A Topic-based Approach for Post-processing Correction of Automatic Translations

Mohamed Morchid, Stéphane Huet, Richard Dufour

Laboratoire Informatique d'Avignon (LIA)
University of Avignon, France
firstname.lastname@univ-avignon.fr

Abstract

We present the LIA systems for the machine translation evaluation campaign of the *International Workshop on Spoken Language Translation* (IWSLT) 2014 for the English-to-Slovene and English-to-Polish translation tasks. The proposed approach takes into account word context; first, it maps sentences into a latent Dirichlet allocation (LDA) topic space, then it chooses from this space words that are thematically and grammatically close to mistranslated words. This original post-processing approach is compared with a factored translation system built with MOSES. While this post-processing method does not allow us to achieve better results than a state-of-the-art system, this should be an interesting way to explore, for example by adding this topic space information at an early stage in the translation process.

1. Introduction

This paper presents an original post-processing approach to correct machine translations using a set of topic-based features. The proposed method proceeds after the use of factored phrase-based machine translation (MT) systems [1]. The post-processed systems were submitted at the IWSLT 2014 MT evaluation campaign for two language directions: English-to-Slovene and English-to-Polish.

The focus and the major contribution of the proposed approach lie on mapping sentences to a topic space learned from a latent Dirichlet allocation (LDA) model [2], in order to replace every word identified as mistranslated with a thematically and grammatically close word. The idea behind this approach is that during the LDA learning process, the words contained in each sentence will retain the grammatical structure. Indeed, a topic space is usually learned from a corpus of documents and each document is considered as a "bag-of-words". Thus, the structure of sentences is lost as opposed to the proposed topic space that is learned from a corpus of sentences instead. This new topic space takes into account word distribution into sentences and is able to infer classes of close words.

In this exploratory study, the topic-based approach is applied in the context of automatic translations of morphologically rich languages. Slovene and Polish are both Slavic languages which are characterized by many inflections for a great number of words to indicate grammatical differences. This introduces many forms for a same lemma and raises many difficulties when translating from morphologically poor languages such as English. To deal with this problem in this study, words identified as erroneous are replaced by the morphological variant form sharing the same lemma and having the highest LDA score.

We summarize in Section 2 the resources used and the main characteristics of our systems based on the MOSES toolkit [3]. Section 3 presents the proposed topic-based approach to correct mistranslated words. Section 4 reports experiments on the use of factored translation models and the proposed approach. Finally, conclusions and perspectives are given in Section 5.

2. MOSES System Based on Factored Translation Models

2.1. Pre-processing

Systems were only built using data provided for the evaluation campaign, *i.e.* the *WIT* and *Europarl* corpora. Texts were pre-processed using an in-house script that normalizes quotes, dashes and spaces. Long sentences or sentences with many numeric or non-alphanumeric characters were also discarded. Each corpus was truecased, *i.e.* all words kept their case, apart from sentence-leading words that may be changed to their most frequent form (*e. g.* "Write" becomes "write" while "Paris" keeps its capital letter). Table 1 summarizes the used data and introduces designations that we follow in the remainder of this paper to refer to these corpora.

Slovene and Polish are morphologically rich languages with nouns, adjectives and verbs inflected for case, number and gender. This property requires to introduce morphological information inside the MT system to handle the lack of many inflectional forms inside training corpora. For this purpose, each corpus was tagged with Part-of-Speech (PoS) tags and lemmatized using OBELIKS [4] for Slovene[1] and TREETAGGER [5] for Polish[2]. These taggers asso-

[1] OBELIKS can be downloaded at http://eng.slovenscina.eu/tehnologije/oznacevalnik.

