Translation Model Adaptation Using Genre-Revealing Text Features

abstract

van der Wees, M.; Bisazza, A.; Monz, C.

Publication date
2015

Document Version
Final published version

Proceedings

Proceedings of the 14th Dutch-Belgian Information Retrieval Workshop

Translation Model Adaptation Using Genre-Revealing Text Features (Abstract)

Marlies van der Wees Arianna Bisazza Christof Monz
Informatics Institute, University of Amsterdam
{m.e.vanderwees,a.bisazza,c.monz}@uva.nl

1. INTRODUCTION AND METHOD

We summarize the findings of van der Wees et al. [6]. Domain adaptation is an active field for statistical machine translation (SMT), and has resulted in various approaches that adapt system components to specific translation tasks. However, the concept of a *domain* is not precisely defined. Different domains typically correspond to different subcorpora, in which documents exhibit a particular combination of *genre* and *topic*. This definition has two major shortcomings: First, subcorpus-based domains depend on *provenance* information, which might not be available, or on manual grouping of documents into subcorpora, which is labor intensive and often carried out according to arbitrary criteria.

Second, the commonly used notion of a domain neglects the fact that topic and genre are two distinct properties of text [5]. While this distinction has long been acknowledged in text classification literature [3, 4, among others], most work on domain adaptation in SMT uses in-domain and out-of-domain data that differs on both the topic and the genre level, making it unclear whether the proposed solutions address topic or genre differences.

In our work, we follow text classification literature for definitions of the concepts topic and genre [4], and we recently studied the impact of both aspects on SMT [7]. Motivated by the observation that translation quality varies more between genres than across topics, we explore in this paper the task of *genre adaptation*. Concretely, we incorporate genre-revealing features, inspired by previous findings in genre classification literature, into a competitive translation model adaptation approach based on phrase pair weighting using a vector space model (VSM) [2].

In this approach, phrase pairs in the training data are represented by a vector capturing specific information about the phrase pair. In addition to the phrase pair vectors, a single vector is created for the development set which is similar to the test set. Next, for each training data phrase pair, we compute a similarity score between its vector and the development vector. This similarity is assumed to indicate the relevance of the phrase pair with respect to the test set's genre and is added to the decoder as a new feature.

We compare a number of variants of the general VSM framework, differing in the way vectors are defined and constructed. First, we adhere to the common scenario in which adaptation is guided by manual subcorpus labels that resemble the training data's provenance. Next, to move away from manual labels, we explore the use of genre-revealing features that have proven successful for distinguishing genres in classification tasks. The features that are most discriminative between the genres in our test sets (newswire (NW) and user-generated (UG) text) are counts of first and second person pronouns, exclamation and question marks, repeating punctuation, emoticons, and numbers. Finally, we use LDA-inferred [1] document distributions as a third vector representation.

2. RESULTS AND CONCLUSIONS

We evaluate different VSM variants on two Arabic-to-English translation tasks, both comprising the genres NW and UG. Besides VSM variants containing only one of the presented feature types (i.e., manual provenance labels, automatic genre features, or LDA distributions), we also explore various combinations in which multiple VSM similarities are added as additional decoder features. Our best performing system includes both genre features and LDA distributions, suggesting that the two vector representations are to some extent complementary.

In a series of experiments we show that automatic indicators of genre can replace manual subcorpus labels, yielding significant improvements of up to 0.9 BLEU over a competitive unadapted baseline. In addition, we observe small improvements when using automatic genre features on top of manual subcorpus labels. We also find that the genre-revealing feature values can be computed on either side of the training bitext, indicating that our proposed features can be robustly projected across languages. Therefore, the advantages of using the proposed method are twofold: (i) manual subcorpus labels are not required, and (ii) the same set of features can be used successfully across different test sets and languages. Finally, we find that our genre-adapted translation models encourage document-level translation consistency (i.e., consistent translation of repeated phrases within a single document) with respect to the unadapted baseline.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[2] B. Chen, R. Kuhn, and G. Foster. Vector space model for adaptation in statistical machine translation. In *Proceedings of ACL*, pages 1285–1293, 2013.
[3] N. Dewdney, C. VanEss-Dykema, and R. MacMillan. The form is the substance: classification of genres in text. In *Proceedings of the Workshop on HLT and KM*, 2001.
[4] M. Santini. State-of-the-art on automatic genre identification. Techn. Report ITRI-04-03, Information Technology Research Institute, University of Brighton, 2004.
[5] B. Stein and S. Meyer Zu Eissen. Distinguishing topic from genre. In *Proceedings of I-KNOW 06*, pages 449–456, 2006.
[6] M. van der Wees, A. Bisazza, and C. Monz. Translation model adaptation using genre-revealing text features. In *Proceedings of DiscoMT'15*, pages 132–141, 2015.
[7] M. van der Wees, A. Bisazza, W. Weerkamp, and C. Monz. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of ACL*, pages 560–566, 2015.