Cache-based Document-level Statistical Machine Translation

Zhengxian Gong[1] **Min Zhang**[2] **Guodong Zhou**[1*]
[1] School of Computer Science and Technology
Soochow University, Suzhou, China 215006
[2] Institute for Infocomm Research, Singapore 138632
{zhxgong, gdzhou}@suda.edu.cn mzhang@i2r.a-star.edu.sg

Abstract

Statistical machine translation systems are usually trained on a large amount of bilingual sentence pairs and translate one sentence at a time, ignoring document-level information. In this paper, we propose a cache-based approach to document-level translation. Since caches mainly depend on relevant data to supervise subsequent decisions, it is critical to fill the caches with highly-relevant data of a reasonable size. In this paper, we present three kinds of caches to store relevant document-level information: 1) a dynamic cache, which stores bilingual phrase pairs from the best translation hypotheses of previous sentences in the test document; 2) a static cache, which stores relevant bilingual phrase pairs extracted from similar bilingual document pairs (i.e. source documents similar to the test document and their corresponding target documents) in the training parallel corpus; 3) a topic cache, which stores the target-side topic words related with the test document in the source-side. In particular, three new features are designed to explore various kinds of document-level information in above three kinds of caches. Evaluation shows the effectiveness of our cache-based approach to document-level translation with the performance improvement of 0.81 in BLUE score over Moses. Especially, detailed analysis and discussion are presented to give new insights to document-level translation.

1 Introduction

During last decade, tremendous work has been done to improve the quality of statistical machine translation (SMT) systems. However, there is still a huge performance gap between the state-of-the-art SMT systems and human translators. Bond (2002) suggested nine ways to improve machine translation by imitating the best practices of human translators (Nida, 1964), with parsing the entire document before translation as the first priority. However, most SMT systems still treat parallel corpora as a list of independent sentence-pairs and ignore document-level information.

Document-level information can and should be used to help document-level machine translation. At least, the topic of a document can help choose specific translation candidates, since when taken out of the context from their document, some words, phrases and even sentences may be rather ambiguous and thus difficult to understand. Another advantage of document-level machine translation is its ability in keeping a consistent translation.

However, document-level translation has drawn little attention from the SMT research community. The reasons are manifold. First of all, most of parallel corpora lack the annotation of document boundaries (Tam, 2007). Secondly, although it is easy to incorporate a new feature into the classical log-linear model (Och, 2003), it is difficult to capture document-level information and model it via some simple features. Thirdly, reference translations of a test document written by human translators tend to have flexible expressions in order to avoid producing monotonous texts. This makes the evaluation of document-level SMT systems extremely difficult.

Tiedemann (2010) showed that the repetition and consistency are very important when modeling natural language and translation. He proposed to employ cache-based language and translation models in a phrase-based SMT system for domain

* Corresponding author.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 909–919,
Edinburgh, Scotland, UK, July 27–31, 2011.

adaptation. Especially, the cache in the translation model dynamically grows up by adding bilingual phrase pairs from the best translation hypotheses of previous sentences. One problem with the dynamic cache is that those initial sentences in a test document may not benefit from the dynamic cache. Another problem is that the dynamic cache may be prone to noise and cause error propagation. This explains why the dynamic cache fails to much improve the performance.

This paper proposes a cache-based approach for document-level SMT using a static cache and a dynamic cache. While such a approach applies to both phrase-based and syntax-based SMT, this paper focuses on phrase-based SMT. In particular, the static cache is employed to store relevant bilingual phrase pairs extracted from similar bilingual document pairs (i.e. source documents similar to the test document and their target counterparts) in the training parallel corpus while the dynamic cache is employed to store bilingual phrase pairs from the best translation hypotheses of previous sentences in the test document. In this way, our cache-based approach can provide useful data at the beginning of the translation process via the static cache. As the translation process continues, the dynamic cache grows and contributes more and more to the translation of subsequent sentences.

Our motivation to employ similar bilingual document pairs in the training parallel corpus is simple: a human translator often collects similar bilingual document pairs to help translation. If there are translation pairs of sentences/phrases/words in similar bilingual document pairs, this makes the translation much easier. Given a test document, our approach imitates this procedure by first retrieving similar bilingual document pairs from the training parallel corpus, which has often been applied in IR-based adaptation of SMT systems (Zhao et al.2004; Hildebrand et al.2005; Lu et al.2007) and then extracting bilingual phrase pairs from similar bilingual document pairs to store them in a static cache.

However, such a cache-based approach may introduce many noisy/unnecessary bilingual phrase pairs in both the static and dynamic caches. In order to resolve this problem, this paper employs a topic model to weaken those noisy/unnecessary bilingual phrase pairs by recommending the decoder to choose most likely phrase pairs according to the topic words extracted from the target-side text of similar bilingual document pairs. Just like a human translator, even with a big bilingual dictionary, is often confused when he meets a source phrase which corresponds to several possible translations. In this case, some topic words can help reduce the perplexity. In this paper, the topic words are stored in a topic cache. In some sense, it has the similar effect of employing an adaptive language model with the advantage of avoiding the interpolation of a global language model with a specific domain language model.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents our cache-based approach to document-level SMT. Section 4 presents the experimental results. Session 5 gives new insights on cache-based document-level translation. Finally, we conclude this paper in Section 6.

