A Wikipedia-based Corpus for Contextualized Machine Translation

Jennifer Drexler*, **Pushpendre Rastogi**†, **Jacqueline Aguilar**‡, **Benjamin Van Durme**‡, **Matt Post**‡
*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
†Center for Language and Speech Processing, Johns Hopkins University
‡Human Language Technology Center of Excellence, Johns Hopkins University
jdrexler@mit.edu, {pushpendre,jacqui.aguilar,vandurme,post}@jhu.edu

Abstract

We describe a corpus for and experiments in *target-contextualized* machine translation (MT), in which we incorporate language models from target-language documents that are comparable in nature to the source documents. This corpus comprises (i) a set of curated English Wikipedia articles describing news events along with (ii) their comparable Spanish counterparts, (iii) a number of the Spanish source articles cited within them, and (iv) English reference translations of all the Spanish data. In experiments, we evaluate the effect on translation quality when including language models built over these English documents and interpolated with other, separately-derived, more general language model sources. We find that even under this simplistic baseline approach, we achieve significant improvements as measured by BLEU score.

Keywords: Machine Translation, Domain Adaptation, Corpus

1. Introduction

We describe a corpus for *target-contextualized* machine translation (MT). Here, the task is to improve the translation of source documents using language models built over presumably related, comparable documents in the target language. As a motivating example, this could be useful in a situation where there is a collection of in-language documents related to a topic, but a specific document of interest is available only in another language. Our corpus comprises (i) a set of curated English Wikipedia articles describing news events, along with (ii) their Spanish counterparts, (iii) a number of the Spanish source articles cited within them, and (iv) human-produced reference translations of all the Spanish documents. In experiments, we translated these Spanish documents using a general translation system built on out-of-domain data. We then built multiple "in-domain" language models on the comparable English documents — one for each pair of documents — and interpolated them with the larger model, evaluating the effect on translation quality, effectively casting this task as one of (hyper) domain adaptation for MT. We find that even under this simplistic baseline approach, we achieve significant improvements as measured by BLEU score (Papineni et al., 2002).

This work relates to previous efforts in domain adaptation in machine translation (Eidelman et al., 2012; Langlais et al., 2000; Langlais and Lapalme, 2002; Brousseau et al., 1995; Dymetman et al., 1994). Many of these previous problem formulations were concerned with recognizing the general domain, or topic, of the material to be translated. In comparison, we are concerned here with adapting the translation via access to information related to the exact events being described, presented in the form of a language model. We hope this might help lead to richer approaches to MT, where coreference and semantics might be measurably brought to bear.

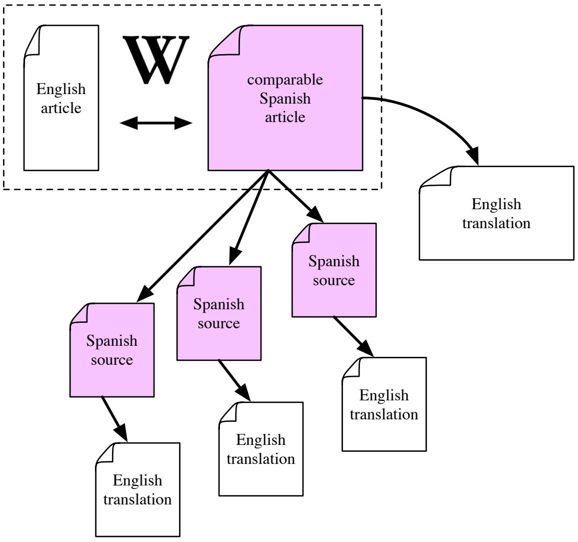

Figure 1: The data configuration. We collected comparable English and Spanish documents along with sources cited within the Spanish articles. Reference translations were then collected for all Spanish documents; in the experiments, we use both the original (comparable) English articles (§3.3.) and the translated articles (§3.1.) to build a language model to aid translation of each set of corresponding source documents.

2. Dataset

Our dataset consists of 9 English articles from Wikipedia on events of interest to the Spanish-speaking portion of the world. We manually extracted and sentence-segmented each article along with the Spanish Wikipedia article on the same topic.[1] It is important to note that these are not parallel documents, but are instead at best "comparable corpora", existing in the form of independently-written article on the same topic. Many such articles on Wikipedia, in fact,

[1] Documents were taken from Wikipedia on July 2, 2013.

