Novel Document Level Features for Statistical Machine Translation

Rong Zhang and Abraham Ittycheriah
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{zhangr,abei}@us.ibm.com

Abstract

In this paper, we introduce document level features that capture necessary information to help MT system perform better word sense disambiguation in the translation process. We describe enhancements to a Maximum Entropy based translation model, utilizing long distance contextual features identified from the span of entire document and from both source and target sides, to improve the likelihood of the correct translation for words with multiple meanings, and to improve the consistency of the translation output in a document setting. The proposed features have been observed to achieve substantial improvement of MT performance on a variety of standard test sets in terms of *TER/BLEU* score.

1 Introduction

Most statistical machine translation (MT) systems use sentence as the processing unit for both training and decoding. This strategy, mainly the result of pursuing efficiency, assumes that each sentence is independent, and therefore suffers the loss of missing many kinds of "global" information, such as domain, topic and inter-sentence dependency, which are particularly important for word sense disambiguation (Chan et al., 2007) and need be learned from the span of entire document.

Table 1 shows the MT output of our sentence level Arabic-to-English translation engine on two sentences excerpted from a news article discussing middle-east politics. The Arabic sentences are displayed in Romanized form. The Arabic word *mrsy* denotes the name of the former Egyptian president *Morsi* in both sentences. In the first sentence it is translated together with prior word *mHmd*(Mohamed) as a phrase and mapped to the name correctly. In the second sentence, where no relevant local context is present, it is incorrectly translated into the word *thank*, which is the most frequent English word aligned to *mrsy* in our training data. This example shows that for ambiguous words like *mrsy*, utilizing only local features is insufficient to find them the correct translation hypotheses. This example also illustrates another weakness of sentence level MT. It has been observed that a word tends to keep same meaning within one document (Gale et al., 1992; Carpuat , 2009). However, such consistency can't be maintained by MT system working on isolated sentences since all decisions are made locally.

AR:	Alr}ys AlmSry AlmEzwl mHmd mrsy ysf nfsh b⟩nh r}ys Aljmhwryp
MT:	The deposed Egyptian president Mohamed Morsi describes himself as the president of the republic
AR:	mrsy ytHdY AlqADy fy mHAkmth bthmp Alhrwb mn Alsjn
MT:	Thank you defy the judge in his trial on charges of escaping from prison

Table 1: Sentence level MT results of two sentences excerpted from same document

To address these issues, this paper investigates document level features to utilize useful information from wider context. Three types of document level features, including source and target side long distance context, and "quasi-topic", are integrated into our MT system via the framework of Maximum Entropy, and lead to substantial improvement of translation performance.

2 A Practical Scheme to Approximate Document Level Machine Translation

Let D_f denote a document in source language f consisting of N sentences: $D_f =< \mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_N >$. The goal of document level MT is to search the best document hypothesis D_e^* in target language e that maximizes the translation probability:

$$D_e^* = \arg\max_{D_e} \Pr(D_e|D_f) \quad (1)$$

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 153–157,
Lisbon, Portugal, 17 September 2015.

We require that the number of sentences in D_f and D_e to be equal: $D_e =< \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N >$. Using chain rule, $\Pr(D_e|D_f)$ is estimated as follows:

$$\Pr(D_e|D_f) = \prod_{i=1}^{N} \Pr(\mathbf{e}_i|\mathbf{f}_i, D_{f,\bar{i}}, D_{e,i-}) \quad (2)$$

where $D_{f,\bar{i}}$ denotes the source document excluding the current sentence $\mathbf{f}_i$, and $D_{e,i-} =< \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{i-1} >$ which is the MT output up to previous sentence. If we keep the i.i.d. assumption of sentence generation that $D_{f,\bar{i}}$ and $D_{e,i-}$ are irrelevant to $< \mathbf{f}_i, \mathbf{e}_i >$, Eq. (2) backs off to standard sentence level translation that

