Statistical Machine Translation with Automatic Identification of Translationese

Naama Twitto-Shmuel Noam Ordan Shuly Wintner

Department of Computer Science
University of Haifa
Haifa, Israel

WMT 2015
Lisbon, 17 September, 2015

Original or Translation?

Example (O or T?)

We want to see countries that can produce the best product for the best price in that particular business . I have to agree with the member that free trade agreements by definition do not mean that we have to be less vigilant all of a sudden .

Example (T or O?)

I would like as my final point to say that we support free trade , but we must learn from past mistakes . Let us hope that negotiations for free trade agreements with the four Central American countries introduce a number of other dimensions absent from these first generation agreements .

Translationese

The language of translated texts

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (Gellerstam, 1986)
- Several reasons:

  SIMPLIFICATION (Blum-Kulka and Levenston, 1978, 1983)
  EXPLICITATION (Blum-Kulka, 1986)
  NORMALIZATION (Chesterman, 2004)
  INTERFERENCE (Toury, 1980, 1995)

IDENTIFYING TRANSLATIONESE

- Automatic identification of translationese (Baroni and Bernardini, 2006; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Popescu, 2011)
- Investigation of translationese features (Volansky et al., 2015)
- Cross-domain evaluation (Koppel and Ordan, 2011; Avner et al., Forthcoming)

Identifying Translationese

Why does it matter?

- Language models for statistical machine translation (Lembersky et al., 2011, 2012b)
- Translation models for statistical machine translation (Ozdowska and Way, 2009; Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)
- Cleaning parallel corpora crawled from the Web (Eetemadi and Toutanova, 2014; Aharoni et al., 2014)
- Inherently depend on data annotated as original vs. translated

Contribution

- Can the predictions of translationese classifiers replace manual annotation?
- When a monolingual corpus in the target language is given for constructing a language model for SMT, automatically identifying the translated portions of the corpus, and using only them for the language model, is as good as using the entire corpus
- When a parallel corpus is given, automatically identifying the portions of the corpus that are translated in the direction of the translation task, and using only them for training the translation model, is again as good as using the entire corpus

Plan

- Train classifiers to tease apart original from translated texts
- Construct SMT systems with language models compiled from the predicted translations, comparing them with similar SMT systems whose language models consist of the entire monolingual corpora.
- Construct SMT systems with translation models compiled from bitexts that are predicted as translated in the same direction as the direction of the SMT task, comparing them with similar SMT systems whose translation models consist of the entire parallel corpora

Outline

1. Introduction
2. Experimental Setup
3. Experiments
4. Conclusion

Datasets and Tools

- Corpora: Europarl, Hansard, the News Commentary Corpus
- **Chunks** of 2,000 tokens
- Sentence detection, tokenization, POS tagging
- Classification with SVM (SMO) using Weka
- SRILM for language models
- Moses for SMT

Outline

1. Introduction
2. Experimental Setup
3. Experiments
4. Conclusion

Classification of Translationese

- Dataset: Europarl (FR→EN, DE→EN, EN→FR)
- Features: **Contextual function words** (Volansky et al., 2015); e.g., $\langle \textit{in,the,Noun} \rangle$
- Intrinsic evaluation
- Perplexity

Accuracy of the Classification

And fitness to the reference set of FR→EN language models compiled from texts predicted as translated

			Perplexity			
Data set	Chunks	Acc. (%)	1-gram	2-gram	3-gram	4-gram
Predicted	1245	98.96	463.51	94.81	71.60	68.76
T	1255		463.58	94.59	71.24	68.37
O	1258		500.56	115.48	91.14	88.31
All	2513		473.00	93.34	67.84	64.47

Accuracy of the Classification

And fitness of language models compiled from texts predicted as translated to the reference set, DE→EN and EN→FR