2 Related work

There are only a few studies on document-level SMT. Representative work includes Zhao et al. (2006), Tam et al. (2007), Carpuat (2009).

Zhao et al. (2006) assumed that the parallel sentence pairs within a document pair constitute a mixture of hidden topics and each word pair follows a topic-specific bilingual translation model. It shows that the performance of word alignment can be improved with the help of document-level information, which indirectly improves the quality of SMT.

Tam et al. (2007) proposed a bilingual-LSA model on the basis of a parallel document corpus and built a topic-based language model for each language. By automatically building the correspondence between the source and target language models, this method can match the topic-based language model and improve the performance of SMT.

Carpuat (2009) revisited the "one sense per discourse" hypothesis of Gale et al. (1992) and gave a detailed comparison and analysis of the "one translation per discourse" hypothesis. However, she failed to propose an effective way to integrate document-level information into a SMT system. For example, she simply recommended some translation candidates to replace some target words in the post-process stage.

In principle, the cache-based approach can be well suited for document-level translation. Basical-

ly, the cache is analogous to "cache memory" in hardware terminology, which tracks short-term fluctuation (Iyer et al., 1999). As the cache changes with different documents, the document-level information should be capable of influencing SMT.

Previous cache-based approaches mainly point to cache-based language modeling (Kuhn and Mori, 1990), which uses a large global language model to mix with a small local model estimated from recent history data. However, applying such a language model in SMT is very difficult due to the risk of introducing extra noise (Raab, 2007).

For cache-based translation modeling, Nepveu et al. (2004) explored user-edited translations in the context of interactive machine translation. Tiedemann (2010) proposed to fill the cache with bilingual phrase pairs from the best translation hypotheses of previous sentences in the test document. Both Nepveu et al. (2004) and Tiedemann (2010) also explored traditional cache-based language models and found that a cache-based language model often contributes much more than a cache-based translation model.

3 Cache-based document-level SMT

Given a test document, our system works as follows:

1) clears the static, topic and dynamic caches when switching to a new test document ;
2) retrieves a set of most similar bilingual document pairs for from the training parallel corpus using the cosine similarity with tf-idf weighting;
3) fills the static cache with bilingual phrase pairs extracted from ;
4) fills the topic cache with topic words extracted from the target-side documents of ;
5) for each sentence in the test document, translates it using cache-based SMT and continuously expands the dynamic cache with bilingual phrase pairs obtained from the best translation hypothesis of the previous sentences

In this way, our cache-based approach can provide useful data at the beginning of the translation process via the static cache. As the translation process continues, the dynamic cache grows and contributes more and more to the translation of subsequent sentences. Besides, the possibility of choosing noisy/unnecessary bilingual phrase pairs in both the static and dynamic caches is wakened with the help of the topic words in the topic cache. In particular, only the most similar document pair is used to construct the static cache and the topic cache unless specified.

In this section, we first introduce the basic phrase-based SMT system and then present our cache-based approach to achieve document-level SMT with focus on constructing the caches (static, dynamic and topic) and designing their corresponding features.

3.1 Basic phrase-based SMT system

It is well known that the translation process of SMT can be modeled as obtaining the best translation **e** of the source sentence **f** by maximizing following posterior probability (Brown et al., 1993):

$$= \arg\max \ (\ | \ ) = \arg\max \ (\ | \ ) \ (\ ) \quad (1)$$

where P(e|f) is a translation model and P_{lm} is a language model.

Our system adopted Moses (a state-of-art phrase-based SMT system) as a baseline, which follows Koehn et al. (2003) and mainly adopts six groups of popular features: 1) two phrase translation probabilities (two directions): $P_{phr}(e|f)$ and $P_{phr}(f|e)$; 2) two word translation probabilities (two directions) : $P_w(e|f)$ and $P_w(f|e)$; 3) one language model (target language): LM(e); 4) one phrase penalty (target language): PP(f); 5) one word penalty (target language):WP(e); 6) a lexicalized reordering model. Besides, the log-linear model as described in (Och and Ney, 2003) is employed to linearly interpolate these features for obtaining the best translation according to the formula (2):

$$= \arg\max \{ \sum_{=1} \lambda \quad (\ , \ ) \} \quad (2)$$

where $h_m(e , f)$ is a feature function, and λ_m is the weight of $h_m(e , f)$ optimized by a discriminative training method on a held-out development data.

