Topic Adaptation for Machine Translation of E-commerce Content

Prashant Mathur prashant@fbk.eu
Marcello Federico* federico@fbk.eu
FBK, Trento, Italy

Selçuk Köprü skopru@ebay.com
Sharam Khadivi skhadivi@ebay.com
Hassan Sawaf hsawaf@ebay.com
eBay Inc., San Jose (CA), USA

Abstract

We describe effort to improve machine translation of item titles found in a large e-commerce inventory through topic modeling and adaptation. Item titles are short texts which typically contain brand names that do not have to be translated, and item attributes whose translation often depends on the context. Both issues call for robust methods to integrate context information in the machine translation process in order to reduce translation ambiguity. We survey both existing topic adaptation approaches and propose novel methods that augment the standard phrase-table models with sparse features and dense features measuring the topic match between each phrase-pair and the input text. We report extensive experiments on the translation of item titles from English into Brazilian Portuguese, and show the impact of topic adaptation both with and without domain adaptation.

1 Introduction

Domain adaptation and topic adaptation can be seen as complementary methods to cope with the variability between training and testing data in machine translation. Under this perspective, domain modeling typically assumes training data partitioned according to human defined labels, while topic modeling builds on fuzzy clustering of the data and automatically learned labels. In statistical machine translation, knowledge about the domain or topics of the input can be leveraged to bias the system towards training data matching the same labels of the input. Potential advantages of topic adaptation over domain adaptation is that fuzzy clustering can better cope with data sparseness than hard clustering, and that automatic labels do not require any manual intervention. On the other side, however, domain adaptation becomes difficult to beat when training data can be effectively and naturally partitioned into in-domain and out-of-domain data.

In this paper we discuss the application of domain and topic adaptation to an e-commerce online MT system (Guha and Heger, 2014), whose target is the translation of user queries and all item titles, descriptions, and specifics shown in the search result pages. In particular, our investigation focuses on the translation of item titles, which consists of concise user-generated texts describing each item put on sale. Item titles differ in several ways from text genres typically considered in machine translation research. Titles are usually short texts, of maximum

*Most of the work was carried out during a visiting period of the first two authors at eBay Inc.

100 characters, with a simple syntactic structure, and containing brand names, feature values, as well as specific jargon. Their translation poses several challenges (Sanchez and Badeka, 2014), such as the correct rendering of proper names, which can often be confused with common names, and the correct translation of product features, which often depends on the context. From a statistical learning perspective, MT of item titles is also hard because of the large variety of content present in eBay's inventory, which are very unevenly populated but also significantly overlapping in terms of linguistic content.

The idea that we follow in this work is to employ topic modeling to better translate English item titles to a foreign language. Since we have a relatively small amount of bilingual in-domain data compared to bilingual out-of-domain data and English in-domain monolingual data, we aim to apply topic adaptation on the bilingual data and to train a topic model on the monolingual data (see Section 2). Then, we enrich in-domain and out-of-domain parallel data with topic information and embed this in the translation model of the MT system. At testing time, we infer the topics of the input and use them to dynamically adapt the MT system. In this work we survey and compare different topic adaptation methods from the recent literature and measure their impact on translation performance with and without domain adaptation. We report experimental result with a Moses-based phrase-based system on the English - Brazilian Portuguese language pair. This particular pair is one of the active language pairs at eBay. While for domain adaptation we deploy a state-of-the-art method, for topic adaptation we investigate the use of new sparse features which we compare against other features proposed in the literature.

The rest of the paper is arranged as follows. We first introduce topic modeling applied to our e-commerce content in Section 2, then we survey previous work on topic adaptation for statistical MT in Section 3, afterwards we present our take on topic adaptation in Section 4, and finally we report on our experimental set-up and results in Sections 5 and 6. We end the paper with some conclusions and future work.

2 Topic Modeling

2.1 Content

In eBay, machine translation plays an important role to facilitate cross-border trade between sellers and buyers with different languages (Guha and Heger, 2014). eBay is a marketplace where sellers can advertise their items on the site and buyers can search for the items and then electronically bid for them. To enable a trade between buyers and sellers with different languages, at least four types of texts need to be translated: queries, item titles, descriptions, and item specifics. This work focuses on the translation of item titles, which are concise and usually very informative descriptions of the items put on sale. For instance, the item title:

new men's white jekyll & hyde jeans winston designer regular fit shirt size s-xxl

specifies, in order, the condition, target gender, color, brand, designer, fit, item type, and size of the product. Common challenges in the translation of eBay's user generated content in general, and of titles (Sanchez and Badeka, 2014) and queries (Picinini, 2014) in particular, are the proper rendering of proper names and the translation of words which can have multiple senses, depending on the context in which they appear. For example, the word "age" might have different meanings if context is "baby", "wine" or "collectibles". Similarly, the word "j-hook" might have different meanings if context is "motors" or "garden".

