Structured and Unstructured Cache Models for SMT Domain Adaptation

Annie Louis
School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB
alouis@inf.ed.ac.uk

Bonnie Webber
School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB
bonnie@inf.ed.ac.uk

Abstract

We present a French to English translation system for Wikipedia biography articles. We use training data from out-of-domain corpora and adapt the system for biographies. We propose two forms of domain adaptation. The first biases the system towards words likely in biographies and encourages repetition of words across the document. Since biographies in Wikipedia follow a regular structure, our second model exploits this structure as a sequence of topic segments, where each segment discusses a narrower subtopic of the biography domain. In this structured model, the system is encouraged to use words likely in the current segment's topic rather than in biographies as a whole. We implement both systems using cache-based translation techniques. We show that a system trained on Europarl and news can be adapted for biographies with 0.5 BLEU score improvement using our models. Further the structure-aware model outperforms the system which treats the entire document as a single segment.

1 Introduction

This paper explores domain adaptation of statistical machine translation (SMT) systems to contexts where the target documents have predictable regularity in topic and document structure. Regularities can take the form of high rates of word repetition across documents, similarities in sentence syntax, similar subtopics and discourse organization. Domain adaptation for such documents can exploit these similarities. In this paper we focus on topic (lexical) regularities in a domain. We present a system that translates Wikipedia biographies from French to English by adapting a system trained on Europarl and news commentaries. This task is interesting for the following two reasons.

Many techniques for SMT domain adaption have focused on rather diverse domains such as using systems trained on Europarl or news to translate medical articles (Tiedemann, 2010a), blogs (Su et al., 2012) and transcribed lectures (Federico et al., 2012). The main challenge for such systems is translating out-of-vocabulary words (Carpuat et al., 2012). In contrast, words in biographies are closer to a training corpus of news commentaries and parlimentary proceedings and allow us to examine how well domain adaptation techniques can disambiguate lexical choices. Such an analysis is harder to do on very divergent domains.

In addition, biographies have a fairly regular discourse structure: a central entity (person who is the topic of the biography), recurring subtopics such as 'childhood', 'schooling', 'career' and 'later life', and a likely chronological order to these topics. These regularities become more predictable in documents from sources such as Wikipedia. This setting allows us to explore the utility of models which make translation decisions depending on the discourse structure. Translation methods for structured documents have only recently been explored in Foster et al. (2010). However, their system was developed for parlimentary proceedings and translations were adapted using separate language models based upon the identity of the speaker, text type (questions, debate, etc.) and the year when the proceedings took place. Biographies constitute a more realistic discourse context to develop structured models.

This paper introduces a new corpus consisting of paired French-English translations of biography articles from Wikipedia.[1] We translate this corpus by developing cache-based domain adaptation methods, a technique recently proposed by Tiede-

[1] Corpus available at http://homepages.inf.ed.ac.uk/alouis/wikiBio.html.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 155–163,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

mann (2010a). In such methods, cache(s) can be filled with relevant items for translation and translation hypotheses that match a greater number of cache items are scored higher. These cache scores are used as additional features during decoding. We use two types of cache—one which encourages the use of words more indicative of the biography domain and another which encourages word repetition in the same document.

We also show how cache models allow for straightforward implementation of structured translation by refreshing the cache in response to topic segment boundaries. We fill caches with words relevant to the topic of the current segment which is being translated. The cache contents are obtained from an unsupervised topic model which induces clusters of words that are likely to appear in the same topic segment. Evaluation results show that cache-based models give upto 0.5 BLEU score improvements over an out-of-domain system. In addition, models that take topical structure into account score 0.3 BLEU points higher than those which ignore discourse structure.

2 Related work

The study that is closest to our work is that of Tiedemann (2010a), which proposed cache models to adapt a Europarl-trained system to medical documents. The system used caching in two ways: a cache-based language model (stores target language words from translations of preceding sentences in the same document) and a cache-based translation model (stores phrase pairs from preceding sentence translations). These caches encouraged the system to imitate the 'consistency' aspect of domain-specific texts i.e., the property that words or phrases are likely to be repeated in a domain and within the same document.

