Cross-Linguistic Topic Classification Using Latent Dirichlet Allocation

Ryan Callihan
ryan.callihan
@student.uni-tuebingen.de

Samantha Tureski
samantha-zoe.tureski
@student.uni-tuebingen.de

Abstract

Unsupervised topic modeling is the sorting of previously unlabeled documents into categories without explicitly defining topics beforehand. Despite a wide variety of existing works in this area, most of the literature surveyed seemed to use solely English documents in their data sets. We were curious to evaluate the consistency of topic modeling with a parallel multilingual corpus.

1 Introduction

Unsupervised topic modeling has already been heavily studied throughout the field of machine learning. It has numerous important applications, including "mining patient notes for diseases" to aid in clinical research (Miller et al., 2016), word sense disambiguation, and categorizing webpages. In contrast with supervised methods, which require relatively expensive and slow human annotators, unsupervised models for topic modeling take advantage of the structure inherent in a given dataset to quickly label documents with "effectiveness levels comparable to those of trained professionals" (Sebastiani, 2002).

The extensively-researched problem of topic classification provided us with ample literature for the theoretical foundations of our project. In his 2002 paper entitled *Machine Learning in Automated Text Categorization*, Sebastiani had already written a thorough summary of techniques and common issues for the task. Ko and Seo (2000) introduce a relatively early and effective form of text categorization that involve a combination of term weighting and Naive Bayes classifiers. However, we opted to use Latent Dirichlet Allocation (LDA) in our project because it has shown to be more promising, and because there are existing Python libraries to implement this type of model. Other research teams, such as Miller et al. (2016), extended the capabilities of LDA by adding informed priors to steer a model toward specific topics of interest. Work done by Rubin et al. (2016) introduced novel modifications to Latent Dirichlet Allocation to account for variations in label frequencies and dependencies between labels. All of the literature collected, however, involved experiments performed on monolingual datasets, especially on texts written in English.

In this term paper, we present our experimental learning approach to the unsupervised topic modeling using Latent Dirichlet Allocation to observe how topic modeling functions across languages. We use the United Nations Parallel Corpus (Ziemski et al., 2016) to train our classifier on a fully aligned subcorpus for the languages of English, Arabic, Russian, and Chinese. With the results produced by our model, we attempt to compare cross-linguistic model alignment through the implementation of a cosine similarity matrix. We also investigate currently-available techniques to visualize our model's output.

2 Background

2.1 The United Nations Parallel Corpus

The United Nations Parallel Corpus (Ziemski et al., 2016) is a collection of documents manually translated and published in the six official languages of the United Nations between 1990 and 2014. The corpus can be downloaded as many documents in XML format, with multiple topic labels featured as an XML tag in each document. The corpus consists of 799,276 individual documents and 1,727,539 aligned document pairs.

Of the six languages offered, we chose to use English, Arabic, Chinese, and Russian in our project. The uniquely different linguistic backgrounds and writing systems of each of these languages made them particularly attractive. Russian, transcribed in our dataset in UTF-8 Cyrillic characters, was the most similar choice to English, as both languages make use of an alphabet. However, the high prevalence of Slavic roots in Russian made it plausible that salient terms would be completely novel, and perhaps even differently distributed, than those in the rejected datasets of French and Spanish. Arabic, with its complex abjadic root-based orthography, was chosen to give insight into what tweaks our model might require to faithfully model topics for such a language. Our last dataset, Chinese, is character-based, but characters are reused and tokens are frequently composed of multiple characters. Again, we wondered if term distributions within topics in Chinese would remain close enough to those of other languages to show any meaningful alignment. If significant alignment could be shown, it would mean that LDA is consistent, that term distribution among unrelated languages is consistent (at least for this content domain), and that a lexically-based topic model can be enough to separate semantic meaning regardless of language.

