IDIAP RESEARCH REPORT

TOPIC AND SENTIMENT IN PHRASE-BASED STATISTICAL MACHINE TRANSLATION

Maryam Habibi Nikolaos Pappas
Andrei Popescu-Belis

Idiap-RR-10-2017

MARCH 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Topic and Sentiment in Phrase-Based Statistical Machine Translation

Maryam Habibi, Nikolaos Pappas and Andrei Popescu-Belis

Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
{mhabibi,npappas,apbelis}@idiap.ch

7 March 2017

Abstract

In this paper, we model two textual properties, topic and sentiment, at the sentence and document levels, with the goal of improving the performance of machine translation by taking into account this information in source and target sentences. In the topical similarity approach, we augment the source sentence with the keywords extracted from its adjacent sentences and re-rank the candidate target sentences (hypotheses of a phrase-based statistical MT system, here Moses) in terms of their topical similarity to the augmented source sentence. The advantage of the model is being independent from the baseline MT system, similarly to IR re-ranking techniques. We model sentiment using the sentiment categories of words and sentences as factors in the same MT system. The results on English-French MT show that topic modeling improves lexical choice with respect to the baseline for about 5% of the lexical items that differ between the two systems. We observe that although the improvement obtained using topical information is not significant in terms of BLEU score, there is an improvement on the choice of terms in the target language based on topical information. As for sentiment information, it leads to an increase in BLEU scores of up to 1.5%.

1 Introduction

State-of-the-art phrase-based or hierarchical statistical machine translation systems (SMT) do not use topic and sentiment information explicitly. Still, especially when entire documents are translated, sentence and document-level semantic properties such as topicality and sentiment are important aspects. Most current SMT systems rely on the implicit assumption that these aspects will be correctly rendered in translation as a result of local decisions. In this paper, we will show that there is

some benefit to be gained by introducing additional semantic features when translating multiple-sentence documents. Specifically, we will model the topic and sentiment properties of sentences and documents and show that valuable information can be extracted from them, in order to improve lexical choices in phrase-based SMT. Intuitively, the translated sentence should not be too far, in terms of topics and sentiment, from the source sentence and its neighboring ones. Topic and sentiment are only two of the document-level properties that contribute to coherence, but they are selected here for their complementarity.

In the study presented here, we will model the topical information from adjacent words, and infer sentiment information at the sentence level. Sentiment will be added as a 'feature' in the Moses phrase-based SMT system, as it is extracted from a sentence and cannot change when considering a larger context. However, we could not add topical information as Moses features, because it varies according to the size of the context that is considered, which would add more computational complexity for learning the model.

The paper is organized as follows. In Section 2 we present our approach for modeling topic and sentiment: for topics, we propose a novel approach using keywords that are extracted based on polylingual topic models, while for sentiment we use an existing lexicon-based analysis tool. Due to this difference, as explained in Section 3, constraints on MT are implemented for topics using re-ranking of translation hypotheses, while for sentiment we use factored models. Based on English-French data from news and TED talks (specified in Section 4), we present in Section 5 our results, showing that topic modeling improves the lexical choice in MT for ambiguous words, while sentiment modeling improves the overall BLEU score. Finally, Section 6 compares our work with previous studies.

2 Topic and Sentiment Models

In this section, we first present the topic modeling approach, and then the sentiment analysis tool.

2.1 Topic-Aware Keyword Extraction

We model topic coherence within the source text and between the source and target sentences using the following approach. We first represent words in both source and target languages using topical information obtained from a polylingual topic model (Mimno et al., 2009). We then extract, for each source sentence, a set of content keywords from the neighboring sentences in the source document, and weigh these keywords by their topic similarity to the source sentence, as explained below. These keywords serve to augment the source sentence with contextual topic information. Finally, as specified in Section 3 below, we re-score and re-rank candidate target sentences based on their topical similarity (according to the polylingual topic model) with the augmented source sentence.