	DE→EN			EN→FR		
Data set	Chunks	Acc. (%)	Ppl	Chunks	Acc. (%)	Ppl
Predicted	1,146	99.08	62.23	1,410	98.47	47.92
T	1,153		62.07	1,413		47.89
O	1,153		76.68	1,411		59.75
All	2,306		57.48	2,824		44.49

LMs Compiled from Predicted Translationese

Evaluation of the FR→EN SMT system built from LMs compiled from predicted translationese

Data set	BLEU↑	MET↑	TER↓	*p*
Predicted	**28.9**	**33.2**	**53.8**	0.16
T	**29.1**	**33.3**	**53.6**	0.58
O	27.8	32.9	54.7	0.00
All	**29.1**	**33.3**	**53.8**	

LMs Compiled from Predicted Translationese

Evaluation of the DE→EN and EN→FR SMT systems built from LMs compiled from predicted translationese

	DE→EN				EN→FR			
Data set	BLEU↑	MET↑	TER↓	*p*	BLEU↑	MET↑	TER↓	*p*
Predicted	**21.9**	**28.6**	**63.8**	0.87	**26.3**	47.8	**58.3**	0.47
T	**21.8**	**28.6**	**63.9**	0.37	26.1	47.7	**58.5**	0.03
O	21.0	28.4	64.6	0.00	25.1	47.0	59.5	0.00
All	**21.9**	**28.6**	**63.7**		**26.3**	**48.0**	**58.7**	

Cross-corpus Experiments

Hansard-based SMT system, Europarl-based classification

Data set	Chunks	Acc. (%)	BLEU↑	MET↑	TER↓	*p*
Predicted	1,321	78.22	**37.8**	**37.7**	**45.9**	0.11
T	2001		**38.0**	**37.8**	**45.7**	0.86
O	2001		37.5	37.6	46.1	0.00
All	4002		**38.0**	**37.7**	**45.8**	

Cross-corpus Experiments

Cross-corpus evaluation: News Commentary corpus

Data set	Chunks	BLEU↑	MET↑	TER↓	p
Predicted	1,470	27.0	**33.0**	**55.2**	0.02
All	2,527	**27.2**	**33.0**	**55.2**	

TMs Compiled from Predicted Translationese

Accuracy of the classification and evaluation of SMT systems

Task	Data set	Chunks	Acc. (%)	BLEU↑	MET↑	TER↓	p
FR→EN	Predicted	1,678	98.93	**31.1**	**34.7**	**52.1**	0.13
	$S \rightarrow T$	1,690		**31.3**	**34.8**	**51.7**	0.94
	$T \rightarrow S$	1,689		28.4	33.3	54.4	0.00
	All	3,379		**31.3**	**34.7**	**51.9**	
DE→EN	Predicted	1,607	99.44	23.7	30.3	61.6	0.00
	$S \rightarrow T$	1,613		**24.0**	30.4	**61.3**	0.05
	$T \rightarrow S$	1,612		21.7	29.0	63.9	0.00
	All	3,225		**24.2**	**30.5**	**61.1**	
EN→FR	Predicted	1,678	98.93	**29.4**	**50.7**	**55.3**	0.11
	$S \rightarrow T$	1,689		**29.3**	**50.8**	**56.1**	0.18
	$T \rightarrow S$	1,690		26.7	48.2	58.2	0.00
	All	3,379		**29.1**	**50.6**	**56.0**	

TMs Compiled from Predicted Translationese

Cross-corpus evaluation: Hansard-based SMT system, Europarl-based classification

Data set	Chunks	Acc. (%)	BLEU↑	MET↑	TER↓	*p*
Predicted	1,840	79.36	36.3	36.9	46.6	0.00
$S \rightarrow T$	3,000		**37.3**	**37.3**	**46.2**	0.94
$T \rightarrow S$	3,000		34.1	35.8	48.9	0.00
All	6,000		**37.3**	**37.4**	**46.0**	

