Bootstrapping Translation Detection and Sentence Extraction from Comparable Corpora

Kriste Krstovski[†,§] **and David A. Smith**[‡]
[†]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA
[§]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA
[‡]College of Computer and Information Science, Northeastern University, Boston, MA
kkrstovski@cfa.harvard.edu, dasmith@ccs.neu.edu

Abstract

Most work on extracting parallel text from comparable corpora depends on linguistic resources such as seed parallel documents or translation dictionaries. This paper presents a simple baseline approach for bootstrapping a parallel collection. It starts by observing documents published on similar dates and the co-occurrence of a small number of identical tokens across languages. It then uses fast, online inference for a latent variable model to represent multilingual documents in a shared topic space where it can do efficient nearest-neighbor search. Starting from the Gigaword collections in English and Spanish, we train a translation system that outperforms one trained on the WMT'11 parallel training set.

1 Introduction

In statistical machine translation (SMT), the quality of the translation model is highly dependent on the amount of parallel data used to build it. Parallel data has usually been generated through the process of human translation, which imposes significant costs when building systems for new languages and domains. To alleviate this problem, researchers have considered comparable corpora—a collection of multilingual documents that are only topically aligned but not necessary translations of each other (Fung and Cheung, 2004). While most previous approaches for mining comparable corpora heavily depend on initializing the learning process with some translation dictionaries or parallel text, we use multilingual topic models to detect document translation pairs and extract parallel sentences with only minimum cross-language prior knowledge: the publication dates of articles and the tendency of some vocabulary to overlap across languages. Processing only four years of Gigaword news stories in English and Spanish, we are able to outperform the WMT'11 baseline system trained on parallel News Commentary corpus (Table 1).

2 Prior Work on Comparable Corpora

Most previous, if not all, approaches for mining comparable corpora heavily depend on bilingual resources, such as translation lexica, bitext, and/or a pretrained baseline MT system. This paper, in contrast, investigates building MT systems from comparable corpora without such resources. In a widely cited early paper, Munteanu and Marcu (2005) use a bilingual dictionary and a collection of parallel sentences to train IBM Model 1 and a maximum entropy classifier to determine whether two sentences are translations of each other. Tillmann and Xu (2009) and Smith et al. (2010) detect parallel sentences by training IBM Model 1 and maximum entropy classifiers, respectively. In later work on detecting sentence and phrase translation pairs, Cettolo et al. (2010) and Hoang et al. (2014) use SMT systems to translate candidate documents; Quirk et al. (2007) use parallel data to train a translation equivalence model; and Ture and Lin (2012) use a translation lexicon to build a scoring function for parallel documents. More recently, Ling et al. (2013) trained IBM Model 1 on bitext to detect translationally equivalent phrase pairs within single microblog posts. Abdul-Rauf and Schwenk (2009), Uszkoreit et al. (2010), and Gahbiche-Braham et al. (2011),

Proceedings of NAACL-HLT 2016, pages 1127–1132,
San Diego, California, June 12-17, 2016.

rather than trying to detect translated sentence pairs directly, translate the entire source language side of a comparable corpus into the target language with a baseline SMT system and then search for corresponding documents.

On the other hand, there exist approaches that mine comparable corpora without any prior translation information or parallel data. Examples of this approach are rarer, and we briefly mention two: Enright and Kondrak (2007) use singleton words (hapax legomena) to represent documents in a bilingual collection for the task of detecting document translation pairs, and Krstovski and Smith (2011) construct a vocabulary of overlapping words to represent documents in multilingual collections. The latter approach demonstrates high precision vs. recall values on various language pairs from different languages and writing systems when detecting translation pairs on a document level such as Europarl sessions. Recently proposed approaches, such as (Klementiev et al., 2012) use monolingual corpora to estimate phrase-based SMT parameters. Unlike our paper, however, they do not demonstrate an end-to-end SMT system trained without any parallel data.

Our approach differs from these and other previous approaches by not relying on any initial translation dictionary or any bitext to train a seed SMT system. Therefore, the primary experimental comparison that we perform is between no bitext at all and a system trained with some bitext.

3 Bootstrapping Approach

Our bootstrapping approach (Figure 1) is a two-stage system that used the Overlapping Cosine Distance (OCD) approach of Krstovski and Smith (2011) as its first step. OCD outputs a ranked list of candidate document pairs, which are then fed through a sentence-alignment system (Moore, 2002). A polylingual topic model (PLTM) (Mimno et al., 2009) is then trained on the aligned portions of these documents. Using the trained model, we infer topics on the whole comparable training set. Once represented as points in the topic space, documents are then compared for similarity using divergence based metrics such as Hellinger (He) distance. Results from these comparisons create a single ranked list of text translation pairs, which are on a sub document length level. From this single ranked list, using thresholding, we again extract the top n candidate translation pairs that are then fed to an aligner for further refinement.