In principle, a phrase-based SMT system can provide the best phrase segmentation and alignment that cover a bilingual sentence pair. Here, a segmentation of a sentence into K phrases is defined as:

$$(f \sim e) \approx \ \sum_{i=1}^{K} (f_i, e_i, \sim) \quad (3)$$

where tuple (f_i, e_i) refers to a **phrase pair**, and ~ indicates corresponding alignment information.

3.2 Dynamic Cache

Our dynamic cache is mostly inspired by Tiedemann (2010), which adopts a dynamic cache to store relevant bilingual phrase pairs from the best translation hypotheses of previous sentences in the test document. In particular, a specific feature is incorporated S_{cache} to capture useful document-level information in the dynamic cache:

$$(\quad | \quad) = \frac{\sum_{=1} (< \quad , \quad >=< \quad , \quad >) \times \quad ^{-\partial}}{\sum_{=1} (\quad = \quad)} \quad (4)$$

where $^{-\partial}$ is a decay factor to avoid the dependence of the feature's contribution on the cache size. Given $<e_c, f_c>$ an existing phrase pair in the dynamic cache and $<e_i,f_i>$ a phrase pair in a new hypothesis, if ($e_i=e_c \wedge f_i=f_c$) is true (i.e. full matching), function I(.) returns 1 , otherwise 0.

One problem with the dynamic cache in Tiedemann (2010) is that it continuously updates the weight of a phrase pair in the dynamic cache. This may cause noticeable computational burden with the increasing number of phrase pairs in the dynamic cache. In addition, as a source phrase (f_c) may occur many times in the dynamic cache, the weights for related phrase pairs may degrade severely and thus his decoder needs a decay factor, which is difficult to optimize. Finally, Tiedemann (2010) only allowed full matching. This largely lowers down the probability of hitting the dynamic cache and thus much affects its effectiveness.

To overcome above problems, we only employ the bilingual phrase pairs in the dynamic cache to inform the decoder whether one bilingual phrase pair exists in the dynamic cache or not, which is slightly similar to (Nepveu et al, 2004) ,thus avoiding extra computational burden and the fine-tuning of the decay factor. In particular, following new feature is incorporated to better explore the dynamic cache:

$$F_d = \sum_{i=1}^{K} \mathrm{dpairmatch}(e_i, f_i) \quad (5)$$

where $\mathrm{dpairmatch}(e_i, f_i)$

$$= \begin{cases} 1 \quad (e_i = e_c \wedge f_i = f_c) \\ \quad \vee \left(\hat{e}_i = e_c \wedge \hat{f}_i = f_c \wedge \| e_c \| > 3\right) \\ \quad \vee \left(e_i = \hat{e}_c \wedge f_i = \hat{f}_c \wedge \| e_i \| > 3\right) \\ 0 \quad \text{other} \end{cases}$$

Here, F_d is called the dynamic cache feature. Assume (e_c,f_c) is a phrase pair in the dynamic cache and (e_i,f_i) is a phrase pair candidate for a new hypothesis. Besides full matching, we introduce a symbol of "^" for sub-phrase, such as $\hat{e}_i$ for a sub-phrase of e_i and $\hat{e}_c$ for a sub-phrase of e_c , to allow partial matching. Finally, F_d measures the overall value of a target candidate f_i by summing over the scores of K phrase pairs.

Obviously, F_d rewards both full matching and partial matching. In order to avoid too much noise, we put some constraints on the number of words in the target phrase of $<e_c,f_c>$ or $<e_i,f_i>$, such as $\| e_i \| > 3$, where " $\|\|$ " measures the number of non-blank characters in a phrase. For example, if phrase pair ", 减少||| and reduced" occurs in the cache, phrase pair ",|||and" is not rewarded because such shorter phrase pairs occur frequently and may largely degrade the effect of the cache. In accordance, the dynamic cache only contains phrase pairs whose target phrases contain 4 or more non-blank characters.

3.3 Static Cache

In Tiedemann (2010), initial sentences in a test document fail to benefit from the dynamic cache due to the lack of contents in the dynamic cache at the beginning of the translation process. To overcome this problem, a static cache is included to store relevant bilingual phrase pairs extracted from similar bilingual document pairs in the training parallel corpus. In particular, a static cache feature F_s is designed to capture useful information in the static cache in the same way as the dynamic cache feature, shown in Formula (5).

For this purpose, all the document pairs in the training parallel corpus are aligned at the phrase level using 2-fold cross-validation. That is, we adopt 50% of the training parallel corpus to train a model using Moses and apply the model to enforce phrase alignment of the remaining training data, and vice versa. Here, the enforcement is done by guaranteeing the occurrence of the target phrase candidate of a source phrase in the sentence pair. Besides, all the words pairs trained on the whole training parallel corpus are included in both folds to ensure at least one possible translation. Finally, the phrase pairs in the best translation hypothesis of a sentence pair is retrieved from the decoder. In this way, we can extract a set of phrase pairs for each bilingual document pairs.

