSAY and POPSCI (both with KNN and SVM). The results for FICTION, INSTR and TOU vary depending on the classifier used. Table 6 indicates which translation method performed better for the given registers depending on the classifier used.

register	KNN	SVM
ESSAY	HU	HU
FICTION	MT	HU
INSTR	HU	MT
POPSCI	HU	HU
TOU	MT	HU

Table 6: Performance for human and machine translation across registers

[6]We perform Pearson's chi-squared test on the evaluation data.

5 Discussion and Outlook

We have shown that translations can be classified according to register features corresponding to the target language conventions. In case of a good classification performance, translations seem to adapt these conventions. However, we also observed misclassification cases, e.g. for tourism texts or those of political essays. We suppose that the reason for this lies in the nature of translated texts which differ from comparable originals. MT systems trained with such human translations result in the same kind of non-correspondence with the register standards of the target language. This might explain the similarities in our classification results for both humans and machines. While human translation characteristics in MT are often considered to be beneficial as they can improve the BLEU scores, we believe that the application of human translation as a reference should be treated with caution. There is a need for a closer approximation of the MT outputs to the original texts in terms of register, which are possible in form of high-level language models capturing register profiles in a target language. One of the ideas here is the application of such profiles (see as conventions of the target language) to rank translated texts, which might serve as basis for new techniques of MT evaluation. However, their implementation, as well as exploitation of such profiles for MT development, need a thorough elaboration of features, which is beyond the aims of the present study. In the area of MT development, we suggest that techniques such as document-wide decoding used for other discourse phenomena in Hardmeier et al. (2012) could be promising in the improvement of register profiles in machine-translated texts.

We believe that the knowledge on the discriminative features resulting from our classification can be beneficial for natural language processing, as they indicate register-specific differences of language means. For example, Petrenz and Webber (2011) show that within a newspaper corpus, the occurrence of the word *states* as a verb

	precision				recall				f-measure			
	HU		MT		HU		MT		HU		MT	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
ESSAY	0.53	0.67	0.53	0.00	0.54	0.12	0.45	0.00	0.53	0.20	0.49	0.00
FICTION	0.68	0.88	0.75	0.80	1.00	1.00	1.00	0.83	0.80	0.93	0.86	0.78
INSTR	0.42	0.28	0.25	0.40	0.65	1.00	0.25	1.00	0.51	0.44	0.25	0.57
POPSCI	0.80	0.88	0.44	0.00	0.50	0.33	0.63	0.00	0.61	0.44	0.52	0.00
TOU	0.32	0.24	0.37	0.19	0.75	0.80	0.70	0.90	0.45	0.37	0.48	0.31
AVERAGE	0.55	0.59	0.47	0.28	0.69	0.65	0.61	0.55	0.58	0.48	0.52	0.33

Table 5: Evaluation of classification results per human and machine translation

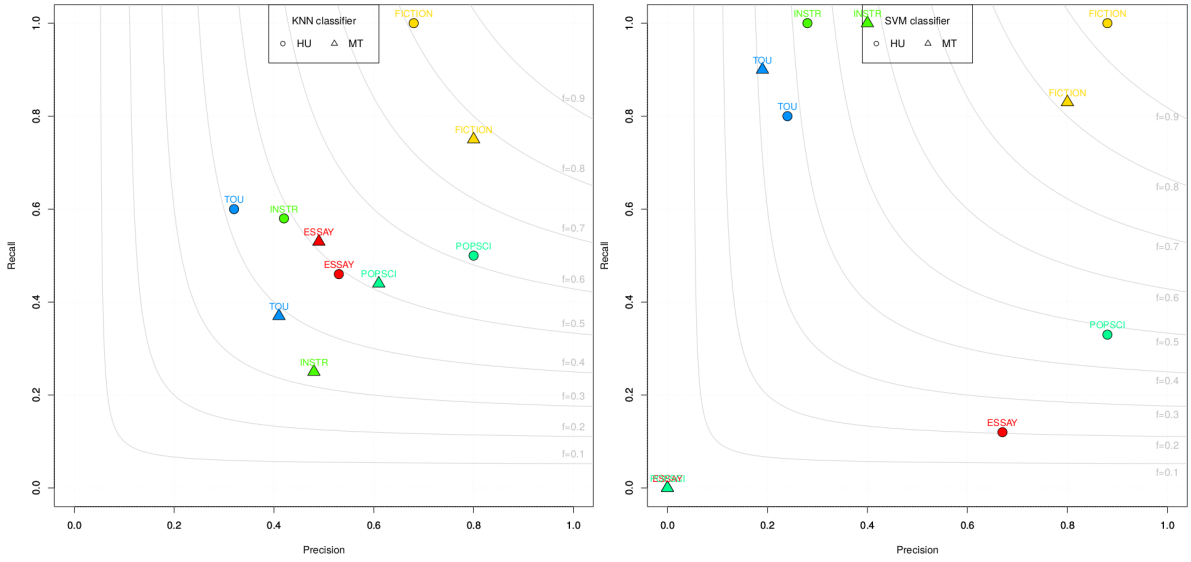

Figure 1: Evaluation of classification results per human and machine translation

is higher in letters than in editorials, and the cues on such specific features correlating with registers may impact system performance. The knowledge from confusion matrices can thus be useful for the decision if we can use an MT system trained on texts of one register and translate texts of another register which was commonly classified as the first one in our experiments. Experiments of this kind are part of our future work, which will also include inspection of the feature weights resulting from classification. The higher the weight of a feature, the more distinctive it is for a class, regardless of its positive or negative sign. A feature ranking will help us to determine the relative discriminatory force of certain features specific for a particular register, as described by (Teich et al., 2015) in their work on register diversification in scientific writing.

