The Features of Translationese

Between Human and Machine Translation

Shuly Wintner

Department of Computer Science
University of Haifa
Haifa, Israel
shuly@cs.haifa.ac.il

ISCOL 2014

University of Haifa, 7 September 2014

Original or Translation?

Example (O or T?)

מאות אלפי אזרחים הפגינו אתמול (ראשון) בפאריס נגד הצעת החוק שמאפשרת נישואי זוגות חד-מיניים ואימוץ ילדים על ידי בני זוג מאותו מין. זוהי ההפגנה האחרונה לפני אימוץ החוק על ידי הסנאט שמתוכנן בחודש אפריל.

ההפגנה התקיימה במיקום סמלי - בין לה דפאנס לבין כיכר אטואל שבמרכזה עומד שער הניצחון.

המפגינים, בהם צעירים רבים, משפחות עם ילדים ואזרחים ותיקים, נשאו שלטים בגנות המהלך שעליהם נכתב בין השאר "רוצים עבודה, לא נישואי הומואים", "מדינת זכויות האדם מגנה על זכויות הילדים", ו"אל תיגעו בנישואים, דאגו לאבטלה".

Example (T or O?)

צרפת הפכה היום (שלישי) אחר הצהריים למדינה ה-14 בעולם המעגנת בחוק נישואים חד מיניים.

ההצבעה הסופית בבית המחוקקים היתה מהירה, מאחר שלמפלגה הסוציאליסטית של הנשיא פרנסואה הולנד, התומכת בשינוי, יש רוב ניכר בפרלמנט.

ואולם, בשבועות האחרונים גברה בצרפת המחאה נגד אישור החוק. ההפגנות גם זימנו למתנגדי הולנד בימה לתקוף את הנשיא עצמו. לדברי התקשורת הצרפתית, יו"ר האסיפה הלאומית קיבל אתמול מכתב המאיים ב"מלחמה" נגד חברי פרלמנט סוציאליסטים, אם החוק אמנם יעבור; על פי הדיווחים, המעטפה הכילה אבק שריפה.

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)
- Toury (1980, 1995) defines two **laws of translation**:

  The law of interference Fingerprints of the source text that are left in the translation product

  The law of growing standardization Effort to standardize the translation product according to existing norms in the target language and culture

Translationese

The language of translated texts

Translation universals (Baker, 1993)

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems"*

simplification (Blum-Kulka and Levenston, 1978, 1983)

explicitation (Blum-Kulka, 1986)

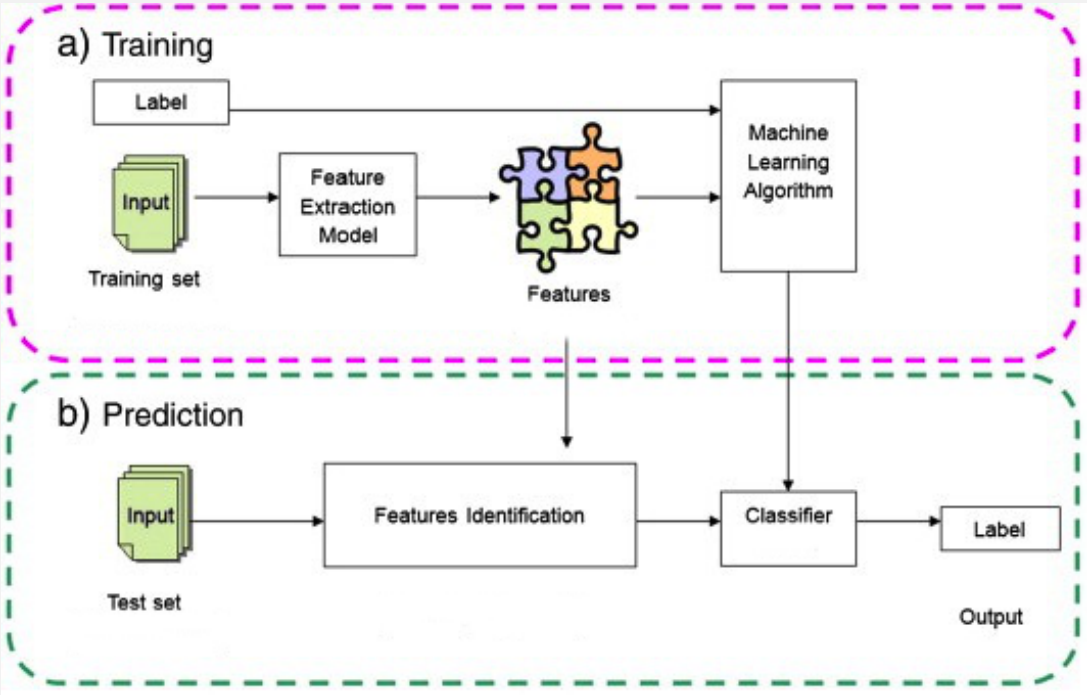
Normalization (Chesterman, 2004)

