Register-Based Machine Translation Evaluation with Text Classification Techniques

Mihaela Vela m.vela@mx.uni-saarland.de
Saarland University, Saarbrücken, Germany

Ekaterina Lapshinova-Koltunski e.lapshinova@mx.uni-saarland.de
Saarland University, Saarbrücken, Germany

Abstract

This paper presents a novel approach to machine translation evaluation by combining register features – characterised by particular distributions of lexico-grammatical features – with text classification techniques. The goal of this method is to compare machine translation output with comparable originals in the same language, as well as with human reference translations. The degree of similarity – in terms of register features – between machine translations and originals, and machine translations and reference translations is measured by applying two text classification methods trained on 1) originals and 2) reference translations, and tested on machine translations. The results from the experiments prove our assumption that machine translations share register features rather with human translations than with non-translated texts produced by humans. This confirms that registers are one of the most important factors that should be integrated into register-based machine translation evaluation.

1 Introduction: Motivation and Goals

The state-of-the-art in evaluating machine translation (MT) nowadays is to measure lexical and eventually syntactic and semantic overlap between a machine translation (called hypothesis translation) and a human-produced reference translation. In this paper, we present a new approach to evaluation, integrating the knowledge on *register*, i.e. language variation according to context, as defined by Halliday and Hasan (1989) and Quirk et al. (1985). The difference in terms of *register* between original and translated texts has been shown by several studies (Hansen-Schirra et al., 2012; Kruger and van Rooy, 2012; Neumann, 2013), proving that translations tend to share a set of lexical, syntactic and/or textual features. More recent investigations by Baroni and Bernardini (2006), Kurokawa et al. (2009) and Lembersky et al. (2012) applied text classification methods to automatically identify these differences.

The aim of the research presented here is twofold: 1) to show that machine translations and the corresponding reference translations are related to each other in terms of register-specific features and as a consequence of this 2) to show that hypothesis translations and human translations share more than the lexical surface. The novel idea introduced here is the notion of register-specific features which relate reference and hypothesis translations, and therefore have implications for MT (evaluation).

We measure the "closeness" between comparable non-translated originals and machine translations, as well as between human and machine translations by applying two different classification methods. The classification is performed on the basis of extracted register-specific features for two data sets. First, we use original non-translated texts as training data and ma-

chine translations as test data. In a second step, human translations are used as training data and machine translations as test data. Our assumption is that in terms of register specificity quantified in the corresponding features, MT output is closer to the corresponding reference translations than to the comparable non-translated originals. We base our hypothesis on the fact that b) translations tend to normalise towards target language conventions and that a) machine translations will adapt more to these conventions than to source texts (Diwersy et al., 2014).

The remainder of the paper is structured as follows. In Section 2, we present related work from the areas of machine translation evaluation and register theory. In Section 3 we present our research questions, describe the selected features and resources used for the experiments, as well as the applied methods. Section 4 demonstrates the results of our analysis, and in Section 5, we discuss the outcome and give an outlook on further analyses.

2 Related Work

2.1 Machine Translation Evaluation

State-of-the-art MT evaluation applies automatic language-independent metrics such as BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) in order to compare MT output (hypothesis translation) with one or more human translations (reference translations). Several studies (Callison-Burch et al., 2006; Vela et al., 2014a,b) have confirmed the fact that BLEU scores should be treated carefully, thus advancing the development of new metrics. New evaluation metrics such as METEOR (Denkowski and Lavie, 2014), Asiya (Gonzàlez et al., 2014) and VERTa (Comelles and Atserias, 2014), are incorporating lexical, syntactic and semantic information into their scores, whereas metrics like BEER (Stanojević and Sima'an, 2014), ReVal (Gupta et al., 2015) and COMET (Vela and Tan, 2015) use machine learning approaches for MT evaluation. The accuracy of the newly introduced evaluation methods is usually proven by human evaluation inputs, more specifically by measuring the correlation of the automatically provided scores with human judgements. Human evaluation is realised by ranking MT outputs (Bojar et al., 2013, 2014; Vela and van Genabith, 2015). In addition, post-editing, which is mainly used for measuring productivity (Guerberof, 2009; Zampieri and Vela, 2014), is also a valid human evaluation method.

