The Effect of Translationese on Tuning for Statistical Machine Translation

Sara Stymne
Department of Linguistics and Philology
Uppsala University
sara.stymne@lingfil.uu.se

Abstract

We explore how the translation direction in the tuning set used for statistical machine translation affects the translation results. We explore this issue for three language pairs. While the results on different metrics are somewhat conflicting, using tuning data translated in the same direction as the translation systems tends to give the best length ratio and Meteor scores for all language pairs. This tendency is confirmed in a small human evaluation.

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 from 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 and 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. However, the domain of the text interacts to a high degree with translationese identification (Rabinovich and Wintner, 2015).

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 noted that the original language of the test sentences influences the Bleu score of translations (Holmqvist et al., 2009).

Besides the data used for the LM and TM, another important text for SMT training is the data used for tuning. The tuning set is used for tuning, or optimizing, the log-linear feature weights of the models, such as TM, LM, and reordering models. 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. It is supposed to be representative of the test set. To the best of our knowledge the effect of translationese has not previously been studied with respect to the tuning set.

We investigate the effect of the translation direction in the tuning text. We explore this for translation between English on one side, and German, French, and Czech on the other side, for the news domain. There is a tendency that tuning in the same direction as the SMT system performs best, especially as measured by length ratio and Meteor.

2 Experimental setup

To facilitate presentation we will use the abbreviations O for original texts and T for translated texts, and the term *foreign* to represent either of the languages German, French, and Czech.

2.1 Data

We use data from the WMT shared tasks of News translation between 2008–2013 (Bojar et

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 241–246,
Gothenburg, Sweden, 23-24 May 2017.

al., 2013).[1] This data includes the 5 languages English, German, Spanish, French, and Czech. The test and tuning sets contains roughly an equal amount of segments, normally a sentence, originally written in each 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 balanced sets we randomly selected 1412 segments from each original language for the test and tuning sets, respectively. We also created a mixed test set with segments from all five original source languages. The mixed and from-English sets are parallel across the language pairs, whereas the from-*foreign* sets are different for each language.

For the test set we follow previous research, that have either used a test set translated in the same direction as the SMT system, which mimics a realistic translation scenario, where we normally have an original text we want to translate, or a mixed test set, which is a common situation in MT evaluation campaigns. For tuning we use tuning texts originally written in English and *foreign*. We also tune systems for all 5 original languages, and create a custom system for the mixed test set, where for each sentence we use the tuning weights that matches the original language of that sentence.

Table 2.1 shows the length ratio of the number of words between the *foreign* and English side of the tuning and test sets. For all languages there is a large ratio difference depending on the direction of translation. The *foreign* texts are always relatively longer when translated from English than compared to being originally authored and translated to English. The actual ratios are different between the language pairs, though, where French has more words than English, Czech has fewer words than English, and for German it depends on the translation direction. The translationese in this news corpus, though, counted in words, is always relatively longer than originally authored texts, which is not a tendency that has been stressed in previous research on translationese. The ratios for the test and tuning corpus are similar in all cases except for Czech→English.

[1]Until 2013 the WMT test and tuning sets were parallel between all languages in the workshop, allowing us to use a five-way parallel corpus. From 2014 texts are parallel only per language pair, with no texts authored in a third language. In addition the language pairs used partly changed from 2014.

Data set	Original	German	French	Czech
Tuning	*Foreign*	0.88	1.07	0.79
	English	1.03	1.16	0.92
Test	*Foreign*	0.90	1.09	0.85
	English	1.03	1.17	0.95
	Mixed	0.98	1.14	0.88

Table 1: Ratio of *foreign* to English words for sets with different original language.

2.2 SMT system

We use Moses (Koehn et al., 2007) to train standard phrase-based SMT systems. For German↔English we use word and POS-tag factors (Koehn and Hoang, 2007) and have LMs for both; for the other language pairs we only use words. 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 over 2M segments for German and French and .77M for Czech. For English→German we used additional data: bilingual Common Crawl (1.5M) and monolingual News (83M).