Cache models developed in later work, Tiedemann (2010b) and Gong et al. (2011), were applied for translating in-domain documents. Gong et al. (2011) introduced additional caches to store *(i)* words and phrase pairs from training documents most similar to a current source article, and *(ii)* words from topical clusters created on the training set. However, a central issue in these systems is that caches become noisy over time, since they ignore topic shifts in the documents. This paper presents cache models which not only take advantage of likely words in the domain and consistency, but which also adapt to topic shifts.

A different line of work very relevant to our study is the creation of topic-specific translations by either inferring a topic for the source document as a whole, or at the other extreme, finer topics for individual sentences (Su et al., 2012; Eidelman et al., 2012). Neither of these granularities seem intuitive in natural discourse. In this work, we propose that tailoring translations to topics associated with discourse segments in the article is likely to be beneficial for two reasons: a) subtopics of such granularity can be assumed with reasonable confidence to re-occur in documents from the same domain and b) we can hypothesize that a domain will have a small number of segment-level topics.

3 System adaptation for biographies

We introduce two types of translation systems adapted for biographies:

General domain models (domain-) that use information about biographies but treat the document as a whole.

Structured models (struct-) that are sensitive to topic segment boundaries and the specific topic of the segment currently being translated.

We implement both models using caches. Since we do not have parallel corpora for the biography domain, our caches contain items in the target language only. We use two types of caches:

Topic cache stores target language words (unigrams) likely in a particular topic. Each unigram has an associated score.

Consistency cache favours repetition of words in the sentences from the same document. It stores target language words (unigrams) from the 1-best translations of previous sentences in the same document. Each word is associated with an age value and a score. Age indicates when a word entered the cache and introduces a 'decay effect'. Words used in immediately previous sentences have a low age value while higher age values indicate words from sentences much prior in the document. Scores are inversely proportional to age.

Both the types of caches are present in both the general domain and structured models, but the cache words and scores are computed differently.

3.1 A general domain model

This system seeks to bias translations towards words which occur often in biography articles.

The topic cache is filled with word unigrams that are more likely to occur in biographies compared to general news documents. We compare the words from 1,475 English Wikipedia biographies articles to those in a large collection (64,875 articles) of New York Times (NYT) news articles (taken from the NYT Annotated Corpus (Sandhaus, 2008)). We use a log-likelihood ratio test (Lin and Hovy, 2000) to identify words which occur with significantly higher probability in biographies compared to NYT. We collect only words indicated with 0.0001 significance by the test to be more likely in biographies. We rank this set of 18,597 words in decreasing order of frequency in the biography article set and assign to each word a score equal to $1/\text{rank}$ of the word. These words with their associated scores form the contents of the topic cache. In the general domain model, these same words are assumed to be useful for the full document and so the cache contents remain constant during translation of the full document.

The consistency cache stores words from the translations of preceding sentences of the same document. After each sentence is translated, we collect the words from the 1-best translation and filter out punctuation marks and out of vocabulary words. The remaining words are assigned an age of 1. Words already present in the cache have their age incremented by one. The new words with age 1 are added to the cache[2] and the scores for all cache words are recomputed as $e^{1/age}$. The age therefore gets incremented as each sentence's words are inserted into the cache creating a decay. The cache is cleared at the end of each document.

During decoding, a candidate phrase is split into unigrams and checked against each cache. Scores for matching unigrams are summed up to obtain a score for the phrase. Separate scores are computed for matches with the topic and consistency caches.

3.2 A structured model

Here we consider topic and consistency at a narrower level—within topic segments of the article.

The topic cache is filled with words likely in individual topic segments of an article. To do this, we need to identify the topic of smaller segments of the article and also store a set of most probable words for each topic. The topics should also have bilingual mappings which will allow us to infer for every French document segment, words that are likely in such a segment in the English language.

We designed and implemented an unsupervised topic model based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to induce such word clusters. In a first step, we induce subtopics from monolingual articles in English and French separately. The topics are subsequently aligned between the languages as explained below.

In the *first step*, we learn a topic model which incorporates two main ideas a) adds sensitivity to topic boundaries by assigning a single topic per topic segment b) allows for additional flexibility by not only drawing the words of a segment from the segment-level topic, but also allows some words to be either specific to the document (such as named entities) or stop words. To address idea b), we have a "switching variable" to switch between document-specific word, stopword or domain-words.