The core idea of this work, hence is to apply topic modeling to efficiently represent the context of single words or expressions in order to improve the accuracy of their translation by a phrase-based statistical MT system. In the following, we describe the monolingual data and the topic model that we used for this purpose.

2.2 Sampling

The amount of monolingual item titles available to eBay is huge and very unevenly distributed across different levels of categories of eBay's inventory. Actually, items in eBay's inventory are classified according to a hierarchical taxonomy. The hierarchy itself contains 34 top-level categories (L1) with varying degrees of depth in each category (L2=400 and L3=4000 categories). For example, the top level category "Books" has eleven second-level categories (L2) and each of those categories have anywhere from four to thirty categories, with many of these having subcategories as each topic becomes more and more specific. All traded items are placed in the leaves of this hierarchy.

For the sake of experimentation, we first perform sampling in order to collect a more balanced and manageable collection of titles. Hence, starting from a collection of billions of item titles, after stratified sampling from each L2 category we end up with a collection of 4.3M item titles. Using a uniform sampling method, we further sub-sampled data (from each L1 category) to around 708K item titles, which is actually used to train the topic model.

2.3 Models

Topic models can be trained from an arbitrary document or sentence collections with different methods, such as probabilistic latent semantic analysis (Hofmann, 1999), hidden topic Markov models (Gruber et al., 2007), and latent Dirichlet allocation (LDA) (Blei et al., 2003). For all methods, there are existing software tools that allow to train topic models after specifying the desired number of topics and a few training options. All tools permit then to use a training model to infer a topic distribution for a given sentence. In our case, we deployed a LDA model trained with the Standford TMT[1] tool. In particular, we worked with a training configuration using 30 topics (similar to size of L1 categories) and 1000 iterations. We also experimented with different number of topics but empirically found 30 to be the optimum number. Moreover, we excluded all item titles with less than 5 words, all words occurring less than 3 times, and all words made of less than 3 characters. The pruning steps were performed to exclude from the model item titles providing too little context, words which are too infrequent, and very frequent and short words that to not bear any topic information.

In the following table, we report the 15 most relevant words of the first 10 topics trained with the LDA model. As can be seen, some of the topics are easily recognizable, e.g. T01 (toys), T04 (electronics), T05 (photo), and T09 (fashion), and T10 (hunting and fishing). The other topics look instead combinations of multiple categories, e.g. T06 seems a combination of pet supplies and fashion.

T01	doll high barbie monster little dolls girl lot fashion dress pony and american hair with
T02	free shipping fish aquarium beads tank diy ship round glass pcs wholesale plastic water craft
T03	screen car baby lcd seat glass replacement touch protector cover cloth diaper safety holder glasses
T04	digital control module with remote power sensor switch board lcd meter system arduino air kit
T05	camera lens canon nikon mount digital video with sony gopro dslr camcorder adapter black hero
T06	dress pet dog clothes coat long winter size sweater women puppy warm apparel jacket fashion
T07	nail set art brush hair color makeup gel eye kit cream polish powder tool skins
T08	figure anime movie poster action disney hot mask toys sex toy series japan prop batman
T09	size black jacket mens shoes leather boots large blue womens nwt medium ski white men's
T10	knife steel tool stainless set blade folding with pocket gun black tools handle hunting fishing

Table 1: LDA topic model of item titles: top 15 relevant words of the first 10 topics.

Finally, by using the same tool, the topic model is applied to annotate with topics all the

[1] http://nlp.stanford.edu/software/tmt/tmt-0.4/

sentences in the source side of all the available parallel data. As a result, each sentence of the parallel training data is associated with a topic vector or distribution.

3 Related Work

Topic models have been investigated in the statistical MT literature to enhance both linear or hierarchical phrase-based models with additional context-topical information derived from the training and testing data.

Previous works have approached this issue under different perspectives, also providing different levels of integration of topic information. For instance, in (Gong et al., 2011) and (Ruiz and Federico, 2011) authors devised and exploited cross-lingual topic models to generate topic relevant target words for an entire document or sentence. This differs from other approaches and our one, which instead exploit monolingual topic models in the source language to directly enhance the selection process of single phrase translations pairs during decoding.