2.1.1 Representing Words Using Polylingual Topic Models

Polylingual topic models are built upon Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and are used here to represent words based on the latent topics that are found. The method uses document pairs, i.e. pairs of source-language and target-language documents that are similar or equivalent to each other, but not necessarily exact translations; it is only supposed that they have the same distribution of topics. The method estimates the probabilities $p(z_T|w_T)$ and $p(z_S|w_S)$, which represent the distributions over a topic z of each word w in the target T and source S languages respectively.

We train a Bilingual LDA (BiLDA) topic model as defined by Mimno et al. (2009), with the implementation provided by Richardson et al. (2013).[1] We set the hyper–parameters to $\alpha = 50/k$ and $\beta = 0.01$ following Vulić et al. (2011), where k is the number of topics, set to 400 following Mimno et al. (2009). The BiLDA topic model is trained using Gibbs sampling with 1,000 iterations.

2.1.2 Keyword Extraction from Sentences

Given a sentence from a source document to translate, we first extract the set of content words C from M sentences before and M sentences after it within the same document. In the experiments presented in this paper, M=5. We apply a recent method for representative and diverse keyword extraction (Habibi and Popescu-Belis, 2013, 2015), which maximizes the coverage of the topics from the set of sentences by a fixed number keywords selected from list C.

Moreover, here, we weigh each representative keyword in proportion to its topical similarity to the source sentence. Specifically, we weigh each extracted keyword $c_i \in C$ with a weight w_i ($0 \leq w_i < 1$) given by the conditional probability of the keyword given the source sentence e, as formulated in the following equation:

$$w_i = p(c_i|e) = \sum_{z_S \in Z_S} p(c_i|z_S) \times p(z_S|e) \quad (1)$$

where $p(z_S|e)$ is the average distribution of topic z_S in the source sentence e, computed by averaging the topic values of all words of the sentence, and $p(c_i|z_S)$ is the topic-word distribution calculated using the topic model.

Finally, we augment the source sentence with the weighted keywords extracted from its context as follows:

$$e_{aug} = \{(e_1, 1), \cdots, (e_{|e|}, 1), (c_1, w_1), \cdots, (c_{|C|}, w_{|C|})\} \quad (2)$$

In other words, e_{aug} contains the words from the source sentence e to be translated with the weight 1, and the keywords from the context with a weight w_i as defined above.

[1] https://bitbucket.org/trickytoforget/polylda/

2.2 Sentiment Analysis

For incorporating sentiment information in the MT process, we extract the sentiment categories of words ($w \in W$) and sentences ($s \in S$) of the English source documents using the Pattern library (De Smedt and Daelemans, 2012), which is based on a lexicon of frequent polar adjectives.[2] Sentiment categories are defined as follows, based on the real-valued sentiment ($sent(w)$) computed by the library for word w: (i) negative (NEG) when $-1 \leq sent(w) < 0$; (ii) neutral (NEU) when $sent(w) = 0$; and (iii) positive (POS) when $0 < sent(w) \leq 1$. Similarly, we compute the sentiment categories of sentences by summing all values of $sent(w)$. If the library does not return a sentiment value for a given word, i.e. it does not carry any sentiment information, we use the empty category (NUL).

3 Integration with Statistical MT

Phrase-based SMT models (Koehn et al., 2003) are frequently used and still have close to state-of-the-art performance, despite recent successes of neural MT (Sennrich et al., 2016). In this paper, we use the Moses toolkit (Koehn et al., 2007) to build an English-French PBSMT system using the training data described hereafter. We take advantage of the following two features of Moses. First, for each sentence in the source language, Moses can provide a lattice of hypotheses with their probabilities, from which a ranked list of sentences with the N-best translation candidates can be derived. Second, Moses implements factored translation models (Koehn and Hoang, 2007), which are a principled way to use word-level linguistic labels in MT (originally morpho-syntactic ones). The factors result in additional feature functions that are used during decoding, combined in a log-linear way with weights that we tune here using MERT (Och, 2003).

