Extracting Multilingual Topics from Unaligned Comparable Corpora

Jagadeesh Jagarlamudi and Hal Daumé III

School of Computing, University of Utah
{jags,hal}@cs.utah.edu

Abstract. Topic models have been studied extensively in the context of monolingual corpora. Though there are some attempts to mine topical structure from cross-lingual corpora, they require clues about document alignments. In this paper we present a generative model called JointLDA which uses a bilingual dictionary to mine multilingual topics from an unaligned corpus. Experiments conducted on different data sets confirm our conjecture that jointly modeling the cross-lingual corpora offers several advantages compared to individual monolingual models. Since the JointLDA model merges related topics in different languages into a single multilingual topic: a) it can fit the data with relatively fewer topics. b) it has the ability to predict related words from a language different than that of the given document. In fact it has better predictive power compared to the bag-of-word based translation model leaving the possibility for JointLDA to be preferred over bag-of-word model for Cross-Lingual IR applications. We also found that the monolingual models learnt while optimizing the cross-lingual copora are more effective than the corresponding LDA models.

1 Introduction

With the increasing amount of text published in varied languages, comparable corpora - documents written in different languages but talking about same topics - are increasingly available. This situation raises the need for novel ways of organizing a multilingual corpus based on common topics/events, which could potentially be useful for many cross-lingual applications like Cross-Lingual Information Retrieval (CLIR) [1] and Cross-Lingual Text Classification [2]. Though there have been many attempts to mine the topical structure from a document corpus [3,4,5] most of these approaches operate in a monolingual scenario.

Topic models like LDA [6] use co-occurrence information to group similar words into a single topic. In case of cross-lingual corpus, two related words in different languages (like English and Spanish) will rarely co-occur in a monolingual document and hence these models fail to group such pair of words into a single topic. As an illustration, we picked a sample of the Europarl [7] English (176777 tokens) and Spanish (227487 tokens) parallel corpus and ran LDA[1] [8]

[1] We used collapsed Gibbs sampler for inference.

C. Gurrin et al. (Eds.): ECIR 2010, LNCS 5993, pp. 444–456, 2010.

Table 1. Few topics that were identified by LDA on Europarl parallel corpus. The language of most probable words (E for English and S for Spanish) in each topic is also indicated.

Topic 3 (E)	Topic 16 (S)	Topic 6 (S)	Topic 18 (E)	Topic 10 (S)	Topic 12 (E)
water	directiva	política	european	consejo	council
food	ambiente	europea	union	kosovo	mr
safety	agua	social	europe	europea	european
environmental	medio	desarrollo	states	unión	kosovo
community	enmiendas	unión	president	pregunta	union
environment	aguas	polticas	policy	señoría	question
fisheries	pesca	pases	mr	situación	peace
disaster	propuesta	mujeres	economic	ayuda	government
fishing	principio	trabajo	countries	usted	situation
states	costes	objetivos	political	sr	cyprus

with 20 topics. Not surprisingly we found ten out of the 20 topics have English words as high probable words and the rest of the topics have Spanish words as high probable words. Table 1 shows six of the 20 topics that were identified.

There is a striking similarity between the topics in different languages. For example, pairs of topics {10,12}, {3,16} and {6,18} are essentially same but realized in different languages. This leads to two primary concerns:

1. Because there are different possible realizations of a topic based on language, similar documents in different languages will have different document-topic probability distributions. This makes the task of finding similar documents across languages harder which is inherent in cross-lingual IR applications.
2. If we can generate a multilingual topic by combining two related monolingual topics then it may be possible to achieve same level of modeling capability with fewer topics.