[2] TREETAGGER and its parameter file for Polish can be downloaded

CORPORA	DESIGNATION	SIZE (SENTENCES)
English-Slovene bilingual training		
Web Inventory of Transcribed and Translated Talks	*WIT*	17 k
Europarl v7	*Europarl*	616 k
English-Slovene development and test		
dev2012	*dev*	1.1 k
tst2012	*test0*	1.4 k
tst2013	*test13*	1.1 k
tst2014	*test14*	0.9 k
English-Polish bilingual training		
Web Inventory of Transcribed and Translated Talks	*WIT*	173 k
Europarl v7	*Europarl*	622 k
English-Polish development and test		
dev2010	*dev*	0.8 k
tst2010	*test0*	1.6 k
tst2013	*test13*	1.0 k
tst2014	*test14*	1.2 k

Table 1: Information on corpora.

ciate each word with a complex PoS including morphological information (e.g. "Ncmsan" for "Noun Type=common Gender=masculine Number=singular Case=accusative Animate=no"), and also its lemma. A description of the Slovene and Polish tagsets can be found on the Web[3].

In order to simplify the use of the two PoS taggers, we applied the tokenizer included in the OBELIKS and TREETAGGER tools to process all the corpora.

2.2. Language Models

Kneser-Ney discounted LMs were built from the Slovene and Polish sides of the bilingual corpora using the SRILM toolkit [6]. 4-gram LMs were trained for words, 7-gram LMs for PoS. A LM was built separately on each corpus: *WIT* and *Europarl*. These LMs were combined through linear interpolation. Weights were fixed by optimizing the perplexity on the *dev* corpus.

2.3. Alignment and Translation Models

All parallel corpora were aligned using MGIZA++ [7]. Our translation models are phrase-based models (PBMs) built with MOSES using default settings on a bilingual corpus made of *WIT* and *Europarl*. Weights of LM, phrase table and lexicalized reordering model scores were optimized on *dev* with the MERT algorithm [8].

at `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`.

[3]See `http://nl.ijs.si/spook/msd/html-en/msd-sl.html` for Slovene and `http://nkjp.pl/poliqarp/help/ense2.html` for Polish.

2.4. Factored Translation Model

The many inflections for Slovene and Polish are problematic for translation since morphological information, including case, gender and number, has to be induced from the English words. Factored translation models can be used to handle morphology and PoS during translations [1], with various setups available to use factors in several decoding or generation steps. In previous experiments conducted on translation into Russian, another morphologically rich language [9], we found that translating English words into (Russian words, PoS) pairs gave the highest improvements. We decided to apply this setup, which disambiguates translated words according to their PoS, for Slovene and Polish.

3. Post-processing Approach Relying on LDA

Classical language models consider words in their context (*n-gram*). Nonetheless, all possible contexts cannot be covered and some n-grams contained in the test corpus may not appear during the training process of the language model. For this reason, we propose to learn a topic space using LDA to associate a word inside a sentence with a set of thematically close words. By thematically, we mean that this word is associated with the context of the words contained in the sentence. Indeed, when a topic space is learned from a corpus of documents with usual LDA, words are associated with a document while grammatical structure is lost. In our case, this structure is preserved. Figure 1 gives an overview of the proposed topic-based approach to correct mistranslated words.

The next sections describe each step of the proposed approach based on a LDA topic space.