$$\Pr(D_e|D_f) = \prod_{i=1}^{N} \Pr(\mathbf{e}_i|\mathbf{f}_i) \quad (3)$$

In our document level MT experiments, the estimate of $\Pr(\mathbf{e}_i|\mathbf{f}_i, D_{f,\bar{i}}, D_{e,i-})$ is divided into three separate modules combined by a normalization function:

$$\Pr(\mathbf{e}_i|\mathbf{f}_i, D_{f,\bar{i}}, D_{e,i-}) = \Theta\{\Pr(\mathbf{e}_i|\mathbf{f}_i), \Pr(\mathbf{e}_i|D_{f,\bar{i}}), \Pr(\mathbf{e}_i|\mathbf{f}_i, D_{e,i-})\} \quad (4)$$

Eq. (4) provides a scheme to integrate document level context features into translation process. In (4), $\Pr(\mathbf{e}_i|\mathbf{f}_i)$ is the standard sentence level translation model, $\Pr(\mathbf{e}_i|D_{f,\bar{i}})$ models how source side long distance features, e.g. feature regarding document topic or trigger word not in current sentence, impact the generation of $\mathbf{e}_i$, and $\Pr(\mathbf{e}_i|\mathbf{f}_i, D_{e,i-})$ can be viewed as a module exploring target side cross-sentence dependency between $\mathbf{e}_i$ and $D_{e,i-}$ to maintain translation consistency. $\Theta(.)$ is a normalized combination function that incorporates the three modules together to generate a probabilistic estimate for each hypothesis. Please note that in consideration of decoding speed, the proposed scheme does not search optimal hypothesis from document space directly, but rather enhance sentence translation by utilizing "global" information not limited to current sentence $\mathbf{f}_i$.

3 Document Level Context Features

The MT system adopted in our experiments is a direct translation model that utilizes the framework of Maximum Entropy to combine multiple types of lexical and syntactic features into translation (Ittycheriah and Roukos , 2007). The model has the following form:

$$\Pr(\mathbf{t}, j|\mathbf{s}) = \frac{\Pr_0(\mathbf{t}, j|\mathbf{s})}{Z} \exp \sum_k \alpha_k \phi_k(\mathbf{s}, \mathbf{t}) \quad (5)$$

where $\mathbf{s}$ is a source side word or phrase, $\mathbf{t}$ is the corresponding word or phrase translation, j is the transition distance from last translated word, $\Pr_0$ is a prior distribution related to phrase to phrase translation model and distortion model, and Z is a normalizing term. In Eq. (5), feature $\phi_k(\mathbf{s}, \mathbf{t})$ can be viewed as a binary question regarding lexical and syntactic attributes of $\mathbf{s}$ and $\mathbf{t}$, e.g. the question can be asked as if $\mathbf{s}$ and $\mathbf{t}$ share same POS class. Weight α_k is estimated using Iterative Scaling algorithm. Testing results from many evaluation tasks have shown that the MaxEnt system performs significantly better than regular phrase system and equally well to hierarchical system.

This section introduces three new types of document level features to model $\Pr(\mathbf{e}_i|D_{f,\bar{i}})$ and $\Pr(\mathbf{e}_i|\mathbf{f}_i, D_{e,i-})$. All the three types of features can be expressed as a triplet that $\phi_k(\mathbf{s}, \mathbf{t}) =< \mathbf{s}, c, \mathbf{t} >$, where c denotes a source or target side context word, identified from the span of entire document, which works as a *bridge* to connect $\mathbf{s}$ and $\mathbf{t}$. Please note that ϕ_k is still a binary feature which indicates if a particular context word c of certain type exists for $\mathbf{s}$ and $\mathbf{t}$.