Computational Investigation of Translationese

- Translated texts exhibit lower lexical variety (type-to-token ratio) than originals (Al-Shabab, 1996)
- Their mean sentence length and lexical density (ratio of content to non-content words) are lower (Laviosa, 1998)
- Corpus-based evidence for the simplification hypothesis (Laviosa, 2002)

Methodology

- Corpus-based approach
- Text classification with machine-learning techniques
- Feature design
- Evaluation



Identifying Translationese

Using text classification

- Baroni and Bernardini (2006)
- van Halteren (2008)
- Kurokawa et al. (2009)
- Ilisei et al. (2010); Ilisei and Inkpen (2011); Ilisei (2013)
- Koppel and Ordan (2011)
- Popescu (2011)

Research Contributions

- Understanding the features of translationese; testing Translation Studies hypotheses (Volansky et al., Forthcoming; Avner et al., Forthcoming)
- Robust classification of translationese (Twitto-Shmuel et al., Forthcoming)
- Language models for statistical machine translation (Lembersky et al., 2011, 2012b)
- Translation models for statistical machine translation (Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)
- Automatic detection of **machine** translated texts (Aharoni et al., 2014)
- Identifying the first language of non-native writers (Tsvetkov et al., 2013)

The Features of Translationese

- Vered Volansky, Noam Ordan, and Shuly Wintner, "On the Features of Translationese", *Literary and Linguistic Computing*, forthcoming
- Goal: test Translation Studies hypotheses using classification as a methodology
- Experimental setup: EUROPARL, 4M tokens in English (O) and 400K tokens translated from each of ten European languages (T)
- After tokenization, the corpus is partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)
- Classification with Weka (Hall et al., 2009), using SVM with a default linear kernel

Hypotheses

Simplification Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

Explicitation The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

Normalization Efforts to standardize texts (Toury, 1995), "a strong preference for **conventional grammaticality**" (Baker, 1993)

Interference The fingerprints of the source language on the translation output (Toury, 1979)

Features Should...

1. Reflect frequent linguistic characteristics we would expect to be present in the two types of text
2. Be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.
3. Be easy to interpret, yielding insights regarding the differences between original and translated texts

Features

Simplification Type-token ratio, Mean word length, Syllable ratio, Mean sentence length, Lexical density, Mean word rank, Most frequent words

Explicitation Explicit naming, Single naming, Mean multiple naming, Cohesive markers

Normalization Repetitions, Contractions, Average PMI, Threshold PMI

Interference POS *n*-grams, Character *n*-grams, Prefixes and suffixes, Contextual function words, Positional token frequency

Miscellaneous Function words, Pronouns, Punctuation, Ratio of passive forms, Token unigrams and bigrams

Results: Sanity Check

Category	Feature	Accuracy (%)
Sanity	Token unigrams	100
	Token bigrams	100

Results: Simplification

Category	Feature	Accuracy (%)
Simplification	TTR (1)	72
	TTR (2)	72
	TTR (3)	76
	Mean word rank (1)	69
	Mean word rank (2)	77
	N most frequent words	64
	Mean word length	66
	Syllable ratio	61
	Lexical density	53
	Mean sentence length	65

Results: Explicitation

Category	Feature	Accuracy (%)
Explicitation	Cohesive Markers	81
	Explicit naming	58
	Single naming	56
	Mean multiple naming	54

Results: Normalization

Category	Feature	Accuracy (%)
Normalization	Repetitions	55
	Contractions	50
	Average PMI	52
	Threshold PMI	66