2.2 Main Notions within Register Theory

Studies related to register theory (Quirk et al., 1985; Halliday and Hasan, 1989; 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 participants involved or the relationship between speaker and addressee(s). These parameters correspond to the variables of (1) *field*, (2) *tenor* and (3) *mode* defined in the framework of systemic functional linguistics (SFL), which describes language variation according to situational contexts; see, for instance, Halliday and Hasan (1989) and Halliday (2004). These variables are associated with the corresponding lexico-grammatical features, e.g. field of discourse is realised in term patterns or functional verb classes, e.g. activity (*approach, supply*, etc.), communication (*answer, inform, suggest*, etc.) and others; tenor is realised in modality expressed e.g. 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*, etc.); 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, e.g. Biber (1988), Biber (1995) or 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.

2.3 Register in Translation

Several studies in systemic functional linguistics are concerned with register settings in human translation (Steiner, 2004; Hansen-Schirra et al., 2012; De Sutter et al., 2012; Neumann, 2013; House, 2014) and their application into translation practice (Vela and Hansen-Schirra, 2006; Vela et al., 2007). To our knowledge, the machine translation (including its evaluation) community has not yet taken into consideration the notion of register, at least according to the definition in the present paper. Studies in the field of MT concerned with translation errors of new domains are covering only the lexical level (Irvine et al., 2013), as the authors operate solely with the notion of domain (field of discourse) and not register (which includes more parameters, as described in Section 2.2 above). Research on adding in-domain bilingual data to the training material of SMT systems (Eck et al., 2004; Wu et al., 2008) or on application of in-domain comparable corpora (Laranjeira et al., 2014; Irvine and Callison-Burch, 2014) consider the notion of domain. However, further register features are mostly ignored.

Domains reflect what a text is about, i.e. 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, e.g. imperative verbal constructions used to change the addressee's opinion, or interrogatives as a rhetorical means, etc. The similarity in domains would cover only the lexical level, in most cases terminology, ignoring the lexico-grammatical patterns specific for the given register, as shown by Lapshinova-Koltunski and Pal (2014) in their discussion on domain vs. register. Although some NLP studies employing web resources are arguing for the importance of register conventions, as by Santini et al. (2010), register remains 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.

3 Methodology and Resources

3.1 Research Questions

Following the assumption that translated language should normalise the linguistic features (like those described in Section 2.2 above) in order to adapt them to target language conventions, we use two different classification methods, KNN and SVM, to prove that in terms of register settings 1) machine translations correspond to human reference translations to a greater extent than 2) to comparable original non-translated texts in the same language. This requires a two-fold experiment design. In the first experiment, we use German original data for training and German machine translations for testing. Based on classification accuracy we can determine the "closeness" of machine translations to comparable non-translated texts in the same language. For the second experiment, we use a different data set applying the same classification methods. Human reference translations are used as training data and machine translations as test data. In this way we can observe the relation between machine translations and comparable non-translated texts in the same language, as well as between machine translations and reference translations.

We also aim at answering the following questions:

(i) Do German machine translations correspond to comparable German non-translated originals?

(ii) Are German machine translations closer to human reference translations than to comparable original German texts?

(iii) What are the main parameters influencing the classification outcome?

Our assumption is that machine translations will comply more with register standards of human-produced translated texts rather than with non-translated texts written by humans, as it was shown by Lapshinova-Koltunski and Vela (2015).