2.2 Topic Modeling with Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is "a probabilistic unsupervised method for grouping tokens into a set of corpus-wide clusters [in which each] document has a probability distribution across K topic indices, and each topic is a global probability distribution across V words in the vocabulary" (Miller et al., 2016). Furthermore, the topic distribution for each document is drawn from a specific type of probability distribution called a Dirichlet distribution. The guiding principle behind LDA is the assumption that each document's content can be described by a specific distribution of topics, and that words that are in the same topic are more likely to co-occur within a document than words selected at random. After a desired number of K topics has been set, the LDA model "learns" by iterating through an algorithm. First, the model randomly assigns a topic to each of the tokens in the bag of words (BoW) generated from each document. Then, for any topic t, LDA computes the proportion of words assigned to that topic in a given document d, as well as the proportion of times a certain word has been assigned that topic within the whole set of documents. From the results of these calculations, the model updates the assignment of the current word accordingly. After an adequate number of iterations, the model reaches a stable state in which each word is generally assigned its "correct" topic. The topic distribution for a given document d can then be calculated by finding the percentage of its words that have been assigned to each topic. One problematic feature of LDA, as noted by Miller et al. (2016) in their experiments, is that separate but similar topics may merge, for example "baseball and hockey, which share quite a bit of terminology (teams, games, scores, etc.)" This research team goes on to remark that "[s]imply increasing the number of topics may solve the problem but will also have the general effect of making categories more specific, which may adversely affect other topics." We took note of this limitation when reviewing the output of our own model.

2.3 Cosine similarity

We used cosine similarity to compare topics LDA predicted for each document across the four languages. Cosine similarity measures the similarity of two vectors. It is based on the solution to the dot product of $cos(\theta)$:

$$cos(\theta) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

The closer $cos(\theta)$ is to 1, i.e. the smaller the angle is between vectors, the more similar they are.

2.4 Principal component analysis (PCA)

Principal component analysis (PCA) is a statistical method for reducing the dimensionality of a dataset. It uses "an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components" (Wei et al., 2017). PCA aims to select features that best represent the variance in the original data.

3 Methods

For the default settings of our project, a separate LDA model was trained on approximately 12,000 documents in each of the four language training sets, and then each languages model was used to categorize approximately 5,000 documents in each languages respective test set (a 70:30 training-to-test-set ratio). What was of interest and the primary purpose of our research was to determine how similar classification would be for documents in four different languages using a basic unsupervised LDA approach. A secondary purpose of our project was to familiarize ourselves with existing tools for the various challenges we encountered, so that they might be of use to us in the future.

3.1 Preprocessing

Using the U.N. parallel corpus, while being an excellent set of texts to compare results across several languages at once, did not come without a considerable amount of preprocessing. While the materials in each language were well-aligned into parallel file systems, it was the case that some documents were merely meetings minutes, and thus unsuitably short for our task. These documents were subsequently weeded out by checking for relevant keyword tags.

It was the case that not all of the documents possessed cross-translations for each of the other languages. To isolate only the files that had versions in all four languages, the Linux "find" command line tool was used to extract all names of absolute file paths in each language directory to a text file. Then language-specific parts of paths (e.g."/en/" or "/fr/") were deleted in the respective lists of files using Find and Replace. For the next step, we found the common files shared by two of the language's directories with the command `comm -12` (the `-12` flags added to suppress the printing of irrelevant data). We performed this action multiple times until we had obtained the completely aligned four-language subset of the corpus which excluded meeting notes. Once a four-way parallel subset had been obtained, the contents of each file were tokenized, stripped of stopwords and lemmatized. The pre-existing tools in the `nltk.corpus` and the `nltk.stem.wordnet` packages sufficed for these tasks in English, Arabic, and Russian. As was expected, Chinese proved more difficult; we discovered that the Stanford Segmenter tool worked well in segmenting the text into relevant groups of characters (i.e. tokens). Chinese stopwords verified by a Mandarin native speaker were procured and removed with a script written by us.