Given a test document, we first find a set of similar source documents by computing the Cosine similarity using the TF-IDF weighting scheme and their corresponding target documents, from the training parallel corpus. Then, the phrase pairs ex-

tracted from these similar bilingual document pairs are collected into the static cache. To avoid noise, we filter out those phrase pairs which occur less than two times in the training parallel corpus.

出口 ||| exports	汇率 ||| exchange
放慢 ||| slowdown	活力 ||| vitality
股市 ||| stock market	加快 ||| speed up the
现行 ||| leading	经济学家 ||| economists
出口 增幅 ||| export growth	
多种 原因 ||| various reasons	
国家 著名 ||| a well-known international	
议会 委员会 ||| congressional committee	
不 乐观 的 预期 ||| pessimistic predictions	
保持 一定 的 增长 ||| maintain a certain growth	
美元 汇率 下跌 ||| a drop in the dollar exchange rate	

Table 1: Phrase pairs extracted from a document pair with an economic topic

Similar to the dynamic cache, we only consider those phrase pairs whose target phrases contain 4 or more non-blank characters to avoid noise. We do not deliberately remove long phrase pairs. It is possible to use these long phrase pairs if our test document is very similar to one training document pair. Table 1 shows some bilingual phrase pairs extracted from a document pair, which reports a piece of news about "impact on slowdown in US economic growth". Obviously, these phrase pairs are closely related to economics.

3.4 Topic Cache

Both the dynamic and static caches may still introduce noisy/unnecessary bilingual phrase pairs even with constraints on the length of phrases and their occurrence frequency in the training parallel corpus. In order to resolve this problem, this paper adopts a topic cache to store relevant topic words and employs a topic cache feature to weaken those noisy/unnecessary phrase pairs.

Given w_t is a topic word in the topic cache, the topic cache feature F_t is designed as follows:

$$F_t = \sum_{i=1}^{K} \text{topicexist}(e_i, f_i) \quad (6)$$

where $\text{topicexist}(e_i, f_i) = \begin{cases} 1 & (w_t \in e_i) \\ 0 & \text{other} \end{cases}$

Here, the target phrase which contains a topic word w_t will be rewarded. w_t is derived by a topic model, LDA (Latent Dirichlet Allocation). This is different from the previous work (Tam, 2007), which mainly interpolated a topic language model with a general language model and added additional two adaptive lexicon probabilities in his phrase table.

In principle, LDA is a probabilistic model of text data, which provides a generative analog of PLSA (Blei et al., 2003), and is primarily meant to reveal hidden topics in text documents. Like most of the text mining techniques, LDA assumes that documents are made up of words and the ordering of the words within a document is unimportant (i.e. the "bag-of-words" assumption).

Figure 1 shows the principle of LDA, where α is the parameter of the uniform Dirichlet prior on the per-document topic distributions, β is the parameter of the uniform Dirichlet prior on the per-topic word distribution, θ is the topic distribution for document i, is the topic for the jth word in document i, and w_{ij} is the specific word. Among all variables, w_{ij} is the only observable variable with all the other variables latent. In particular, K denotes the number of topics considered in the model and φ is a K*V (V is the dimension of the vocabulary) Markov matrix each line of which denotes the word distribution of a topic. The inner plate over and illustrates the repeated sampling of topics and words until words have been generated for document . The plate surrounding θ illustrates the sampling of a distribution over topics for each document for a total of documents. The plate surrounding φ illustrates the repeated sampling of word distributions for each topic until topics have been generated.

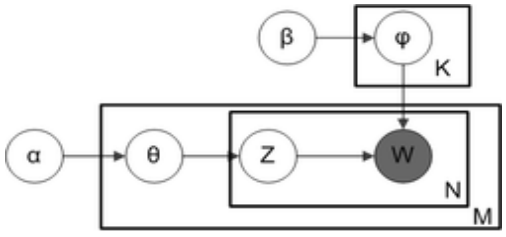

Figure 1: LDA

We use a LDA tool[1] to build a topic model using the target-side documents in the training parallel corpus. Using LDA, we can obtain the topic distribution of each word , namely p(|) for topic ϵK. Moreover, using the obtained word topic distributions, we can infer the topic distribution of a new document, namely p(|) for each topic ϵK.

Given a test document, we first find the most similar source document from the training data in

[1] http://www.arbylon.net/projects/

the same way as done in the static cache. After that, we retrieve its corresponding target document. Then, the topic of the target document is determined by its major topic, with the maximum p(|). Finally, we load some topic words corresponding to this topic into the topic cache. In particular, our LDA model deploy the setting of K=15, α=0.5 and β=0.1. Besides, only top 1000 topic words are reserved for each topic. Table 2 shows top 10 topic words for five topics.