In (Eidelman et al., 2012), topic dependent lexical probabilities are directly integrated in a hierarchical phrase-based system. The authors start with topic distributions inferred at the document level on the source side of the parallel data. After extraction of the translation rules, topic-conditional translation probabilities are inferred by computing the expectation over the topic vectors observed in all the sentences where each translation rule was extracted from. At decoding time, these probabilities are weighted by the topic prior inferred on the test document. This results in a set of sparse features, one per topic, which are tuned on a development set. In this work, we implement and evaluate a similar set of sparse features for a phrase-based decoder.

In (Su et al., 2012) topic models are used to *off-line* adapt a phrase table trained on out-of-domain parallel data by using in-domain monolingual data. Topic distributions are inferred through *hidden topic Markov* models trained on monolingual sentences. Two distinct topic models are trained, one in-domain and one out-of-domain, which are mapped via a mixture model. Finally, similar to (Eidelman et al., 2012) topic-conditioned phrase translation probabilities trained on out of domain data are weighted with the in-domain topic prior probabilities. Contrary in our case, topic information is integrated just in the training phase to bias the translation model toward the in-domain data. Our purpose, instead, is to dynamically adapt the translation model at testing time.

In (Xiao et al., 2012) monolingual topic models are trained on the source side with LDA and topic vectors are associated to hiero rules. Differently from (Eidelman et al., 2012), during decoding topic posterior distribution on phrase-pairs are matched against the topic distribution of the test sentence. Matching is performed with the Hellinger divergence. In addition, a feature measuring topic sensitivity of each phrase-pair is included based on the entropy function. In our work, we integrate and compare several dense features that compute the match between the topic distributions of the input and of each phrase-pair candidate. Among them we also include the two features proposed by this work.

In (Hewavitharana et al., 2013) topic adaptation is performed in context of machine translation of task-driven conversations. Topics vectors are inferred via LDA on the source side of in-domain parallel training data. At test time, the topic vector of the conversation is incrementally updated at each turn. During decoding, with a Moses-like phrase-based system, each candidate phrase-pair activates a feature function measuring the highest similarity between the current topic vector and all topic vectors associated to the occurrences of the phrase-pair in the training corpus. Similarity is computed by taking the complement of the Jensen-Shannon divergence. In our work, we also deploy this divergence measure to measure the match between the topic vector of the input and the topic vector of each candidate phrase-pair.

Finally, in (Hasler et al., 2014a) the authors combine and compare domain adaptation and

topic adaptation in phrase-based statistical MT for the translation of texts from three different domains. Concerning topic adaptation, the standard Moses phrase-based feature functions associated to the phrase-table are augmented three sets of dense feature functions: (i) two translation probabilities, (ii) one language model score and (iii) three topic distribution similarity feature functions. The first set of features introduce source-to-target phrase probabilities that account for topic information, the second set scores unigrams of the target phrase according to their topic relevance, and the latter measures the similarity between the input topic distribution and the topic distribution associated, respectively, to the whole phrase-pair, to the target phrase, and to the most representative target of the phrase. Topics distributions are inferred similarly to (Eidelman et al., 2012) at the level of whole speeches. Particular care is taken about sampling data of different domains, to avoid domain biases, as well as sampling the same amount of context data for each phrase-pair, to not avoid context bias. The authors main conclusion is that topic adaptation helps especially if there is a high divergence between training and testing domains. Moreover, topic vectors can be helpful also to predict the domain of test data when topic domain vectors are used as proxy of data. As is the case with Hasler et al. (2014a), we also compare topic adaptation with domain adaptation, but with respect to topic adaptation our main emphasis is on the combination of sparse and dense features.

4 Topic Adaptation

4.1 Approach

We apply the topic model discussed in Section 2 to infer the topic distribution on the source side of all bilingual training data of our statistical MT system (details will follow in Section 5), on the development set and on the evaluation set. Notice that we inferred the topic distribution on out-of-domain training data at the sentence level rather than at paragraph or document level. This choice is to keep annotation consistent with the training and testing conditions of the topic model, which are performed on item titles.