LDA requires documents to be represented as a bag of words. Using gensim's `dictionary` module, we created a term dictionary of our corpus, where every unique term was assigned an index. Each document was then converted into a fully numerical BoW format, which consisted of a list of each token in that document listed as a `(term_index, term_frequency)` two-tuple. From this it followed to create a list of such BoW-documents, which could finally be passed into gensim's `LdaModel` tool.

3.2 Building the Model

The gensim `LdaModel` tool helped us to speed up the implementation of our project greatly. It allows one to train on a corpus, test the model on unseen documents, and additionally to save and load previously-trained models. The tool is also helpful in that it allows for the setting of parameters such as number of desired topics and one's own integer to word mapping scheme. Further helpful facets of the tool include the ability to print a list of the words most strongly associated with each topic, as well as the ability to print a full list of the predicted topic distribution on an unseen document. We determined the optimal number of LDA topics by running the model with 5, 10, 20, and 50 topics for each language. We set the model to iterate only at 5 passes, because the documentation recommended this, because we did not observe a significant change in results after more iterations. Time constraints also made adding more iterations to the model less feasible.

3.3 Grouping documents by topic

We then briefly attempted to utilize these lists of predicted topic distributions as features for K-Means clustering. An LDA model for 10 topics, for example, returned a predicted distribution for each document in the test set with 10 values between 1 and 0, with each index corresponding to a different topic. The cluster with whose midpoint was closest to each documents features was determined to be the cluster for that document. Initially, we thought the cluster for each document would be used to provide a rough comparison across the languages using cosine similarity. When it was discovered that no meaningful clusters were being found from the output of our model, we decided to calculate cosine similarity for language-topic vectors, the results of which are described in the next section.

4 Evaluation

4.1 Aligning clusters across languages

The most important part of testing the consistency of LDA cross-linguistically is comparing the models. This is not a straightforward task because these unsupervised models will form topic distributions differently. In other words, even if the model forms similar topics for parallel documents in different languages, it will not necessarily be organized in the same way. In fact, it most likely will not look similar.

4.2 Language-topic vectors

In order to test the model alignment between languages, we had to get a little creative. First, in order to create a topic order independent data to use for comparison, a sparse vector was created for each topic and each language.

LDA models output a topic distribution for each input document. Either the most likely topic can be chosen, or the distribution can be used.

The topic distribution for each document was added to the corresponding language-topic vector.

For example: Let's say that we want to test 4 documents with a model with 3 topics. We retrieve the topic distributions using our model and the results are as follows:

$d_1 = [.1, .15, .75]$
$d_2 = [0, .95, .05]$
$d_3 = [.3, .65, .5$
$d_4 = [.92, .5, .3]$

This would mean that our three language-topic vectors would be the following:

$ltv_1 = [.1, 0, .3, .92]$
$ltv_2 = [.15, .95, .65, .5]$
$ltv_3 = [.75, .05, .5, .3]$

Whereas, our three language-topic vectors with only the maximum topic chosen would be the following:

$ltv_1\_max = [0, 0, 0, 1]$
$ltv_2\_max = [0, 1, 1, 0]$
$ltv_3\_max = [1, 0, 0, 0]$

We could then use these vectors to visualize our model and measure comparisons using cosine similarity.

For each model, two types of language-topic vectors were obtained. First, vectors made using the topic distributions and second, multi-hot vectors made using the most likely topic for each document.

4.3 Measuring Alignment with Cosine Similarity

To measure the model alignment using the language-topic vectors, a confusion matrix was created, which obtained the cosine similarity of each vector pair with dimensions of $(num\_languages * num\_topics) * (num\_languages * num\_topics)$. Theoretically, if there is a topic which is similar between languages, their similarity score would be close to 1.

Therefore, for each topic in each language, the top similarity score was obtained from each language and added to a matrix with dimensions of $(num\_languages * num\_topics) * num\_languages$.