The generative story to create a monolingual dataset of biographies is as follows:

- Draw a distribution η for the proportion of the three word types in the full corpus (domain subtopic, document-specific, stopwords) $\sim$ Dirichlet(γ)
- For each domain subtopic ϕ_l, $1 \leq l \leq T$, draw a distribution over word vocabulary $\sim$ Dirichlet(β)
- Draw a distribution ψ over word vocabulary for stopwords $\sim$ Dirichlet(ϵ)
- For each document D_i:
  - Draw a distribution π_i over vocabulary for document-specific words $\sim$ Dirichlet(μ)
  - Draw a distribution θ_i giving the mixture of domain subtopics for this document $\sim$ Dirichlet(α)
  - For each topic segment M_{ij} in D_i:
    - ∗ Draw a domain subtopic $z_{ij} \sim$ Multinomial(θ_i)
    - ∗ For each word w_{ijk} in segment M_{ij}:
      - · Draw a word type $s_{ijk} \sim$ Multinomial(η)
      - · Depending on the chosen switch value s_{ijk}, draw the word from the subtopic of the segment $\phi_{z_{ij}}$ or document-specific vocabulary π_i, or stopwords ψ

We use the section markings in the Wikipedia articles as topic segment boundaries while learning the model. We use symmetric Dirichlet priors

[2] If the word already exists in the cache, it is first removed.

for the vocabulary distributions associated with domain subtopics, document-specific words and stopwords. The concentration parameters are set to 0.001 to encourage sparsity. The distribution θ_i for per-document subtopics is also drawn from a symmetric Dirichlet distribution with concentration parameter 0.01. We use asymmetric Dirichlet priors for η set to (5, 3, 2) for (domain topic, document-specific, stopwords). The hyperparameter values were minimally tuned so that the different vocabulary distributions behaved as intended.

We perform inference using collapsed Gibbs sampling where we integrate out many multinomials. The sampler chooses a topic z_{ij} for every segment and then samples a word type s_{ijk} for each word in the segment. We initialize these variables randomly and the assignment after 1000 Gibbs iterations are taken as the final ones. We create these models separately for English and French, in each case obtaining T domain subtopics.

The *second step* creates an alignment between the source and target topics using a bilingual dictionary[3]. For each French topic, we find the top matching English topic by scoring the number of dictionary matches. It is unlikely for every French topic to have a closely corresponding English topic. Based on observations about the quality of topic alignment, we select the top 60% (out of T) pairs of French-English aligned topics only.

Note that our method uses two steps to learn bilingual topics in contrast to some multilingual topic models which learn aligned topics directly from parallel or comparable corpora (Zhao and Xing, 2006; Boyd-Graber and Blei, 2009; Jagarlamudi and Daumé III, 2010). These methods induce topic-specific translations of words. Rather we choose a less restrictive pairing of word clusters by topic since (i) we have monolingual biographies in the two languages which could be quite heterogenous in the types of personalities discussed, (ii) we seek to identify words likely in a topic segment for example 'career-related' words rather than specific translations for source words.

During translation, for each topic segment in the source document, we infer the French topic most likely to have produced the segment and find the corresponding English-side topic. The most probable words for that English topic are then loaded into the topic cache. The score for a word is its probability in that topic. When a topic segment boundary is reached, the topic cache is cleared and the topic words for the new segment are filled.

The consistency cache's contents are computed similarly to the general domain case. However, the cache gets cleared at segment boundaries.

4 Training and test data

We distinguish two resources for data. The out-of-domain system is trained using the WMT'12 datasets comprising Europarl and news commentary texts. It has 2,144,820 parallel French-English sentence pairs. The language model is trained using the English side of the training corpus. The tuning set has 2,489 sentence pairs.

Our test set is a corpus of French to English translations of biographies compiled from Wikipedia. To create the biography corpus, we collect articles which are marked with a "Translation template" in Wikipedia metadata. These markings indicate a page which is translated from a corresponding page in a different language and also contains a link to the source article. (Note that these article pairs are **not** those written on the same topic separately in the two languages.) We collect pairs of French-English pages with this template and filter those which do not belong to the Biography topic (using Wikipedia metadata).