3.2 Feature Selection

For our analysis, we select a set of features derived from register studies described in Section 2.2 above. These features represent 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*, etc.) and others. The selected features reflect linguistic characteristics of all texts under analysis, are content-independent (do not contain terminology or keywords), and are easy to interpret yielding insights on the differences between the analysed variables. For instance, we use groupings of specific types of phrases (e.g. nominal, verbal, etc.) instead of part-of-speech n-grams, as they are easier to interpret as n-grams. The set of selected features for our 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 2.2.

The number of nominal and verbal parts-of-speech, chunks and nominalisations (*ung*-nominalisations) reflects 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 (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 words 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 representing subtypes of the categories given in Table 1.

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

However, for the final interpretation, a reduced number of features is used, which results from the validation step described in Section 3.4.1 below.

3.3 Data

In the first experimental setting, the training data set consists of German non-translated texts (GO=German originals) 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 shareholders (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 totals number of tokens in GO is 252711.

The corresponding test data set is smaller and includes 50 texts translated from English into German with a rule-based (RBMT) and a statistical (SMT) machine translation system. The rule-based machine translations were produced with the rule-based system SYSTRAN6 (Systran, 2001). The statistical machine translations were produced with a Moses-based system[1], trained with EUROPARL (Koehn, 2005), a parallel corpus containing texts from the proceedings of the European parliament. The total number of tokens in RBMT and SMT comprise 127865 and 124462 respectively. Both variants contain translations of the same texts belonging to the same registers as in the originals (training data).

In the second setting, English to German human translations (HT) – extracted from the CroCo corpus – are used for training, and comprise 100 texts (262655 tokens). For testing we use the same data as in the previous setting – machine translated texts (RBMT and SMT). Both training and test data are translations of the same source texts, which corresponds a common setting in MT evaluation. Table 2 gives an overview of the number of texts, sentences and tokens in both experiment settings.

	Setting 1			**Setting 2**		
	Train	**Test**		**Train**	**Test**	
	GO	RBMT	SMT	HT	RBMT	SMT
Texts	108	50	50	100	50	50
Sentences	15736	6195	6131	13077	6195	6131
Tokens	252711	127865	127865	262655	127865	127865

Table 2: Statistics for the data sets used in experiments

To extract the occurrences of register features described in Section 3.2, we annotate both training and test sets with information on token, lemma, part-of-speech, syntactic chunks and sentence boundaries using Tree Tagger (Schmid, 1994). The availability of these annotation levels in both corpora allows us to analyse certain lexico-grammatical patterns (see Section 3.2) required for register-sensitive analysis of translation. The features are then defined as linguistic patterns and modelled as regular expressions for the Corpus Query Processor (Evert, 2005), available within the CWB tools (CWB, 2010).

3.4 Classification Approaches

For our classification task, we train two models by using two different classifiers for each experiment setting: *k-nearest-neighbors* (KNN), a non-parametric method, and *support vector*

[1] We did not perform tuning for the SMT system.

machines (SVM) with a linear kernel, a supervised method, both commonly used in text classification. The models are then tested on German translations, and in this way, we obtain classification performance scores for:

- KNN and SVM, having German originals as training data and machine translations (English-German) as test data;
- KNN and SVM, having human translations (English-German) and again machine translations (English-German) as test data[2].

The performance scores are then judged in terms of *precision*, *recall* and *f-measure*. These scores are class-specific and indicate the results of automatic assignment of register labels to certain machine-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 10 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 understood as the harmonic mean of both.

3.4.1 K-Nearest Neighbors (KNN)

When using KNN, the input consists of the K closest training examples in the feature space[3], 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, see (Aha et al., 1991; Witten et al., 2011).