This motivated us to explore techniques to identify multilingual topic-word distributions from an unaligned cross-lingual corpora. The main desirable property of any such approach is to identify topics that distribute their probability mass on related words from different languages. Thus two similar documents, irrespective of their language, will have similar topical distributions. In addressing this task we also explore some interesting questions that arise because of the availability of cross-lingual corpora. For example, [9] shows that bursty patterns can be effectively mined by using cross-lingual documents when compared to mining only from monolingual documents. We would like to see if a similar phenomenon happens in the topic models as well, i.e. "does the availability of related information in different language, i.e. in a completely different style, help in mining any better topical structure?" Another question, related to the ability to compress the data, is "does the additional, but related, data in different language require twice the number of topics to achieve the same level of accuracy (in terms of predictability on an unseen data)?"

There have been some attempts to mine topical structure from cross-lingual corpus, but those approaches assume either explicit or some indirect clues about document alignment. In one of the early approaches for CLIR [10], the authors form an artificial document by concatenating the aligned documents in different languages. A term by document matrix of these new documents is used to learn the lower dimensional representation using Latent Semantic Indexing. Documents across language are compared in this subspace. [9] propose a generative model to mine correlated bursty topic patterns from news articles of different languages. In their approach authors use time index to link documents in different languages. In CorrLDA [11] authors propose an asymmetric model to match words and pictures, even in this model both the image and its corresponding words are generated simultaneously. Recently [12] propose an extension of LDA to mine multilingual topics from Wikipedia articles by forcing aligned articles to share at least one topical distribution. All these approaches critically require alignments at the document level to mine the multilingual topic models and hence can't be applied to a comparable corpora.

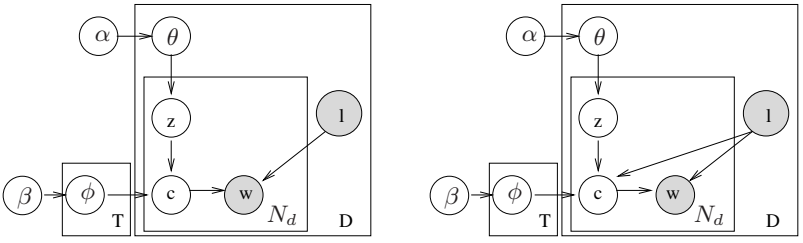
In this paper we explore the use of bilingual dictionary to identify the common structure and hence our model *does not* require document alignments. We propose an extension of the LDA model, called JointLDA, which uses bilingual dictionary to generate documents in different languages.

2 Joint Model of Cross-Lingual Corpora

In this section, we describe the details of JointLDA model for cross-lingual corpora. First we propose a model assuming every word is found in the dictionary and then extend it to handle out-of-dictionary words. Neither of these models needs document alignments.

Similar to LDA model [6], a document is assumed to be a mixture over T topics where the mixture weights (θ_d) is drawn from a Dirichlet distribution with symmetric prior (α). But we introduce an additional layer of hidden variables, called *concepts*, in defining topic distributions. Each topic is now a mixture over these concepts rather than words. The topic distribution (ϕ_k) is also drawn from a Dirichlet distribution with a different symmetric prior (β). Finally, a concept can be realized in different ways depending on the choice of the document language (l_d). This additional layer of language independent abstraction over the words allows the model to capture common topics in different languages effectively. In this paper we use bilingual dictionary entries[2] as substitute for these concepts. To understand the process consider generating an English document, first choose a topic mixture say 70% of sports and 30% of entertainment. Now choose a topic for the first word say 'sports' and then choose a concept from the sports topic, let it be 'player:jugador'. Since we are generating an English document we will pick the word 'player' from this concept and discard the Spanish

[2] Bilingual dictionary entry (or simply dictionary entry) is used to refer to a pair of words from different language that are possible translations of each other.



(a) For Complete Dictionary (b) For Partial Dictionary

Fig. 1. The graphical representation of JointLDA model

word. If we were to generate Spanish document we would pick 'jugador'. This process repeats as many times as the number of words in the document.

Formally the model is described as follows (Fig. 1(a)):

1. For each topic k=1...T, choose $\phi_k \sim \text{Dir}(\beta)$.
2. For each document d, choose $\theta_d \sim \text{Dir}(\alpha)$ and language $l_d \sim \text{Binomial}(\frac{1}{2})$.
   - For each token $i = 1 \cdots N_d$:
     (a) Select a topic $z_i \sim \text{Multinomial}(\theta_d)$.
     (b) Select a concept (dictionary entry) $c_i \sim \text{Multinomial}(\phi_{z_i})$.
     (c) Select a word from $p(w_i|c_i, l_d)$.

