Dynamic Topic Adaptation for SMT using Distributional Profiles

Eva Hasler[1] **Barry Haddow**[1] **Philipp Koehn**[1,2]
[1]School of Informatics, University of Edinburgh
[2]Center for Language and Speech Processing, Johns Hopkins University
e.hasler@ed.ac.uk, {bhaddow,pkoehn}@inf.ed.ac.uk

Abstract

Despite its potential to improve lexical selection, most state-of-the-art machine translation systems take only minimal contextual information into account. We capture context with a topic model over distributional profiles built from the context words of each translation unit. Topic distributions are inferred for each translation unit and used to adapt the translation model dynamically to a given test context by measuring their similarity. We show that combining information from both local and global test contexts helps to improve lexical selection and outperforms a baseline system by up to 1.15 BLEU. We test our topic-adapted model on a diverse data set containing documents from three different domains and achieve competitive performance in comparison with two supervised domain-adapted systems.

1 Introduction

The task of lexical selection plays an important role in statistical machine translation (SMT). It strongly depends on context and is particularly difficult when the domain of a test document is unknown, for example when translating web documents from diverse sources. Selecting translations of words or phrases that preserve the sense of the source words is closely related to the field of *word sense disambiguation* (WSD), which has been studied extensively in the past.

Most approaches to WSD model context at the sentence level and do not take the wider context of a word into account. Some of the ideas from the field of WSD have been adapted for machine translation (Carpuat and Wu, 2007b; Carpuat and Wu, 2007a; Chan et al., 2007). For example, Carpuat and Wu (2007a) extend word sense disambiguation to phrase sense disambiguation and show improved performance due to the better fit with multiple possible segmentations in a phrase-based system. Carpuat (2009) test the "one sense per discourse" hypothesis (Gale et al., 1992) for MT and find that enforcing it as a constraint at the document level could potentially improve translation quality. Our goal is to make correct lexical choices in a given context without explicitly enforcing translation consistency.

More recent work in SMT uses latent representations of the document context to dynamically adapt the translation model with either monolingual topic models (Eidelman et al., 2012; Hewavitharana et al., 2013) or bilingual topic models (Hasler et al., 2014), thereby allowing the translation system to disambiguate source phrases using document context. Eidelman et al. (2012) also apply a topic model to each test sentence and find that sentence context is sufficient for picking good translations, but they do not attempt to combine sentence and document level information. Sentence-level topic adaptation for SMT has also been employed by Hasler et al. (2012). Other approaches to topic adaptation for SMT include Zhao and Xing (2007) and Tam et al. (2008), both of which use adapted lexical weights.

In this paper, we present a topic model that learns latent distributional representations of the context of a phrase pair which can be applied to both local and global contexts at test time. We introduce similarity features that compare latent representations of phrase pair types to test contexts to disambiguate senses for improved lexical selection. We also propose different strategies for combining local and global topical context and show that using clues from both levels of contexts is beneficial for translation model adaptation. We evaluate our model on a dynamic adaptation task where the domain of a test document is unknown and hence the problem of lexical selection is harder.

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 445–456,
Baltimore, Maryland USA, June 26–27, 2014.

2 Related work

Most work in the WSD literature has modelled disambiguation using a limited window of context around the word to disambiguate. Cai et al. (2007), Boyd-graber and Blei (2007) and Li et al. (2010) further tried to integrate the notion of latent topics to address the sparsity problem of the lexicalised features typically used in WSD classifiers. The most closely related work in the area of sense disambiguation is by Dinu and Lapata (2010) who propose a disambiguation method for solving lexical similarity and substitution tasks. They measure word similarity in context by learning distributions over senses for each target word in the form of lower-dimensional distributional representations. Before computing word similarities, they contextualise the global sense distribution of a word using the sense distribution of words in the test context, thereby shifting the sense distribution towards the test context. We adopt a similar distributional representation, but argue that our representation does not need this disambiguation step because at the level of phrase pairs the ambiguity is already much reduced.

Our model performs adaptation using similarity features which is similar to the approach of Costa-jussà and Banchs (2010) who learn a vector space model that captures the source context of every training sentence. In Banchs and Costa-jussà (2011), the vector space model is replaced with representations inferred by Latent Semantic Indexing. However, because their latent representations are learned over training sentences, they have to compare the current test sentence to the latent vector of every training instance associated with a translation unit. The highest similarity value is then used as a feature value. Instead, our model learns latent distributional representations of phrase pairs that can be directly compared to test contexts and are likely to be more robust. Because context words of a phrase pair are tied together in the distributional representations, we can use sparse priors to cluster context words associated with the same phrase pair into few topics.

Recently, Chen et al. (2013) have proposed a vector space model for domain adaptation where phrase pairs are assigned vectors that are defined in terms of the training corpora. A similar vector is built for an in-domain development set and the similarity to the development set is used as a feature during translation. While their vector representations are similar to our latent topic representations, their model has no notion of structure beyond corpus boundaries and is adapted towards a single target domain (*cross-domain*). Instead, our model learns the latent topical structure automatically and the translation model is adapted *dynamically* to each test instance.

