LDA Based Similarity Modeling for Question Answering on a Turkish Wikipedia

İsmet Burak Kadron and Çağıl Uluşahin Sönmez
Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
{ismet.kadron, cagil.ulusahin}@boun.edu.tr

Abstract

We present a generative model for a Question Answering system applied to a Turkish corpora. The problem is "answering a natural language question using relevant documents". Topic modeling using LDA (Latent Dirichlet Allocation) to rank the similarities between a question and different passages in the document is a known method to solve this kind of problem as shown in the recent work of Celikyilmaz et. al. Their approach is multilingual however a morphologically rich language like Turkish might benefit a lot from a model that is fine tailored for such languages.

We propose a model that can handle the sparsity in MRLs (Morphological Rich Languages) better when compared to the original model. This model also takes syntactic properties (stems, inflectional affixes) into account because in MRLs substantial grammatical information, i.e., information concerning the arrangement of words into syntactic units or cues to syntactic relations, is expressed at word level. In the end, we apply both models to a Turkish corpora and compare the results.

Motivation

Question Answering (QA) is the task of automatic retrieval of an answer given a question. Usually processing the question, system receives several search phrases and uses those to find relevant documents from a corpora or a knowledge base.
Hazırcevap [1] is a Turkish question answering system designed for high-school students to assist their education. It is a rule based QA system that extracts parts of the question as a focus and the type and the character of the answer then is chosen based on these linguistically calculated focus words in the question.
In this project we aim to assist Hazırcevap in finding better answers by building a topic model relation between those focus words and the relevant documents.

Original Model

A passage in retrieved documents (document collection) is represented as a mixture of fixed topics, with topic z getting weight $\theta_z^{(s)}$ in passage s and each topic is a distribution over a finite vocabulary of words, with word w having a probability $\phi_w^{(z)}$ in topic z. Placing symmetric Dirichlet priors on $\theta^{(s)}$ and $\phi^{(z)}$, where α and β, the generative model is given by:

$$w_i|z_i, \phi_{wi}^{(z_i)} \sim Discrete(\phi^{(z_i)}), \quad i = 1, ..., W$$
$$\phi(z) \sim Dirichlet(\beta), \quad z = 1, ..., K$$
$$z_i|\theta^{(s_i)} \sim Discrete(\theta^{(s_i)}), \quad i = 1, ..., W$$
$$\theta^{(s)} \sim Dirichlet(\alpha), \quad s = 1, ..., S$$

S is the number of passages, K is the total number of topics, W is the total number of words in the document collection, and s_i and z_i are passage and topic of the ith word w_i, respectively.

Our goal is to calculate the expected posterior probabilities of a word in a candidate passage given a topic and expected posterior probability of topic mixings of a given passage, using the count matrices. $n_{w_i k}^{WK}$ is the count of w_i in topic k and n_{sk}^{SK} is the count of topic k in passage s.

$$\hat{\phi}_{wi}^{(z_i)} = \frac{n_{w_i k}^{WK} + \beta}{\sum_{j=1}^{W} n_{w_j k}^{WK} + W\beta} \qquad \hat{\theta}^{(s)} \frac{n_{sk}^{SK} + \alpha}{\sum_{j=1}^{K} n_{sj}^{SK} + K\alpha}$$

For comparison metric, we have used Information Retrieval Metric (IR) which is represented as follows. The advantage of IR over KL is that it is symmetric and there is no problem with infinite values because $\frac{p_s^{(z)} + p_q^{(z)}}{2} \neq 0$ if either $p_s^{(z)} \neq 0$ or $p_q^{(z)} \neq 0$

$$IR(p_q^{(z)}, p_s^{(z)}) =$$
$$KL(p_q^{(z)} || \frac{p_s^{(z)} + p_q^{(z)}}{2}) + KL(p_s^{(z)} || \frac{p_s^{(z)} + p_q^{(z)}}{2})$$

Proposed Model

Recently the effects of applying different stemming techniques such as fixed-length word truncation and morphological analysis are explored in the area of document summarisation [2, 3]. Following those studies we aim to overcome the data sparsity problem of morphologically rich languages by using different stemming approaches.

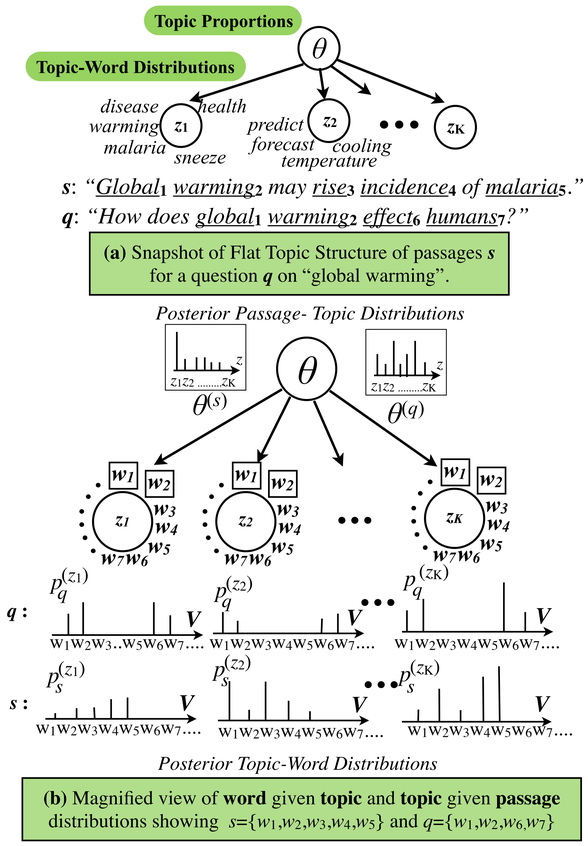

Figure 1: Topic-Word and Passage-Topic Distributions

Dataset and Results

We have used the Turkish Wikipedia dataset as the corpus. Our dataset contains 50000 sentences and 50 questions to train the topics. We have processed the original dataset to obtain the dataset with inflectional affixes removed, and trained those two different datasets using LDA, and compared the questions and sentences using IR and posterior document-topic distributions.

After comparison of questions and sentences and obtaining 50 candidate passages for each question, we compare if the top 50 passages for each question contain the answers to corresponding question and define our success metric as # of successful findings / # of questions.

	Original Model	Inflectional Model	Prefix-N Model
Success Ratio	20%		

Acknowledgements

We are grateful to Caner Derici and Yavuz Nuzumlalı for their help and support. We also want to thank Tunga Güngör and the team working with him on TÜBİTAK Project 113E036 for supplying us the data set and the opportunity to work further on their problem.

References

[1] C. Derici, K. Celik, A. Ozgur, T. Gungor, E. Kutbay, Y. Aydin, and G. Kartal, "Rule-based focus extraction in turkish question answering systems," in *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pp. 1604–1607, IEEE, 2014.
[2] E. Galiotou, N. Karanikolas, and C. Tsoulloftas, "On the effect of stemming algorithms on extractive summarization: a case study.," in *Panhellenic Conference on Informatics*, pp. 300–304, 2013.
[3] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, and O. M. Vursavas, "Information retrieval on turkish texts," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 3, pp. 407–421, 2008.