The Effect of Translationese on SMT Tuning

Sara Stymne

Uppsala University
sara.stymne@lingfil.uu.se

1. Introduction

Translationese is a term that is used to describe the special characteristics of translated texts, as opposed to originally authored tests (Gellerstam, 1986). Translations are different than original texts, which can be due both to influences from the source language and as a result of the translation process itself. For instance, texts that are translated tends to have shorter sentences, have a lower type/token ratio than original texts, and explicitate information, for instance by using more cohesive markers than in original texts. (Lembersky, 2013). Several studies have shown that it is possible to use text classification techniques to distinguish between original and translated texts with high accuracy (Baroni and Bernardini, 2006; Volansky et al., 2015), further supporting that there is a clear difference between original and translated texts.

Translationese has been shown to have an effect in relation to the training of statistical machine translation (SMT) systems, where the best results are seen when the texts used for training the SMT system have been translated in the same direction as that of the SMT system. This has been shown both for the translation model (TM) (Kurokawa et al., 2009; Lembersky et al., 2012; Joelsson, 2016) and for the language model (LM) for which it is better to use translated than original texts (Lembersky et al., 2011). It works nearly as well to use predicted translationese as known translationese, both for the LM and TM (Twitto et al., 2015). It has also been shown that adding features related to translationese to the TM improves translation quality (Lembersky et al., 2012).

Besides the data used for training the LM and TM, another important text for SMT training is the text used for tuning. The tuning text is used for tuning, or optimizing, the weights of the different models used in the SMT system. It is small compared to the other training data, and usually contains a couple of thousands of sentences, as opposed to millions of sentences for the LM and TM. To the best of our knowledge the effect of translationese has not previously been studied with respect to the tuning text.

In this paper we perform a small pilot study investigating the effect of the translation direction in the tuning text. We explore this for translation between German (de) and English (en). Our expectation was that using a tuning text translated in the same direction as the SMT task would improve the translation quality. The results are somewhat conflicting, however, with different evaluations showing different results. It is clear, though, that the length ratio between the tuning texts is different depending on translation direction, which largely influences the SMT results.

2. Experimental setup

In this section we will describe the data used in the system, the SMT system we used, and the evaluation.

2.1 Data

We use data from the WMT shared tasks of News translation between 2008–2013 (Bojar et al., 2013). The workshop includes the 5 languages English, German, Spanish, French and Czech. The test and tuning sets contains roughly an equal amount of segments originally written in each of these languages. Each aligned segment usually consists of one sentence, but in cases where translators have merged or split sentences, there could be two or three sentences in one language. We collected all test and tuning data from 2008–2013, a total of 17093 segments, and split it based on the original language of each text. The lowest number of segments for any source language is 2825. To have a balanced set we thus randomly picked 1412 segments from each source language for the test and tuning sets, respectively. We also created a mixed test set with segments from all original source languages.

In this study we focus on translation between English and German. For tuning we thus use tuning texts originally written in English and German. In addition we wanted to explore the effect of using a tuning text written in a third language, for which we chose Spanish. This means that both the English and German text in this set are translated from Spanish. For the test set we follow previous research, that have either used a test set translated in the same direction as the SMT system, or a mixed test set. To facilitate presentation we will use the abbreviations O for original texts and T for translated texts.

Tables 1 and 2 show the length in number of words in the tuning and test texts. It is quite clear that the length of the texts, and thus the average segment lengths are quite different. In particular, the German original texts are shorter than all the others. What is more interesting is the relation in length between the English and German side of the bitexts. For texts translated from German the English translation is much longer than the German original. For English originals, the German translations are slightly longer. This means that the length ratio between the texts, as counted in words, have a marked difference. The length ratio can be compared to that in the training data, where the original texts are from different languages, which is 0.96. For the Spanish and mixed originals, the differences in length between the German and English sides are smaller. The number of sentences per text is between 1475–1600 in each text, and there is usually a higher number of sentences in the

Original	German	English	Ratio
German	26926	30703	0.88
English	35513	34630	1.03
Spanish	38509	37997	1.01

Table 1: Number of words in the tuning sets. Each text consists of 1412 segments.

Original	German	English	Ratio
German	25761	28649	0.90
English	33746	32907	1.03
Mixed	35301	35917	0.98

Table 2: Number of words in the test sets used. Each text consists of 1412 segments.

translated than original text. The average sentence length is always slightly longer in the translated than original texts.

2.2 SMT system

We use Moses (Koehn et al., 2007) to train a factored phrase-based SMT system (Koehn and Hoang, 2007) that outputs both words and POS-tags, and have LMs for both. KenLM (Heafield, 2011) was used to train a 5-gram word LM and SRILM (Stolcke, 2002) was used to train a 7-gram POS LM. Tagging was performed using Tree Tagger (Schmid, 1994). For training we used Europarl and News commentary, provided by WMT, with a total of 2.1M segments.