Topic 1	Topic 2	Topic 3	Topic4	Topic5
company	army	party	bush	election
corporation	armed	represents	united	olympic
limited	military	study	adminis-tration	games
manager	officers	theory	policy	votes
board	forces	leadership	president	bid
branch	units	political	clinton	gore
companies	troops	cadres	office	presi-dential
ltd	force	speech	secretary	party
business	soldiers	comrade	powell	won
personnel	police	central	relations	speech

Table 2: Topic words extracted from target-side documents

4 Experimentation

We have systematically evaluated our cache-based approach to document-level SMT on the Chinese-English translation task.

4.1 Experimental Setting

Here, we use SRI language modeling toolkit to train a trigram general language model on English newswire text, mostly from the Xinhua portion of the Gigaword corpus (2007) and performed word alignment on the training parallel corpus using GIZA++(Och and Ney,2000) in two directions. For evaluation, the NIST BLEU script (version 13) with the default setting is used to calculate the Bleu score (Papineni et al. 2002), which measures case-insensitive matching of n-grams with n up to 4. To see whether an improvement is statistically significant, we also conduct significance tests using the paired bootstrap approach (Koehn, 2004)[2]. In this paper, '***', '**', and '*' denote p-values less than or equal to 0.01, in-between (0.01, 0.05), and bigger than 0.05, which mean significantly better, moderately better and slightly better, respectively.

[2] http://www.ark.cs.cmu.edu/MT

In this paper, we use FBIS as the training data, the 2003 NIST MT evaluation test data as the development data, and the 2005 NIST MT test data as the test data. Table 3 shows the statistics of these data sets (with document boundaries annotated).

Corpus		Sentences	Documents
Role	Name		
Train	FBIS	239413	10353
Dev	NIST2003	919	100
Test	NIST2005	1082	100

Table 3: Corpus statistics

In particular, the sizes of the static, topic and dynamic caches are fine-tuned to 2000, 1000 and 5000 items, respectively. For the dynamic cache, we only keep those most recently-visited items, while for the static cache; we always keep the most frequently-occurring items.

4.2 Experimental Results

Table 4 shows the contribution of various caches in our cache-based document-level SMT system. The column of "BLEU_W" means the BLEU score computed over the whole test set and "BLEU_D" corresponds to the average BLEU score over separated documents.

System	BLEU on Dev(%)	BLEU on Test(%)		
		BLEU_W	NIST	BLEU_D
Moses	29.87	25.76	7.784	25.08
Fd	29.90	26.03 (*)	7.852	25.39
Fd+Fs	30.29	26.30 (**)	7.884	25.86
Fd+Ft	30.11	26.24 (**)	7.871	25.74
Fd+Fs+Ft	30.50	26.42 (***)	7.896	26.11
Fd+Fs+Ft with merging	-	26.57 (***)	7.901	26.32

Table 4: Contribution of various caches in our cache-based document-level SMT system. Note that significance tests are done against Moses.

Contribution of dynamical cache (Fd)

Table 4 shows that the dynamic cache slightly improves the performance by 0.27 (*) in BLEU_W. This is similar to Tiedemann (2010). However, detailed analysis indicates that the dynamic cache does have negative effect on about one third of documents, largely due to the instability of the dynamic cache at the beginning of translating a document. Figure 2 shows the distribution of the

BLEU_D difference of 100 test documents (sorted by BLEU_D). It shows that about 55% of test documents benefit from the dynamic cache.

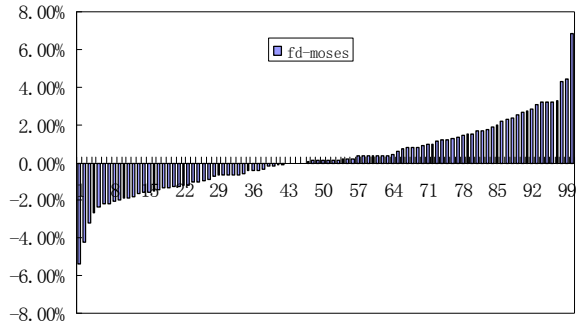

Figure 2: Contribution of employing the dynamic cache on different test documents

Contribution of static cache (Fs)

Table 4 shows that the combination of the static cache with the dynamic cache further improves the performance by 0.27(*) in BLEU_W. This suggests the effectiveness of the static cache in eliminating the instability of the dynamic cache when translating first few sentences of a test document. Together, the dynamic and static caches much improve the performance by 0.54 (**) in BLEU_W over Moses. Figure 3 shows the distribution of the BLEU_D difference of 100 test documents (sorted by BLEU_D), with more positive effect on those borderline documents, compared to Figure 2.