Note that given a dictionary entry and language there is only one possibility for a word and hence $p(w_i|c_i, l_d) = 1$. Note that the model doesn't require translation probability for a pair of words[3].

2.1 Handling Out-of-Dictionary Words

Since the coverage of bilingual dictionary is limited, new words will always appear. The model as described above, does not describe the generation of such words. Neglecting these words will leave a major portion of the document unexplained, especially when the dictionary is small. As a result the model will not learn good topic distributions. In order to overcome this, we will handle out-of-dictionary words by adding some artificial dictionary entries to the dictionary. For each of the out-of-dictionary source[4] (target) word we create an artificial dictionary entry of the form w : _NA_ (_NA_ : w). The only difference between an artificial entry and an actual bilingual dictionary entry is that the former is restricted to generate a word in only one language while the latter can generate both source and target language words. Note that if there is any common word between the vocabulary of both these languages that is not found in the dictionary then we create two unrelated artificial entries. In the extreme case

[3] Hence techniques like [13] can be used when the dictionary is not available

[4] For clarity, one of the languages is referred as source and the other as target language.

where the dictionary has only artificial entries, the one-to-one relationship between artificial entries and words forces the topic distribution to a distribution over words. In this case JointLDA model reduces to LDA model.

Although artificial entries explain the generation of out-of-dictionary words they lead to deficient topic-word probability distributions. To understand this, consider $p(w|k, l; \theta, \phi)$

$$= \sum_{c \in C} p(w, c|k, l) = \sum_{c \in C} p(w|c, l)p(c|k) = \sum_{c \in C_b \cup C_s \cup C_t} p(w|c, l)p(c|k)$$

where C_b, C_s and C_t are dictionary entries that can generate both language words, only source language and only target language words respectively. Now with out loss of generality fix the language to be source. Then, for any dictionary entry $c \in C_t$ and $\forall w,\ p(w|c, l{=}\text{src}) = 0$ (because it can not generate a source language word) and hence

$$p(w|k, l{=}\text{src}) = \sum_{c \in C_b \cup C_s} p(w|c, l_s)p(c|k) \Rightarrow \sum_{w} p(w|k, l_s) = \sum_{c \in C_b \cup C_s} p(c|k) \ \leq 1$$

This is because of our assumption that choosing a dictionary entry is independent of the document language, which is a reasonable assumption in the absence of artificial entries. But in the presence of them, while generating a source (target) language word the model should not choose a dictionary entry that can generate only target (source) language word otherwise it fails to generate source (target) language word.

Here we propose a refined model called JointLDA model (Fig. 1(b)) which carefully chooses a dictionary entry based on (document) language.

1. For each topic k =1...T, choose $\phi_k \sim \text{Dir}(\beta)$.
2. For each document d, choose $\theta_d \sim \text{Dir}(\alpha)$ and language $l_d \sim \text{Binomial}(\frac{1}{2})$.
   - For each token $i = 1 \cdots N_d$:
     (a) Select a topic $z_i \sim \text{Multinomial}(\theta_d)$.
     (b) Select a concept (dictionary entry) $c_i \sim \text{Multinomial}(\phi_{z_i}) \cdot \psi(c_i, l_d)$.
     (c) Select a word from $p(w_i|c_i, l_d)$.

Where the function $\psi(c_i, l_d)$ is 1 if the dictionary entry c_i can generate a word from language l_d and 0 otherwise. Note that the effect of language variable in sampling dictionary entry is only to constrain the model to choose a dictionary entry that can generate a given language word. Intuitively, once language variable is observed, this is same as renormalizing the probability mass across a subset of dictionary entries and sampling a dictionary entry from that set.