For our experiments we have to determine the final number for K and the most appropriate number of features used for classification. By measuring the distribution of errors during training (by performing 10-fols cross-validation) we determined the best K=11[4] and the final number of features for our setting. For our classification analysis we work with the tuple (*numberOfFeatures=17, K=11*), performing classification on the German translation (test) data by using the *knn* library in Weka (Hall et al., 2009). The final list of features include:

→ total words – words per text
→ content words – content (lexical) words
→ NP chunks – nominal chunks
→ VP chunks – verbal chunks
→ chunks – chunks per text
→ nominal – nominal part-of-speech categories
→ verbal – verbal part-of-speech categories
→ adversative – adversative conjunctive relations
→ causal – causal conjunctive relations
→ temporal – temporal conjunctive relations
→ modal – modal conjunctive relations
→ ung-nominalisations – nominalisations formed with *ung*-suffix
→ pron – personal pronouns
→ dempron – demonstrative pronouns
→ pronnp – nominal phrases filled with pronouns
→ gnouns – general nouns
→ modals denoting permission

[2]Note that human and machine translations are translation variants of the same English source text.
[3]In our experiments, the features are quantified by their frequency in the corresponding data set.
[4]The final value for K was chosen from an interval between 3 and 19

3.4.2 Support Vector Machines (SVM)

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 German labelled training data, the algorithm outputs an optimal hyperplane which categorises new instances, here German translations[5]. We use the same list of features for the classification with SVM as in the classification with KNN. We perform SVM classification with a 10-fold cross-validation.

The cross-validation in the training phase has shown that registers SPEECH and SHARE show low accuracy, especially in the first experiment setting (when German original data is used). For this reason, we we exclude these registers from further analyses.

3.5 Experimental Setup

In the first setting, both classifiers are supposed to store all cases from German originals (108 data points) with the corresponding register labels available. New cases from test data, which are machine translations in this case (50+50 data points), are then classified, i.e. assigned register labels. Classification is performed on the basis of distance function measure, for which Euclidean distance is used. The results of automatic assignment are indicated with scores (precision, recall and f-measure), which are used to measure the class agreement of the data with the positive labels given by the classifiers.

In the second setting, both classifiers store all labelled cases available in human translations (100 data points). The trained model is then applied on the same machine translations as in the first setting (50+50 data points). And again, we use precision, recall and f-measure to judge the class agreement of the data with the positive labels given by the classifiers.

4 Classification Results

As already mentioned above, the results of both classification algorithms are analysed in terms of *precision*, *recall* and *f-measure*. In case of precision, we measure the class agreement of the data with the positive labels given by the classifier. Our assumption is that precision values would indicate if the test data correspond to the training data in terms of the register settings. Hence, in the first experiment setting, the higher the precision, the better a machine translation corresponds to comparable originals, whereas in the second setting, the higher the precision, the better a machine translation corresponds to human translations.

4.1 Setting 1: Originals vs. Machine Translations

	ESSAY		**FICTION**		**INSTR**		**POPSCI**		**TOU**	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
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

Table 3: F-measure scores for classification per machine translation variant and register in the first setting

Table 3 provides the confusion matrix for the five registers and the two MT outputs split on the two classification methods. Concerning both the classification method and the MT system, we notice that FICTION is the register performing best, achieving an f-measure of 0.86% for the combination RBMT-SVM and SMT-SVM. Based on the f-measure we observe that

[5]One of the reasons why SVM are often used is their robustness towards overfitting, as well as their ability to map to a high-dimensional space.

ESSAY performs well for the classification with KNN but fails with SVM. INSTR and TOU perform similar in terms of f-measure for both KNN and SVM, the lowest f-measure values being measured for INSTR in the combination RBMT-KNN. POPSCI fails for the combinations RBMT-SVM and SMT-SVM, but performs well for KNN (0.53% for RBMT and 0.46% for SMT). These results imply that overall, FICTION performs best for both translation types and both classifiers, which indicates that translated fictional texts obey best to original fictional texts in terms of register settings.

Obviously, the data used for developing/training MT systems as well as the classification method play an important role. Registers like ESSAY (political essays) and INSTR (manuals) are usually used for training SMT systems, whereas registers like FICTION (fictional texts) and TOU (tourism leaflets) are less likely used for training MT systems. The more surprising are here the results for FICTION. The big difference in the results for both classifiers can be explained in the distinction between these two classification techniques. KNN uses all training data (predefined neighbours) in classification, while for SVM the maximised distance (margin) to the nearest example of each class plays a crucial role (all non-support vectors being discarded), thus influencing the difference in the classification outcome.