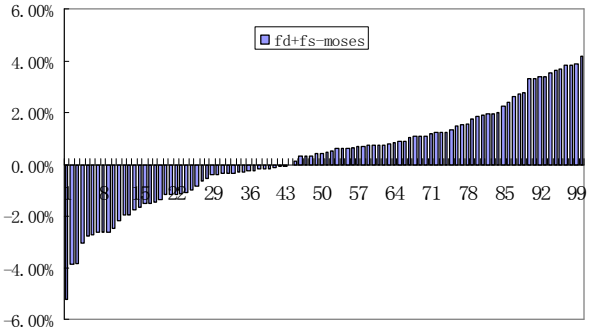

Figure 3: Contribution of combining the dynamic and static cache on different test documents

Contribution of topic cache (Ft)

Table 4 shows that the topic cache has comparable effect on improving the performance as the static cache when combined with the dynamic cache (0.48 vs. 0.54 in BLEU_W). Figure 4 shows the effectiveness of combining the dynamic and topic caches (sorted by BLEU_D).

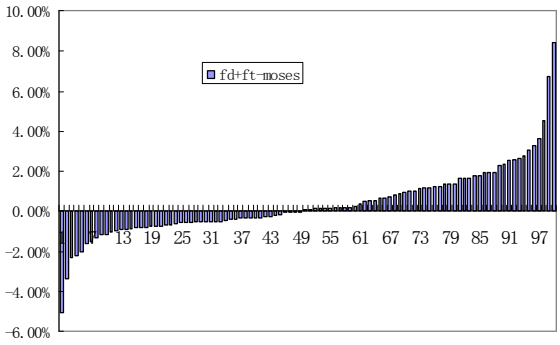

Figure 4: Contribution of combining the dynamic and topic caches

However, detailed analysis shows that the topic cache and the static cache are quite complementary by contributing on different test documents, largely due to that while the static cache tends to keep translation consistent, the topic cache plays like a document-specific language model. This is justified by Table 4 that the combination of the dynamic, static and topic caches significantly improve the performance by 0.66 (***) in BLEU_W, and by Figure 5 that about 75% of test documents benefit from the combination of the three caches (sorted by BLEU_D).

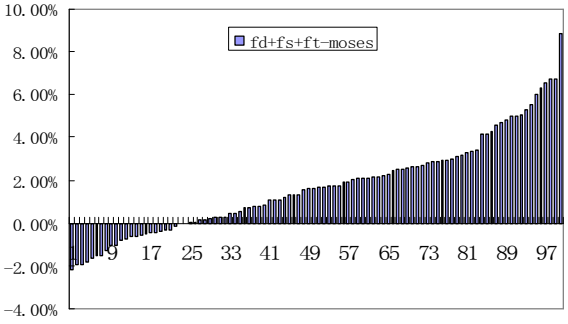

Figure 5: Contribution of combining the three caches

Contribution of merging phrase pairs of similar document pairs

Here, the number of similar documents we adopt is different from previous experiments. In the previous experiments, we only cache bilingual phrase pairs extracted from the most similar document. Here, we merge phrase pairs for several most similar documents (5 at most) which have the same topic.

Table 4 shows that employing this trick can further improve the performance by 0.15 in BLEU_W. As a result, the cache-based approach significantly improve the performance by 0.81 (***) in BLEU_W over Moses.

5 Discussion

In this section, we explore in more depth why the static cache can help the dynamic cache, some constrained factors which impact the effectiveness of our cache-based approach.

Effectiveness of the static cache

We investigate why the static cache affects the performance. Basically, it is difficult for the dynamic cache to capture such similar information in the static cache.

In principle, the static cache can both influence the initial and subsequent sentences; however subsequent ones can be affected by multiple caches. In order to give an insight of the static cache, we evaluate its effectiveness on the first sentence for each test document. Figure 6 shows the contribution of the static cache on translating these first sentences (y-axis shows BLEU value of the first sentence for each test document). It notes that the most BLEU scores of them are zeros because of the length limitation of first sentences.

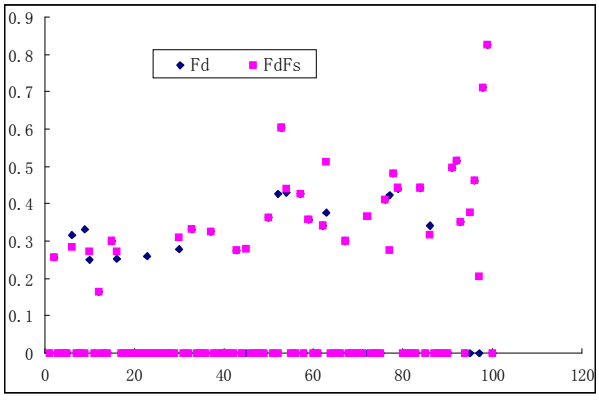

Figure 6: Contribution of the static cache on the first sentence of each test document (i.e. with empty dynamic cache)

Furthermore, we count the hit (matching) frequency of the static cache for each test documents. Since we use 1 or 0 for the static cache feature, it is easy to retrieve its effect for each test document. Our statistics shows that the hit frequency on static cache fluctuates between 5 and 18 for each test document. Without the static cache, the hit frequency of the dynamic cache is 504 on whole test sets, this figure increases to 685 with the static cache. This means that the static cache significantly enlarges the effectiveness of the dynamic cache by including more relevant phrase pairs to the dynamic cache, largely due to the positive impact of the static cache on the initial sentences of each test document.