We use collapsed Gibbs Sampling [8] for estimating the parameters (θ, ϕ). In each iteration the topic and dictionary entry assignments for each token are sampled from the probability distribution given by:

$$p(z_i = k, c_i = j|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{c}_{-i}, \mathbf{l}) \propto \frac{n^{d_i}_{-i,k} + \alpha}{n^{d_i}_{-i,(\cdot)} + T\alpha} \cdot \frac{n^{j}_{-i,k} + \beta}{n^{(\cdot)}_{-i,k} + C\beta} \cdot p(w_i|c = j, l_d)$$

Where $n^{j}_{-i,k}$ ($n^{(\cdot)}_{-i,k}$) denote the number of times the dictionary entry $c = j$ (any dictionary entry) is used along with topic k for sampling any word excluding the token w_i. Similarly, $n^{d_i}_{-i,k}$ ($n^{d_i}_{-i,(\cdot)}$) is the number of tokens in document d_i that are assigned to topic k (any topic) excluding the token w_i. Note that the above probability is non-zero only for dictionary entries that can generate the word w_i[5] and hence this is a very small subset compared to the total number of dictionary entries. As a result the running time complexity of the joint model is comparable to that of LDA model.

3 Experiments

We ran our model on cross-lingual corpora from two language pairs: English-Spanish (datasets with prefix ENES-) and English-German (prefix ENDE-). We collected two types of data sets for each language pair. The first one is a subset of articles from Europarl corpus (denoted by ENES-P and ENDE-P with 529707 and 386648 tokens respectively). The second one consists of a set of aligned Wikipedia articles in both the pairs of languages (ENES-W and ENDE-W with 282446 and 489840 tokens). Though the first data set is parallel, the Wikipedia articles are related only at the topic level and aligned articles differ in document lengths. The article alignments are used only to facilitate comparison with other models and are hidden to JointLDA model. The dictionaries required for JointLDA are also generated from Europarl corpus using GIZA++ [14]. For language pairs with similar script (like English and Spanish) the common script can be exploited to get initial dictionary [13]. But for generality of our results we ignore this in our experiments. In all our experiments the vocabularies of each language are disjoint, i.e. a common word in different languages is treated differently.

Table 2 shows four out of 20 topical dictionary entries (ranked according to $p(c|k)$ within each topic) that were identified by JointLDA on Wikipedia articles (ENES-W). Since a dictionary entry can generate either of the words depending on language variable, a multilingual topic (as shown in the table) is essentially merged version of two monolingual topics into a single topic. The dictionary entries within each topic are related and as a result a topic-word distribution will have related words from both the languages. The word "speer" in topic 1 occurred in the vocabulary of both the languages and the dictionary doesn't provide any evidence about them being translations. Yet JointLDA model grouped the artificial entries corresponding to these words into the same topic. Also notice that JointLDA is able to group related words in different languages (aramaic & arameo in topic 16 and comunión & communion in topic 17) into a single topic though they are not directly related by any dictionary entry.

[5] For this reason, both $\psi(c = j, l_d)$ and $p(w_i|c = j, l_d)$ terms can be omitted during sampling.

Table 2. Few topics that were identified by JointLDA on Wikipedia articles (ENES-W). Entries with _NA_ are artificial entries (Sec. 2.1).