Note, however, that these article pairs are not very close translations. During translation an editor may omit or add information and also reorganize parts of the article. So we filter out the paired documents which differ significantly in length. We use LFAligner[4] to create sentence alignments for the remaining document pairs. We constrain the alignments to be within documents but since section headings were not maintained in translations, we did not further constrain alignments within sections. We manually corrected the resulting alignments and keep only documents which have good alignments and have manually marked topic segments (Wikipedia section headings). Unaligned sentences were filtered out. Table 1 shows a summary of this data and the split for tuning and test. The articles are 12 to 87 sentences long and contain 5 topic segments on average.

We also collect a larger set of monolingual French and English Wikipedia biographies to create the domain subtopics. We select only articles that have at least 10 segments (sections) to ensure

[3] A filtered set of 13,400 entries from `www.dict.cc`

[4] `http://sourceforge.net/projects/aligner/`

	Tuning	Test
No. of article pairs	15	30
Total sentences pairs	430	1008
Min. article size (in sentences)	13	12
Max. article size (in sentences)	59	85
Average no. of segments per article	4.7	5.3

Table 1: Summary of Wikipedia biographies data

that they are comprehensive ones. This collection contains 1000 French and 1000 English articles.

5 Experimental settings

We use the Moses phrase-based translation system (Koehn et al., 2007) to implement our models.

5.1 Out-of-domain model

This baseline model is trained on the WMT 2012 training sets described in the previous section and uses the six standard features from Koehn et al. (2003). We build a 5-gram language model using SRILM. The features were tuned using MERT (Och, 2003) on the WMT 2012 tuning sets. This system does not use any data about biographies.

5.2 Biography-adapted models

First we perform experiments using the manually marked sections in Wikipedia as topic segments. We also report results with automatic segmentation in Section 7.

The domain and structured models have two extra features 'topic cache' and 'consistency cache'. For the structured model, topic segment boundaries and inferred topic is passed as XML markup on the source documents. For the consistency cache, we use a wrapper which passes the 1-best translation (also using XML markup) of the preceding sentence and updates the cache before translating every next sentence.

We tune the weights for these new cache features as follows. The weights for the baseline features from the out-of-domain model are kept constant. The weights for the new cache features are set using a grid search. This tuning uses the biographies documents listed in Table 1 as tuning data. We run the decoding using the baseline feature weights and a weight for a cache feature and compute the (case-insensitive) BLEU (Papineni et al., 2002) scores of each tuning document. The weight for the cache feature which maximizes the average BLEU value over the tuning documents is chosen. We have not tuned the features using MERT in this study since a grid search allowed us to quantify the influence of increasing weights on the new features directly. Previous work has noted that MERT fails to find good settings for cache models (Tiedemann, 2010b). In future work, we will explore how successful optimization of baseline and cache feature weights could be done jointly. We present the findings from our grid search below.

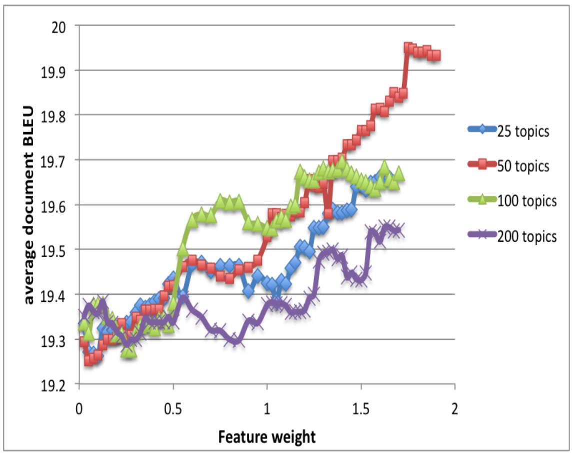

Figure 1: Effect of feature weights and number of topics on accuracy for structured topic cache

The struct-topic cache has two parameters, the number of topics T and the number of most probable words from each topic which get loaded into the cache. We ran the tuning for $T = 25$, 50, 100 and 200 topics (note that 60% of the topics will be kept after bilingual alignment, see Section 3.2). We also varied the number of topic words chosen—50, 100, 250 and 500.

The performance did not vary with the number of topic words used and 50 words gave the same performance as 500 words for topic models with any number of topics. This interesting result suggests that only the most likely and basic words from each topic are useful. The top 50 words from two topics (one capturing early life and the other an academic career) taken from the 50-topic model on English biographies are shown in Table 2.