A final confusion matrix is obtained from this matrix which has the average cosine similarity score for each language pair along with the standard deviation. If there is a good alignment between language models, the average similarity score should be high with little deviation.

4.4 Qualitative Assessment of Model Alignment

In order to have a more tangible metric of model alignment, PCA was used to reduce the large language-topic vectors to two dimensions, and then were to be plotted using a scatter plot.

This approach, while more qualitative than quantitative, can still lead to some interesting insights.

5 Results

In order to test the reliability and consistency of lexical based LDA topic models across languages, we used a completely parallel corpus and examined the the results to see if the model results followed a similar pattern.

5.1 Effects of the data on the model

The results were not exactly what we were expecting. One very influential, yet unforeseen, factor which had a large effect on the end results was related to the data itself. After the time-consuming process of training 16 different models (4 different topic models for each language) with 5 epochs apiece, we found that the models did not fit the data very well. In each language, a large proportion of the test documents were predicted to be in one single category, with only a very small number being in other topics.

This was surprising because there was a wide number of topics to cover in the corpus, and each document had a seemingly diverse set of pre-annotated topic keywords in its metadata.

The lopsided topic distributions can be clearly seen in the plotted PCA results in Figure 1.

Each language is represented by a unique color and shape. In this figure for 5 topics, we can see there is a cluster of one language-topic vectors from each language clearly separated from the rest. Then there is a second, rather unclear cluster. And finally, the rest of the vectors are c3entered around 0, meaning that these vectors contained little to no documents.

This pattern is the same for each language and for each number of topics per model. This is not an ideal situation for our project, but the fact that the results are so consistent leads us to conclude that the data was not ideal for our project. I am not sure that altering the hyperparameters would make a large difference in results for this data. Our hyperparameters were similar to those used for the pretrained model available publicly created by Gensim's creator using the very large Wikipedia corpus.

Furthermore, we can see the topic distribution of the 14,439 test documents in the following figures. In Figure 2(a), we can see the probability distribution for the documents with 5 categories for English. They are all relatively similar. When only using the top-scoring topic for each document in English, the data becomes much more imbalanced, as is exemplified by Figure 2(b).

5.2 Interlingual alignment

The main challenge of our project was to find a system of aligning results of the models from different languages. As mentioned above, our main tool in this process was using pairwise cosine similarity scores, and this method was found to be effective.

To obtain aligned vectors, we took the num_topics top scoring vector pairs for each language pair and then took the mean score of the num_topics top scoring pairs for each topic in each language. This can be exemplified in the confusion matrix in Figure 3(a).

We can see that for a model with 5 topics and language-topic vectors using probability distributions, the mean alignment scores for each language pair are quite high, ranging from between 0.86 to 0.93. For 5 topic language-topic vectors using max category vectors, the numbers were much lower, ranging from between 0.21 to 0.39, as can be seen in Figure 3(b).

This pattern follows for the models with 10, 20, and 50 topics but with one difference. The mean scores become increasingly lower, which makes sense because, as mentioned earlier, the topic distributions are quite sparse and not very well distributed. Therefore, we believe that the lower scores were a result of the data itself rather than the method.

5.3 Discussion

As noted earlier, standard LDA is plagued by an inability to make fine-grained distinctions among topics that require similar terminologies. The broad range of politically-related topics in the U.N. Corpus may have rendered it an unsuitable training set. One of the most potentially helpful features of the United Nations Parallel Corpus, the pre-annotated topic keywords for each document, went unused in our project. Had we used the keywords as initial topic "seeds" to perform semi-supervised LDA, the topic alignment among languages may have been much closer. LDAvis, a sophisticated visualization tool that includes such interactive features as a "Intertopic Distance Map" and "Most Salient Terms" chart, was implemented in our project, but continuous errors and failure to produce the desired output caused us to abandon the attempt. With more time and resources, we may have been able to fix the bugs that caused this tool to fail within our project.