For tuning we used MERT (Och, 2003) as implemented in Moses, optimized towards the Bleu metric (Papineni et al., 2002). For each tuning condition we ran tuning three times and show the mean result, in order to account for optimizer instability (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. Meteor is also tuned for different target languages, to increase correlation with human

evaluation scores. TER is an extension of the Levenshtein distance, with the addition of a shift operation to account for movement. Like Bleu, TER only considers exact word form matches. We also give the length ratio (LR), counted as the number of words, of the translation hypothesis relative to the reference text.

In addition we perform a small human evaluation on a sample of segments for German→English translation. For each setting, we randomly picked 100 segments of length 10–15 words. 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.

3 Results

Table 2 shows the results on the O→T test set. The scores are obviously different for the different language pairs, which are due to both differences between the languages, differences in the use of training data and factors in the SMT systems, and for from-*foreign*, different test sets.

The differences between O→T and T→O are often large, with up to 1.5 Bleu points difference for English–German. This is quite notable since the actual models in the SMT systems are identical; the only difference is in the weights balancing the models and features of the SMT system. For all language pairs, except Czech–English, the length ratio for O→T tuning is around 1, which is desired, and much lower for T→O tuning. That this is not the case for Czech–English is most likely due to the fact that the length ratios in the O→T tuning and test sets were different. On the metrics, however, the scores are somewhat conflicting. In most cases Bleu and Meteor have the best scores for O→T tuning, whereas the scores for TER are the worst. For Czech–English the two systems have the same Bleu score, which probably is due to the long length ratio with O→T tuning. For French–English O→T tuning gives a better TER score than T→O tuning. This is an exception to the pattern, for which we do not yet have an explanation.

Table 3 shows the results on the mixed test set. For all language pairs, the pattern is the same on this test set as regards Meteor, which is higher for O→T tuning, and TER which is lower for O→T tuning. The length ratio is always low with T→O tuning. For O→T tuning, it is around 1 for from-English, but always high for from-*foreign*. Bleu is better on O→T than T→O tuning for four out of the six translation directions.

Table 3 also includes a custom system, where the tuning direction was chosen separately for each sentence based on its original language. We would expect this system to give the best results on this test set, since it is optimized for each language direction, but again the results are conflicting. It overall gives a good length ratio, though, and has the best or (near)-equal Bleu score to the O→T tuning. The TER score is always between that of T→O and O→T tuning. The Meteor score, however, is always lower for the custom system than for O→T tuning, which might indicate that there is some advantage with O→T tuning that shows up when using the flexible matching in Meteor.

To get some further insight we performed a small, thus quite limited, human evaluation for German→English. A comparison on the O→T test set, between O→T and T→O tuning is shown in Table 4. The O→T system is preferred more often than the T→O system, even though the segments were often of equal quality. The difference is significant at the 0.01-level, using a two-sided sign test. 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. Table 5 shows a comparison between custom and O→T tuning on the mixed test set. In this case the translations are similar to an even larger extent, and we can find no difference between the systems. This might indicate that Bleu punishes the longer O→T system too harshly. In both cases there is no agreement between TER and the human evaluation.

Overall it seems that TER rewards very short translations; the shortest translation for each setting always has the best TER score. According to our, very limited, human evaluation, short translations should not be rewarded. On the other hand the longest system in each setting always has the best Meteor score, which is in contrast to Bleu, which generally prefers translations with a length ratio around 1. This is likely because Meteor takes recall into account, as opposed to Bleu, which is only based on precision and a brevity penalty. Long translations might be good, if they explici-

	English–German				English–French				English–Czech			
Tuning	Bleu↑	Meteor↑	TER↓	LR	Bleu↑	Meteor↑	TER↓	LR	Bleu↑	Meteor↑	TER↓	LR
O→T	**21.0**	**42.0**	61.4	**1.00**	**22.3**	**51.3**	57.6	**0.99**	**13.6**	**20.8**	70.0	**0.97**
T→O	19.5	39.3	**59.0**	0.89	21.8	50.4	**56.3**	0.94	12.5	19.7	**68.3**	0.88
	German–English				French–English				Czech–English			
O→T	**20.5**	**28.4**	62.4	**1.00**	**26.9**	**35.7**	**50.1**	**1.02**	**18.8**	**28.9**	66.2	1.06
T→O	19.8	27.8	**59.2**	0.90	25.8	33.9	51.0	0.95	**18.8**	28.1	**62.1**	**0.95**

Table 2: Metric scores and length ratio on the O→T test set.