Topic	Articles		Sources
	Eng.	Spa.	Spa.
Bolivia gas war	236	114	21
Catalan autonomy protest	31	45	87
Chavez death	122	52	49
Chilean mining accident	474	140	144
Comayagua prison fire	53	11	50
H1N1 chile	151	92	66
Papal election	147	70	82
Pemex explosion	29	82	117
Venezuelan presidential election	110	44	79
Total	1,343	650	695

Table 1: The topics and sentence counts of the Spanish Wikipedia articles used as the basis for data collection. For each topic, we have the comparable Wikipedia articles (both English and Spanish) and the sources cited within the Spanish articles.

λ	1.0	0.7	0.3	0.0
Articles	28.97	**31.45**	31.27	23.69
Sources	26.76	**27.89**	27.50	18.01

Table 2: BLEU scores (averaged across each article or group of source articles) from changing the interpolation weights λ, which denotes the weight assigned to the large, general-purpose language model. *Articles* denotes the average score from translating the Spanish articles, while *sources* denotes the source documents cited within those articles.

are written by authors writing independently in their native language on the same topic, but from their own perspective, in contrast to simply filling out the inter-language resources of the encyclopedia by translating documents from high- to low-resource languages. Our dataset leverages this fact, providing a relatively small corpus with high information content on the Spanish side, and comparable but impoverished versions on the target (English side). This is pictured in Figure 1.

In addition to the comparable document pairs, we reached into the Spanish documents and collected many of the sources cited within them. Many of these citations linked outside of Wikipedia, often to news websites. These sources were then also manually extracted and sentence-segmented. In total, we collected 695; Table 1 contains a list of the Wikipedia articles along with sentence-level extract statistics.

All Spanish data was then manually translated by a single bilingual speaker to produce a reference translation used for translation.

3. Experiments

We used the Joshua Machine Translation Toolkit[2] (Post et al., 2013). We include both baseline results translating the Spanish source articles as well as method of adapting the baseline machine translation system to the domain of a specific document. The baseline models were trained on grammars extracted from the Spanish–English portion of version 7 of the Europarl Corpus (Koehn, 2005). From these, we extracted Hiero grammars (Chiang, 2005) using the standard extraction settings (Chiang, 2007). We then build a 5-gram Kneser-Ney-smoothed language model from the English side of this data using KenLM (Heafield, 2011). The parameters of the decoder's linear model were tuned with Z-MERT (Och, 2003; Zaidan, 2009), which are included in the Joshua toolkit.

We describe three experiments. In the first, we use the reference translations of the articles as our set of target knowledge. Since they were produced from the Spanish versions, these reference translations can be thought of as *parallel* Wikipedia documents, in contrast to the comparable ones that we extracted. In the second, the English Wikipedia articles serve this role. We explore a range of interpolation weights in order to determine whether any setting will result in some improvement. The purpose of these two experiments is to determine whether in-domain language models built even on very small, targeted pieces of data might be useful in improving machine translation quality: first, in a parallel setting, and second, in a comparable setting. Finally, in a third set of experiments, we test whether these ideas can generalize when we have very small amounts of data, using a proxy setting to determine the interpolation weights.

3.1. Experiment 1: Can the contextual data help (parallel setting)?

The contextualized target-side language models are built on very small amounts of data, numbering at most a few hundred sentences (Table 1). This made them too unreliable to use as separate language models during decoding (with their own weight assigned in the decoder's linear model). Instead, our approach was to interpolate the large, general language model trained on millions of sentences (Europarl) with the domain-specific one.

For this first set of experiments, instead of using the (comparable) English Wikipedia documents to build the contextualized language model, we looked at what should be an easier setting: using the translations of the articles to build a language model when translating the sources, and vice versa. Table 2 presents the results four settings for the interpolation weight λ, which determines how much weight is given to the large LM. With a weight of $\lambda = 0.7$, we see improvements in BLEU score from including the in-domain LM. To be clear, when translating the articles, the contextualized LM is built on the sources, and when translating the sources, the LM is built on the articles. Note that the setting $\lambda = 1.0$ is the baseline setting, with all the weight assigned to the large, out-of-domain language model.