Size of topic cache

Table 5 shows the impact of the topic cache when the number of the retained topic words for each topic increases from 500 to 2000. It shows that too more topic words actually harm the performance, due to the increase of noise. 1000 topic words seem a lot largely due to that we didn't do stemming for our topic modeling since we hope to introduce some tense information of them in the future.

Number of topic words	BLEU_W
500	26.27
700	26.31
1000	26.42
1500	26.23
2000	26.19

Table 5: Impact of the topic cache size

Influenced translations

In order to explore how our cache-based system impacts on translation results, we manually inspected 5 documents respectively which is improved or degraded in translation quality compared to the baseline Moses output. Those documents have 107 sentences in sum.

The good effectiveness of each kind of cache can be observed by the example 1 and 2 showed in Table 6. Both the example 1 and 2 come from the same document whose "BLEU_D" score exceeds Moses with 8.4 point. The example 1 benefits from the topic cache which contains the item of "action". The example 2 benefits from the static cache which contains a phrase pair of "承诺||| promised to" while Moses use "commitment" for "承诺", which may be the reason for missing the part of "prime minister" in Moses output. Furthermore, due to the phrase pair of "停火 协议||| the ceasefire agreement" existing in our static cache, our decoder keeps using "ceasefire" to translate "停火" in the whole document while Moses randomly use "ceasefire" or "cease-fire" for this translation.

1	官员 预测 “ 准备 工作 将 会 进行 到 七月 ， 然后 再 展开 政治 动作 ”
	Moses: official forecasts said that preparatory work will be carried out in july and then launched a political **maneuver** .
	Ours: official forecasts said that preparatory work will be carried out in july , then began a political **action** .
	Reference: officials expected that "preparations would take place until July, after which political action will begin".
2	关于 这 一 点 ， 中东新闻社 说 ， 以色列 总理 夏隆 承诺 “ 只 要 巴勒斯坦 当局 尊重 停火 协议 ， 控制 好 它们 的 地方 ， 以色列 将 会 停止 对 巴勒斯坦 人 的 军事 行动 ” 。
	Moses: on this point , said that israeli **commitment** to the palestinian authorities to respect the **cease-fire** agreement , where they are well under control , israel will stop its military actions against palestinians .
	Ours: on this point , said that israeli **prime minister promised to** respect the **ceasefire** agreement , the palestinian authorities to properly control their areas and israel will stop its military actions against palestinians .
	Reference:For this point , MENA said Israeli Prime Minister Sharon has promised to " stop Israeli military operations against the Palestinians insofar as they continue to respect the ceasefire deal and control their territory . "
3	17 日 晚 ， 近 3000 多 名 市民 在 市中心 的 武器 广场 观看 了 由 市政府 举办 的 精彩纷呈 的 歌舞 晚会 ， 五颜六色 的 灯光 装扮 着 广场 周围 的 古老 建筑 ， 著名 歌舞 艺术家 们 表演 了 不同 地区 的 民族 歌舞 。
	Moses: on the evening , nearly 3,000 residents in the downtown square of the weapons held by the municipal government , watched a song and dance soiree , having colorful lighting disguise of ancient buildings around the square , singing and dancing famous artists staged different regions of ethnic song and dance .
	Ours: later on, nearly 3,000 residents in the downtown square to watch the government of having a song and dance performances were held under the disguise of colorful lighting around the square , a famous ancient buildings and local artists of different ethnic song and dance .
	Reference: On the night of the 17th , nearly 3,000 residents watched a wonderful gala of songs and dances , organized by the municipal government , at Plaza da Armas . Colorful lights lighted up ancient architecture around the plaza . Famous artists including singers and dancers staged performances of national songs and dances of different regions .
4	利马 的 城市 面积 已 从 建城 之 初 的 2.14 平方公里 发展 到 2600 多 平方公里 ， 而 人口 也 增加 到 800 万 左右 ， 约 占 全国 总人口 的 31% 。
	Moses: at lima 's urban area from the beginning of 2600 **square** to 2.14 million **square kilometers** , while the population has increased to 8 percent of the country 's total , about 31% .
	Ours: lima , the urban area from the beginning of 2600 **square kilometers** to 2.14 million **square kilometers** , but also increased to about 8 million population , the country 's total population of about 31% .
	Reference: The area of Lima city has expanded to more than 2,600 square kilometers from the original 2.14 square kilometers when the city was founded , while the population has increased to around 8 million , roughly accounting for 31% of the nation's total .

Table 6: Positive and negative examples

The example 3 and 4 also come from the same document however whose performance degrades with 2.17 point. We don't think the translation quality for example 4 in our system is worse than Moses. However, the translation quality for example 3 in our system is very bad and especially showed on "re-ordering". We found this sentence did not match any item in our static cache and topic cache. Although this phenomenon also happens in other documents, but this is the most typical negative example among these documents.