6 Conclusion

This project was an excellent introduction to various techniques in unsupervised topic classification using Latent Dirichlet Allocation. While admittedly a number of improvements could have been made to enhance the accuracy of our results, and that the nature of the U.N. corpus may have made it unsuitable, we feel well-prepared to tackle other unsupervised computational linguistics tasks in the future.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993-1022, January 2003.

Dorado, Ruben, and Sylvie Ratte. 2016. Semisupervised Text Classification Using Unsupervised Topic Information. FLAIRS Conference, 2016.

Ko, Youngjoong, and Jungyun Seo. 2000. Automatic text categorization by unsupervised learning. *Proceedings of the 18th conference on Computational linguistics, Volume 1. Association for Computational Linguistics*, 2000.

Miller, Timothy, Dmitriy Dligach, and Guergana Savova. 2016. Unsupervised document classification with informed topic models. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 28(1):114–133.

Rubin, Timothy N., et al. 2012. Statistical topic models for multi-label document classification. *Machine learning*, 88.1 (2012): 157-208.

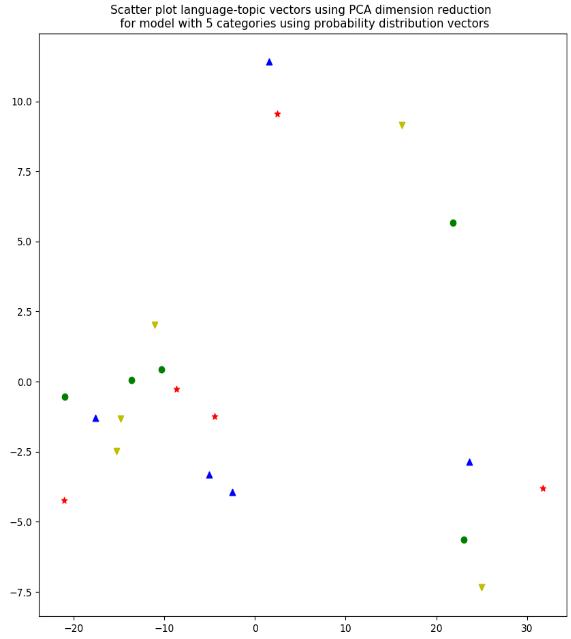
Wei, Xiu-Shen, et al. 2017. Deep Descriptor Transforming for Image Co-Localization. *IJCAI*, 2017.

Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM COMPUTING SURVEYS*, 34(2002):1–47.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. 2016. The United Nations Parallel Corpus, Language Resources and Evaluation. *LREC '16*, Portoroz, Slovenia, May 2016.

Figure 1: Plotted PCA Results for Five Topics

(a)



(b)

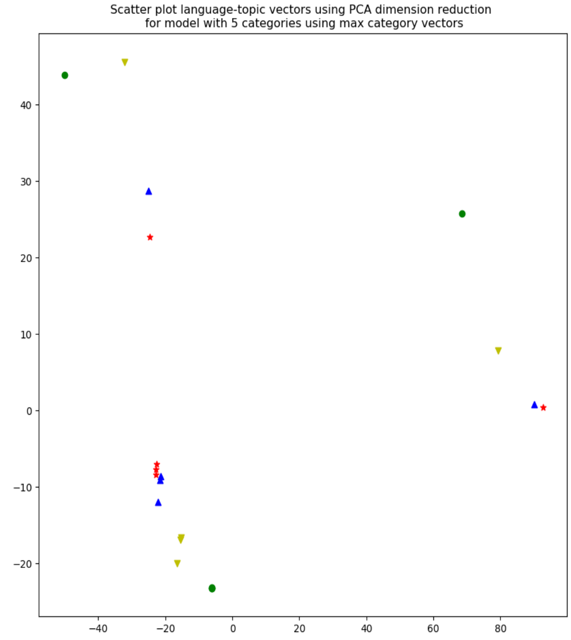
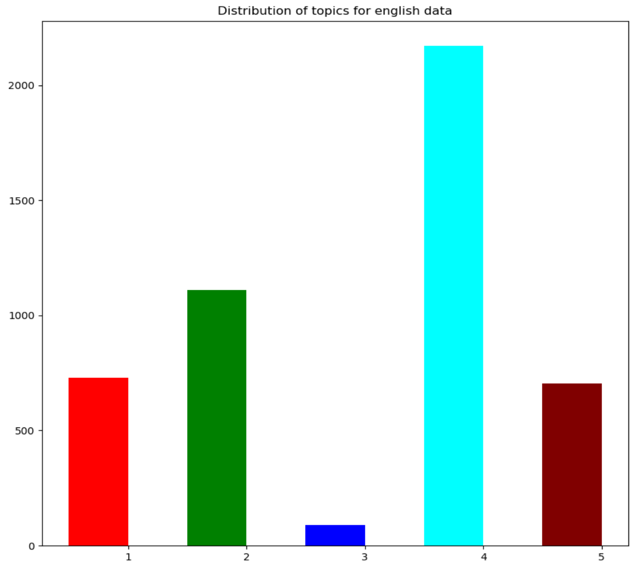