3.2. Experiment 2: Can contextualized LMs help (comparable setting)?

For the second set of experiments, we use the comparable English Wikipedia articles to build the in-domain language model. Table 3 contains the results. Note that the scores are all slightly lower than when using the parallel setting, which makes sense, since the comparable data are not certain to be as close. Importantly, however, the best setting

[2] joshua-decoder.org

λ	1.0	0.7	0.3	0.0
Sources	26.76	**27.32**	26.30	16.18

Table 3: Averaged BLEU scores from varying λ, this time using an LM built on the comparable target-language corpora (English Wikipedia articles, instead of translated Spanish articles).

remains unchanged.

3.3. Experiment 3: Can they generalize?

The above experiment demonstrates that there are interpolation weights that result in an improved BLEU score. However, how can we automatically find that weight? The conventional approach is to use tuning data of the same sort and use that to set it. However, our sources, articles, and their corresponding in-domain LMs are small enough that this is difficult. We could reserve some of the data as development data for tuning this parameter. Instead, we used a proxy approach that uses unrelated (and abundant) data set up in an analogous situation to learn the interpolation weights.

This analogous data was created as follows: we generated development data from the Spanish–English portion of the News Commentary Corpus, released as part of WMT 2013.[3] We split this data into its constituent news stories, and selected the longest ones. We split each of these stories in half, treating the English side of the first half as context for the construction of a language model and using the second half to tune the interpolation weight against. This language model was then interpolated with the much larger Europarl language model, searching over interpolation weights with grid search. The best interpolation weight is then used to interpolate the same Europarl LM with the in-domain LM built over the English Wikipedia article associated with whatever Spanish source document was being translated. A single interpolation weight is thus learned for all of the articles.

The best weight found in these "proxy experiments" over the weights presented in Table 2 was also 0.7. For these experiments, this proxy situation allows us to discover the best interpolation weight, providing nice gains in BLEU score over the baseline.

4. Summary

The domain adaptation method used here is based on a simple idea: we perform optimization using models tailored to the development data, and then replace those models at test time with models adapted to the test domain. Here, the news commentary corpus language model used during tuning is replaced with the target-contextualized language model at test time. In this way, our methods can be used for documents for which no matching development data is available and can easily scale to the translation of large document collections.

Along with these experiments, we release the associated data collected from Wikipedia and translated.[4] This data and enables research in this specific domain-adaptation scenario, and the results suggest that further research in this area could be productive. We hope this collection will help spur discussion on the task of contextualized machine translation, and are willing to extending the resource based on community interest.

Acknowledgments The authors thank the Human Language Technology Center of Excellence for supporting this work via its summer workshop program.

5. References

Brousseau, J., Drouin, C., Foster, G. F., Isabelle, P., Kuhn, R., Normandin, Y., and Plamondon, P. (1995). French speech recognition in an automatic dictation system for translators: the transtalk project. In *Eurospeech*.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Dymetman, M., Brousseau, J., Foster, G., Isabelle, P., Normandin, Y., and Plamondon, P. (1994). Towards an automatic dictation system for translators: The TransTalk Project. *arXiv preprint cmp-lg/9409012*.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 115–119. Association for Computational Linguistics.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Workshop on Statistical Machine Translation*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Langlais, P. and Lapalme, G. (2002). Trans type: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 17(2):77–98.

Langlais, P., Foster, G., and Lapalme, G. (2000). Transtype: a computer-aided translation typing system. In *Proceedings of the 2000 NAACL-ANLP Workshop on Embedded machine translation systems-Volume 5*, pages 46–51. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*, Sapporo, Japan, July.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July.

Post, M., Ganitkevitch, J., Orland, L., Weese, J., Cao, Y., and Callison-Burch, C. (2013). Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 206–212, Sofia, Bulgaria, August. Association for Computational Linguistics.

[3] statmt.org/wmt13/

[4] hltcoe.jhu.edu/publications/data-sets-and-resources/

Zaidan, O. F. (2009). Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.