Similar to previous works, we trickle-down topic information from the sentence level to the phrase-pair level. By borrowing the notation from (Hasler et al., 2014b), we estimate the probability $p(t|s,k)$ of a target phrase t given a source phrase s and the topical information k ($k \in 1 \dots K$ where K is the total number of topics), through the formula:

$$p(t|s,k) = \frac{\sum_d p(k|d) \cdot c(t,s;d)}{\sum_t \sum_d p(k|d) \cdot c(t,s;d)} \quad (1)$$

Where $c(t,s;d)$ is the count of target phrase t and source phrase s being extracted from sentence (our proxy for document) d in the training data, and $p(k|d)$ is the probability of topic k in sentence d in the training data.

In addition to the probabilities computed with (1), we also infer topic vector (distribution) $\phi_{s,t}$ for each phrase-pair (s, t) extracted from the training data. Each component of the vector is the probability $p(k|s,t)$, which is inferred by averaging over all topic vectors corresponding to the sentences (d) in which the phrase pair was extracted from.

For each phrase pair we only keep the *relevant* topics and set the probability of the other topics to zero. In particular, we compute the perplexity[2] (PP) of the topic distribution and keep only the most probable $\lceil PP \rceil$ topics. In general, the perplexity tells how many equally likely topics can be represented with the number of bits of an optimal encoding of a topic distribution. We have empirically observed that the ceiling of the PP typically identifies the point in the ranked list of topics after which there is a significant drop in probability. Moreover, PP gives us also a measure of the topic specificity of a phrase pair. In particular, the lower

[2]Definition and properties of the perplexity function can be found for instance in (Federico and De Mori, 1998).

the perplexity of the topic distribution is, the more topic-specific is the translation represented by the phrase-pair phrase. On the contrary, phrase-pairs with high perplexity values reflect translations observed in sentences from many different topics, and thus identify translations not weakly depending on their context. Hence, we expect that the translation model will mostly benefit from topic information only for phrase-pairs with low perplexity. For this reason, we add topic information $p(t \mid s, k)$ and $p(k \mid s, t)$ to every entry s, t of the phrase-table only when the perplexity of $\phi_{s,t}$ is below a given threshold PP_{max} and up to PP topics.

The information entered in the phrase-table is exploited to activate feature functions at test time, by combining them with topic information inferred on the input sentence at test time. Similarly to (Hasler et al., 2014b) we denote the input topic information with the vector ϕ_c, where c stands for context representation. Notice, that we apply the same topic pruning strategy explained before also for the input topic vector.

In the following, we present the list of feature functions that we have explored in this work.

4.2 Features

1. **Joint Probability**: The topic probability $p(t|s, k)$ in itself is not enough to disambiguate between the context. As an example, assume that our English-Portuguese phrase table contains the following two phrase pairs with corresponding probabilities of target given source and topic:

   (a) age ||| idade ||| topic14 0.6 topic04 0.2

   (b) age ||| era ||| topic25 0.5 topic14 0.3

   The two entries show two different translations of the English word **age**, whose probabilities vary according to the context they have been observed with. In particular, the translation **idade** has been observed with *topic14* and *topic04*, while the translation **era** has been observed with *topic25* and *topic14*. However, this information can be properly exploited only once the topic information of the input sentence is known.

   Let us assume the following two input sentences and (pruned) context vectors:

   (a) *early english bronze* **age** *period blade c. 1600 bc* <topic25 0.9>

   (b) *supre hempz* **age** *defying moisturizer- 1 bottle* <topic14 0.7>

   Our first feature function activates for each translation option a sparse feature for each topic that *occurs in both the context and the phrase-table* with value:

   $$f_1^k(t, s, c) \;=\; -\log p(k|c) \cdot p(t|k, s)$$

   Going back to the first input sentence of our example, translation of **age** with **idade** would not activate any sparse feature, while translation of **age** with **era** would activate sparse features $topic25$ with score $-log(0.45)$. For the second input sentence, both translations of **age** with **idade** and **era** would activate sparse features $topic14$ with scores $-log(0.42)$ and $-log(0.21)$, respectively. The plain interpretation of this example is that our sparse feature function only rewards the translation **era** for the first sentence, while for the second input sentence it rewards both translations but gives a higher score to the translation **idade**.

   Notice, that our sparse feature is derived from the feature that Eidelman et al. (2012) proposed for hierarchical phrase-based decoding. This feature has a clear probabilistic interpretation: the product $p(k|c) \cdot p(t|s, k)$ computes the joint translation-topic probability $p(k, t \mid s)$ by multiplying the translation probability conditioned to topic with the prior topic probability coming from the context.