Figure 2: Distribution of Topics for English for Five Topics

(a) topic distributions for each document



(b) highest-scoring topic for each document

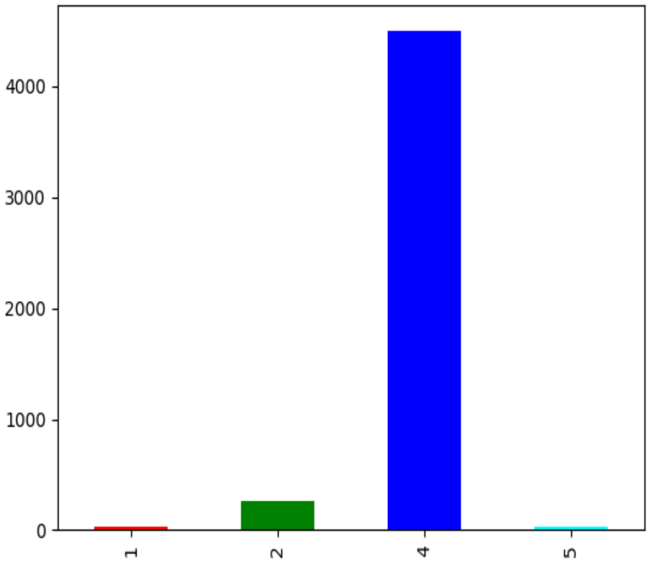
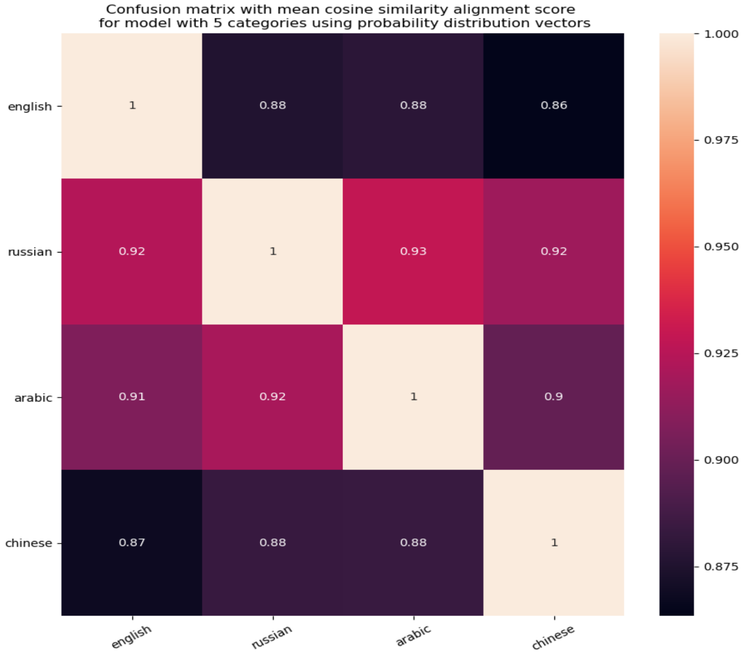
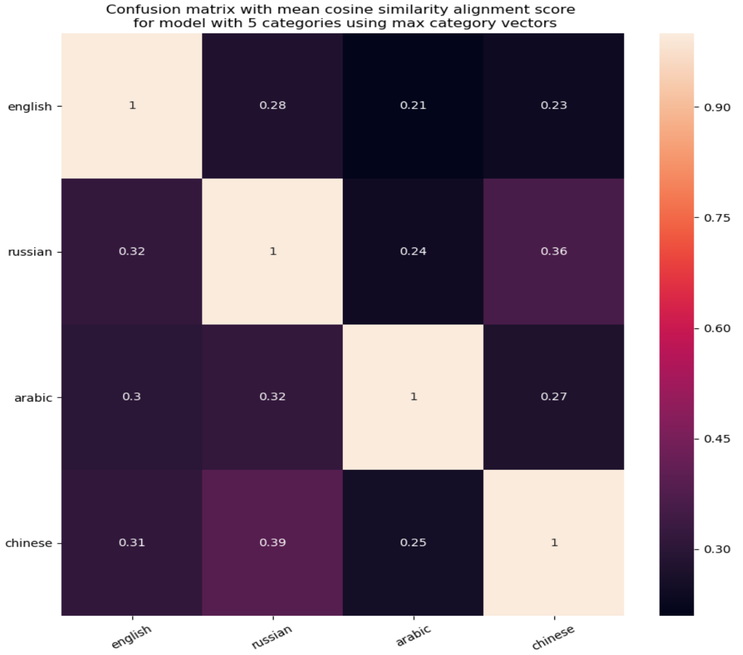