We are not aware of prior work in the field of MT that investigates combinations of local and global context. In their recent work on neural language models, Huang et al. (2012) combine the scores of two neural networks modelling the word embeddings of previous words in a sequence as well as those of words from the surrounding document by averaging over all word embeddings occurring in the same document. The score of the next word in a sequence is computed as the sum of the scores of both networks, but they do not consider alternative ways of combining contextual information.

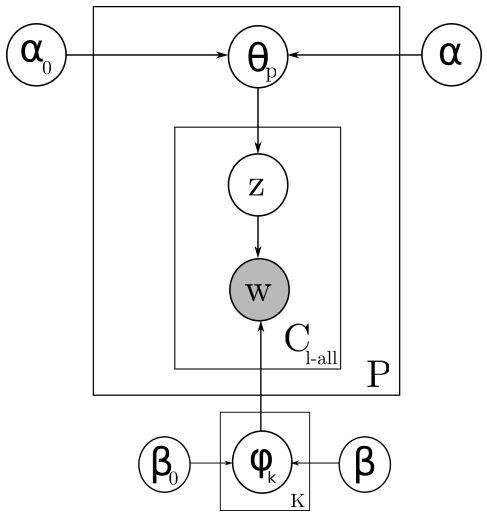
3 Phrase pair topic model (PPT)

Our proposed model aims to capture the relationship between *phrase pairs* and *source words* that frequently occur in the local context of a phrase pair, that is, context words occurring in the same sentence. It therefore follows the *distributional hypothesis* (Harris, 1954) which states that words that occur in the same contexts tend to have similar meanings. For a phrase pair, the idea is that words that occur frequently in its context are indicative of the sense that is captured by the target phrase translating the source phrase.

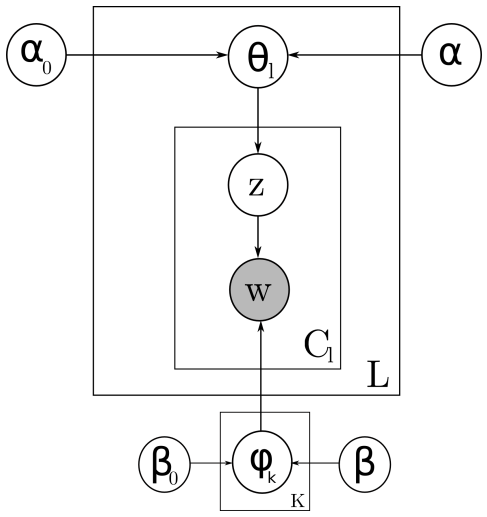
We assume that all phrase pairs share a global set of topics and during topic inference the distribution over topics for each phrase pair is induced from the latent topic of its context words in the training data. In order to learn topic distributions for each phrase pair, we represent phrase pairs as documents containing all context words from the source sentence context in the training data. These distributional profiles of phrase pairs are the input to the topic modelling algorithm which learns topic clusters over context words.

Figure 1a shows a graphical representation of the following generative process for training. For each of P phrase pairs pp_i in the collection

1. Draw a topic distribution from an asymmetric Dirichlet prior, $\theta_p \sim \text{Dirichlet}(\alpha_0, \alpha \ldots \alpha)$.
2. For each position c in the distributional profile of pp_i, draw a topic from that distribution, $z_{p,c} \sim \text{Multinomial}(\theta_p)$.



(a) Inference on phrase pair documents (training).



(b) Inference on local test contexts (test).

Figure 1: Graphical representation of the phrase pair topic (PPT) model.

3. Conditioned on topic $z_{p,c}$, choose a context word $w_{p,c} \sim$ Multinomial($\psi_{z_{p,c}}$).

α and β are parameters of the Dirichlet distributions and ϕ_k denotes topic-dependent vocabularies over context words. Test contexts are generated similarly by drawing topic mixtures θ_l for each test context[1] as shown in Figure 1b, drawing topics z for each context position and then drawing context words w for each z. The asymmetric prior on topic distributions (α_0 for topic 0 and α for all other topics) encodes the intuition that there are words occurring in the context of many phrase pairs which can be grouped under a topic with higher a priori probability than the other topics. Figure 1a shows the model for training inference on the distributional representations for each phrase pair, where C_{l-all} denotes the number of context words in all sentence contexts that the phrase pair was seen in the training data, P denotes the number of phrase pairs and K denotes the number of latent topics. The model in Figure 1b has the same structure but shows inference on test contexts, where C_l denotes the number of context words in the test sentence context and L denotes the number of test instances. θ_p and θ_l denote the topic distribution for a phrase pair and a test context, respectively.