3.1 Source Side Long Distance Context Feature

The first type of document level feature is motivated by the example shown in Table 1. The ambiguous Arabic word *mrsy* in the second sentence is mistranslated to English word *thank* because there is no local evidence to suggest it is a person name rather than a verb which is more common in training data and thus has higher translation probability in prior phrase model $\Pr_0$. In this case, if the words co-occurring with *mrsy* in the first sentence, i.e. *mHmd*(Mohamed), can be identified and passed to subsequent sentences, the probability of *mrsy* in the same document being translated into *Morsi* is likely to be increased.

ϕ_{LDC}, the long distance context (LDC) feature, is implemented as follows in training stage. Suppose the questioned source word w_f occur in sentence *i* with translation w_e. To identify the relevant LDC word c_f, the entire document excluding current sentence i is analyzed to find if the alignment (w_f, w_e) also occurs in other sentences. If yes, the source words within a window centered by w_f at that place are collected as the candidates for c_f. For instance, if the two Arabic sentences of Table 1 are in training data, the words

mHmd(Mohamed) and *AlmSry*(Egyptian) in the first sentence will be viewed as the LDC word for *mrsy* in the second sentence, which results in two ϕ_{LDC} features i.e. $<$ *mrsy*, $mHmd$, *Morsi* $>$ and $<$ *mrsy*, $AlmSry$, *Morsi* $>$. As illustrated in Eq. (5), the two features can boost the translation probability of $\Pr(\textit{Morsi}|\textit{mrsy})$ for entire document if their weights are properly learned.

In training stage the check of aligned target word w_e is to ensure that only words with same meaning can be grouped together to share context features. In decoding where true w_e is unknown, we only use w_f instead of (w_f, w_e) to identify LDC word c_f. In our experiment function words are not allowed to be c_f, and $tf * idf$ score is used to filter out irrelevant context word.

3.2 Target Side Long Distance Context Feature

In order to improve the consistency of word choice in hypothesis generation, ϕ_{LDC} can be extended to target side to utilize the correlation between $D_{e,i-}$, the translation up to previous sentences, and $\mathbf{e}_i$, the translation of current sentence.

In training stage, ϕ_{tLDC}, the target side long distance context (tLDC) feature, is implemented in the following way. For a questioned source word w_f which is aligned to target word w_e in sentence i, we search their occurrence in all previous sentences from 1 to i-1. If exists, the *target* side words within the window centered by w_e in that sentence are identified as the candidates of tLDC word c_e for w_f. For the example used before, the English side words *Mohamed* and *president* are expected to make ϕ_{tLDC} features for the word *mrsy* in the second sentence.

The feature in decoding stage is implemented similarly by remembering previous translation $D_{e,i-}$ and its alignment to source words. Please note we don't use the hypothesized translation w_e itself as the tLDC word for w_f. This is because if it is an incorrect translation, such error can be spread to subsequent sentences to cause duplicated errors.

3.3 LSA based Quasi-Topic Feature

LDC and tLDC features are effective for repeated words. For words occurring once in a document, quasi-topic (QT) feature ϕ_{QT} is proposed as a back-off model which utilizes underlying topic information to eliminate ambiguity for these words.

In training stage, Latent Semantic Analysis (LSA) is performed on bilingual corpus consisting of a large set of documents with parallel sentences. Both source and target side words are mapped to vectors locating in a unified high dimensional space. For a questioned word w_f, its QT feature words are selected as follows. First all source side content words in the same document are calculated $tf * idf$ score, and sorted by their values from high to low. The top *L* words are then collected as the indicators of the underlying document topic. Next semantic similarity is measured between w_f and each of the *L* candidates based on Cosine metric. Only words showing strong correlation are selected as the QT feature c_t for w_f.

In decoding stage, MaxEnt model, as shown in Eq. (5), is utilized to estimate the probability of a hypothesis w_e being generated from w_f and QT feature words c_t. Our preliminary experiments found MaxEnt model performs better than commonly used vector based similarity metric. Generally speaking the QT features provide an implicit way for topic adaptation. When applied to translation, it changes the lexical distribution of target words to prefer the one more relevant to the hidden topic represented by c_t.