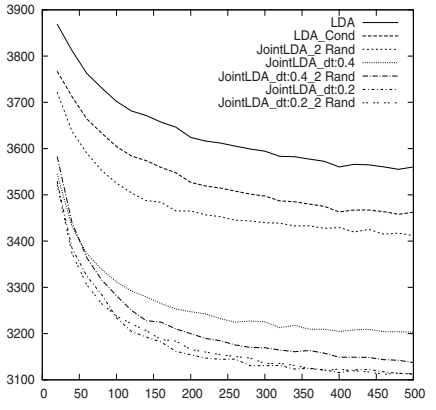
Topic 1	Topic 16	Topic 17	Topic 13
_NA_:speer	arabic:árabe	church:iglesia	aol:aol
hitler:hitler	art:arte	anglican:anglicano	apple:apple
archery:archery_NA_	words:palabras	churches:iglesias	ii:ii
arc:arco	word:palabra	english:inglés	language:lenguaje
attack:ataque	form:forma	ad:ad	assembly:asamblea
speer:_NA_	language:lengua	prayer:oraciṕn	games:juegos
arrow:flecha	aramaic:_NA_	sick:enfermos	software:software
racing:carreras	arabic:árabes	_NA_:comunión	code:código
german:alemán	dialects:dialectos	communion:_NA_	amway:_NA_
hand:mano	forms:formas	roman:romano	atari:_NA_
target:objetivo	letter:letra	catholic:católica	amd:_NA_
allosaurus:_NA_	_NA_:arameo	regular:regulares	users:usuarios

3.1 Perplexity Evaluation

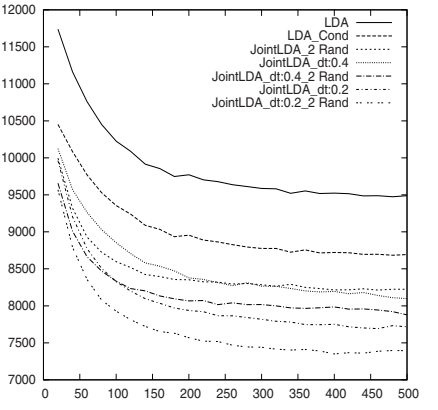
Perplexity is a standard way to evaluate the predictive power of a generative model on an unseen data. We compare our model with LDA and CorrLDA[11] models in terms of perplexity scores. In each data set 75% of document tokens are randomly chosen for training while the rest of the tokens are used for computing the perplexity. For all the models, Collapsed Gibbs Sampling [8] is used to estimate the parameters on the training data and the parameter estimates for testing are obtained from a single sample of Gibbs iteration. The article alignments in each of the data sets are available only for CorrLDA model and are hidden to JointLDA model.

For JointLDA, the perplexity is given by $\exp(-\frac{1}{N}\sum_{w_i} p(w_i|d_i, l_d))$ where $p(w|d, l_d) = \sum_k p(w|k, l_d)p(k|d)$ and $p(w|k, l_d)$ is the sum of $p(c|k, l_d)$ over all the dictionary entries that can generate the word w. While computing the perplexity values for the LDA, we have used the normal $p(w|d) = \sum_k p(w|k)p(k|d)$ (run labelled as LDA) as well as the probability of test word conditioned on its language: $p(w|l_d, d) = \sum_k p(w|k, l_d)p(k|d)$ where $p(w|k, l_d)$'s are obtained by renormalizing topic word probabilities specific to the given language (LDA_Cond run). The results are shown in Fig. 2, the set of figures in first column report perplexity scores on the Europarl data sets while the second column report the scores on the Wikipedia articles. In all the cases, LDA_Cond model results in a better perplexity scores than the normal LDA model which is intuitive as the uncertainty in the possible words decrease dramatically when language is known.

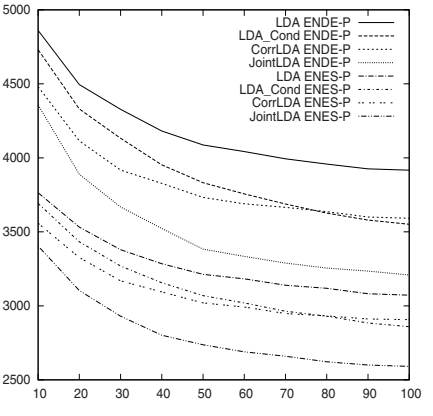
Figures 2(a), 2(b) show the effect of jointly modeling the cross-lingual corpus versus individual models (with 20 topics). We run JointLDA with different initializations of dictionary: a) for every source language word two target language words are selected at random and are added as translations ('JointLDA_2 Rand') b) with different levels of threshold on the conditional translation