3.1 Inference for PPT model

We use collapsed variational Bayes (Teh et al., 2006) to infer the parameters of the PPT model. The posterior distribution over topics is computed as shown below

$$P(z_{p,c} = k|\mathbf{z}^{-(p,c)}, \mathbf{w_c}, p, \alpha, \beta) \propto \frac{(\mathbb{E}_{\hat{q}}[n_{.,k,w_c}^{-(p,c)}] + \beta)}{(\mathbb{E}_{\hat{q}}[n_{.,k,.}^{-(p,c)}] + W_c \cdot \beta)} \cdot (\mathbb{E}_{\hat{q}}[n_{d,k,.}^{-(p,c)}] + \alpha) \quad (1)$$

where $z_{p,c}$ denotes the topic at position c in the distributional profile p, $\mathbf{w_c}$ denotes all context word tokens in the collection, W_c is the total number of context words and $\mathbb{E}_{\hat{q}}$ is the expectation under the variational posterior. $n_{.,k,w_c}^{-(p,c)}$ and $n_{p,k,.}^{-(p,c)}$ are counts of topics occurring with context words and distributional profiles, respectively, and $n_{.,k,.}^{-(p,c)}$ is a topic occurrence count.

Before training the topic model, we remove stop words from all documents. When inferring topics for test contexts, we ignore unseen words because they do not contribute information for topic inference. In order to speed up training inference, we limit the documents in the collection to those corresponding to phrase pairs that are needed to translate the test set[2]. Inference was run for 50 iterations on the distributional profiles for training and for 10 iterations on the test contexts. The output of the training inference step is a model file with all the necessary statistics to compute posterior topic distributions (which are loaded before running test inference), and the set of topic vectors for all phrase pairs. The output of test inference is

[1] A local test context is defined as all words in the test sentence excluding stop words, while contexts of phrase pairs in training do not include the words belonging to the source phrase. The naming in the figure refers to local test contexts L, but global test contexts will be defined similarly.

[2] Reducing the training contexts by scaling or sampling would be expected to speed up inference considerably.

the set of induced topic vectors for all test contexts.

3.2 Modelling local and global context

At training time, our model has access to context words only from the local contexts of each phrase pair in their distributional profiles, that is, other words in the same source sentence as the phrase pair. This is useful for reducing noise and constraining the semantic space that the model considers for each phrase pair during training. At test time, however, we are not limited to applying the model only to the immediate surroundings of a source phrase to disambiguate its meaning. We can potentially take any size of test context into account to disambiguate the possible senses of a source phrase, but for simplicity we consider two sizes of context here which we refer to as local and global context.

Local context Words appearing in the sentence around a test source phrase, excluding stop words.

Global context Words appearing in the document around a test source phrase, excluding stop words.

4 Similarity features

We define similarity features that compare the topic vector θ_p assigned to a phrase pair[3] to the topic vector assigned to a test context, The feature is defined for each source phrase and all its possible translations in the phrase table, as shown below

$$sim(pp_i, test\ context) = cosine(\theta_{p_i}, \theta_c), \quad \forall pp_i \in \{pp_i | \bar{s} \rightarrow \bar{t_i}\} \qquad (2)$$

Unlike Banchs and Costa-jussà (2011), we do not learn topic vectors for every training sentence which results in a topic vector per phrase pair token, but instead we learn topic vectors for each phrase pair type. This is more efficient but also more appealing from a modelling point of view, as the topic distributions associated with phrase pairs can be thought of as expected latent contexts. The application of the similarity feature is visualised in Figure 2. On the left, there are two applicable phrase pairs for the source phrase *noyau*, *noyau → kernel* and *noyau → nucleus*, with their distributional representations (words belonging to the *IT* topic versus the *scientific* topic) and assigned topic vectors θ_p. The local and global test contexts are similarly represented by a document containing the context words and a resulting topic vector θ_l or θ_g. The test context vector θ_c can be one of θ_l and θ_g or a combination of both. In this example, the distributional representation of *noyau → kernel* has a larger topical overlap with the test context and will more likely be selected during decoding.

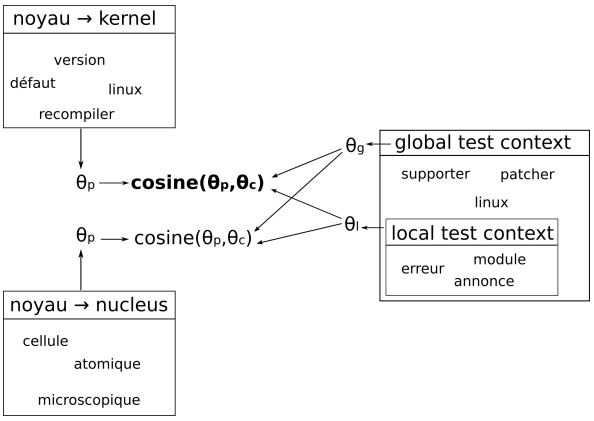


Figure 2: Similarity between topic vectors of two applicable phrase pairs θ_p and the topic vectors θ_l and θ_g from the local and global test context during test time.