4 Related Work

Recent years witness a growing interest in exploiting long distance dependency to improve MT performance (Wong and Kit , 2012; Hardmeier et al., 2013). Domain adaptation and topic adaptation have attracted considerable attentions (Eidelman et al., 2012; Chen et al., 2013; Hewavitharana et al., 2013; Xiong and Zhang, 2013a; Hasler et al., 2014). There are also efforts that explore lexical cohesions with the help of WordNet to describe semantic co-occurrence from document span (Ben et al., 2013; Xiong et al., 2013b). Translation consistency, related to the observation of *one sense per discourse* (Gale et al., 1992; Carpuat , 2009; Guillou , 2013), has been discussed recently as an additional metric to evaluate translation quality (Xiao et al., 2011; Ture et al., 2012). There are also efforts for Arabic proper name disambiguation (Hermjakob et al., 2008). This paper investigates novel document level features to utilize lexical and semantic dependencies between sentences. In contrast to (Ben et al., 2013; Xiong et al., 2013b), our work doesn't need external resources e.g. WordNet or human efforts to identify word cohesion and isn't limited to certain word type. The advantage makes the proposed

Model	MT03		MT04		MT05		MT06		MT08		MT09	
Baseline	39.19	57.01	37.15	56.12	35.46	58.71	41.16	51.79	42.60	50.62	42.27	51.53
+LDC	38.99	57.58	37.19	56.43	35.41	59.20	41.29	52.27	42.45	**51.45**	42.18	52.02
+tLDC	39.02	57.32	37.16	56.41	35.32	58.91	41.12	52.03	42.56	51.14	42.16	51.77
+QT	39.10	57.16	37.03	56.23	35.37	58.79	41.14	51.80	42.52	50.75	42.22	51.68
+LDC+tLDC	**38.46**	**57.83**	36.83	56.50	34.97	59.64	40.91	52.23	42.25	51.26	41.92	52.26
+L+tL+QT	38.54	57.73	**36.73**	**56.75**	**34.84**	**59.75**	**40.68**	**52.40**	**42.05**	51.42	**41.82**	**52.47**

Table 2: MT performance on MT03-MT09 in terms of TER and BLEU.

features more suited to low-resource languages.

5 Experiments

Our system is primarily built for an Arabic dialect to English MT task. The training data contains LDC-released parallel corpora for the BOLT project. There are totally 6.9M sentence pairs with 207M Arabic ATB tokens and 201M English words, respectively. Three types of word alignments, maximum entropy, GIZA++ and HMM alignment, are used to generate phrase pairs as the prior model in Eq. (5). Approximately 1.8M in-domain sentence pairs distributed in 106K documents, consisting of 36M Arabic ATB tokens and 38M English words, are selected to learn sentence and document level MaxEnt features. The tuning set contains 3700 sentences in 350 documents which are mainly weblog and dialect data. Module weights for prior model, sentence and document level features, LM, and other components are tuned with PRO algorithm (Hopkins and May , 2011) to minimize the score of *(TER-BLEU)*.

We select NIST Arabic MT03-MT09 as the test sets. Results are shown in Table 2. The two numbers in each score column are TER followed by BLEU. The best performance is illustrated in bold. The result of MT system using only sentence level features is listed as the baseline. The integrations of the three document features are denoted as +LDC, +tLDC and +QT, respectively. Table 2 shows that substantial improvements of translation quality, measured by both TER and BLEU, are achieved for most of the test sets.

To understand the effectiveness of document features on different type of data, we further split MT09 set into newswire and weblog, and conduct test on them. Table 3 shows that long distance context features, ϕ_{LDC} and ϕ_{tLDC}, perform better on newswire than on weblog respecting to the relative improvement of TER and BLEU. One reason to explain this is that the rate of content word repetition is different on the two types of data. According to our calculation, about 19% content words in newswire repeat themselves while the ratio on weblog is about 13%.