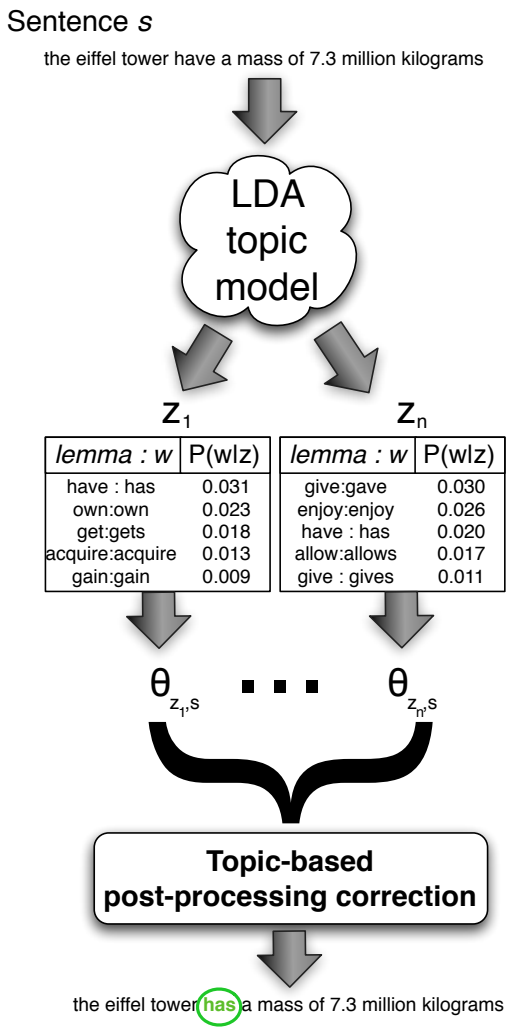

Figure 1: General overview of the proposed post-processing topic-based correction approach.

3.1. Latent Dirichlet Allocation (LDA)

Previous studies proposed to consider a document as a mixture of latent topics. The developed methods, such as Latent Semantic Analysis (LSA) [10, 11], Probabilistic LSA (PLSA) [12] or Latent Dirichlet Allocation (LDA) [2] build a high-level representation of a document in a topic space. Documents are then considered as "bags-of-words" [13] where the word order is not taken into account.

LDA is presented in its plate notation in Figure 2. These methods demonstrated their performance on various tasks, such as sentence [14] or keyword [15] extraction. Contrary to multinomial mixture models, LDA considers that a topic is associated with each occurrence of a word composing the document, rather than with the complete document. Thereby, a document can switch topic at any given word. Word occurrences are connected by a latent variable which controls the global distribution of topics inside a document. These latent topics are characterized by words and their corresponding distribution probability. PLSA and LDA models have been shown to generally outperform LSA on information retrieval tasks [16]. Moreover, LDA provides a direct estimate of the relevance of a topic, given a word set.

The generative process corresponds to the hierarchical Bayesian model shown in Figure 2. Several techniques, such as variational methods [2], expectation-propagation [17] or Gibbs sampling [18], have been proposed to estimate the parameters describing a LDA hidden space. Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) [19] and gives a simple algorithm to approximate inference in high-dimensional models such as LDA [20]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood defined as:

$$P(W|\overrightarrow{\alpha}, \overrightarrow{\beta}) = \prod_{w \in W} P(\overrightarrow{w}|\overrightarrow{\alpha}, \overrightarrow{\beta}) \quad (1)$$

for the whole data collection W given the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

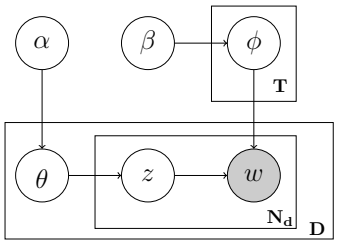

Figure 2: Generative models in plate notation for LDA model.

LDA estimation through Gibbs sampling was firstly reported in [18]; a more detailed description can be found in [20]. This method is used both to estimate the LDA parameters and to infer an unseen document with a hidden space of n topics. According to LDA, topic z is drawn from a multinomial over θ which is drawn itself from a Dirichlet distribution ($\overrightarrow{\alpha}$). In our context, topic space is learned from a lemmatized corpus where each word is associated with its lemma. Thus, a sentence can be inferred from a set of (word, lemma) pairs.

3.2. Topic-based Translation Correction

The first step of the proposed translation correction approach is to spot words that are likely to be mistranslated. For this purpose, a confidence score is computed for each word occurring in a sentence s using n-gram probabilities for each target word computed by the language model. Words with the smallest scores are assumed to be mistranslated and have to be corrected. In this paper, we propose to use a LDA topic space to find out relevant concurrent words w' to replace these suspected mistranslated words w. In order to do so,