For tuning we used MERT (Och, 2003) as implemented in Moses. For each tuning text we ran tuning three times and show the mean result, in order to account for optimizer instability, as suggested by Clark et al. (2011). For the manual analysis we use the system with the median Bleu score.

2.3 Evaluation

In much of the work on translationese, with the exception of Lembersky (2013), only Bleu (Papineni et al., 2002) has been used for evaluation. Bleu has its limitations though, and to give a somewhat more thorough evaluation we also show results on Meteor (Denkowski and Lavie, 2010) and TER (Snover et al., 2006). These metrics capture somewhat different aspects of MT quality. Bleu is mainly based on the precision of n-grams up to length 4, and thus rewards local fluency highly. Meteor is based on a weighted F-score on unigrams, with a matching step that consider word forms, stems, synonyms (for English), and paraphrases with different weights for content and function words, and a fragmentation score. It is thus less sensitive than Bleu to allowable linguistic variation. TER is an extension of the Levenshtein distance, with the addition of a shift operation to account for movement. To gain further understanding than just using these metrics we also show the scores on a subset of their individual components; 1-gram and 4-gram precision from Bleu, 1-gram precision and recall with flexible matching from Meteor, and the average number per segment of each TER operation: insertion, deletion, substitution and shift. We also give the length ratio of the translation hypothesis relative to the reference text.

Dir	Type	Bleu↑	Meteor↑	TER↓
en-de	O→T	**21.0**	**42.0**	61.4
	T→O	19.5	39.3	**59.0**
	T→T	20.8	41.7	62.1
de-en	O→T	**20.5**	**28.4**	62.4
	T→O	19.8	27.8	**59.2**
	T→T	19.6	27.4	59.7

Table 3: Metric scores for the O→T test set

Dir	Type	Bleu↑	Meteor↑	TER↓
en-de	O→T	17.2	38.1	67.9
	T→O	16.5	36.1	**64.3**
	T→T	**17.4**	**38.2**	68.3
de-en	O→T	17.2	**28.1**	69.1
	T→O	**18.4**	27.7	**63.3**
	T→T	**18.4**	27.4	63.8

Table 4: Metric scores for the mixed test set

In addition we perform a small human evaluation on a sample of segments for German–English translation. We randomly picked 100 segments of length 10–15, and one annotator compared the output from two systems for overall quality. Using only short segments can introduce a bias, since they might not be representative for all segments (Stymne and Ahrenberg, 2012), but it has the trade-off of being much faster and more consistent. We used the Blast tool for the evaluation (Stymne, 2011).

3. Results

Table 3 shows the results for the test sets with the same direction as the SMT system. The top two rows for each translation direction are of the highest interest, as they compare texts originally written in German or English. The T→T system is the variant where the tuning text was originally written in Spanish. The results are as expected when measured by Bleu and Meteor. Using O→T texts is much better than T→O. For TER, on the other hand, the results are reversed, and tuning using T→O gives better results. This pattern is the same in both translation directions.

For the T→T tuning, the results are different in the two directions. When translating from English it is similar to O→T, being slightly worse on all metrics. For the other direction it is instead similar to T→O, having slightly worse scores on all metrics. One possible reason for this difference might be that Spanish is more similar in structure to English than German, but this hypothesis needs to be further investigated in future work.

Table 4 shows the results on the mixed test set. Here the results are even more varying. For English–German the pattern is quite similar to the O→T test set, with the difference that T→T is slightly better rather than slightly worse than O→T on Bleu and Meteor. Both O→T and T→T are still clearly better than T→O on Bleu and Meteor, and worse on TER, like on the other test set. For German–English, T→T is again similar to T→O. However, the scores on Bleu and Meteor are different here, with O→T best on Meteor and T→O and T→T best on Bleu.

These results are not easy to interpret. The expectation

Dir	Type	Length ratio	1-gram P	4-gram P	Meteor P	Meteor R	insertions	deletions	substitutions	shifts
en-de	O→T	1.00	0.56	0.087	0.58	0.58	1.99	2.04	8.73	1.92
	T→O	0.89	0.59	0.092	0.61	0.54	3.56	0.86	7.99	1.69
	T→T	1.01	0.56	0.087	0.57	0.58	1.97	2.17	8.77	1.93
de-en	O→T	1.00	0.57	0.080	0.62	0.59	1.80	1.86	7.11	1.89
	T→O	0.90	0.61	0.087	0.64	0.57	2.97	0.90	6.50	1.65
	T→T	0.90	0.60	0.086	0.64	0.56	2.96	0.96	6.51	1.73

Table 5: Details of the different SMT systems for the O→T test set (P: precision, R:recall)