2. **Geometric Mean**: The same principle of 1. applies with our second basic feature, but this time with the phrase table annotated with posterior probabilities $p(k|s,t)$. Again, for each phrase-pair and for topics present both in the phrase table and in the context, our sparse feature takes the product of the topic probability on the input and the topic probability of the phrase pair.

$$f_2^k(t,s,c) \;=\; -\log p(k|c) \cdot p(k|t,s)$$

The intuitive interpretation behind this feature is to measure the level of each matched topic with the geometric mean between the probabilities in the context and phrase-table. Notice that the the above expression is equivalent to the log of the geometric mean but a constant factor 0.5 which we assume to be absorbed by the feature weight.

While the previous features are sparse, in the sense that they are computed over single topics thus resulting in K distinct features, the following feature are dense, i.e. it is a single feature is computed for all topics. In particular, all features try to measure the similarity between the topic distributions on the input and on each phrase pair.

3. **Crude-Count**: In this feature, we simply count the number of topics active in both input and the phrase pair and normalize this count over the total number of topics in both the context and the phrase-pair. Formally:

$$f_3(t,s,c) \;=\; -\log \tfrac{(|\{k:\, p(k|c)\cdot p(k|t,s)>0\}|)}{(|\{k:\, p(k|c)+p(k|t,s)>0\}|)}$$

Notice that this feature as well as the remaining features are activated for phrase-pair only if there is at least one common topic between context and the phrase-pair.

4. **Cosine Similarity**: This feature computes the cosine similarity between the input topic vector (ϕ_c) and the phrase-pair topic vector ($\phi_{t,s}$) .

$$f_4(t,s,c) \;=\; -\log cos(\phi_c, \phi_{t,s})$$

This feature was proposed in (Hasler et al., 2014b).

5. **JS Divergence**: This feature computes Jensen-Shannon divergence between input topic vector (ϕ_c) and phrase-pair topic vector ($\phi_{t,s}$).

$$f_5(t,s,c) \;=\; -\log JS(\phi_c, \phi_{t,s})$$

This feature was proposed in Hewavitharana et al. (2013).

6. **Hellinger Divergence**: Finally, this feature computes the similarity with Hellinger's divergence.

$$f_6(t,s,c) \;=\; -\log HD(\phi_c, \phi_{t,s})$$

$$HD(\phi_c, \phi_{t,s}) = \frac{1}{\sqrt{2}}\sqrt{\Sigma_{i=1}^{k}(\sqrt{\phi_{c,i}} - \sqrt{\phi_{t,s,i}})} \tag{2}$$

This feature was proposed in (Xiao et al., 2012).

7. **Sensitivity:** We measure how sensitive the phrase pairs are to the topics. Here, the idea is to penalize the phrase pairs with the topic vectors of high entropy. High entropy of a topic vector denotes that the phrase pair is susceptible to multiple topics, which means it occurs in multiple context and hence topic vectors are not useful for any disambiguation.

$$f_7(t,s,c) \;=\; -\log H(\phi_{t,s})$$

This feature which can be computed offline, as it does not depend on the context, was also proposed by Xiao et al. (2012).

The first two sparse features and the other dense features are linearly combined with standard features used in phrase based decoding. Hence, with the notation $f_1 + f_3 + f_4 + f_5 + f_6 + f_7$ we indicate that the standard translation score computed by the decoder is augmented with:

$$\sum_{k=1}^{K} \lambda_{1,k} f_1^k(s,t) + \sum_{i=3}^{7} \lambda_i f_i(s,t) \tag{3}$$

where the $K + 5$ λ-weights are tuned together with the other weights of the translation model.

5 Experiments

5.1 Task and Data

We evaluated our topic adaptation approach on the translation of item titles from English to Portuguese (Brazilian). Parallel data used to train, tune and evaluate MT systems comes from various publicly available collections, proprietary repositories and in house translated item titles. In particular, in-house translated items, descriptions, and specifics are here considered as in-domain data while all the rest is regarded as out of domain data. For development and testing purposes we use manually translated item titles for which two reference translations are available. Statistics on the amount of parallel data for each category are given in Table 2.

	Train (Out-Domain)	Train (In-Domain)	Dev	Test
Segments	5.28M	336K	1631	1000
Tokens EN	69M	2M	27K	10K
Tokens PT	70M	2M	31K	11.6K

Table 2: Statistics of English-Portuguese parallel data.