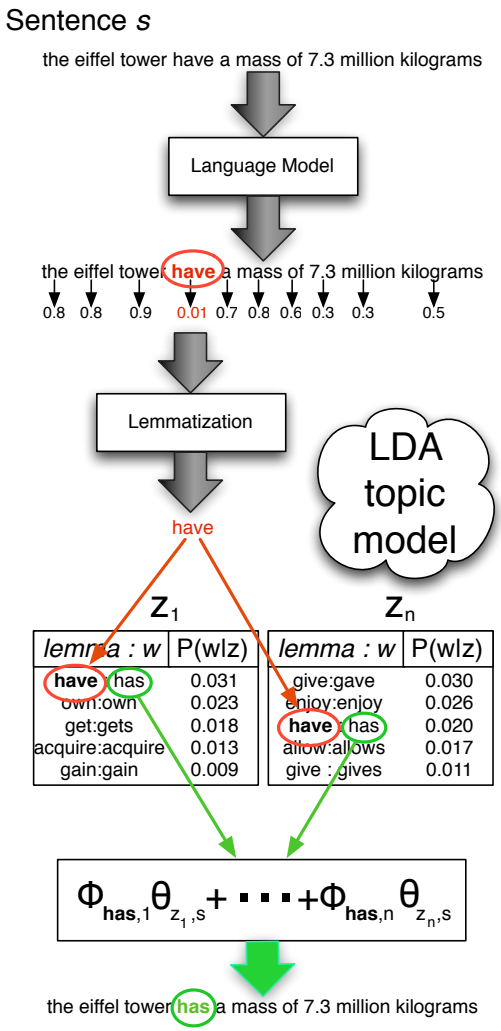

Figure 3: Details about the post-processing correction approach based on a LDA topic space.

Gibbs sampling is used to represent a new sentence s within the topic space of size n ($n = 100$ in our experiments) as shown in Figure 1, and to obtain the topic distribution:

$$\theta_{z_j,s} = P(z_j|s) \ . \tag{2}$$

The next step is to find out a relevant word w' that should replace the erroneous one w. Alternate words are searched among the words having a different inflection but satisfying the constraint:

$$\text{lemma}(w') = \text{lemma}(w) \ .$$

Each topic z is a distribution $P(w|z)$ over the vocabulary. Thus, a thematic confidence score is estimated for a concurrent word w' by:

$$\begin{aligned} \delta(w', s) &= P(w'|s) \\ &= \sum_{j=1}^{n} P(w'|z_j)P(z_j|s) \\ &= \sum_{j=1}^{n} \phi_{w',z_j}\theta_{z_j,s} \end{aligned} \tag{3}$$

where $\phi_{w',z_j} = P(w'|z_j)$ are computed during the training process of the LDA topic space. Each word w' contained in the training corpus is associated with a thematic confidence score δ. Finally, the hypothesis w' with the highest score δ is selected as shown in Figure 3.

4. Experiments

The proposed approach is based on a topic space learned with the LDA MALLET Java implementation[4]. This topic space contains 100 classes and the LDA hyper-parameters are chosen empirically as in [18] ($\alpha = \frac{50}{100} = 0.5$ and $\beta = 0.1$). During the learning process, the MALLET package requires to lowercase input text. For this reason, the results considered for the post-processing step are computed on lowercased sentences.

The effectiveness of the proposed approach is evaluated in the IWSLT benchmark. Table 2 reports case-sensitive BLEU and TER scores measured on the *test0*, *test13* and *test14* corpora, with two factored phrase-based TM model setups: a first one ($w \rightarrow w$) where only words are considered on the source and target sides, and a second one ($w \rightarrow (w, p)$) where English words are translated into (word, PoS) pairs. Disambiguating words with their PoS by the second factored model improves BLEU and TER over the first model for the three test corpora and both studied language pairs. For example, an absolute increase of BLEU (between 0.85 and 1.2) is observed for Slovene; a more limited but still consistent improvement of BLEU (between 0.1 and 0.5) happens for Polish.

Translation produced by the second TM models were used as entry of the LDA post-processing step. Table 3 shows results measured this time in terms of case-insensitive BLEU and TER, since sentences are lowercased before the post-processing step. The thresholds to consider a word as mistranslated from LM-based confidence scores were optimized in terms of BLEU on *test0*. These thresholds lead to change 1.2 % of words for Slovene and around 3 % for Polish (Table 3, columns 3 and 6). Unfortunately, using the proposed LDA-based approach did not translate into an observed gain in terms of BLEU or TER (line 1 vs line 2 and line 3 vs line 4).