We also need to have a closer look at the features contributing to misclassification, as they might also serve as translation error indicators. For this, human assessments of quality is required, which involves manual evaluation of our translation data. The manual effort would also allow us to evaluate the performance of the automatic feature extraction, which might be erroneous, as stated in Section 3.3.

6 Acknowledgement

We thank Elke Teich, Erich Steiner and all anonymous reviewers for their constructive comments. We also gratefully acknowledge the help of Heike Przybyl in preparing the final version of this article. All remaining errors and misconceptions are our own.

References

Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243.

Marco Baroni 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.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Douglas Biber. 1995. *Dimensions of Register Variation. A Cross Linguistic Comparison*. Cambridge University Press, Cambridge.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-Evaluation the Role of Bleu in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

2010. The IMS Open Corpus Workbench. accessed February 2015.

Gert De Sutter, Isabelle Delaere, and Koen Plevoets. 2012. Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. In Michael P Oakes and Ji Meng, editors, *Quantitative Methods in Corpus-based Translation Studies: a Practical Guide to Descriptive Translation Research*, volume 51, pages 325–345. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Isabelle Delaere and Gert De Sutter. 2013. Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics*, 27:43–60.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics-2011 Conference*, Birmingham, UK.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

M.A.K. Halliday and Ruqaiya Hasan. 1989. *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford University Press, Oxford.

M.A.K. Halliday. 2004. *An Introduction to Functional Grammar*. Arnold, London.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, pages 1179–1190, Stroudsburg, PA, USA. Association for Computational Linguistics.

Juliane House. 2014. *Translation Quality Assessment. Past and Present*. Routledge.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Stefan Munteanu. 2013. Measuring machine translation errors in new domains. *TACL*, 1:429–440.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

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

Haidee Kruger and Bertus van Rooy. 2012. Register and the Features of Translated Language. *Across Languages and Cultures*, 13(1):33–65.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*.

Ekaterina Lapshinova-Koltunski and Santanu Pal. 2014. Comparabilty of corpora in human and machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Seventh Workshop on Building and Using Comparable Corpora*, Reykjavik, Iceland, May. European Language Resources Association (ELRA). LREC-2014.

Ekaterina Lapshinova-Koltunski. 2013. VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski. *in press.* Linguistic features in translation varieties: Corpus-based analysis. In G. De Sutter, I. Delaere, and M.-A. Lefer, editors, *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. Mouton de Gruyter. TILSM series.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Stella Neumann. 2013. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. De Gruyter Mouton, Berlin, Boston.

Sylwia Ozdowska and Andy Way. 2009. Optimal bilingual data for french-english pb-smt. In *EAMT 2009 – 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain, May.

Kishore Papineni, Salim Roukus, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37:385–393.

Philipp Petrenz. 2014. *Cross-Lingual Genre Classification*. Ph.D. thesis, School of Informatics, University of Edinburgh, Scotland.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Brian D. Ripley. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Marina Santini, Alexander Mehler, and Serge Sharoff. 2010. Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 3–30. Springer.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Erich Steiner. 2004. *Translated Texts. Properties, Variants, Evaluations*. Peter Lang Verlag, Frankfurt/M.

Systran. 2001. Past and present. Technical report.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2015. The linguistic construal of disciplinarity: A data mining approach using register features. *Journal of the Association for Information Science and Technology*, pages n/a–n/a.

Elke Teich. 2003. *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Vladimir N. Vapnik and Alexey J. Chervonenkis. 1974. *Theory of pattern recognition*. Nauka.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014a. Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 47–56, April.

Mihaela Vela, Anne-Kathrin Schumann, and Andrea Wurm. 2014b. Human Translation Evaluation and its Coverage by Automatic Scores. In *Proceedings of the LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE)*, pages 20–30, May.

William N. Venables and Brian D. Ripley. 2002. *Modern Applied Statistics with S*. Statistics and Computing. Springer.

Ian H Witten, Eibe Frank, and Mark A Hall. 2011. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, Burlington, Massachusetts.

Marcos Zampieri and Ekaterina Lapshinova-Koltunski. 2015. Investigating genre and method variation in translation using text classification. In Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue - 18th International Conference, TSD 2015, Plzen, Czech Republic, Proceedings*, Lecture Notes in Computer Science. Springer.