5.2 MT Systems

This section describes the topic adapted MT systems and the two baseline MT systems developed for comparison purposes. All MT systems are built using the Moses toolkit (Koehn et al., 2007) and the linear weights for all systems are optimized using the k-best batch MIRA implementation provided in the Moses toolkit (Cherry and Foster, 2012). Performance of all systems are reported in terms of BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores. Statistical significance tests were conducted using approximate randomization tests (Clark et al., 2011).

Baseline System In-domain and out-domain parallel data were taken in 10:1 ratio for training the word alignments. Translation models along with operation sequence models (Durrani et al., 2011) were trained using the standard pipeline of Moses. Due to the nature of the item titles, we did not use any lexicalized reordering model in the MT system. The distortion limit was set to 6. On the target side, we built a trigram LM, using KenLM (Heafield, 2011) trained with modified Kneser-Ney smoothing (Chen and Goodman, 1996).

Domain Adapted System The domain adapted system was built on top of the baseline system. An additional translation model was built using the in-domain data and then the fill-up adaptation method (Bisazza et al., 2011) was applied to combine the in-domain and out-domain phrase tables. Fill-up simply adds a provenance feature in the phrase table with a score of 1 if the phrase pair is present in in-domain phrase table and 0 if it is from out-domain phrase table.

Topic Adapted Systems To evaluate the performance of the topic adapted systems using the features functions presented in Section 4.2 we followed a component analysis approach. Each basic sparse feature was added to the domain adapted system separately, shortly (1) DA+f1, (2) DA+f2. We built two separate systems because when we tag the topics in the phrase table, we can either set them to a distribution of topics over phrase pair ($P(k|s,t)$) or a target-phrase translation probability given the source phrase and the topic ($P(t|s,k)$). Then we added the other dense features one by one, resulting in 10 distinct systems i.e. 1) DA+f1+f3 2) DA+f1+f4 3) DA+f1+f5 4) DA+f1+f6 5) DA+f1+f7 6) DA+f2+f3 7) DA+f2+f4 8) DA+f2+f5 9) DA+f2+f6 10) DA+f2+f7. Finally, we also combined all dense features together on top of basic sparse features to build 1) DA+f1+f3+f4+f5+f6+f7 2) DA+f2+f3+f4+f5+f6+f7. To evaluate the impact of topic adaptation independently from domain adaptation we performed the same analysis also with the baseline (BA) system.

6 Results and Discussion

6.1 Topic Model Analysis

Since, our method of topic adaptation depends on the quality of topic labels inferred on training data, development and evaluation sets; first, we analyze the distribution of topic labels. Figure 1 shows the average probability mass for each topic in three sets. The plot shows that topics represented in the training are not always very well covered in the development and evaluation sets. Topics 14, 15 and 26 are very frequent in the training data, as a result when we project the topic distribution to phrases in the phrase table these topics occur in more phrase pairs than any other topic. Thus, as a consequence, these topics should have less discrimination power than other topic features. To validate this hypothesis we also plotted the weights of the topic features f1 after tuning them with k-batch MIRA. In fact, we can observe that the algorithm weights them less compared to other topic features as shown by green lines in the figure.

An interesting note is that for some topic labels the tuning algorithm assigns negative weights. A possible reason is that there is a topic-translation mismatch between training and development data. This can be explained by the fact that the training data contains both in-domain and out-domain data while the development set contains only in-domain data. Hence, the possibility is that topics mostly occurring in phrase-pairs extracted from out-domain data are being penalized to the advantage of topics mostly occurring in in-domain phrase-pairs.

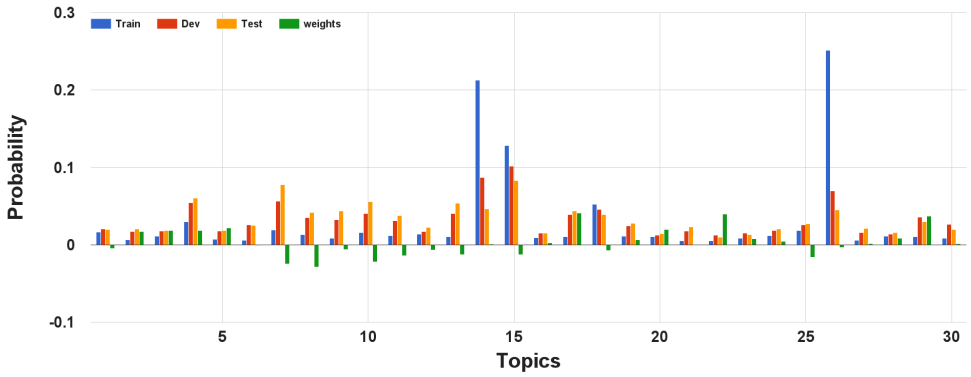

Figure 1: Topic distribution on training, development, and test data set. Green bar shows the weight of each topic feature when tuned with MIRA.