While this work focuses on exploring vector space similarity for adaptation, mostly for computational ease, it may be possible to derive probabilistic translation features from the PPT model. This could be a useful addition to the model and we leave this as an avenue for future work.

Types of similarity features

We experiment with local and global phrase similarity features, *phrSim-local* and *phrSim-global*, to perform dynamic topic adaptation. These two similarity features can be combined by adding them both to the log-linear SMT model, in which case each receive separate feature weights. Whenever we use the + symbol in our results tables, the additional features were combined with existing features log-linearly. However, we also experimented with an alternative combination of local and global information where we combine the local and global topic vectors for each test context before computing similarity features.[4] We were

[3] The mass of topic 0 is removed from the vectors and the vectors are renormalised before computing similarity features.

[4] The combined topic vectors were renormalised before computing their similarities with each candidate phrase pair.

motivated by the observation that there are cases where the local and global features have an opposite preference for one translation over another, but the log-linear combination can only learn a global preference for one of the features. Combining the topic vectors allows us to potentially encode a preference for one of the contexts that depends on each test instance.

For similarity features derived from combined topic vectors, $\oplus$ denotes the additive combination of topic vectors, $\otimes$ denotes the multiplicative combination of topic vectors and $\circledast$ denotes a combination that favours the local context for longer sentences and backs off incrementally to the global context for shorter sentences.[5] The intuition behind this combination is that if there is already sufficient evidence in the local context, the local topic mixture may be more reliable than the global mixture.

We also experiment with a combination of the phrase pair similarity features derived from the PPT model with a document similarity feature from the pLDA model described in Hasler et al. (2014). The motivation is that their model learns topic mixtures for documents and uses phrases instead of words to infer the topical context. Therefore, it might provide additional information to our similarity features.

5 Data and experimental setup

Our experiments were carried out on a mixed French-English data set containing the TED corpus (Cettolo et al., 2012), parts of the News Commentary corpus (NC) and parts of the Commoncrawl corpus (CC) from the WMT13 shared task (Bojar et al., 2013) as described in Table 1. To ensure that the baseline model does not have an implicit preference for any particular domain, we selected subsets of the NC and CC corpora such that the training data contains 2.7M English words per domain. We were guided by two constraints in chosing our data set in order to simulate an environment where very diverse documents have to be translated, which is a typical scenario for web translation engines: 1) the data has document boundaries and the content of each document is assumed to be topically related, 2) there is some degree of topical variation within each data set. This setup allows us to evaluate our dynamic topic adaptation approach because the test documents are from different domains and also differ within each domain, which makes lexical selection a much harder problem. The topic adaptation approach does not make use of the domain labels in training or test, because it infers topic mixtures in an unsupervised way. However, we compare the performance of our dynamic approach to domain adaptation methods by providing them the domain labels for each document in training and test.

In order to abstract away from adaptation effects that concern tuning of length penalties and language models, we use a mixed tuning set containing data from all three domains and train one language model on the concatenation of the target sides of the training data. Word alignments are trained on the concatenation of all training data and fixed for all models. Table 2 shows the average length of a document for each domain. While a CC document contains 29.1 sentences on average, documents from NC and TED are on average more than twice as long. The length of a document could have an influence on how reliable global topic information is but also on how important it is to have information from both local and global test contexts.

Data	Mixed	CC	NC	TED
Train	354K (6450)	110K	103K	140K
Dev	2453 (39)	818	817	818
Test	5664 (112)	1892	1878	1894

Table 1: Number of sentence pairs and documents (in brackets) in the data sets.

Data	CC	NC	TED
Test documents	65	31	24
Avg sentences/doc	29.1	60.6	78.9

Table 2: Average number of sentences per document in the test set (per domain).

5.1 Unadapted baseline system

Our baseline is a phrase-based French-English system trained on the concatenation of all parallel data. It was built with the Moses toolkit (Koehn et al., 2007) using the 14 standard core features including a 5-gram language model. Translation quality is evaluated on a large test set, using the average feature weights of three optimisation runs with PRO (Hopkins and May, 2011). We use the

[5] The interpolation weights between local and global topic vectors were set proportional to sentence lengths between 1 and 30. The length of longer sentences was clipped to 30.