Figure 3: Confusion Matrices of Cross-Linguistic Vector Alignment for Five Topics

(a)



(b)



Language	Top 10 words for LDA model with 5 topics
English	1) state party committee international article paragraph convention session law report 2) right human woman child state law person committee national government 3) development country united nation programme international organization economic national policy 4) united nation service office staff support per report mission committee 5) security united council resolution nation international republic government state peace
Russian	1) year security united nations organization council resolution also which weapons 2) organization united nations year USA dollars issues period l also 3) the of and to in a for on that de [all of these terms were in English] 4) development field also organization women united nations year countries activities 5) rights year human also rights Committee respect rights conventions individuals
Arabic	1) rights Human Commission Subject Law Country In all Public about 2) United Nations Nations And public Commission General Administration Dollars desk 3)Nations And board Office of the United Nations Commission General Countries public Association In 4) the of and to in a for on that de [all of these terms were in English] 5) United Nations Nations the countries Public about Development the work For Strengthen field
Chinese	1) council Human rights Article Rights Convention organization 2) development Country Meeting Problems Society organization 3) Safety Resolution Meeting International United Nations Council Committee 4) People jobs United Nations Council report USD 5) the of and to in a for on that by [all of these terms were in English]

Table 1: The terms here were roughly translated using an online translation tool. At a glance, the words selected by our model tend to belong to a broad set of politically-related terms, which, given the nature of the data, was to be expected. One important fact to note is that some languages, like Arabic and Chinese, would encode in one token a term that might have required several in English, such as ”Office of the United Nations.” Another notable observation is the presence of English prepositions in every non-English language. These were overlooked during pre-processing, but interestingly they fell into their own nicely-aligned topic across languages. Furthermore, the Chinese topics have fewer terms due to a pre-processing error in which language specific punctuation was not accounted for. While the generated topic sets have some promising keywords, these keywords in the U.N. Corpus may have not have been distributed in such a way to provide our model with sufficient information to create starkly contrasting topics.