In Figure 1, we show the performance of systems using different number of topics. In each case, the same number of topic words (50) was added to the cache. We find that 50 topics model performs best confirming our hypothesis that only a small number of domain subtopics is plausible. We choose the 50 topic model with top 50 words for each topic for the structured topic cache.

The best weights and average document level BLEU scores on the tuning set are given in Table 3. The scores were computed using the *mteval-v13a.pl* script in Moses. BLEU scores for the

his	a	s	family	on	life	She	child	St	mother
of	in	married	children	They	death	became	whom	friends	attended
and	had	He	that	daughter	son	marriage	lived	later	work
to	was	born	died	wife	years	met	couple	I	age
he	her	at	she	father	home	moved	about	husband	house
of	is	He	received	has	included	National	original	Academy	French
and	The	by	its	used	works	study	list	book	College
his	work	Award	Medal	award	His	Institute	life	contributed	Year
in	he	are	awarded	also	title	Arts	Royal	edition	awards
s	University	Prize	Society	A	honorary	Library	include	Sciences	recognition

Table 2: Top 50 words from 2 topics of the $T = 50$ topic model

Cache type	**weight**	**BLEU-doc**
Domain-topic	0.075	19.79
Domain-consistency	0.05	19.70
Domain-topic + consis.	0.05, 0.05	19.80
Struct-topic (50 topics)	1.75	**19.94**
Struct-consistency	0.125	19.70
Struct-topic + consis.	0.4, 0.1	19.84
Domain-consis. + struct-topic	0.1, 0.25	19.86
Out-of-domain		19.33

Table 3: Best weights for cache features and BLEU scores (averaged for tuning documents).

out-of-domain model are shown on the last line. Note that these scores are overall on a lower scale for a French-English system due to out-of-domain differences and because the reference translations from Wikipedia are not very close ones.

These numbers show that cache models have the potential to provide better translations compared to an out-of-domain baseline. The structured topic model system is the best system outperforming the out-of-domain system and also the domain-topic system. Hence, treating documents as composed of topical segments is a useful setting for automatic translation.

The domain and structured versions of the consistency cache however, show no difference. This result could arise due to the decay factor incorporated in the consistency cache. Higher scores are given to words from immediately previous sentences compared to those far off. This decay implicitly gives lower scores to words from earlier topic segments than those from recent ones. Explicitly refreshing the cache in the structured model does not give additional benefits.

When consistency and topic caches are used together in both general domain and structured settings, the combination is not better than individual caches. We also tried a setting where the consistency cache is document-range and the topic cache works at segment level (domain-consis. + struct-topic). This combination also does not outperform using the structured topic cache alone.

Model	**BLEU-doc**	**BLEU-sent**
Domain-topic	17.63	17.61
Domain-consistency	17.70	17.75
Domain-topic + consis.	17.63	17.63
Struct-topic (50 topics)	**17.76**	**17.84**
Struct-consistency	17.33	17.34
Struct-topic + consis.	17.47	17.51
Struct-topic + dom-consis.	17.29	17.25
Out-of-domain	17.37	17.43

Table 4: BLEU scores on the test set. 'doc' indicates BLEU scores averaged over documents, 'sent' indicates sentence-level BLEU

6 Results on the test corpus

The best weights chosen on the tuning corpus are used to decode the biographies test corpus (summarized in Table 1). Table 4 reports the average BLEU of documents as well as sentence level BLEU scores of the corpus. We used the paired bootstrap resampling method (Koehn 2004) to compute significance.

The struct-topic model gives the highest improvement of 0.4 sentence level BLEU over the out-of-domain model. Struct-topic is also 0.23 BLEU points better compared to the domain-topic model confirming the usefulness of modeling structure regularities. These improvements at significant at 95% confidence level.

The second best model is the domain-consistency model (significantly better than out-of-domain model at 90% confidence level). But the performance of this cache decreases in the structured setting. Moreover, combinations of caches fail to improve over individual caches. One hypothesis for this result is that biography subtopic words which give good performance in the topic cache differ from the words which provide benefits in the consistency cache. For example, words related to named entities and other document-specific content words could be ones that are more consistent within the document. Then clearing the consistency cache at topic boundaries would remove such words from the cache leading to low performance of the 'structured' version. In our current model, we do not distinguish between words making up the consistency cache. In future, we plan to experiment with consistency caches of different ranges and which hold different types of words. This approach would require identifying named entities and parts of speech on the automatic translations of previous sentences, which is likely to be error-prone and so require methods for associating a confidence measure with the cache words.