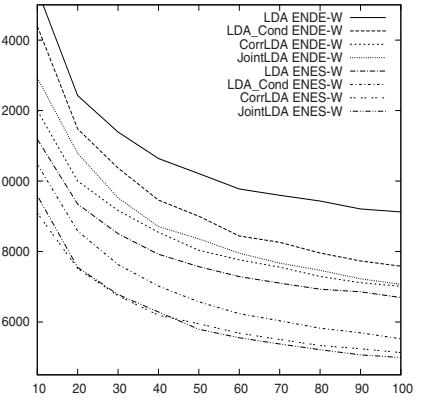
(a) Perplexity on ENES-P with iterations



(b) Perplexity on ENES-W with iterations



(c) Perplexity on ENES(DE)-P vs. Topics



(d) Perplexity on ENES(DE)-W vs. Topics

Fig. 2. Perplexity scores on both data sets, the first column being Europarl data set and the second column is the Wikipedia articles

probabilities[6] given by GIZA++ ('JointLDA_dt:0.4/0.2'- dictionary threshold of 0.4 and 0.2 respectively) c) combine both the dictionary translations and random translations ('JointLDA_dt:0.4/0.2_2 Rand'). The fact that 'JointLDA_2 Rand' run performed better than the 'LDA_Cond' model indicates that having bilingual information helps. From the rest of the curves (for example, 'JointLDA_2 Rand' vs. 'JointLDA_dt:0.4') it is very evident that the quality of translations does effect and aid the model in identifying better multilingual topics. But, note that there is an increase in performance when the translation probability threshold is decreased from 0.4 to 0.2. This is because of the increased number of bilingual

[6] Notice that JointLDA doesn't use translation probability and hence all translations with probability greater than the threshold are treated equally likely.

Table 3. Number of bilingual and total (including artificial) dictionary entries vs. size of of vocabulary

	Bilingual	Total	Vocab Size
ENES-P	16922	32731	38605
ENDE-P	14976	38585	40979
ENES-W	22400	53638	70843
ENDE-W	26515	88854	92086

dictionary entries as the threshold decreased. In general, we observed that as the number of dictionary entries increase, number of free parameters increase and hence model finds a better fit for the document corpus. But, the reader should not attribute the lower perplexity scores of JointLDA (compared to LDA_Cond) to this fact, because in all our data sets we found that the total number of free parameters per topic when the dictionary is loaded with translation threshold of 0.2 (third column of table 3) is less than that of LDA (the vocabulary size – last column of table 3). In rest of the experiments it is assumed that a threshold of 0.2 is used while loading the dictionary unless explicitly mentioned. With a closer look, we found that JointLDA efficiently uses dictionaries in predicting infrequent words and out-of-training words more accurately compared to other models. From figures 2(a), 2(b) it is clear that jointly modeling cross-lingual corpora is better than individually modeling. For brevity we don't include the graphs for English-German data set but they look similar.

Figures 2(c), 2(d) show the ability of the models to fit the data with respect to the number of topics required. When the data is parallel, JointLDA is able to achieve the same modeling capability with nearly half of the number topics as needed by the other models. This is completely justifiable because in any parallel data nearly half of the information is redundant and is simply expressed in different form. If a model can identify this redundancy it needs fewer topics. As the data set becomes comparable (less parallel) it needs more than half of topics, but significantly less than the number of topics required by LDA_Cond. Though CorrLDA performs competitively with JointLDA on Wikipedia data set, it estimates different topic-word distributions for each language and fails to identify the relatedness between topics of different language. It also uses the alignment information between training documents in different languages, which is not required for JointLDA.

One of the hoped advantages of modeling the cross-lingual corpus together is that by using the extra information written in another language, the model will learn better monolingual models. Here we compare the monolingual models learnt by the JointLDA while optimizing the cross-lingual corpus to the monolingual models that LDA learn only on the monolingual data. Fig. 3 shows the perplexity values on monolingual part of each test set (indicated by EN, ES and DE). When the data is parallel JointLDA efficiently uses the cross-lingual corpora to mine better monolingual models and when the data is not parallel (e.g. Wikipedia article) its monolingual models are not as effective.