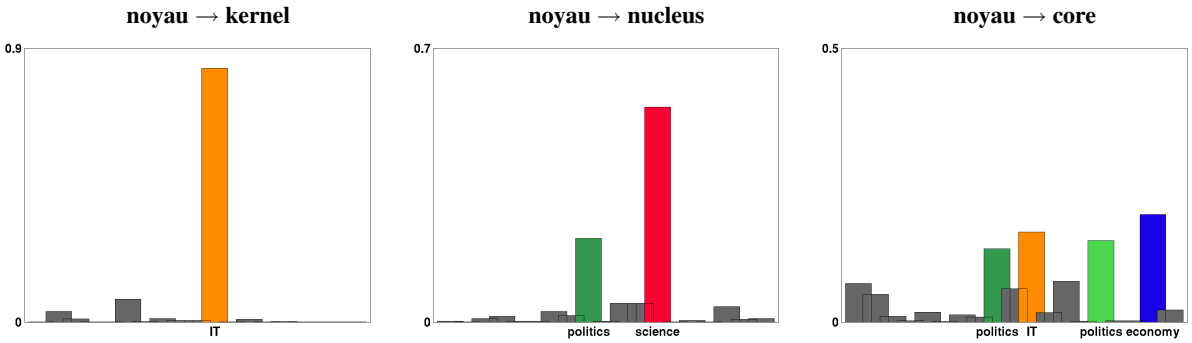


Figure 3: Topic distributions for source phrase *noyau* and three of its translations (20 topics without topic 0). Colored bars correspond to topics *IT*, *politics*, *science*, *economy* with topic proportions $\geq$10%.

mteval-v13a.pl script to compute case-insensitive BLEU scores.

5.2 Domain-adapted benchmark systems

As domain-aware benchmark systems, we use the linear mixture model (DOMAIN1) of Sennrich (2012) and the phrase table fill-up method (DOMAIN2) of Bisazza et al. (2011) (both available in the Moses toolkit). For both systems, the domain labels of the documents are used to group documents of the same domain together. We build adapted tables for each domain by treating the remaining documents as out-of-domain data and combining in-domain with out-of-domain tables. For development and test, the domain labels are used to select the respective domain-adapted model for decoding. Both systems have an advantage over our model because of their knowledge of domain boundaries in the data. This allows for much more confident lexical choices than using an unadapted system but is not possible without prior knowledge about each document.

5.3 Implementation of similarity features

After all topic vectors have been computed, a feature generation step precomputes the similarity features for all pairs of test contexts and applicable phrase pairs for translating source phrases in a test instance. The phrase table of the baseline model is filtered for every test instance (a sentence or document, depending on the context setting) and each entry is augmented with features that express its semantic similarity to the test context. We use a wrapper around the Moses decoder to reload the phrase table for each test instance, which enables us to run parameter optimisation (PRO) in the usual way to get one set of tuned weights for all test sentences. It would be conceivable to use topic-specific weights instead of one set of global weights, but this is not the focus of this work.

6 Qualitative evaluation of phrase pair topic distributions

In order to verify that the topic model is learning useful topic representations for phrase pairs, we inspect the inferred topic distributions for three phrase pairs where the translation of the same source word differs depending on the topical context: *noyau* → *kernel*, *noyau* → *nucleus* and *noyau* → *core*. Figure 3 shows the topic distributions for a PPT model with 20 topics (with topic 0 removed) and highlights the most prominent topics with labels describing their content (politics, IT, science, economy)[6]. The most peaked topic distribution was learned for the phrase pair *noyau* → *kernel* which would be expected to occur mostly in an IT context and the topic with the largest probability mass is in fact related to IT. The most prominent topic for the phrase pair *noyau* → *nucleus* is the science topic, though it seems to be occurring in with the political topic as well. The phrase pair *noyau* → *core* was assigned the most ambiguous topic distribution with peaks at the politics, economy and IT topics. Note also that its topic distribution overlaps with those of the other translations, for example, like the phrase pair *noyau* → *kernel*, it can occur in IT contexts. This shows that the model captures the fact that even within a given topic there can still be ambiguity about the correct translation (both target phrases *kernel* and *core* are plausible translations in an IT context).

[6]Topic labels were assigned by inspecting the most probable context words for each topic according to the model.

Ambiguity of phrase pair topic vectors
The examples in the previous section show that the level of ambiguity differs between phrase pairs that constitute translations of the same source phrase. It is worth noting that introducing bilingual information into topic modelling reduces the sense ambiguity present in monolingual text by preserving only the intersection of the senses of source and target phrases. For example, the distributional profiles of the source phrase *noyau* would contain words that belong to the senses *IT*, *politics*, *science* and *economy*, while the words in the context of the target phrase *kernel* can belong to the senses *IT* and *food* (with source context words such as *grain*, *protéines*, *produire*). Thus, the monolingual representations would still contain a relatively high level of ambiguity while the distributional profile of the phrase pair *noyau* → *kernel* preserves only the *IT* sense.

7 Results and discussion

In this section we present experimental results of our model with different context settings and against different baselines. We used bootstrap resampling (Koehn, 2004) to measure significance on the mixed test set and marked all statistically significant results compared to the respective baselines with asterisk (*: $p \leq 0.01$).