6.2 MT experiments

Table 3 presents the results for the baseline, domain adapted and all topic adapted systems. We split the results in three blocks: the first block of results shows the impact of adding sparse topic features ($f1$ and $f2$) on top of the baseline (BA) system. The second block shows results with the *Joint Probability* feature and the third shows results with the *GeoMean* feature. In the first block, both features improve performance consistently (at least 0.3 gain in BLEU points) over the baseline system and we even observe statistical significant improvements in BLEU score with the Joint Probability feature (f1). Since, our goal is to improve over domain adaptation we tested the rest of the features only on top of the domain adapted (DA) system.

System	BLEU	TER
Baseline (BA)	**36.99**	**49.15**
BA + f_1	37.52*	49.09
BA + f_2	37.28	48.93
Domain Adapted (DA)	**39.42**	**48.12**
DA + f_1	39.56	48.11
DA + $f_1 + f_3$	39.66	47.77†
DA + $f_1 + f_4$	39.70†	47.75†
DA + $f_1 + f_5$	39.56	47.84
DA + $f_1 + f_6$	39.66	47.77†
DA + $f_1 + f_7$	39.60	47.84
DA + $f_1 + f_3 + f_4 + f_5$	39.67	47.77†
DA + $f_1 + f_3 + f_4 + f_5 + f_6 + f_7$	39.57	47.72†
Domain Adapted (DA)	**39.42**	**48.12**
DA + f_2	39.60	47.86
DA + $f_2 + f_3$	39.61	47.79
DA + $f_2 + f_4$	39.60	47.76†
DA + $f_2 + f_5$	39.62	47.77†
DA + $f_2 + f_6$	39.60	47.78
DA + $f_2 + f_7$	39.66	47.68†
DA + $f_2 + f_3 + f_4 + f_5$	39.71†	47.77†
DA + $f_2 + f_3 + f_4 + f_5 + f_6 + f_7$	39.53	47.68†

Table 3: BLEU and TER scores of systems on English to Portuguese data set showing impact of sparse topic features against a weak (BA) and a strong (DA) baseline system (in bold). System performance marked with $*$ and † show significant different results w.r.t baseline (BA) and domain adapted system (DA) with p-value < 0.05.

The second block shows results with the topic adapted system using the *Joint Probability* feature. Individually, the best features which give significant improvements in TER scores over the DA system are the *crude-count*, *cosine similarity* and the *Hellinger's divergence* feature, i.e. DA+f1+f3, DA+f1+f4, DA+f1+f6. Concerning the BLEU scores, only the *cosine similarity* feature is significantly better than the domain adapted but on an average we observe 0.2 BLEU points improvement per feature. We also observe significant gains in TER over the domain adapted system when we combine all dense features together (see systems DA+f1+f2+f3+f4+f5 and DA+f1+f2+f4+f5+f6+f7).

The results in the third block are using the *GeoMean* sparse feature. The component wise analysis shows that the *cosine similarity*, *Hellinger's divergence* and the *Sensitivity* features give statistically significant improvements ($p < 0.05$) over the DA system in terms of TER scores. Average improvements of 0.2 BLEU points are observed across individual feature systems. In

terms of BLEU, our best system is the one which combines all the features together without *Hellinger's divergence* achieved statistically significant gains ($p < 0.05$) over the DA system (DA+f2+f3+f4+f5). In terms of TER, we observe statistically significant gains ($p < 0.05$) of 0.34 TER points in the best systems. These systems are the combination of *GeoMean* feature with *Sensitivity* feature (DA+f2+f7) and the one which combines all dense features together with the *GeoMean* feature (DA+f2+f3+f4+f5+f6+f7).

In Table 4 we show examples from the evaluation set where our system solved the problems of context disambiguation and proper rendering of proper names. In the first example the source title contains the words **endurance** which is a brand and **colander** which is the name of a cooking tool. Domain adapted system (DA) incorrectly translates **endurance** as **resistência** while the topic adapted system (TA) correctly took the verbatim translation. In the same sentence, DA translates **colander** as **escorredor** while TA correctly picked the more specific translation **escorredor de macarrão** (colander for pasta). In the second example the source contains: **columbia river crkt** which is a brand name. The DA system erronously translates **river** as **rio** while the TA system produced the correct verbatim translation of the brand name.