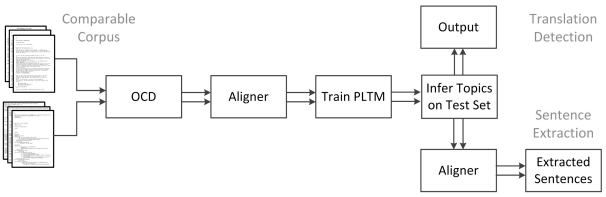

Figure 1: The bilingual collection processing pipeline.

3.1 Discovering Document Translation Pairs

For a given comparable corpus, OCD assumes that there is a set of words that exist in both languages that could be used as features in order to discriminate between documents that are translations of each other, documents that carry similar content, and documents that are not related. Firstly, for each language in the collection a vocabulary is created which consists of all word types seen in the corpora of that language. Words found in both source (s) and target (t) languages are extracted and the overlapping list of words are then used as dimensions for constructing a feature vector template. Documents in both languages are then represented using the template vector whose dimensions are the tf·idf values computed on the overlapping words which we now consider as features. While the number of overlapping words is dependent on the families of the source and target languages and their orthography, Krstovski and Smith (2011) showed that this approach yields good results across language pairs from different families and writing systems such as English-Greek, English-Bulgarian and English-Arabic where, as one would expect, most shared words are numbers and named entities.

We compare these vector representations efficiently using Cosine (Cos) distance and locality sensitive hashing (Charikar, 2002). This results in a single ranked list of all document pairs. Compared to the traditional cross-language information retrieval (CLIR) task where a set of document queries is known in advance, there is no prior information on the documents in the source language that may or may not have translation documents in the target language of the collection. Due to the length in-

variance of Cos distance, the single ranked list may contain document pairs with high similarity value across all documents in the target language. This issue in OCD is resolved by applying length and diversity filtering. Length filtering removes translation pairs where the length of the target document t is not within $\pm 20\%$ of the source document s length, $lf : 0.8 \leq |s| / |t| \leq 1.2$. For a given source document, diversity filtering is done by allowing only the top five ranked target document pairs to be considered in the single ranked list. Limiting the number of target documents for a given source document may discard actual document translation pairs such as in a comparable corpus of news stories where documents in the target language originate from large number of news source. While it may restrict more document translation pairs to be discovered, the diversity filtering, on the other hand prevents from limiting the number of discovered similar and translation documents to be from the same topic and domain and thus introduces diversity on another, domain or topic based, level.

3.2 Representing Multilingual Collections with Topics

Latent topic models are statistical models of text that discover underlying hidden topics in a text collection. We use PLTM (Mimno et al., 2009), a multilingual variant of LDA, which assumes that document tuples in multilingual parallel and comparable corpora are drawn from the same tuple-specific multinomial distribution over topics θ. For each document in the tuple, PLTM assumes that words are generated from a language L specific topic distribution over words β_L. Using this generative model we represent documents in multiple languages in a common topic space which allows us to perform similarity comparisons across documents in different languages.

The original PLTM posterior inference is approximated using collapsed Gibbs sampling (Mimno et al., 2009). While more straightforward to implement, this inference approach requires iterating over the multilingual collection multiple times to achieve convergence. This incurs a computational cost that could be significant for large collections such as Gigaword. Moreover, detecting and retrieving document translation pairs requires all-pairs comparison across documents in both languages with a worst case time complexity of $O(N^2)$ which is impractical for large comparable corpora. One solution to this problem is to parallelize the brute-force approach through the MapReduce framework (Ture et al., 2011; Ture and Lin, 2012) but this approach requires special programming methods.

In order to use the PLTM on large collections and avoid the bottleneck introduced by Gibbs sampling, we use the online variational Bayes (VB) approach originally developed by (Hoffman et al., 2010) for LDA model to develop a fast, online PLTM model. As in the regular VB approach, online VB approximates the hidden parameters θ, z and β using the free variational parameters: γ, ϕ and λ. Rather than going over the whole collection of documents to bring the variational parameters to a convergence point, Krstovski and Smith (2013) perform updates of the variational parameters γ and ϕ_L on document batches and update the λ_L variational parameter as a weighted average of its stochastic gradient based approximation and its value on the previous batch. The approximation is done through Expectation-Maximization (EM).