3.1 MT with Keyword-based Topic Modeling

Considering the translation hypotheses obtained from the Moses SMT system for the source sentence e, we compute a score s_n for each target sentence hypothesis f_n as follows:

$$s_n = p(f_n|e) = \sum_{z_T \in Z_T, z_S \in Z_S} (p(f_n|z_T) \times p(z_T|z_S) \times p(z_S|e_{aug})) \quad (3)$$

In this equation, $p(z_T|z_S)$ is considered to be $1_{(z_T=z_S)}$, and $p(f_n|z_T)$ is the average distribution of topic z_T in relation to the target sentence f_n. We compute $p(z_S|e_{aug})$ as follows:

$$p(z_S|e_{aug}) = \frac{1}{|e| + \sum_{i=1}^{|C|} w_i} \times \left(\sum_{a \in e} p(z_S|a) + \sum_{i=1}^{|C|} w_i \cdot p(z_S|c_i) \right) \quad (4)$$

[2] http://www.clips.ua.ac.be/pages/pattern-en#sentiment

Factor	Sentence
Words	It can be a **very complicated** thing , the ocean .
Word-level sentiment	NUL NUL NUL NUL **NEG** **NEG** NUL NUL NUL NUL NUL
Sentence-level sentiment	NUL NUL NUL NUL **NEG** **NEG** NUL NUL NUL NUL NUL
Words	And those **simple** themes aren 't **really** themes about the **complex** science of what 's going on , but things that we all **pretty** well know .
Word-level sentiment	NUL NUL **NEU** NUL NUL NUL **POS** NUL NUL NUL **NEG** NUL NUL NUL NUL NUL NUL NUL NUL NUL NUL NUL NUL **POS** NUL NUL NUL
Sentence-level sentiment	NUL NUL **POS** NUL NUL NUL **POS** NUL NUL NUL **POS** NUL NUL NUL NUL NUL NUL NUL NUL NUL NUL NUL NUL **POS** NUL NUL NUL

Table 1: Examples of word-level and sentence-level factors for a negative and positive detected sentence.

Finally, the translation $\hat{f}$ for the source sentence e is computed by selecting the translation hypothesis with the highest s_n score:

$$\hat{f} = \arg\max_{f_n} s_n \tag{5}$$

3.2 Sentiment Labels as MT Factors

To provide sentiment information to MT, we use the above-mentioned factored translation models from Moses as they are particularly adapted to model word-level information such as polarity. As with topic models, we run translation experiments from English to French. Examples of word-level and sentence-level categories (used as factors) extracted from the English corpus are displayed in Table 1 for a negative and positive detected sentence.

4 Data, Setup and Evaluation Methods

4.1 Training Data

We used the European Parliament Proceedings Parallel Corpus (Europarl, 2011 release, Koehn (2005)) to train the bilingual topic models of BiLDA as well as the translation models for the English-French language pair and the French target language model. For learning the bilingual topic models we used document aligned texts, and for learning the translation model we utilized sentence aligned texts. Although we trained topic models using a parallel corpus, any comparable corpus can be used, such as Wikipedia articles.

We extracted five keywords from a total of five sentences before and five sentences after the source sentence, and augmented the source sentence with these

keywords. We obtained the N-best translation hypotheses for each source sentence from the Moses SMT system (version 2.1.1 released in March 2014), with N = 200. We re-ranked these hypotheses based on the method proposed above.

As our method only contributes to improving the semantics of the translation sentences, and not their morphological information, and also to avoid skipping words which do not fit the topic models, we selected for re-ranking only the sentences that have the same length as the 1-best translation obtained by Moses.