Dir	Type	Length ratio	1-gram P	4-gram P	Meteor P	Meteor R	insertions	deletions	substitutions	shifts
en-de	O→T	1.02	0.52	0.064	0.53	0.54	2.05	2.58	10.14	2.20
	T→O	0.91	0.55	0.068	0.56	0.51	3.53	1.24	9.29	2.01
	T→T	1.03	0.52	0.066	0.53	0.54	2.03	2.73	10.10	2.21
de-en	O→T	1.09	0.53	0.062	0.58	0.60	1.60	3.93	9.50	2.55
	T→O	0.97	0.57	0.070	0.61	0.58	2.77	2.01	8.96	2.35
	T→T	0.98	0.57	0.069	0.61	0.58	2.73	2.11	9.02	2.36

Table 6: Detailed evaluation of the different SMT systems for the mixed test set (P: precision, R:recall)

that O→T text would be best also for tuning is not overall met, and the difference between the different metrics needs to be further explained. Tables 5 and 6 show the results for the sub-components of the metrics and length ratio. Overall the T→T system is again quite similar to O→T for English-German and T→O for German-English.

It is quite striking that the length ratio, counted in number of words, are different depending on the tuning direction. O→T has considerably longer translations than T→O. In most cases the T→O translations are considerably shorter than the reference, whereas the O→T translations have similar length. The only exception is German–English for the mixed test set, which is the case where the metrics disagreed most, where O→T is longer than the reference, and T→O is only slightly shorter. The fact that the length differs is not surprising based on the length ratio for the tuning texts, where O→T has a longer target than source text, and T→O a shorter target than source text, in both translation directions. This also holds true for the O→T test text. For the T→T tuning text, see Table 1, the length ratio is similar to that of English–German, which potentially can explain the differing performance in the two translation directions.

The length difference is also related to other differences between the systems. There is a precision/recall trade-off in Meteor, where the short translations have higher precision than recall, and the longer translations have equal or higher recall than precision. The shorter translations also has a higher number of insertions and fewer deletions than the longer translations in TER. However, the longer translations, including O→T, which we thought would be best, also has a higher number of substitutions and shifts than T→O. For the Bleu precision the short T→O translation always has better precision then O→T. However, when combined with the brevity penalty, which compensates for the lack of recall in Bleu, the full Bleu score is usually higher for O→T.

To get some further insight we performed a small human evaluation where we compared the T→O and O→T systems for German-English, as detailed in section 2.3. As shown in Table 7 the O→T system is preferred more often than the T→O system, even though the segments were often of equal quality. This gives at least some indication that O→T is indeed the preferred system, as Bleu, Meteor and the length ratio suggests in most cases.

Equal	Equal quality	O→T better	T→O better
28	37	26	9

Table 7: Human comparison of O→T and T→O for German-English translation

Overall it seems that TER is biased towards short translations and rewards them, which Bleu and Meteor do not do. According to our, very limited, human evaluation, short translations should not be rewarded in this case. We also think that this is a situation which is very difficult for automatic metrics to handle, when the lengths of the two systems to be compared are very different.

4. Conclusion

In this initial study we have explored the effect of translationese on SMT tuning for translation between English and German. We expected that tuning on texts that were translated in the same direction as the SMT system would be preferable to texts translated in the opposite direction or from a third language. Our evaluation gave conflicting results on different metrics, however. Most strikingly the length was more similar to the reference length when tuning on O→T texts. This was partly due to different length ratios in the tuning texts we used for O→T and T→O translation. Using texts translated from a third language, Spanish, gave similar results to using O→T texts for English–German and as T→O texts for German–English.

This study was quite small, and we only investigated translation between two languages, German and English. We would specifically like to explore the issue of the differing length ratio in the tuning texts, where translations are longer than the originals. We want to investigate if this is typical for News texts, and if possible, to use tuning texts that are better balanced for length, to see if there are still other differences remaining. We also plan to explore other tuning algorithms than MERT, which could possibly have an effect on length. We would also like to extend the study to more language pairs, to see if we find the same patterns. This would allow us to further investigate the issue with length ratio, since it might be different for other language pairs and original languages.

References

M. Baroni and S. 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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 176–181, Portland, Oregon, USA.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden.

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

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland.

Jakob Joelsson. 2016. Translationese and swedish-english statistical machine translation. Bachelor thesis, Uppsala University.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

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*, pages 81–88, Ottawa, Canada.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the EACL*, pages 255–265, Avignon, France.

Gennadi Lembersky. 2013. *The Effect of Translationese on Statistical Machine Translation*. Ph.D. thesis, University of Haifa, Israel.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human notation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies, Demonstration session*, Portland, Oregon, USA.

Naama Twitto, Noam Ordan, and Shuly Wintner. 2015. Statistical machine translation with automatic identification of translationese. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 47–57, Lisbon, Portugal.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.