7.1 Local context

In Table 3 we compare the results of the concatenation baseline and a model containing the *phrSim-local* feature in addition to the baseline features, for different numbers of latent topics. We show results for the mixed test set containing documents from all three domains as well as the individual results on the documents from each domain. While all topic settings yield improvements over the baseline, the largest improvement on the mixed test set (+0.48 BLEU) is achieved with 50 topics. Topic adaptation is most effective on the TED portion of the test set where the increase in BLEU is 0.59.

7.2 Global context

Table 4 shows the results of the baseline plus the *phrSim-global* feature that takes into account the whole document context of a test sentence. While the largest overall improvement on the mixed test set is equal to the improvement of the local feature, there are differences in performance for the individual domains. For Commoncrawl documents, the results vary slightly but the largest improvement is still achieved with 50 topics and is almost the same for both. For News Commentary, the scores with the local feature are consistently higher than the scores with the global feature (0.20 and 0.22 BLEU higher for 20 and 50 topics). For TED, the trend is opposite with the global feature performing better than the local feature for all topics (0.28 and 0.40 BLEU higher for 10 and 20 topics). The best improvement over the baseline for TED is 0.83 BLEU, which is higher than the improvement with the local feature.

Model	Mixed	CC	NC	TED
Baseline	26.86	19.61	29.42	31.88
10 topics	*27.15	19.87	29.63	32.36
20 topics	*27.19	19.92	**29.76**	32.31
50 topics	***27.34**	**20.13**	29.70	**32.47**
100 topics	*27.26	20.02	29.75	32.40
>Baseline	+0.48	+0.52	+0.34	+0.59

Table 3: BLEU scores of baseline system + *phrSim-local* feature for different numbers of topics.

Model	Mixed	CC	NC	TED
Baseline	26.86	19.61	29.42	31.88
10 topics	*27.30	20.01	29.61	32.64
20 topics	***27.34**	20.07	29.56	**32.71**
50 topics	*27.27	**20.12**	29.48	32.55
100 topics	*27.24	19.95	**29.66**	32.52
>Baseline	+0.48	+0.51	+0.24	+0.83

Table 4: BLEU scores of baseline system + *phrSim-global* feature for different numbers of topics.

7.3 Relation to properties of test documents

To make these results more interpretable, Table 5 lists some of the properties of the test documents per domain. Of the three domains, CC has the shortest documents on average and TED the longest. To understand how this affects topic inference, we measure topical drift as the average divergence (cosine distance) of the local topic distributions for each test sentence to the global topic distribution of their surrounding document. There seems to be a correlation between document length and topical drift, with CC documents showing the least topical drift and TED documents showing the most. This makes sense intuitively

because the longer a document is, the more likely it is that the content of a given sentence diverges from the overall topical structure of the document. While this can explain why for CC documents using local or global context results in similar performance, it does not explain the better performance of the local feature for NC documents. The last row of Table 5 shows that sentences in the NC documents are on average the longest and longer sentences would be expected to yield more reliable topic estimates than shorter sentences. Thus, we assume that local context yields better performance for NC because on average the sentences are long enough to yield reliable topic estimates. When local context provides reliable information, it may be more informative than global context because it can be more specific.

For TED, we see the largest topical drift per document, which could lead us to believe that the document topic mixtures do not reflect the topical content of the sentences too well. But considering that the sentences are on average shorter than for the other two domains, it is more likely that the local context in TED documents can be unreliable when the sentences are too short. TED documents contain transcribed speech and are probably less dense in terms of information content than News commentary documents. Therefore, the global context may be more informative for TED which could explain why relying on the global topic mixtures yields better results.

Property	**CC**	**NC**	**TED**
Per document			
Avg number of sentences	29.1	60.6	78.9
Avg topical divergence	0.35	0.43	0.49
Avg sentence length	26.2	31.5	21.7

Table 5: Properties of test documents per domain. Average topical divergence is defined as the average cosine distance of local to global topic distributions in a document.

7.4 Combinations of local and global context

In Table 6 we compare a system that already contains the global feature from a model with 50 topics to the combinations of local and global similarity features described in Section 4.

Of the four combinations, the additive combination of topic vectors ($\oplus$) yields the largest improvement over the baseline with +0.63 BLEU on the mixed test set and +0.88 BLEU on TED. The improvements of the combined model are larger than the improvements for each context on its own, with the only exception being the NC portion of the test set where the improvement is not larger than using just the local context. A possible reason is that when one feature is consistently better for one of the domains (local context for NC), the log-linear combination of both features (tuned on data from all domains) would result in a weaker overall model for that domain. However, if both features encode similar information, as we assume to be the case for CC documents, the presence of both features would reinforce the preference of each and result in equal or better performance. For the additive combination, we expect a similar effect because adding together two topics vectors that have peaks at different topics would make the resulting topic vector less peaked than either of the original vectors.