The sentiment-aware system was trained on English-French parallel data from WIT corpus[3]. For our evaluation we used the provided official development and test sets from 2010 and 2012, as well as the provided evaluation metric (BLEU score) to assess the quality of the translations. The tuning of all the factored models was performed on the 2010 development set which was the one provided in both years.

4.2 Evaluation Data and Metrics

For evaluation, we first consider the BLEU scores, by comparing the MT output against a reference translation in terms of precision and a brevity penalty, at the document level (Papineni et al., 2002). However, BLEU shows almost no variation across the 200-best hypotheses. Therefore, to assess more precisely the merits of our proposal, we performed subjective evaluation using one expert proficient in both French and English, as explained hereafter.

We used five news articles from the WMT 2013 test set[4] for the subjective evaluation. The documents are, respectively, about the following subjects: voting rights and ID documents in the USA; taking or not the test for prostate cancer; the discovery of Higgs' boson; palliative care institutions in Canada; and an interview about the Paris Saint-Germain football team. They contain respectively 47, 51, 42, 66, 38 and 99 sentences in the English sources and the aligned reference translations into French.

For evaluation, the expert looked at the sentences obtained by the three SMT systems presented below, considering only the triples where at least one sentence differed from the others. The expert examined only the content (not the morphosyntactic inflection) of the words which are different across the sentences, by comparing them with the words from the reference sentence and with those from the source sentence. While in most cases the differences are observed on 1:1 alignments, we also considered 1:n or n:1 alignments, which can fall under the second and third following cases. The outcome of the evaluation for each version is coded as follows:

1. If the word is identical to the reference word, then it obtains a score of 2.

2. If the word(s) are correct but not identical to the reference, then it obtains 1.

[3] `https://wit3.fbk.eu/`

[4] `http://www.statmt.org/`

	USA politics	Cancer test	Higgs' boson	Palliative care	PSG football team	Average
Sentences	47	51	42	66	99	343
Changed:						
sentences	27	30	19	33	53	178
words	31	32	25	41	66	213
EC > M	43% (3)	67% (4)	22% (2)	50% (5)	52% (13)	46%
EC < M	57% (4)	33% (2)	78% (7)	50% (5)	48% (12)	54%
KC > M	50% (3)	67% (4)	40% (2)	50% (5)	50% (11)	52%
KC < M	50% (3)	33% (2)	60% (3)	50% (5)	50% (11)	48%
KC > EC	57% (4)	50% (4)	100% (3)	62% (5)	50% (12)	64%
KC < EC	43% (3)	50% (4)	0% (0)	38% (3)	50% (12)	36%

Table 2: Comparison scores obtained using subjective evaluation over five different documents from test set. The compared methods are noted KC (the proposed method), M (the Moses baseline), and EC (method using all words of adjacent sentences as context). The numbers show the number of times one system is correct while the other is wrong, and the proportions for the opposite comparisons (e.g. EC>M and EC<M), which sum up to 100%. The results indicate that EC < M < KC.

3. Otherwise the words receive a 0 score.

This approach rewards the most the translations that are identical to the reference, and with half the score those that are considered to be correct but differ from the reference. One option to compute a total score is to simply sum up all individual scores. Another option is to group the values of '2' and '1', and count how many non-zero values were assigned by the expert. Moreover, we favor a comparative approach: we count for each system, in comparison to another one, the number of sentences which are correct (score of 2 or 1) when the corresponding sentence from the other system is wrong (score of 0). We compare these two numbers as a proportion of their total, noted as $S\%$ and $(1-S)\%$. This allows us to assess the improvement brought by a method in comparison to another one.

5 Experimental Results

5.1 MT with Topic Modeling

To assess the merits of topic modeling using keyword-based context representation, we compare using the method described above, i.e. in pairs, three SMT systems:

1. The proposed re-ranking approach, noted KC, for *Keyword-based Context representation*.

Example 1: KC outperforms M	
Source sentence	When, in fact, a particle having an electric **charge** accelerates or changes direction, ...
Reference sentence	Quand , en effet , une particule ayant une **charge** électrique accélère ou change de direction , ...
M translation	Lorsque , en fait , une **accusation** électriques particules avoir une légère modification direction , ...
KC translation	Lorsque , en fait , une **charge** électriques particules avoir une légère modification direction , ...