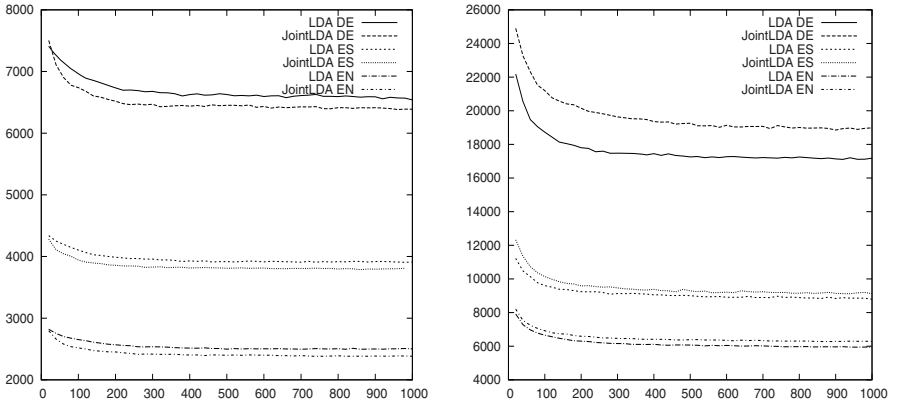


Fig. 3. Comparison of monolingual models learnt by JointLDA vs. the monolingual models of LDA on parallel (left figure) and comparable (right figure) corpora

Table 4. Test set perplexity given an aligned article in different language

	JointLDA	WordTrans
ENES-P	5732.503	3244.35
ENDE-P	4936.483	3771.34
ENES-W	7867.091	11930.3
ENDE-W	12750.12	18078.42

3.2 Perplexity of the Aligned Test Set

The traditional perplexity measures only the ability to predict a test word given a document of same language. Apart from this, a cross-lingual model should also be able to predict related words from different languages. In order to measure this aspect we compute a modified perplexity score using topic distribution of corresponding aligned document. We also report $\exp(-\frac{1}{N}\sum_{w_i} p(w_i|d_i^a, l_{d_i}))$ where d_i^a denote the aligned document (of d_i) in other language. For comparison, we use bag-of-word based translation model (referred as WordTrans) smoothed using appropriate unigram language model [15] which is proved to give good results in CLIR [1]. Under this model:

$$p(w_t|d_s) = (1-\lambda)\sum_{w_s} p(w_t|w_s)p(w_s|d_s) + \lambda p(w_t|C_t)$$

where $p(w_t|C_t)$ is the unigram probability of the word in the target language corpus. Table 4 shows the perplexity scores of JointLDA (with 100 topics and 1000 iterations) in comparison with WordTrans model. The better performance of WordTrans model on first two data sets is due to the fact that the dictionary is also learnt from Europarl data set. Also note that WordTrans model uses the

translation probabilities given by GIZA++, where as JointLDA model does not. But on the Wikipedia articles, JointLDA model achieves lower perplexity scores which indicate better predictability than a bag-of-word translation model. This leaves a possibility for JointLDA to be preferred over bag-of-word translation for applications like CLIR and Cross-lingual Text Categorization [2].

4 Discussion

As discussed in section 2, the JointLDA model is not limited to cross-lingual scenario. We claim that the model is applicable in a wide range of situations where some initial matching is available between the observations. For example, we can apply the JointLDA model to monolingual data by using synonyms (extracted from WordNet) as concepts. The generative story for the document corpus remains same and the probability of a word is given by:

$$p(w|d;\theta,\phi) = \sum_{k} p(k|d)p(w|k,d) = \sum_{k,c} p(k|d)p(c|k)p(w|c)$$

But, unlike cross-lingual situation, a synonym can generate both words so the parameters $p(w|c)$'s also need to be estimated during the inference process. When we tested this model on the English corpus of Wikipedia articles we found that JointLDA not only achieves lower perplexity scores (compared to LDA) on the whole test set but it also models infrequent words very well, which are typically excluded during the preprocessing stage of topic modeling algorithms.