The overview of the performance of both MT systems in Table 4 (split on register and classification method) reveals that SMT performs better than RBMT in certain combinations for the registers ESSAY, FICTION and INSTR. The classification failure of the RBMT and SMT produced POPSCI translations for SVM indicates that SVM is not the appropriate classification method for this kind of texts. The fact that we observe contradictory results with both classifiers for ESSAY and POPSCI prevents us to claim that certain registers are generally more difficult to be identified in translated data than others.

register	**KNN**	**SVM**
ESSAY	SMT	RBMT=SMT
FICTION	RBMT=SMT	SMT
INSTR	SMT	SMT
POPSCI	RBMT	RBMT=SMT
TOU	RBMT	RBMT

Table 4: Performance of KNN and SVM across registers based on f-measure in the first setting

	Precision		**Recall**		**F-Measure**	
	KNN	SVM	KNN	SVM	KNN	SVM
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 5: Classification accuracy per machine translation variant in the first setting

In the last step, we analyse the performance of the MT systems disregarding registers, see Table 5. We notice that results do not differ much in the MT systems under analysis, which means that register agreement between non-translated and machine-translated texts is not dependent on the method involved in translation. This is proven by Pearson's chi-square test, which confirms our observation, as for both the KNN and the SVM results, p-value is higher than 0.05 (0.85 and 0.58 respectively).

4.2 Setting 2: Reference Translations vs. Machine Translations

The same experiments and analysis steps as in Section 4.1 are performed for the second setting, where human reference translations are used as training data and hypothesis translations as test

data. However, we observe a strong improvement in the results presented in Table 6. The registers ESSAY and FICTION achieve the best performance, showing an f-measure of up to 100%. Over all, f-measure remains over 50% for all registers, which means that classifiers perform also well for TOU, POPSCI and INSTR. These results can, in fact, serve as indicators that human and machine translations have more similarities, sharing register features.

	ESSAY		**FICTION**		**INSTR**		**POPSCI**		**TOU**	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
RBMT	0.93	1.00	1.00	1.00	0.54	0.73	0.71	0.77	0.57	1.00
SMT	0.93	1.00	0.50	0.80	0.62	0.67	0.60	0.73	0.75	0.83

Table 6: F-measure scores for classification per machine translation variant and register in the second setting

In this setting, ESSAY matches register features of human translations best, leading to the assumption that the texts used for developing and training MT systems play a key role. In contrast to the results in the first setting, the difference between the results of the two classification methods for ESSAY is minor. The lowest value is scored by INSTR with an f-measure of 0.54 for RBMT-KNN.

Table 7 demonstrates that, different from the results in the first experiment in Section 4.1, RBMT performs better then SMT when compared to human reference translations, with some exceptions. RBMT-INSTR and RBMT-TOU are better classified than the same registers translated with the SMT system, if the results from KNN are taken into account.

register	**KNN**	**SVM**
ESSAY	RBMT=SMT	RBMT=SMT
FICTION	RBMT	RBMT
INSTR	SMT	RBMT
POPSCI	RBMT	RBMT
TOU	SMT	RBMT

Table 7: Performance of the classification methods across registers based on the f-measure in the second setting

Similar tendencies are observed if we average the scores for registers.

	Precision		**Recall**		**F-Measure**	
	KNN	SVM	KNN	SVM	KNN	SVM
RBMT	0.81	0.90	0.75	0.90	0.75	0.90
SMT	0.82	0.84	0.65	0.80	0.68	0.80

Table 8: Classification accuracy per translation variant in the second setting

We observe in Table 8 that, in terms of MT system, RBMT performs almost always better than SMT, and in terms of classification method applied, SVM performs always better than KNN. However, significance test show that the difference between both machine translation systems is not significant, as the computed p-values exceed 0,05 for both KNN and SVM scores (0.78 and 0.97 respectively), confirming not only our assumption that reference and hypothesis translations are similar if register features are considered, but also that both machine-translated variants are similar as well.