Example 2: KC outperforms EC	
Source sentence	The new election laws **require** voters to show a photo ID card and proof of US citizenship.
Reference sentence	Les nouvelles lois électorales **exigent** que les électeurs présentent une carte, d'identité avec photo et une preuve de citoyenneté américaine .
EC translation	Les nouvelles lois électorales **ont besoin** d'électeurs de montrer une photo de carte d'identité et la preuve de la citoyenneté américaine.
KC translation	Les nouvelles lois électorales **exiger** que les électeurs de montrer une photo carte d'identité et la preuve de la citoyenneté américaine .

Table 3: Two examples of machine translation results: example (1) shows our method (KC) outperforming the Moses baseline (M), and example (2) shows the superiority of our method (KC) over the method using all words as a context (EC).

2. An alternative re-ranking approach similar to the proposed one, but which augments the source sentence using all words in the five sentences before and after the source sentence instead of using keywords only. This method is noted EC for *Entire Context*.

3. The baseline 1-best translation obtained directly from the *Moses system*, noted M.

We compare our method, KC, with the Moses baseline M, but also with EC, in order to study the contribution brought by keyword extraction in comparison to an unfiltered use of context. First, we consider the BLEU scores, which are very similar across all the N-best hypotheses. Still, in terms of BLEU, the baseline system appears to outperform the others, with a score of 27.30, while KC reaches 26.19 and EC 26.18. As Moses is trained and tuned to optimize BLEU, it is perhaps of no surprise that a re-ranking of the 200-best hypotheses tends to slightly decrease BLEU.

The results of the human comparative evaluations are provided for each of the five test documents in Table 2. The numbers show the number of times one system

gets a correct translation (scored 2 or 1) while the other gets it wrong (scored 0). The proportions for opposite comparisons such as "EC > M" versus "EC < M" sum up to 100% as only different translations are counted. The comparison excludes the first and last five sentences of the documents, as topic modeling is less reliable.

The average comparison values are: 52% for KC vs. 48% for M; 46% for EC vs. 54% for M; and 36% for EC vs. 64% for KC. These results indicate the following ranking: EC < M < KC, though the differences between the systems are actually very small. The ranking shows that words from minor topics added by EC from the adjacent sentences appear to degrade the results of SMT, while relevant keywords selected from the context by our method have the potential to improve the translation output.

There are a few cases in which the scores assigned to the sentences are zero for all three compared methods, i.e. all systems are wrong. In these cases, we examined the 200 best candidate translations and found out that among them there were no better translations to be selected. While it is possible that a better translation could be found below the 200 best ones, it is also likely that in many cases the translation model did not learn an appropriate phrase pair to use.

We provide two examples of results in Table 3. In the first example, KC outperforms M. In this example, the English word "charge" from the source sentence has two possible translations in French, corresponding to the different English meanings in "electric charge" versus "criminal charge" (a formal accusation). The correct translation in this example is by the French word "charge", and this is indeed correctly selected by our method when re-ranking the N-best list. In the second example, we compare the translation results of KC and EC. In this example, the word "require" in the English source should be translated into "exigent" in French, which is the third person plural of transitive verb "exiger". However, the EC method translated it into "ont besoin" which means "need", which has a similar meaning but reversing the agent and the patient, hence it cannot be used here. The KC method translated it into "exiger", which is the correct translation, but not with the correct mode/number/person.

5.2 MT with Sentiment Models

For sentiment-aware MT, we compare the following systems. They all make use of factored models, but differ in the factors used on the source vs. target side:

1. *POS(target only)*: two factors on the source side and one factor on the target side (word + part-of-speech → word).