[4] http://mallet.cs.umass.edu/

	TM MODELS	*test0*		*test13*		*test14*	
		BLEU	TER	BLEU	TER	BLEU	TER
English → Slovene	$w \rightarrow w$	12.27	69.58	13.20	67.70	10.92	69.66
	$w \rightarrow (w, p)$	13.35	68.64	14.05	66.32	12.16	68.59
English → Polish	$w \rightarrow w$	10.36	77.61	10.78	79.04	9.16	86.68
	$w \rightarrow (w, p)$	10.45	75.70	11.29	76.59	9.63	83.88

Table 2: Case-sensitive BLEU and TER (in %) measured to evaluate the use of a PoS factor inside the TM model.

	TM MODELS	*test0*			*test14*		
		BLEU	TER	% modified words	BLEU	TER	% modified words
English → Slovene	$w \rightarrow (w, p)$	13.68	67.78	-	12.69	67.90	-
	+ post-processing	13.42	68.03	1.16	12.23	68.17	1.29
English → Polish	$w \rightarrow (w, p)$	11.09	74.20	-	10.12	82.51	-
	+ post-processing	10.66	74.95	2.81	9.63	83.39	3.53

Table 3: Case-insensitive BLEU and TER (in %) measured before and after the LDA post-processing step.

5. Conclusions and Perspectives

In this paper, we propose an original post-processing approach to automatically correct translated texts. Our method takes advantage of a latent Dirichlet (LDA) model that provides thematically and grammatically close forms of mistranslated words. Experiments were conducted in the framework of the IWSLT machine translation evaluation campaign on the English-to-Polish and English-to-Slovene tasks. The proposed system was compared to a more classical factored translation system.

Results showed that the original proposed system does not improve results obtained with the baseline one, but we think that this preliminary work should lead to further investigations in the future. For example, we would like to use this model at an early stage, during the decoding process of the MT system, and not only at a post-processing stage. Furthermore, other features than n-gram probabilities should be exploited to identify mistranslated translations [21]. Finally, the low results observed with the topic-based correction approach are obtained with a topic space which still considers sentences as "bag-of-words" and ignore their internal grammatical structure. For this reason, a promising future work is to embed the position of the word in the sentence or n-gram containing the word.

6. References

[1] P. Koehn and H. Hoang, "Factored translation models," in *Proc. of EMNLP-CoNLL*, 2007, pp. 868–876.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. of ACL, Companion Volume*, 2007, pp. 177–180.

[4] M. Grčar, S. Krek, and K. Dobrovoljc, "Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik," in *Proc. of the 15th International Multiconference (IS)*, 2012, pp. 89–94.

[5] H. Schmid, "Improvements in part-of-speech tagging with an application to German," in *Proc. of the ACL SIGDAT Workshop*, 1995, pp. 47–50.

[6] A. Stolcke *et al.*, "SRILM—an extensible language modeling toolkit." in *Proc. of Interspeech*, 2002.

[7] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Proc. of the ACL Workshop: Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 2008, pp. 49–57.

[8] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of ACL*, vol. 1, 2003.

[9] S. Huet, E. Manishina, and F. Lefèvre, "Factored machine translation systems for Russian-English," in *Proc. of WMT*, 2013.

[10] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[11] J. R. Bellegarda, "A latent semantic analysis framework for large-span language modeling," in *Proc. of Eurospeech*, 1997.

[12] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*, 1999.

[13] G. Salton, "Automatic text processing: the transformation," *Analysis and Retrieval of Information by Computer*, 1989.

[14] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[15] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi, "Keyword extraction using term-domain interdependence for dictation of radio news," in *Proc. of Coling*, vol. 2. ACL, 1998, pp. 1272–1276.

[16] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[17] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proc. of the Conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.

[18] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[19] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.

[20] G. Heinrich, "Parameter estimation for text analysis," Fraunhofer IGD, Tech. Rep., 2009, version 2.9. [Online]. Available: http://www.arbylon.net/publications/text-est.pdf

[21] N. Bach, F. Huang, and Y. Al-Onaizan, "Goodness: A method for measuring machine translation confidence," in *Proc. of ACL*, 2011.