Another line of approach to mine multilingual topics would be to use LDA to find monolingual topics in one language and use the dictionary to translate the topics into the other language. The disadvantage of this strategy is its inherent bias towards one language. It forces the topics in second language to be consistent with the identified topics in first language rather than letting them to evolve from the data. Comparison with WordTrans model in Sec. 3.2 confirms that, such a translation of topics would fail to predict unseen data when the data becomes less parallel.

Recently [16] has proposed MuTo model to extract multilingual topics from cross-lingual corpora. At any stage MuTo considers a matching between vocabularies of both languages and hence it doesn't allow any source word to pair up with multiple target language words. This underlies a strong assumption that a word is used in only one sense in the entire corpus. Where as JointLDA model deals with sense ambiguity by allowing a word to be paired with multiple target language words. Another major difference is that, in MuTo all unmatched words come from a single topic distribution. Which implies that when the dictionary size is small MuTo reduces to a simple unigram model while JointLDA reduces to the LDA model. Thus JointLDA can be seen as a generalization of the MuTo model.

5 Conclusion and Future Work

In this paper we have proposed generative model called JointLDA, which can extract multilingual topics from an unaligned cross-lingual corpora. Unlike other models, JointLDA model doesn't require document alignments among training documents for inference. It needs parallel data only to learn dictionaries and these dictionaries can be used again for a different document corpus. In order to facilitate comparison with other models and to compute the perplexity on the aligned test set we used aligned documents. The experiments conducted on different data sets showed that jointly modeling the cross-lingual corpus has several advantages compared to modeling the individual monolingual corpora.

It may appear that the model relies heavily on the availability of dictionary but the topics mined by JointLDA (Table 2) do contain translations that are not part of the initial dictionary. So we believe that it may be possible to start with a small but good quality translations and learn pairs of related words to be added to the dictionary at regular intervals. We leave this for future work.

References

1. Xu, J., Weischedel, R., Nguyen, C.: Evaluating a probabilistic model for cross-lingual information retrieval. In: SIGIR 2001, pp. 105–110. ACM, New York (2001)
2. Bel, N., Koster, C.H.A., Villegas, M.: Cross-lingual text categorization. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 126–139. Springer, Heidelberg (2003)
3. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. Annals of Applied Statistics, 17–35 (August 2007)
4. Blei, D.M., Lafferty, J.: Topic models. Text Mining: Theory and Applications. Taylor and Francis, Abington (2009)
5. Steyvers, M., Griffiths, T.: Probabilistic topic models. Latent Semantic Analysis: A Road to Meaning (2005)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Maching Learning Research 3, 993–1022 (2003)
7. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit (2005)
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of National Academy of Sciences USA 101(suppl. 1), 5228–5235 (2004)
9. Wang, X., Zhai, C., Hu, X., Sproat, R.: Mining correlated bursty topic patterns from coordinated text streams. In: KDD 2007: Proceedings of the 13th ACM SIGKDD, pp. 784–793. ACM, New York (2007)
10. Dumais, S.T., Landauer, T.K., Littman, M.L.: Automatic cross-linguistic information retrieval using latent semantic indexing. In: Working Notes of the Workshop on Cross-Linguistic Information Retrieval, SIGIR, Zurich, Switzerland, pp. 16–23. ACM, New York (1996)
11. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: SIGIR 2003, pp. 127–134. ACM, New York (2003)
12. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Mining multilingual topics from wikipedia. In: 18th International World Wide Web Conference, April 2009, pp. 1155–1155 (2009)

13. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of the ACL 2002 workshop on Unsupervised lexical acquisition, Morristown, NJ, USA, pp. 9–16. Association for Computational Linguistics (2002)
14. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
15. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: SIGIR 2001, pp. 334–342. ACM Press, New York (2001)
16. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Uncertainty in Artificial Intelligence (2009)