2. *POS (target + source)*: two factors on the source and target language sides (word + part-of-speech → word + part-of-speech).

3. *SEN(word-level)*: two factors on the source side and one factor on the target side (word + word-sentiment → word).

Model / Test set	2010	2012
POS(target only)	29.92	35.24
POS (target + source)	30.20	36.20
SEN(word)	30.31	36.15
SEN(word + sentence)	**30.42**	**36.69**

Table 4: Performance of factored models on the 2010 and 2012 test sets of English-French parallel data from WIT corpus in terms of BLEU score. The best scores are marked in **bold**.

4. *SEN(word + sentence)*: three factors on the source side and one factor on the target side (word + word-sentiment + sentence-sentiment $\rightarrow$ word).

The performance of the above models in terms of BLEU score on the 2010 and 2012 test sets are displayed in Table 4. Overall, from the results we observe that the factored models which incorporate sentiment information perform better than the ones that incorporate part-of-speech information, which suggests that sentiment factors bring additional useful information for translation that is not captured by the other factored models.

The best performance among the examined factored models is the *SEN(word + sentence)* model: it achieved 0.5 and 1.4 higher BLEU scores than the best factored model which uses part-of-speech information on the 2010 and 2012 test sets respectively; this result can be clearly attributed to the sentiment factors used by the SEN model. Furthermore, the *SEN(word + sentence)* model performs better than the *SEN(word)* model because it leverages both word-level and sentence-level factors as opposed to the one that uses word-level factors only.

The *POS(target + source)* model, which takes into account part-of-speech information in both target and source languages, performs better than the *POS(target only)* model. Therefore, we hypothesize that a model which would capture sentiment information in both source and target languages would perform even better than our *SEN(word + sentence)* model, but we could not test the hypothesis from lack of a French sentiment detector. Such variants of the *SEN* factored model should be investigated in the future, including also a more fine-grained definition of sentiment (e.g., very positive, positive, neutral, negative, very negative).

To investigate the effect of sentiment information on unfactored translation models, we ran an experiment with such a model by training Moses on words concatenated with the word-level sentiment categories, e.g., happy_POS for 'positive'. The performance achieved by this unfactored model was 30.77 and 36.75 BLEU score on the 2010 and 2012 test sets respectively, which outperformed the WIT official baseline scores for the 2010 and 2012 test sets by 1.32 and 1.86 points respectively. However, when we trained our own version of the unfactored model without the concatenation with sentiment categories, the difference was much smaller – which may be related to a different version of Moses used by us and the WIT orga-

nizers. This experiment indicates that incorporating sentiment information simply with concatenated labels in unfactored translation models does not clearly improve the translation results in the examined datasets.

6 Related Work

Phrase-based statistical MT (Zens et al., 2002; Koehn et al., 2003) does not capture document-level constraints on the meanings of the words across sentences of a coherent text, although this would help to ensure that the translated document preserves the coherence of the source one. The use of discourse-level information to improve MT has been recently attempted in a variety of approaches, as surveyed in three recent PhD theses (Hardmeier, 2014; Meyer, 2014; Guillou, 2016). The discourse-level phenomena studied in relation to MT cover discourse connectives (Meyer and Popescu-Belis, 2012; Meyer et al., 2015), verb tenses (Meyer et al., 2013; Loaiciga et al., 2014), and more recently pronouns and noun phrases (as reviewed by Luong et al. (2017) or Pu et al. (2017)), including shared tasks on pronoun translation or prediction (Hardmeier et al., 2015; Guillou et al., 2016).

Representing topical distance across languages and implementing such a constraint in an SMT system remains, to a large extent, an open problem. Several studies have addressed the issue of document-level topic coherence by representing words or phrases using monolingual topic models (Su et al., 2012), or multilingual topic models obtained from parallel or comparable corpora aligned at the document level (Zhao and Xing, 2008; Tam et al., 2007). The latter approach estimates the topical similarity between the topics of a target sentence and the entire source document.