Model	**Mixed**	**CC**	**NC**	**TED**
Baseline	26.86	19.61	29.42	31.88
+ global	27.27	20.12	29.48	32.55
+ local	*27.43	20.18	29.65	**32.79**
$\oplus$ local	***27.49**	**20.30**	**29.66**	32.76
$\otimes$ local	27.34	20.24	29.61	32.50
$\circledast$ local	*27.45	20.22	29.51	**32.79**
$\oplus$ >BL	+0.63	+0.69	+0.24	+0.88

Table 6: BLEU scores of baseline and combinations of phrase pair similarity features with local and global context (significance compared to baseline+global). All models were trained with 50 topics.

The additive topic vector combination is slightly better than the log-linear feature combination, though the difference is small. Nevertheless, it shows that combining topic vectors before computing similarity features is a viable alternative to log-linear combination, with the potential to design more expressive combination functions. The multiplicative combination performs slightly worse than the additive combination, which suggests that the information provided by the two contexts is not always in agreement. In some cases, the global context may be more reliable while in other cases the local context may have more accurate topic estimates and a voting approach does not take advantage of complementary information. The combination of topic vectors

Source: Le **noyau** contient de nombreux pilotes, afin de fonctionner chez la plupart des utilisateurs.
Reference: The precompiled **kernel** includes a lot of drivers, in order to work for most users.

Source: Il est prudent de consulter les pages de manuel ou les faq spécifiques à votre **os**.
Reference: It's best to consult the man pages or faqs for your **os**.

Source: Nous fournissons nano (un petit éditeur), vim (vi amélioré), qemacs (clone de emacs), **elvis**, joe .
Reference: Nano (a lightweight editor), vim (vi improved), qemacs (emacs clone), **elvis** and joe.

Source: Elle a introduit des politiques [..] à coté des **relations** de gouvernement à gouvernement traditionnelles.
Reference: She has introduced policies [..] alongside traditional government-to-government **relations**.

Figure 4: Examples of test sentences and reference translations with the ambiguous source words and their translations in bold.

depending on sentence length ($\circledast$) performs well for CC and TED but less well for NC where we would expect that it helps to prefer the local information. This indicates that the rather ad-hoc way in which we encoded dependency on the sentence length may need further refinement to make better use of the local context information.

Model	**noyau** →	**os** →
Baseline	nucleus	bones
global	kernel*	os*
local	nucleus	bones
global$\oplus$local	kernel*	os*

Table 7: Translations of ambiguous source words where global context yields the correct translation (* denotes the correct translation).

Model	**elvis** →	**relations** →
Baseline	elvis*	relations*
global	the king	relationship
local	elvis*	relations*
global$\oplus$local	the king	relations*

Table 8: Translations of ambiguous source words where local context yields the correct translation (* denotes the correct translation).

7.5 Effect of contexts on translation

To give an intuition of how lexical selection is affected by contextual information, Figure 4 shows four test sentences with an ambiguous source word and its translation in bold. The corresponding translations with the baseline, the global and local similarity features and the additive combination are shown in Table 7 for the first two examples where the global context yields the correct translation (as indicated by *) and in Table 8 for the last two examples where the local context yields the correct translation.[7] In Table 7, the additive combination preserves the choice of the global model and yields the correct translations, while in Table 8 only the second example is translated correctly by the combined model. A possible explanation is that the topical signal from the global context is stronger and results in more discriminative similarity values. In that case, the preference of the global context would be likely to have a larger influence on the similarity values in the combined model. A useful extension could be to try to detect for a given test instance which context provides more reliable information (beyond encoding sentence length) and boost the topic distribution from that context in the combination.

7.6 Comparison with domain adaptation

Table 9 compares the additive model ($\oplus$) to the two domain-adapted systems that know the domain label of each document during training and test. Our topic-adapted model yields overall competitive performance with improvements of +0.37 and +0.25 BLEU on the mixed test set, respectively. While it yields slightly lower performance on the NC documents, it achieves equal performance on TED documents and improves by up to +0.94 BLEU on Commoncrawl documents. This can be explained by the fact that Commoncrawl is the most diverse of the three domains with documents crawled from all over web, thus we expect topic adaptation to be most effective in comparison to domain adaptation in this scenario. Our dynamic approach allows us to adapt the similarity features to each test sentence and test document individually and is therefore more flexible than

[7] For these examples, the local model happens to yield the same translations as the baseline model.

Type of adaptation	Model	Mixed	CC	NC	TED
Domain-adapted	DOMAIN1	27.24	19.61	29.87	32.73
	DOMAIN2	27.12	19.36	29.78	32.71
Topic-adapted	global $\oplus$ local	***27.49**	**20.30**	29.66	**32.76**
	>DOMAIN1	+0.25	+0.69	-0.21	+0.03
	>DOMAIN2	+0.37	+0.94	-0.12	+0.05

Table 9: BLEU scores of translation model using similarity features derived from PPT model (50 topics) in comparison with two (supervised) domain-adapted systems.

