Assessing Genre and Method Variation in Translation Using Computational Techniques

Ekaterina Lapshinova-Koltunski[1], Marcos Zampieri[1,2]

[1]Saarland University
[2]German Research Center for Artificial Intelligence (DFKI)

Abstract

In this paper, we propose the use of automatic text classification techniques to analyse variation in English-German translations. Our experiments work under the assumption that text classification methods can level out discriminative features of different translation varieties that intuition alone cannot grasp; thus enabling researchers to investigate in more detail the properties of each of them. We call different types of translations distinguished by genre, translation method and further influencing factors *translation varieties*, see (Lapshinova-Koltunski, 2015).

Text classification is an important area of research in Natural Language Processing (NLP) and it has been applied to a wide range of tasks such as spam detection (Medlock, 2008), language identification (Lui and Baldwin, 2012) and temporal text classification(Abe and Tsumoto, 2010). From a pure engineering perspective, researchers are interested in how well classification methods can distinguish between two or more classes and what kind of features and algorithms deliver the best performance in each task. In recent works (Zampieri et al., 2013; Diwersy et al., 2014), state-of-the-art text classification methods were proposed (e.g. Naive Bayes, Support Vector Machines), which operate with linguistically motivated features to investigate language variation across corpora. These methods were successfully applied in the identification of languages, their varieties and dialects, as well as genres. In the present approach, we intend to automatically classify translated texts distinguished by genre and method of translation. Our main goal is to level out features diversifying genres vs. translation method.

The application of text classification methods requires annotated corpora containing morphosyntactic information, e.g. part-of-speech tags. For our analysis, we use VARTRA-SMALL (Lapshinova-Koltunski, 2013), a corpus of multiple translations from English into German. The multiple translations of the same texts were produced with different translation methods (e.g. human translation and machine translation), and the texts contained in the corpus belong to different genres, i.e. political speeches and essays, fictional texts, instruction manuals, popular-scientific articles and tourism leaflets.

In light of the aforementioned recent studies, we propose to use Naive Bayes classifier to discriminate between (1) different genres of the translation corpus (2) different translation methods, and investigate what are the discriminative features in this classification

task. These procedures will provide us with the information on which linguistic features are responsible for the two-fold variation in translation: (1) across genres, and (2) across translation method as in Volansky et al. (2011).

In our presentation, we will show the methods and resources used in the present study, as well as the classification results and their interpretation.

References

Abe, H. and Tsumoto, S. (2010). Text categorization with considering temporal patterns of term usages. In *Proceedings of ICDM Workshops*, pages 800–807. IEEE.

Diwersy, S., Evert, S., and Neumann, S. (2014). A semi-supervised multivariate approach to the study of language variation. *Linguistic Variation in Text and Speech, within and across Languages.*

Lapshinova-Koltunski, E. (2013). VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria. Association for Computational Linguistics.

Lapshinova-Koltunski, E. (2015). Linguistic features in translation varieties: Corpus-based analysis. In De Sutter, G., Delaere, I., and Lefer, M.-A., editors, *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*, TILSM. Mouton de Gruyter (to appear).

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Meeting of the ACL.*

Medlock, B. (2008). Investigating classification for natural language processing tasks. Technical report, University of Cambridge - Computer Laboratory.

Volansky, V., Ordan, N., and Wintner, S. (2011). More human or more translated? original texts vs. human and machine translations. In *Proceedings of the 11th Bar-Ilan Symposium on the Foundations of AI With ISCOL (Israeli Seminar on Computational Linguistics).*

Zampieri, M., Gebre, B. G., and Diwersy, S. (2013). N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN2013*, pages 580–587, Sable d'Olonne, France.