Another idea is to perform domain adaptation at the document level (Sennrich, 2013) or dynamically at the sentence level using topic models (Eidelman et al., 2012). The latter study also compared domain-adaptation performance using sentence vs. document-level topics. Similarly, Hasler et al. (2014) investigated the combination of local and global topics, and attempted to model the evolution of topics throughout a document. Instead of topics, Ture et al. (2012) encouraged consistency for Arabic-English MT by simply introducing cross-sentence consistency features in the translation model, building upon the hypothesis of one translation per discourse (Gale et al., 1992; Carpuat, 2009). The best results on translation domain adaptation with topic modeling have been obtained by Hu et al. (2014), to our best knowledge.

The work of Xiao et al. (2012) assumes that all the sentences in a document share the same topic at the entire document, and uses this topic to constrain the meanings of words in each sentence. Gong et al. (2011) re-ranked the target candidate sentences obtained by a phrase-based SMT by scoring them based on their topical similarity with the entire source document, using a multilingual topic model extending LSA (Tam et al., 2007).

However, these methods are limited by the coarse granularity of the document-

level topics, which do not set precise constraints on sentence-level topics, which may differ from the former. Xiong and Zhang (2013) have indeed found that around 40% of sentences in their data (NIST MT 2003 and 2005 datasets) had topics that were different from those of their document. Consequently, Xiong and Zhang replaced the document topics with the topics of neighboring sentences when translating a given sentence. However, modeling these topics from parallel corpora was not accurate enough given their somewhat limited size for this task.

Comparable corpora, i.e. documents in two languages with similar content, though not exact translations of each other, have appeared to provide a solution for robust multilingual topic modeling. Several studies have used comparable corpora aligned at document level (Ni et al., 2009; Mimno et al., 2009) to identify and constrain the potential translations of specific words in a document, but not for translating entire sentences, because phrase-level information was not combined with them. For instance, Vulić et al. (2011) ranked the potential word candidates in the target language directly based on their topical similarity with words in the source language. In subsequent studies, Vulić and Moens augmented each word using the semantic information from the entire source and target vocabulary (Vulić and Moens, 2013a), and the contextual information defined by the co-occurrence words in a predefined context window (Vulić and Moens, 2013b), and then measured the similarity of words based on these semantic contextual information.

Few articles have aimed at improving MT with sentiment analysis – unlike the reverse task, i.e. using MT to perform multilingual sentiment analysis (Balahur and Turchi, 2014). In one of the few works known to us, Pal et al. (2014) showed that aligning sentiment phrases, and post-editing the output of Moses to align sentiment holders, sentiment words, and their objects, improved English-Bengali phrase-based SMT, Bengali being an under-resourced language.

7 Conclusion

In this paper, we have proposed two models of the context of a source sentence in terms of topic and sentiment. To preserve topics in translation, we extract keywords from adjacent sentences, and then re-rank the translation hypotheses from Moses by scoring their multilingual topic similarity with the source sentence augmented with the keywords. This outperforms a Moses baseline by about 5% in the cases where a difference was observed, and outperforms by a larger margin a re-ranking method which does not rely on keywords. To preserve sentiment, we encode polarity as source-side factors, and achieve a small but measurable increase in BLEU scores (0.22–1.45).

Unlike previous methods which require document-aligned parallel corpora to utilize contextual information for MT, our modeling of topics can use comparable corpora as well. Our method uses sentence-aligned parallel corpora for training the translation model, as usual, but needs only document-aligned comparable corpora for learning the multilingual topic models.

However, we observed that there is sometimes no proper translation candidate for a source sentence in the list of N-best hypotheses obtained from the Moses SMT system. In the future, we plan to integrate the topical and sentiment information with the translation and language model by directly adding the information obtained from the adjacent sentences of a source sentence as new features, instead of using them for re-ranking only.