Unlike the usual metric spaces where two vectors are compared using distance metrics such as Euclidean (Eu) or Cos distance, in the probability simplex similarity is computed using information-theoretic measurements such as Kullback-Leibler, Jensen-Shannon divergence and He distance. We alleviate the $O(N^2)$ worst case time-complexity in the probability simplex by utilizing approximate nearest-neighbor (NN) search techniques proven in the metric space. More specifically, we use the formulaic similarity between He and Eu: $He(p,q) \equiv Eu(x,y)$, when $\forall i : i = 1, n$ of x_i and y_i, $x_i = \sqrt{p_i}$ and $y_i = \sqrt{q_i}$, and compute He distance using Eu based, approximate NN computation approaches such as k-d trees[1] (Bentley, 1975).

4 Experiments and Results

We demonstrate the performance of the bootstrapping approach on the task of extracting parallel sentences to train a translation system. We evaluate MT systems trained on extracted parallel sentences and

[1] We use k-d tree implementation in the ANN library (Mount and Arya, 2010).

compare their performance against MT systems created using clean parallel collections. MT systems were evaluated with the standard BLEU metric (Papineni et al., 2002) on two official WMT test sets that cover different domains: News (WMT'11) and Europarl (WMT'08). We trained the Moses SMT system (Koehn et al., 2007) following the WMT shared task guidelines for building a baseline system with one of two parallel training collections from WMT'11: English-Spanish News Commentary (v6) and Europarl (v6). MT systems were trained using test-domain specific language models (LM) — English News Commentary for News test and English Europarl for the Europarl test. Our comparable corpus consists of news stories from the English (LDC2011T07) and Spanish (LDC2011T12) Gigaword collections.

We perform the following processing in each step of the pipeline. We run OCD on days of news originating from multiple news agencies or more specifically on news stories originating from the same day which we consider as the "minimal supervision" in initiating the bootstrapping process. Since the OCD approach generates a single list of ranked document translation pairs, for the second stage of our pipeline we consider the top n document translation pairs. We define n to be all document translation pairs whose Cos similarity is between the range of the max (i.e. the top 1 scored document translation pair in the single ranked list) and $\frac{max}{2}$. Unlike previous thresholding based on absolute values (Ture et al., 2011), this approach allows us to utilize threshold values that are automatically adjusted to the dynamic range of the Cos distance of a particular corpus. Sentences from the top n news stories are extracted and are further aligned. The output of the aligner is then used as a training set for the PLTM model. We represent each of the news stories using the per story aligned sentences. Once trained, we use the PLTM model to infer topics back on to the news stories. We then again create a single ranked list of translation news story pairs by computing divergence based similarity using He distance (§3.2). Keeping the top n ranked news story pairs, we obtain a list of what we believe are parallel documents which we then use to extract sentence pairs. Sentences are finally processed through an aligner and then used as the training corpus to our MT system.

Training Source	Bitext	Extr.	Test Set
News Comm. (NC)	131k	0	23.75
Europarl (EP)	1,750k	0	23.91
Gigaword (GW)	0	926k	24.28*
NC+GW	131k	926k	24.92*
EP+GW	1,750k	926k	25.90*

Table 1: BLEU score values computed over the WMT'11 News test set with MT systems developed using extracted and parallel sources of training data. * denotes statistical significance level (p-value$\leq$0.001) above NC.

The Gigaword collection contains news stories generated from various agencies in different languages. On any given day, a news story in English may or may not cover the same topic as one in a different language. To perform a fair evaluation with the WMT'11 News test, we considered stories published in non-overlapping years[2]: 2010, 2009, 2005 and 2004. Table 1 shows the performance comparison, on the News test set (WMT'11), of the MT system trained on extracted parallel sentences from four years of Gigaword data (GW) with a MT system trained on two WMT'11 baseline parallel collections: Europarl (EP) and News Commentary (NC). While over 10 times bigger than NC, EP is out of domain and thus performs only slightly better. On the News test set, parallel sentences automatically extracted from only four years of Gigaword data outperform systems trained on clean NC or EP bitext.

In order to determine statistically significant differences between the results of different MT systems we ran the randomization test (Smucker et al., 2007) on the News test set with 10k iterations. In each iteration we performed permutations across the translation sentences obtained from the two MT systems whose statistical difference in performance we evaluate.

Table 2 shows the performance comparison on the Europarl test set (WMT'08) between the MT system trained on the extracted parallel sentences and the two MT baseline systems. On this test set, unsurprisingly, EP training performed very well.

Table 3 gives a summary of ablation experiments that we performed across the two stages of our bootstrapping approach. More specifically, we ex-

[2]We did not consider news stories from 2006-2008 due to a known issue with diacritic marks in the Spanish collection.