Outline

1 Introduction

2 Experimental Setup

3 Experiments

4 Conclusion

CONCLUSION

- Direction matters
- Translationese matters
- Import for less-resourced languages

Future Directions

- Utilize parallel corpora for classification (Eetemadi and Toutanova, 2014, 2015)
- Improve feature set
- Reduce chunk size
- **Un**supervised classification (Rabinovich and Wintner, 2015): Sunday 13:30–13:55, Session 5A

Thank You

Diolch
Grazie
Gracias
Arigatō
Tack
Tak
Shukrān
Dík
תודה רבה!
Danke
Merci
THANK YOU
ありがとう
Dank u
Дякую
謝謝
Mersi
Dzięki
Ευχαριστώ
Takk
Спасибо
Qujanaq
Köszi
Teşekkürler
Asante
Obrigado
Kiitos
감사합니다
Hvala
شكراً
Ďakujem
どうも
Ta

Bibliography I

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. Automatic detection of machine translated text and translation quality estimation. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**, pages 289–295, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-2048.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. Identifying translationese at the word and sub-word level. **Digital Scholarship in the Humanities**, Forthcoming. doi: http://dx.doi.org/10.1093/llc/fqu047. URL http://dx.doi.org/10.1093/llc/fqu047.

Marco Baroni and Silvia Bernardini. 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, September 2006. URL http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1.

Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, **Interlingual and intercultural communication Discourse and cognition in translation and second language acquisition studies**, volume 35, pages 17–35. Gunter Narr Verlag, 1986.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. **Language Learning**, 28(2):399–416, December 1978.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, **Strategies in Interlanguage Communication**, pages 119–139. Longman, 1983.

Andrew Chesterman. Beyond the particular. In A. Mauranen and P. Kujamäki, editors, **Translation universals: Do they exist?**, pages 33–50. John Benjamins, 2004.

Sauleh Eetemadi and Kristina Toutanova. Asymmetric features of human generated translation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 159–164. Association for Computational Linguistics, October 2014. URL http://www.aclweb.org/anthology/D14-1018.

Sauleh Eetemadi and Kristina Toutanova. Detecting translation direction: A cross-domain study. In **NAACL Student Research Workshop**. ACL Association for Computational Linguistics, June 2015. URL http://research.microsoft.com/apps/pubs/default.aspx?id=249114.

Bibliography II

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

Iustina Ilisei and Diana Inkpen. Translationese traits in Romanian newspapers: A machine learning approach. **International Journal of Computational Linguistics and Applications**, 2(1-2), 2011.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, **Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing**, volume 6008 of **Lecture Notes in Computer Science**, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL `http://dx.doi.org/10.1007/978-3-642-12116-6`.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1132`.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In **Proceedings of MT-Summit XII**, pages 81–88, 2009.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 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, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1034`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting translation models to translationese improves SMT. In **Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics**, pages 255–265, Avignon, France, April 2012a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E12-1026`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. **Computational Linguistics**, 38(4):799–825, December 2012b. URL `http://dx.doi.org/10.1162/COLI_a_00111`.

Bibliography III

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. **Computational Linguistics**, 39(4):999–1023, December 2013. URL http://dx.doi.org/10.1162/COLI_a_00159.

Sylwia Ozdowska and Andy Way. Optimal bilingual data for French-English PB-SMT. In **Proceedings of EAMT-2009, the 13th Annual Conference of the European Association for Machine Translation**. European Association for Machine Translation, May 2009.

Marius Popescu. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, **Proceedings of RANLP-2011**, pages 634–639, 2011.

Ella Rabinovich and Shuly Wintner. Unsupervised identification of translationese. **Transactions of the Association for Computational Linguistics**, 3:419–432, 2015. ISSN 2307-387X. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/618.

Gideon Toury. **In Search of a Theory of Translation**. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.

Gideon Toury. **Descriptive Translation Studies and beyond**. John Benjamins, Amsterdam / Philadelphia, 1995.

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