Document-specific characteristics

It seems that using the same weight for the whole test sets (all documents) is not very reasonable. Actually, if we can determine those negative documents which are not suitable for the cache-based approach, our cache-based approach may gain much improvement. Tiedemann (2010) explored the correlation to document length, baseline performance and source document repetition. However, it seems that there are no obvious rules to filter out those negative documents. Besides, there may be two more document-specific factors: repetition of the reference text and document style.

Tiedemann (2010) only considered the repetition of the test text in the source side. Since BLEU score is computed against the reference text, the repetition in the reference text may greatly influence the performance of our cache-based approach to document-level SMT. As for document style, it is quite possible that a document may contain several topics. Therefore, it may be useful to track such change over topics and refresh various caches when there is a topic change. We will leave the above issues to the future work.

6 Conclusion

We have shown that our cache-based approach significantly improves the performance with the help of various caches, such as the dynamic, static and topic caches, although the cache-based ap-

proach may introduce some negative impact on BLEU scores for certain documents.

In the future, we will further explore how to reflect document divergence during training and dynamically adjust cache weights according to different documents.

There are many useful components in training documents, such as named entity, event and coreference. In this experiment, we only adopt the flat data in our cache. However, the structured data may improve the correctness of matching and thus effectively avoid noise. We will explore more effective ways to pick up various kinds of useful information from the training parallel corpus to expand our cache-based approach. Besides, we will resort to comparable corpora to enlarge our cache-based approach to document-level SMT.

Acknowledgments

This research was supported by Projects 90920004, 60873150, and 61003155 under the National Natural Science Foundation of China, Project 20093201110006 under the Specialized Research Fund for the Doctoral Program of Higher Education of China.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of machine Learning Research 3, pages 993–1022.

Francis Bond. 2002. Toward a Science of Machine Translation. Asian Association of Machine Translation (AAMT) Journal, N0.22, Tokyo, Japan, pages 12-20.

PF Brown, SA Della Pietra, VJ Della Pietra, RL Mercer.1992.The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics. 19(2):263-309.

Marine Carpuat. 2009. One Translation per Discourse. In Proc. of the NAACL HLT workshop on Semantic Evaluation, pages 19-26.

John DeNero, Alexandre Buchard-Cˆot´e, and Dan Klein. 2008. Sampling Alignment Structure under a Bayesian Translation Model. In Proc. of EMNLP 2008, pages 314–323, Honolulu, October.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense per Discourse. In Proceedings of the workshop on Speech and Natural Language, Harriman, NY.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel,and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. Proceedings of EAMT 2005:133-142.

Rukmini M. Iyer and Mari Ostendorf. 1999. Modeling Long Distance Dependence in Language:Topic Mixtures Versus Dynamic Cache Models. IEEE Transactions on speech and audio processing, 7(1).

Philipp Koehn, Franz Josef Och ,and DanielMarcu.2003. Statistical Phrase-Based Translation. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 48-54.

Philipp Koehn.2004.Statistical Significance Tests for Machine Translation Evaluation. In Proc. of EMNLP 2004, pages 388–395.

Roland Kuhn and Renato De Mori. 1990. A Cache-based Natural Language Model for Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,12(6):570-583.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In Proc. of EMNLP 2007,pages 343–350, Prague, Czech Republic, June.

Daniel Marcu and William Wong. 2002. A phrase-based Joint Probability Model for Statistical Machine Translation. In Proc. of EMNLP 2002, July.

Laurent Nepveu, Guy Lapalme, Philippe Langlais and George Foster.2004. Adaptive Language and Translation Models for Interactive Machine Translation. In Proc. of EMNLP 2004, pages 190-197.

Eugene A. Nida. 1964. Toward a Science of Translating. Leiden, Netherlands:E.J.Brill.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proc. of ACL, pages 160–167.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proc. of ACL, pages 440–447.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proc. of ACL02, pages 311–318.

Martin Raab. 2007. Language Modeling for Machine Translation. VDM Verlag, Saarbrucken, Germany.

G. Salton and C. Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. Information Processing and management,24(5):513-523,1988.

Yik-Cheung Tam, Ian Lane and Tanja Schultz. 2007. Bilingual ISA-based Adaptation for Statistical Machine Translation. , 28:187-207.

Jorg Tiedemann. 2010. Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In Proc. of the 2010 workshop on domain adaptation for Natural Language Processing, ACL 2010, pages 8-15.

Joern Wuebker and Arne Mauser and Hermann Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. In Proc. of ACL, pages 475-484.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004.Language Model Adaptation for Statistical Machine Translation with Structured Query Models. In COLING 2004, Geneva, August.

Bing Zhao and Eric P. Xing .2006. BiTAM:Bilingual Topic Ad-Mixture Models for Word Alignment. In Proc. of ACL2006.