Results: Interference

Category	Feature	Accuracy (%)
Interference	POS unigrams	90
	POS bigrams	97
	POS trigrams	98
	Character unigrams	85
	Character bigrams	98
	Character trigrams	100
	Prefixes and suffixes	80
	Contextual function words	100
	Positional token frequency	97

Results: Reduced Parameter Space

(300 most frequent features)

Category	Feature	Accuracy
	POS bigrams	96
	POS trigrams	96
Interference	Character bigrams	95
	Character trigrams	96
	Positional token frequency	93

Results: Miscellaneous

Category	Feature	Accuracy (%)
Miscellaneous	Function words	96
	Punctuation (1)	81
	Punctuation (2)	85
	Punctuation (3)	80
	Pronouns	77
	Ratio of passive forms to all verbs	65

Conclusion

- Machines can accurately identify translated texts
- The best performing features are those that attest to the 'fingerprints' of the source on the target
- Interference by its nature is a pair-specific phenomenon
- Translation "universals" should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- Ideally, such claims should be studied using a **comparable corpus**

Other Contributions

- Ehud Alexander Avner, Noam Ordan, and Shuly Wintner, "Identifying Translationese at the Word and Sub-word Level", *Literary and Linguistic Computing*, forthcoming
- Gennadi Lembersky, Noam Ordan, and Shuly Wintner, "Language models for machine translation: Original vs. translated texts", *Computational Linguistics* 38(4):799-825, 2012
- Gennadi Lembersky, Noam Ordan, and Shuly Wintner, "Improving statistical machine translation by adapting translation models to translationese", *Computational Linguistics* 39(4):999-1023, 2013
- Naama Twitto, Noam Ordan, and Shuly Wintner, "Statistical Machine Translation and Automatic Identification of Translationese", in preparation

Other Contributions

- Roee Aharoni, Moshe Koppel, and Yoav Goldberg, "Automatic detection of machine translated text and translation quality estimation", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 289-295, 2014
- Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer, "Identifying the L1 of non-native writers", *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 2013

Future Directions

- Identification of translationese at the sentence-pair level
- The features of **machine** translationese
- More applications to machine translation

Acknowledgements

- Noam Ordan, Vered Volansky, Ehud Alexander Avner, Naama Twitto, Gennadi Lembersky, Moshe Koppel
- Israel Ministry of Science and Technology

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`.

Omar S. Al-Shabab. **Interpretation and the language of translation: creativity and conventions in translation**. Janus, Edinburgh, 1996.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. Identifying translationese at the word and sub-word level. **Literary and Linguistic Computing**, Forthcoming.

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, **Text and technology: in honour of John Sinclair**, pages 233–252. John Benjamins, Amsterdam, 1993.

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.

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.

Bibliography II

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. **SIGKDD Explorations**, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL http://dx.doi.org/10.1145/1656274.1656278.

Iustina Ilisei. **A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models**. PhD thesis, University of Wolverhampton, Wolverhampton, UK, February 2013. URL http://clg.wlv.ac.uk/papers/ilisei-thesis.pdf.

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.

Sara Laviosa. Core patterns of lexical use in a comparable corpus of English lexical prose. **Meta**, 43(4):557–570, December 1998.

Sara Laviosa. **Corpus-based translation studies: theory, findings, applications**. Approaches to translation studies. Rodopi, 2002. ISBN 9789042014879.

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.

Bibliography III

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.

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.

Lin Øverås. In search of the third code: An investigation of norms in literary translation. **Meta**, 43(4):557–570, 1998.

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.

Gideon Toury. Interlanguage and its manifestations in translation. **Meta**, 24(2):223–231, 1979.

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.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. Identifying the L1 of non-native writers: the CMU-Haifa system. In **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pages 279–287. Association for Computational Linguistics, June 2013. URL http://www.aclweb.org/anthology/W13-1736.

Naama Twitto-Shmuel, Noam Ordan, and Shuly Wintner. Statistical machine translation and automatic identification of translationese. Under review, Forthcoming.

Bibliography IV

Hans van Halteren. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, editors, **COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK**, pages 937–944, 2008. ISBN 978-1-905593-44-6. URL `http://www.aclweb.org/anthology/C08-1118`.

Ria Vanderauwerea. **Dutch novels translated into English: the transformation of a 'minority' literature**. Rodopi, Amsterdam, 1985.

Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. **Literary and Linguistic Computing**, Forthcoming.