Polylingual Topic Models (PLTM)

1 Introduction

In this project we are going to explore Polylingual Topic Models (PLTM) [Mimno et al., 2009] which are a subset of extensions of Latent Dirichlet Allocation (LDA). They model multilingual collections that contain pairs of documents that are translations of each other. As part of this project we are going to learn about PLTM and use them to model speeches from the European parliament (Europarl) sessions [Koehn, 2005] . In particular we are going to focus on a subset of English-Spanish Europarl speeches. We'll cover two aspects of PLTM: the topic-word distributions across the two languages and the per document-topic distributions for each of the languages in the collection. Furthermore we'll use the topical representation of English and Spanish speeches in the shared topic space, i.e. the probability simplex, to find speeches that are translations of each other.

2 Polylingual Topic Models

Figure 1 shows a graphical representation of the PLTM. Given a set of document translation pairs, the PLTM assumes that across an aligned multilingual document pair, there exists a single, pair-specific, distribution across topics. In addition, PLTM assumes that for each language–topic pair, there exists a distribution over words in that language φ_l. Using these underlying assumptions, PLTM models the creation of a multilingual corpus through a generative process where first a document tuple is generated by drawing a tuple-specific distribution over topics θ which, as it is the case with LDA, is drawn from a Dirichlet prior $\theta \sim Dir\left(\alpha\right)$. In the traditional LDA model (shown in Figure 2) θ is used to specify the document specific distribution over topics.
For each of the languages l in the tuple and for each of the N words w_n^l in the document PLTM assumes the following generative process: first chooses a topic assignment $z_n^l \sim Multinomial\left(\theta\right)$ which is then followed by choosing a word w_n^l from a multinomial distribution conditioned on the

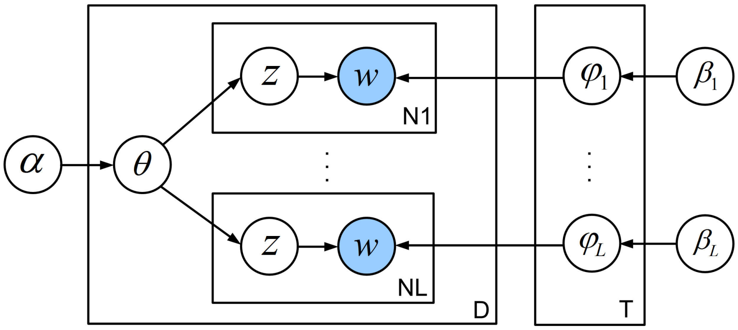

Figure 1: Graphical representation of the Polylingual topic model (PLTM).

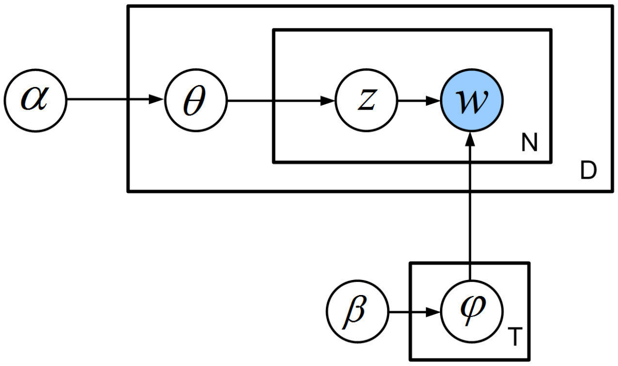

Figure 2: Graphical representation of the Latent Dirichlet Allocation (LDA).

topic assignment and the language specific topics distribution over words:

$$\varphi_l \sim Dir\left(\beta_l\right) \tag{1}$$

Both α and $\beta_{1,\ldots,L}$ are symmetric priors, i.e. the priors are exchangeable Dirichlet distributions. Finally, each word is generated from a language- and topic-specific multinomial distribution φ_t^l as selected by the topic assignment variable z_n^l:

$$w_n^l \sim p\left(w_n^l \mid z_n^l, \varphi_n^l\right) \tag{2}$$

For more information on PLTM please refer to [Mimno et al., 2009]. In this project we are going to use online PLTM (oPLTM) [Krstovski and Smith, 2013]. Unlike PLTM which use Gibbs sampling to obtain posterior estimates, oPLTM uses online Variational Bayes inference [Hoffman et al., 2010] which is more efficient, especially on large document collections, while offering similar performance.

3 Project Tasks

3.1 Infer Topics on Unseen Documents

In the first part of the project we are going to focus on the per document-topic distributions θ. We'll use the trained per topic-word distributions to infer topics on a test set of Europarl speeches. The test set is located in the *test* folder. In order to perform inference we'll use the online VB implementation in *infer_PLTM.py.* Please go over the code in order to familiarize yourself with the required input arguments. We'll start with inferring 50 topics and continue with the remaining 3 topic configurations. We'll run the inference step for each language - English (en) and Spanish (es). For each language we would need to provide the set of test speeches in that language (file_list), language specific vocabulary (vocab), the topic-word distribution for the specific topic configuration (topic_word_dist), number of topics (T) which should correspond with the topic configuration, hyperparameter value (alpha) and a stopping criterion (threshold) for the online VB.
Once you are done inferring topics using different topic configurations you should pick several English speeches (for each topic configuration) and visualize its topic distribution using a histogram plot (for example). See if you could notice a topic or a set of topics (or some sort of a pattern) that dominate certain set of Europarl speeches.

3.2 Find Document Translation Pairs

PLTM represent bilingual documents in the probability simplex, and thus the task of finding document translation pairs is formulated as finding similar probability distributions. In the last part of this project you are going to use the inferred topic distributions in order to find document translations pairs. Given that for each English test speech we have its Spanish translation equivalent you could measure the performance of different topic configurations by computing precision at rank one (P@1). The test set contains ∼1300 speeches. Depending on your available time you could either use all speeches in the test collection to compute P@1 or a subset of them (e.g. 100 speeches). In order to compute P@1 for each English speech you would need to calculate its topical similarity with all the Spanish speeches. Unlike the metric space, in the probability simplex the similarity between two probability distributions is computed using information-theoretic measures such as Hellinger distance. Once you compute Hellinger distance across all Spanish speeches use the distance values to create a ranked list. Computing P@1 is simply done by counting the number of times the true translation speech is at rank one and dividing that number by the total number of English speeches in your test collection which serve as queries. In case you are having difficulties computing the Hellinger distance, as an alternative you could use the Euclidean distance which is implemented in many Python packages.

References

M. Hoffman, D. Blei, and F. Bach. Online learning for latent Dirichlet allocation. In *NIPS*, pages 856–864, 2010.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, pages 79–86, 2005.

K. Krstovski and D. A. Smith. Online polylingual topic models for fast document translation detection. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 252–261, 2013.

D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP*, pages 880–889, 2009.