5 Conclusion

The results of the presented experiments have proven our assumption that machine translations share their register features rather with human produced translations than with human produced non-translated texts, regardless the method involved in translation.

Setting1: Original vs. Machine Translation The results of the first experiment show that register settings of most German machine translations do not comply with the register settings of non-translated German, this being shown also by Lapshinova-Koltunski and Vela (2015). This is especially valid for ESSAY and POPSCI, which show low performance in most classification scenarios, not adapting the target language register conventions. The good performance for FICTION is an indicator that fictional texts adapt to the conventions of the target language. However, as known from Neumann (2013), German and English fictional texts share a lot of features, which might also influence the performance of the classifiers. We suppose that the influence of the source language register conventions[6] might also have an impact on the outcome. To prove this, we would need to perform additional experiments, in which English source texts should be used as training data.

Furthermore, as the performed significance tests have shown that the difference between RBMT and SMT is not meaningful, our suggestions apply regardless the translation method involved. In case of SMT, the results are apparently influenced by the training data used, as classification performs better for registers which are commonly used for SMT training, i.e. like ESSAY and INSTR. Off-the-shelf RBMT systems, like the one used here, are developed to cover a more general aspect of language which also essentially complicate the adaptation of register features.

Setting 2: Reference vs. Machine Translation The results of the second experiment presented in Section 4 have proven our assumption that the overlap between hypothesis and reference translations is higher than between hypothesis translations and comparable non-translated texts. On the one hand, this corresponds to our intuition in Lapshinova-Koltunski and Vela (2015), where we show that both human and machine translations do not correspond to comparable German originals, suggesting that both machine and manual translations should have more in common. In fact, this is also shown by Lembersky et al. (2012), who demonstrate 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 our claim that machine translations comply more with translated rather than with non-translated texts produced by humans. This results in the improvement of the BLEU score, but not necessary leading to a better quality of machine translation.

Following the results from both experimental setting, we argue that register features should be integrated into MT evaluation process, as an additional layer to the already existing automatic metrics. As future work, we would like to test this hypothesis by combining and correlating the results presented here with state-of-the-art evaluation metrics.

References

Aha, D. W., Kibler, D., and Albert, M. (1991). Instance-based Learning Algorithms. *Machine Learning*, 6(1):37–66.

Baroni, M. and Bernardini, S. (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.

[6]Here, we mean "shining through" of the source language as defined by Teich (2003).

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Bojar, O., Buck, C., Callison-Burch, C., Haddow, B., Koehn, P., Monz, C., Post, M., Saint-Amand, H., Soricut, R., and Specia, L., editors (2013). *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT)*. ACL.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., and Specia, L., editors (2014). *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Callison-Burch, C., Osborne, M., and Koehn, P. (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.

Comelles, E. and Atserias, J. (2014). VERTa Participation in the WMT14 Metrics Task. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 368–375, Baltimore, Maryland, USA. Association for Computational Linguistics.

CWB (2010). The IMS Open Corpus Workbench. accessed February 2015.

De Sutter, G., Delaere, I., and Plevoets, K. (2012). Lexical Lectometry in Corpus-based Translation Studies: Combining Profile-based Correspondence Analysis and Logistic Regression Modeling. In Oakes, M. P. and Meng, J., 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.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Diwersy, S., Evert, S., and Neumann, S. (2014). A Semi-supervised Multivariate Approach to the Study of Language Variation. In Szmrecsanyi, B. and Wälchli, B., editors, *Linguistic variation in text and speech, within and across languages*, pages 174–204. De Gruyter, Berlin.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technologies (HLT)*, pages 138–145.