Training Source	Bitext	Extr.	Test Set
News Comm. (NC)	131k	0	25.43
Europarl (EP)	1,750k	0	32.06
Gigaword (GW)	0	926k	23.88
NC+GW	131k	926k	25.61
EP+GW	1,750k	926k	31.59

Table 2: BLEU score values computed over the WMT'08 Europarl test set with MT systems developed using extracted and parallel sources of training data.

Pipeline Configuration	Extr.	Test Set	
		News	Europarl
OCD	684k	$24.00^{\ddagger}$	23.84
OCD (dedup.)	469k	23.84	23.75
GW	926k	$24.28^{*,\dagger}$	23.88
GW (dedup.)	588k	$24.20^{*,\S}$	24.67

Table 3: Summary of ablation experiments: BLEU score values of MT systems trained on extracted bitext by OCD alone and with PLTM reestimation along with the deduplication (dedup.) effect. * denotes statistical significance level (p-value$\leq$0.001) above NC. ‡ denotes statistical significance level (p-value$\leq$0.05) above NC. † denotes statistical significance level (p-value$\leq$0.001) above OCD. § denotes statistical significance level (p-value$\leq$0.03) above OCD.

plored using bitext extracted by OCD alone, without PLTM reestimation, to train a MT system. Both extracted bitext sets also contained many duplicate sentence pairs. In this set of experiments we also explored the effect of deduplicating them, i.e. going over the extracted set of English-Spanish sentence pairs and removing the duplicate ones. Bitext extracted by OCD alone without PLTM reestimation performed only slightly worse on WMT'11. The OCD-only data, however, only showed 70% overlap with OCD+PLTM (GW). Deduplicating the two bitexts (dedup.) hurts OCD somewhat more than OCD+PLTM. On the Europarl test set, however, deduplicating OCD+PLTM bitext caused a significant boost from 23.88 to 24.67, while causing slight performance drop for OCD (cf. NC-trained 25.43). These interactions of test domain, redundancy, and model settings leave room for further studies of the performance of our bootstrapping approach.

5 Conclusion

We introduced a bootstrapping approach for detecting document translations and extracting parallel sentences through latent topic models that are trained with minimal prior knowledge and no lexical resources. The proposed approach is able to extract parallel sentences from comparable corpora to train MT models that outperform a baseline model trained on a parallel collection.

Acknowledgments

This work was supported in part by the Harvard-Smithsonian CfA predoctoral fellowship, in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *EACL*, pages 16–23.

Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *CACM*, 18(9):509–517.

Mauro Cettolo, Marcello Federico, and Nicola Bertoldi. 2010. Mining parallel fragments from comparable texts. In *IWSLT*, pages 227–234.

Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 308–388.

Jessica Enright and Grzegorz Kondrak. 2007. A fast method for parallel document identification. In *NAACL/HLT*, pages 29–32.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *EMNLP*, pages 57–63.

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, and François Yvon. 2011. Two ways to use a noisy parallel news corpus for improving statistical machine translation. In *the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 44–51.

Cuong Hoang, Anh-Cuong Le, Phuong-Thai Nguyen, Son Bao Pham, and Tu Bao Ho. 2014. An efficient

framework for extracting parallel sentences from non-parallel corpora. *Fundamenta Informaticae - Computing and Communication Technologies*, 130(2):179–199.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent Dirichlet allocation. In *NIPS*, pages 856–864.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without parallel corpora. In *EACL*, pages 130–140.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Kriste Krstovski and David A. Smith. 2011. A minimally supervised approach for detecting and ranking document translation pairs. In *WMT*, pages 207–216.

Kriste Krstovski and David Smith. 2013. Online polylingual topic models for fast document translation detection. In *WMT*, pages 252–261.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *ACL*, pages 176–186.

David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*, pages 880–889.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *AMTA*, pages 135–144.

David M. Mount and Sunil Arya, 2010. *ANN: A Library for Approximate Nearest Neighbor Searching*. `http://www.cs.umd.edu/~mount/ANN`.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Chris Quirk, Raghavendra Udupa U, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *MT Summit*, pages 321–327.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *NAACL/HLT*, pages 403–411.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, pages 623–632.

Christoph Tillmann and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *NAACL/HLT, Companion Volume: Short Papers*, pages 93–96.

Ferhan Ture and Jimmy Lin. 2012. Why not grab a free lunch?: mining large corpora for parallel sentences to improve translation modeling. In *NAACL/HLT*, pages 626–630.

Ferhan Ture, Tamer Elsayed, and Jimmy Lin. 2011. No free lunch: Brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In *SIGIR*, pages 943–952.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *COLING*, pages 1101–1109.