Acknowledgments

We are grateful for their support to the Swiss National Science Foundation (SNSF) under the Sinergia MODERN project (`www.idiap.ch/project/modern/`, grant n. 147653) and to the European Union under the Horizon 2020 SUMMA project (`www.summa-project.eu`, grant n. 688139).

References

Balahur, A. and Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Carpuat, M. (2009). One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27, Boulder, CO, USA.

De Smedt, T. and Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13:2031–2035.

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

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237.

Gong, Z., Zhou, G., and Li, L. (2011). Improve SMT with source-side topic-document distributions. In *Proceedings of the MT Summit*, pages 496–501.

Guillou, L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, University of Edinburgh, UK.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016

WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin, Germany.

Habibi, M. and Popescu-Belis, A. (2013). Diverse keyword extraction from conversations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 651–657, Sofia, Bulgaria.

Habibi, M. and Popescu-Belis, A. (2015). Keyword extraction and clustering for document recommendation in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):746–759.

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Sweden.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Hasler, E., Haddow, B., and Koehn, P. (2014). Combining domain and topic adaptation for SMT. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1, pages 139–151.

Hu, Y., Zhai, K., Eidelman, V., and Boyd-Graber, J. (2014). Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1166–1176, Baltimore, Maryland.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, pages 868–876, Prague, Czech Republic.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54.

Loaiciga, S., Meyer, T., and Popescu-Belis, A. (2014). English-French verb phrase alignment in Europarl for tense translation modeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Luong, N. Q., Popescu-Belis, A., Rios Gonzalez, A., and Tuggener, D. (2017). Machine translation of Spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain.

Meyer, T. (2014). *Discourse-level Features for Statistical Machine Translation*. PhD thesis, EPFL, Lausanne.

Meyer, T., Grisot, C., and Popescu-Belis, A. (2013). Detecting narrativity to improve English to French translation of simple past verbs. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 33–42, Sofia, Bulgaria.

Meyer, T., Hajlaoui, N., and Popescu-Belis, A. (2015). Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(7):1184–1197.

Meyer, T. and Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 129–138, Avignon, France.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889.

Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2009). Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International Conference on the World Wide Web*, pages 1155–1156.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Pal, S., Gopal Patra, B., Das, D., Kumar Naskar, S., Bandyopadhyay, S., and van Genabith, J. (2014). How sentiment analysis can help machine translation. In *Proceedings of the 11th International Conference on Natural Language Processing (ICON)*, Goa, India.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Pu, X., Mascarell, L., and Popescu-Belis (2017). Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain.

Richardson, J., Nakazawa, T., and Kurohashi, S. (2013). Robust transliteration mining from comparable corpora with bilingual topic models. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 261–269, Nagoya, Japan.

Sennrich, R. (2013). *Domain adaptation for translation models in statistical machine translation*. PhD thesis, University of Zurich, Faculty of Arts.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Su, J., Wu, H., Wang, H., Chen, Y., Shi, X., Dong, H., and Liu, Q. (2012). Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 459–468.

Tam, Y.-C., Lane, I., and Schultz, T. (2007). Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.

Ture, F., Oard, D. W., and Resnik, P. (2012). Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 417–426, Montréal, Canada.

Vulić, I., De Smet, W., and Moens, M.-F. (2011). Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 479–484.

Vulić, I. and Moens, M.-F. (2013a). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 106–116.

Vulić, I. and Moens, M.-F. (2013b). A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1613–1624.

Xiao, X., Xiong, D., Zhang, M., Liu, Q., and Lin, S. (2012). A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 750–758.

Xiong, D. and Zhang, M. (2013). A topic-based coherence model for statistical machine translation. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence*, pages 18–32. LNCS 2479, Springer-Verlag.

Zhao, B. and Xing, E. P. (2008). HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems*, pages 1689–1696.