Eck, M., Vogel, S., and Waibel, A. (2004). Improving Statistical Machine Translation in the Medical Domain using the Unified Medical Language system. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 792–798, Geneva, Switzerland.

Evert, S. (2005). *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.

Gonzàlez, M., Barrón-Cedeño, A., and Màrquez, L. (2014). IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 394–401, Baltimore, Maryland, USA. Association for Computational Linguistics.

Guerberof, A. (2009). Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation. *International Journal of Localization*, 7(1).

Gupta, R., Orăsan, C., and van Genabith, J. (2015). ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Halliday, M. (2004). *An Introduction to Functional Grammar*. Arnold, London.

Halliday, M. and Hasan, R. (1989). *Language, Context and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford University Press, Oxford.

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

House, J. (2014). *Translation Quality Assessment. Past and Present*. Routledge.

Irvine, A. and Callison-Burch, C. (2014). Using Comparable Corpora to Adapt MT Models to New Domains. In *Proceedings of the ACL Workshop on Statistical Machine Translation (WMT)*.

Irvine, A., Morgan, J., Carpuat, M., III, H. D., and Munteanu, D. S. (2013). Measuring Machine Translation Errors in New Domains. *TACL*, 1:429–440.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

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

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic Detection of Translated Text and its Impact on Machine Translation. In *Proceedings of MT-Summit XII*.

Lapshinova-Koltunski, E. and Pal, S. (2014). Comparabilty of Corpora in Human and Machine Translation. In *Proceedings of the Seventh Workshop on Building and Using Comparable Corpora*, Reykjavik, Iceland. LREC.

Lapshinova-Koltunski, E. and Vela, M. (2015). Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation (DiscoMT)*, pages 000–000, Lisbon, Portugal. Association for Computational Linguistics. fr.

Laranjeira, B., Moreira, V., Villavicencio, A., Ramisch, C., and Finatto, M. J. (2014). Comparing the Quality of Focused Crawlers and of the Translation Resources Obtained from them. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lembersky, G., Ordan, N., and Wintner, S. (2012). Language Models for Machine Translation: Original vs. Translated Texts. *Computational Linguistics*, 38(4):799–825.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Petrenz, P. (2014). *Cross-Lingual Genre Classification*. PhD thesis, School of Informatics, University of Edinburgh, Scotland.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, London.

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

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

Stanojević, M. and Sima'an, K. (2014). BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

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

Systran (2001). Past and Present. Technical report.

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

Vapnik, V. N. and Chervonenkis, A. J. (1974). *Theory of Pattern Recognition*. Nauka.

Vela, M. and Hansen-Schirra, S. (2006). The Use of Multi-level Annotation and Alignment for the Translator. In *Proceedings of Translating and the Computer 28 (ASLIB)*.

Vela, M., Neumann, S., and Hansen-Schirra, S. (2007). Querying Multi-layer Annotation and Alignment in Translation Corpora. In *Proceedings of the Corpus Liguistics Conference (CL)*.

Vela, M., Schumann, A.-K., and Wurm, A. (2014a). Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 47–56, Gothenburg, Sweden. Association for Computational Linguistics.

Vela, M., Schumann, A.-K., and Wurm, A. (2014b). Human Translation Evaluation and its Coverage by Automatic Scores. In *Proceedings of MTE Workshop at LREC 2014*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Vela, M. and Tan, L. (2015). Predicting Machine Translation Adequacy with Document Embeddings. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 382–390, Lisboa, Portugal. Association for Computational Linguistics.

Vela, M. and van Genabith, J. (2015). Re-assessing the WMT2013 Human Evaluation with Professional Translators Trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, Massachusetts.

Wu, H., Wang, H., and Zong, C. (2008). Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora. In Scott, D. and Uszkoreit, H., editors, *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pages 993–1000, Manchester, UK.

Zampieri, M. and Vela, M. (2014). Quantifying the Influence of MT Output in the Translators Performance: A Case Study in Technical Translation. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 93–98.