7 Understanding factors that influence structured cache models

The documents in our test corpus have varying lengths, number of segments and segment sizes. This section explores the behavior of structured models on these different document types. For this analysis, we compare the BLEU scores from the *domain* and the *structured* versions of the two caches. We do not consider the out-of-domain system here since we are interested in quantifying gains from using document structure.

For each document in our test corpus, we compute *(i)* the difference between the BLEU scores of struct-topic and domain-topic systems (*BLEU-gain-topic*), and *(ii)* the difference in BLEU scores between the struct-consistency and domain-consistency systems (*BLEU-gain-consis*). Table 5 reports the average BLEU gains binned by a) the document length (in sentences) b) number of topic segments in the document and c) the average size of topic segments in a document (in sentences).

The numbers clearly indicate that performance is not uniform across different types of documents. The struct-topic cache performs much better on longer documents of over 30 sentences giving 0.3 to 0.4 BLEU points increase compared to the general domain model. On the other hand, the performance worsens when the structured cache is applied on documents with less than 20 sentences. Similarly, the struct-topic cache is beneficial for documents where the average segment size is larger than 5 sentences and when the number of topic segments is around 5 to 7.

The struct-consistency cache generally performs worse than the unstructured version and there does not appear to be a niche set according to any of the properties—document length, number of segments and segment size.

(a) Average BLEU gains and document length

doc. length	no. docs	gain-topic	gain-consis
12 to 19	7	-0.41	-0.20
20 to 29	10	0.17	-0.63
30 to 49	8	0.44	-0.16
50 to 85	5	0.34	-0.45

(b) Average BLEU gains and no. of topic segments

no. segments	no. docs	gain-topic	gain-consis
3 to 4	9	-0.09	-0.21
5	13	0.24	-0.37
6 to 7	5	0.34	-0.74
9	3	-0.03	-0.26

(c) Average BLEU gains and topic segment size

avg. segment size	no. docs	gain-topic	gain-consis
< 5	10	-0.23	-0.41
5 to 10	18	0.33	-0.37
11 to 17	2	0.39	-0.24

Table 5: Average BLEU score gains from a structured cache (compared to domain caches) split by different properties of documents in the test set

Given these findings, it is possible that the struct-topic cache can benefit by modifying the document structure to match that handled better by the structured model. We test this hypothesis by segmenting all test documents with an ideal segment size. The model seems to perform better when each segment has around 5 to 10 sentences (longer segments are also preferred but we have few very long documents in our corpus), so we try to re-segment the articles to contain approximately 7 sentences in each segment. We use an automatic topic segmentation method (Eisenstein and Barzilay, 2008) to segment the source articles in our test corpus. For each article we request (document length)/7 segments to be created.[5]

We then run the structured topic and consistency models on the automatically segmented corpus using the same feature weights as before. The results are shown in Table 6.

Model	BLEU (doc)	BLEU (sent)
Struct-topic	**17.94**	**17.94**
Struct-consistency	17.51	17.46

Table 6: Translation performance on automatically segmented test corpus

The struct-topic cache now reaches our best result of 0.5 BLEU improvement over the out-of-domain model and 0.3 improvement over the unstructured domain model. The consistency cache is also slightly better using the automatic segmentation than the manual sections. Choosing the right granularity appears to be important for structured caches and coarse section headers may not be ideal. This result also shows automatic segmen-

[5] Note that we only specify the number of segments, but the system could create long or short segments.

of (42)	he (36)	his (36)	the (22)	to (11)	in (9)	was (7)	one (6)	a (3)	at (3)
head (3)	that (3)	construction (3)	empire	office	french	bases	reconstruction	only	such
all	ban	marseille	main	charged	have	well	researchers	openness	retreat
an	two	mechanical	events	army	iron	class	surrender	order	thirty
and	black	objectives	factory	disciple	largest	close	budget	part	time
as	who	ceremony	figure	majority	level	even	sentence	project	trained
on	seat	diplomatic	wheat	working	winner	life	archaeological	9	during

Table 7: Impact words computed on the test corpus. The number of times each word was found in the impact list is indicated within parentheses. Words listed without parentheses appeared once in the list.