Model	MT09-nw		MT09-wl	
Baseline	33.85	61.15	50.46	41.35
+LDC+tLDC	33.38	61.98	50.23	42.02

Table 3: MT performance on MT09 newswire and weblog in terms of TER and BLEU.

Table 4 shows the new MT output of the two example sentences. Three LDC features are fired for *mrsy* in the 2nd sentence: ⟨*mrsy,AlmSry,Morsi*⟩, ⟨*mrsy,mHmd,Morsi*⟩ and ⟨*mrsy,ySf,Morsi*⟩ where the 3rd one is a false alarm. Items in the triplets correspond to source word, context word and hypothesized word, respectively. Three tLDC features are also fired including ⟨*mrsy,Egyptian,Morsi*⟩, ⟨*mrsy,Mohamed,Morsi*⟩ and ⟨*mrsy,describes,Morsi*⟩ where the 3rd one is also a false alarm. To our surprise, word *Alr}ys* and its translation *president* aren't fired as context feature. Analysis found that this is due to the fact that our LDC training data was collected before Dr. Morsi was elected as president in 2012. Therefore no relevant feature is learned into MaxEnt model.

AR:	Alr}ys AlmSry AlmEzwl mHmd mrsy ysf nfsh b⟩nh r}ys Aljmhwryp
MT:	The deposed Egyptian president Mohamed Morsi describes himself as the president of the republic
AR:	mrsy ytHdY AlqADy fy mHAkmth bthmp Alhrwb mn Alsjn
MT:	Morsi defies the judge in his trial on charges of escaping from prison

Table 4: New MT results using document level features

References

Guosheng Ben, Deyi Xiong, Zhiyang Teng, Yajuan Lu and Qun Liu. 2013. Bilingual Lexical Cohesion Trigger Model for Document-Level Machine Translation. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria.

Marine Carpuat. 2009. One Translation per Discourse. in *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Boulder, USA.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech.

Boxing Chen, Roland Kuhn and George Foster. 2013. Vector Space Model for Adaptation in Statistical Machine Translation. in *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria.

Vladimir Eidelman, Jordan Boyd-Graber and Philip Resnik. 2012. Topic Models for Dynamic Translation Model Adaptation. in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. in *Proceedings of the workshop on Speech and Natural Language, HLT-91*.

Liane Guillou. 2013. Analysing Lexical Consistency in Translation. in *ACL Proceedings of the Workshop on Discourse in Machine Translation*.

Christian Hardmeier, Sara Stymne, Jrg Tiedemann and Joakim Nivre. 2013. Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. Dynamic Topic Adaptation for Phrase-based MT. in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden.

Ulf Hermjakob, Kevin Knight and Hal Daume III. 2008. Name Translation in Statistical Machine Translation Learning When to Transliterate. in *Proceedings of ACL-08: HLT*. Columbus, Ohio, USA.

Sanjika Hewavitharana, Dennis N. Mehay, Sankaranarayanan Ananthakrishnan and Prem Natarajan. 2013. Incremental Topic-Based Translation Model Adaptation for Conversational Spoken Language Translation. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria.

Mark Hopkins and Jonathan May. 2011. Tuning as Ranking. in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK.

Abraham Ittycheriah and Salim Roukos. 2007. Direct Translation Model 2. in *Proceedings of NAACL HLT 2007*. Rochester, NY.

Ferhan Ture, Douglas W. Oard and Philip Resnik. 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*. Montral, Canada.

Billy T. M. Wong and Chunyu Kit. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion To Document Level. in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level Consistency Verification in Machine Translation. in *Machine Translation Summit XIII*. Xiamen, China.

Deyi Xiong and Min Zhang. 2013. A Topic-Based Coherence Model for Statistical Machine Translation. in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. Bellevue, USA.

Deyi Xiong, Ding Yang, Min Zhang and Chew Lim Tan. 2013. Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation. in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, USA.