Source	rsvp international 5-qt . endurance colander 1024
DA	rsvp international 5-qt . resistência escorredor 1024
TA	rsvp international 5-qt . endurance escorredor de macarrão 1024
Ref1	rsvp international 5-qt . coador endurance 1024
Ref2	escorredor de macarrão endurance de 5 quartos de galão da rsvp international 1024
Source	columbia river crkt crawford kasper lawks zytel knife !
DA	rio columbia crkt crawford kasper lawks zytel faca !
TA	columbia river crkt crawford kasper lawks zytel faca !
Ref1	faca columbia river crkt crawford kasper lawks zytel !
Ref2	faca canivete columbia river crkt crawford kasper lawks zytel !

Table 4: Examples from the domain adapted (DA) and topic adapted (TA) systems.

7 Conclusion

An open problem in machine translation is how to effectively handle and incorporate the context information in the translation models that can help the system to properly disambiguate between competing translation alternatives. In this paper we presented methods for topic adaptation for phrase-based machine translation that have been experimented in an e-commerce application scenario. In particular, starting from state-of-the-art LDA topic modeling, we present several feature functions (some of them new and others derived from the literature) that combine topic information integrated in the phrase-table with topic information inferred on-the-fly on the input text. We report results on an English-Portuguese translation task of item titles that show consistent and statistical significant improvements through topic adaptation over both a generic baseline MT system and a domain adapted MT system. Our work shows that the use of sparse features permits to identify and cope with topic-translation inconsistencies in the training data, which from one side cannot be avoided when data from multiple and diverse sources are pooled together, but from the other side calls for more refined methods to label such training data with topic labels. In the future we plan to investigate feature regularization methods in order to select topic-features with high discrimination power.

Acknowledgements

The first author received support through a financial gift by eBay Inc. to FBK. The second author received support by the EU H2020 funded MMT project (grant agreement No 645487),

by eBay Inc., and by FBK's Mobility programme.

References

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. In *IWSLT*, pages 136–143.

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

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada. Association for Computational Linguistics.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Durrani, N., Schmid, H., and Fraser, A. M. (2011). A joint sequence translation model with integrated reordering. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1045–1054.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 115–119, Jeju Island, Korea. Association for Computational Linguistics.

Federico, M. and De Mori, R. (1998). Language modelling. In Mori, R. D., editor, *Spoken Dialogues with Computers*, pages 199–230. Academy Press, London, UK.

Gong, Z., Zhang, M., and Zhou, G. (2011). Cache-based Document-level Statistical Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Gruber, A., Weiss, Y., and Rosen-Zvi, M. (2007). Hidden topic Markov models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 163–170.

Guha, J. and Heger, C. (2014). Machine Translation for Global E-Commerce on eBay. In *Proceedings of the AMTA*, volume 2: MT Users, pages 31–37.

Hasler, E., Blunsom, P., Koehn, P., and Haddow, B. (2014a). Dynamic Topic Adaptation for Phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–337, Gothenburg, Sweden. Association for Computational Linguistics.

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

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Hewavitharana, S., Mehay, D., Ananthakrishnan, S., and Natarajan, P. (2013). Incremental Topic-Based Translation Model Adaptation for Conversational Spoken Language Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 697–701, Sofia, Bulgaria. Association for Computational Linguistics.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

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

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

Picinini, S. (2014). Challenges of Machine Translation for User Generated Content: Queries from Brazilian users. In *Proceedings of the AMTA*, volume 2: MT Users, pages 55–65.

Ruiz, N. and Federico, M. (2011). Topic Adaptation for Lecture Translation Through Bilingual Latent Semantic Models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 294–302, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sanchez, J. and Badeka, T. (2014). Linguistic QA for MT of user-generated content at eBay. In *Proceedings of the AMTA*, volume 2: MT Users, pages 1–24.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Su, J., Wu, H., Wang, H., Chen, Y., Shi, X., Dong, H., and Liu, Q. (2012). Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 459–468, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiao, X., Xiong, D., Zhang, M., Liu, Q., and Lin, S. (2012). A Topic Similarity Model for Hierarchical Phrase-based Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 750–758, Stroudsburg, PA, USA. Association for Computational Linguistics.