(1) **(S)** Pendant la Première Guerre mondiale, mobilisé dans les troupes de marine, il combat dans les Balkans et les Dardanelles.
(R) During the First World War, conscripted into the navy, **he** fought in the Balkans and the Dardanelles.
(B) During World War I, mobilized in troops navy, it fight in the Balkans and Dardanelles.
(C) During World War I, mobilized troops in the navy, **he** fight in the Balkans and the Dardanelles.

(2) **(S)** À l'âge de 15 ans, elle a été choisie par la troupe d'opéra de l'armée chinoise pour être formée au chant.
(R) At the age of 15, she was selected by the Chinese Armys Operatic troupe to be **trained** as a singer.
(B) In the age of 15 years, she was chosen by the pool of opera of the Chinese military to be formed the call.
(C) In the age of 15 years, she was chosen by the pool of opera of the Chinese military to be **trained** to call.

(3) **(S)** La figure de la Corriveau n'a cessé, depuis, d'inspirer romans, chansons et pièces de théâtre et d'alimenter les controverses.
(R) The **figure** of Corriveau still inspires novels, songs and plays and is the subject of argument.
(B) The perceived the Corriveau has stopped, since, inspire novels, songs and parts of theater and fuel controversies.
(C) The **figure** of the Corriveau has stopped, since, inspire novels, songs and parts of theater and fuel controversies.

Table 8: Three examples of impact words in test translations. Abbreviations: S - source sentence, R - reference translation, B - baseline translation, C - structured topic cache translation

tation can be successfully used in these models.

8 Changes made by the cache models

Here we examine the kinds of changes made by the cache models which have lead to the improved BLEU scores. We focus on the the topic cache since its changes are straightforward to compute compared to consistency. We analyze the struct-topic cache translations on automatically segmented documents as that provided the best performance overall.

To do this analysis, we define the notion of an *impact word*. An *impact word* is one which satisfies three conditions: (i) the word **is not present** in the out-of-domain translation of a sentence, (ii) it **is present** in the translation produced by the topic cache model (iii) the word **matches the reference** translation for the sentence.

These impact words provide a simple (albeit approximate) way to analyze useful changes made by the topic cache over the out-of-domain system.

On the test corpus (30 documents), 231 impact word tokens were found and they come from 70 unique word types. So topic cache model significantly affects translation decisions and over 200 useful word changes were made in the 30 documents. The impact word types and counts are shown in Table 7. Several of these changes relate to function words and pronouns. For example, the pronoun 'he' and the past tense verb 'was' were correctly introduced in several sentences such as Example (1) in Table 8. A content word change is indicated in examples (2) and (3). These changes appear to be appropriate for biographies.

9 Conclusions

We have introduced a new corpus of biography translations which we propose as suitable for examining discourse-motivated SMT methods. We showed that cache-based techniques which also take the topic organization into account, make more appropriate lexical choices for the domain. In future work, we plan to explore how other domain similarities such as sentence syntax and entity reference, for example biographies have a central entity (person), can be used to improve translation performance. We also plan to take advantage of recent methods to do document level decoding (Hardmeier et al., 2012).

Acknowledgements

The first author was supported by a Newton International Fellowship (NF120479) from the Royal Society and The British Academy. We also thank the NLP group at Edinburgh for their comments and feedback on this work.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of UAI*, pages 75–82.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel Rudinger. 2012. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of ACL*, pages 115–119.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, pages 334–343.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stueker. 2012. Overview of the IWSLT 2012 evaluation campaign. *Proceedings of IWSLT*.

George Foster, Pierre Isabelle, and Roland Kuhn. 2010. Translating structured documents. In *Proceedings of AMTA*.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of EMNLP*, pages 909–919.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the EMNLP-CoNLL*, pages 1179–1190.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 444–456. Springer Berlin Heidelberg.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL-HLT*, pages 48–54.

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

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

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

Evan Sandhaus. 2008. The New York Times Annotated Corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia.*

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of ACL*, pages 459–468.

Jörg Tiedemann. 2010a. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*.

Jörg Tiedemann. 2010b. To cache or not to cache?: experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 189–194.

Bing Zhao and Eric P. Xing. 2006. Bitam: bilingual topic admixture models for word alignment. In *Proceedings of the COLING-ACL*, pages 969–976.