Model	Mixed	CC	NC	TED
Baseline	26.86	19.61	29.42	31.88
+ docSim	27.22	20.11	29.63	32.40
+ phrSim-global $\oplus$ phrSim-local	*27.58	20.34	**29.71**	32.96
+ phrSim-global $\circledast$ phrSim-local	***27.60**	**20.35**	29.70	**33.03**
global$\circledast$local>BL	+0.74	+0.74	+0.38	+1.15

Table 10: BLEU scores of baseline, baseline + document similarity feature and additional phrase pair similarity features (significance compared to baseline+docSim). All models were trained with 50 topics.

cross-domain adaptation approaches while requiring no information about the domain of a test instance.

7.7 Combination with an additional document similarity feature

To find out whether similarity features derived from different types of topic models can provide complementary information, we add the *phrSim* features to a system that already includes a document similarity feature (*docSim*) derived from the pLDA model (Hasler et al., 2014) which learns topic distributions at the document level and uses phrases instead of words as the minimal units. The results are shown in Table 10. Adding the two best combinations of local and global context from Table 6 yields the best results on TED documents with an increase of 0.63 BLEU over the baseline + *docSim* model and 1.15 BLEU over the baseline. On the mixed test set, the improvement is 0.38 BLEU over the baseline + *docSim* model and 0.74 BLEU over the baseline. Thus, we show that combining different scopes and granularities of similarity features consistently improves translation results and yields larger gains than using each of the similarity features alone.

8 Conclusion

We have presented a new topic model for dynamic adaptation of machine translation systems that learns topic distributions for phrase pairs. These latent topic representations can be compared to latent representations of local or global test contexts and integrated into the translation model via similarity features.

Our experimental results show that it is beneficial for adaptation to use contextual information from both local and global contexts, with BLEU improvements of up to 1.15 over the baseline system on TED documents and 0.74 on a large mixed test set with documents from three domains. Among four different combinations of local and global information, we found that the additive combination of topic vectors performs best. We conclude that information from both contexts should be combined to correct potential topic detection errors in either of the two contexts. We also show that our dynamic adaptation approach performs competitively in comparison with two supervised domain-adapted systems and that the largest improvement is achieved for the most diverse portion of the test set.

In future work, we would like to experiment with more compact distributional profiles to speed up inference and explore the possibilities of deriving probabilistic translation features from the PPT model as an extension to the current model. Another avenue for future work could be to combine contextual information that captures different types of information, for example, to distinguish between semantic and syntactic aspects in the local context.

Acknowledgements

This work was supported by funding from the Scottish Informatics and Computer Science Alliance (Eva Hasler) and funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE) and grant agreement 288769 (ACCEPT). Thanks to Annie Louis for helpful comments on a draft of this paper and thanks to the anonymous reviewers for their useful feedback.

References

Rafael E Banchs and Marta R Costa-jussà. 2011. A Semantic Feature for Statistical Machine Translation. In *SSST-5 Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 126–134.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proceedings of IWSLT*.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of WMT 2013*.

Jordan Boyd-graber and David Blei. 2007. A Topic Model for Word Sense Disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 1024–1033.

Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. 2007. Improving Word Sense Disambiguation Using Topic Features. In *Proceedings of EMNLP*, pages 1015–1023.

Marine Carpuat and Dekai Wu. 2007a. How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In *International Conference on Theoretical and Methodological Issues in MT*.

Marine Carpuat and Dekai Wu. 2007b. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of EMNLP*, pages 61–72.

Marine Carpuat. 2009. One Translation per Discourse. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of EAMT*.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of ACL*.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector Space Model for Adaptation in Statistical Machine Translation. In *Proceedings of ACL*, pages 1285–1293.

Marta R. Costa-jussà and Rafael E. Banchs. 2010. A vector-space dynamic feature for phrase-based statistical machine translation. *Journal of Intelligent Information Systems*, 37(2):139–154, August.

Georgiana Dinu and Mirella Lapata. 2010. Measuring Distributional Similarity in Context. In *Proceedings of EMNLP*, pages 1162–1172.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of ACL*.

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

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2012. Sparse Lexicalised Features and Topic Adaptation for SMT. In *Proceedings of IWSLT*.

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.

Sanjika Hewavitharana, Dennis N Mehay, and Sankaranarayanan Ananthakrishnan. 2013. Incremental Topic-Based Translation Model Adaptation for Conversational Spoken Language Translation. In *Proceedings of ACL*, pages 697–701.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of ACL*, pages 873–882.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for SMT. In *Proceedings of ACL: Demo and poster sessions*.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of EMNLP*.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic Models for Word Sense Disambiguation and Token-based Idiom Detection. In *Proceedings of ACL*, pages 1138–1147.

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of EACL*.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2008. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207, November.

Yee Whye Teh, David Newman, and Max Welling. 2006. A collapsed variational Bayesian inference algorithm for LDA. In *Proceedings of NIPS*.

B Zhao and E P Xing. 2007. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. *Neural Information Processing*.