	English–German				English–French				English–Czech			
Tuning	Bleu↑	Meteor↑	TER↓	LR	Bleu↑	Meteor↑	TER↓	LR	Bleu↑	Meteor↑	TER↓	LR
O→T	17.2	**38.1**	67.9	1.02	**20.4**	**49.7**	60.3	**0.99**	**13.0**	**20.6**	71.9	**1.00**
T→O	16.5	36.1	**64.3**	0.91	20.0	48.9	**59.1**	0.95	12.4	19.9	**69.5**	0.92
Custom	**17.7**	38.0	66.7	**1.00**	**20.4**	49.6	59.8	0.98	**13.0**	20.5	70.5	0.97
	German–English				French–English				Czech–English			
O→T	17.2	**28.1**	69.1	1.09	**18.6**	**31.3**	62.2	1.05	18.0	**28.9**	68.0	1.08
T→O	18.4	27.7	**63.3**	0.97	18.0	30.0	**60.9**	0.98	18.2	28.1	**63.9**	0.96
Custom	**18.5**	27.8	64.8	**1.01**	18.5	30.7	61.1	**1.02**	**18.6**	28.7	65.5	**1.02**

Table 3: Metric scores and length ratio on the mixed test set.

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

Table 4: Human comparison of O→T and T→O tuning for German-English O→T test set.

Equal	Equal quality	O→T better	Custom better
51	26	12	11

Table 5: Human comparison of O→T and custom tuning for German-English mixed test set.

cate information in a good way. We doubt, however, that this is what Meteor rewards, since it, like the other metrics, is based on matching towards one reference translation. We believe that a situation like this, when the lengths of the two systems to be compared are very different, is very difficult for automatic metrics to handle in a fair way.

4 Conclusion

In this paper we have investigated the effect of translationese on SMT tuning for three language pairs. We found that across language pairs, using tuning texts translated in the same original direction as the SMT system tended to give a better length ratio, Meteor score, and often a better Bleu score. However, the very short translations that were the result of tuning with a text translated in the opposite direction were preferred by the TER metric. We also explored a custom system, with tuning in the same direction as each test sentence, which overall performed on par with the system with tuning in the same direction. A small human evaluation confirmed that tuning in the same direction was preferable to the opposite direction, but performed on par with custom tuning.

As the study was relatively small we think there is a need for extending it to more language pairs, other domains than news, and other tuning algorithms than MERT. We also think it would be important to do a more large-scale human evaluation. Especially we want to find out if there are other differences than length ratio, based on tuning direction, which we could not find in this small study. We would also like to extend the study of translationese to other types of MT than SMT. Specifically, we want to focus on neural MT, which have given very good translation results recently, but for which no studies of the relation to translationese have been attempted.

For most SMT research the translation direction of neither test sets nor tuning sets have been taken into account. The data from the WMT workshops, for instance, contains data sets translated from many different languages or in both directions between a pair of languages. It is well-known that tuning sets should be representative of the type of text the SMT system should be used for, but this has mostly been considered for content or domain. This study shows that at least the length ratio of the tuning set, and possibly also the translation direction, is important. This study also indicates that automatic MT metrics may not be reliable for situations where the hypotheses have very different lengths and that different metrics favor different length ratios. However, this needs to be further explored in future work. The interactions with domain should also be further investigated.

Acknowledgments

This work was supported by the Swedish strategic research programme eSSENCE.

References

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.

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: Proceedings from The Scandinavian Symposium on Translation Theory II*, 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.

Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving alignment for SMT by reordering and augmenting the training corpus. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 120–124, Athens, Greece.

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.

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.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

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.

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.