Persian Metaphor Frequency Prediction through Latent Dirichlet Allocation Model

Hadi Abdi Ghavidel
Sharif University of Technology
hadi_stlt@yahoo.com

Parvaneh Khosravizadeh
Sharif University of Technology
khosravizadeh@sharif.ir

Afshin Rahimi
University of Melbourne
arahimi@student.unimelb.edu.au

Persian Metaphor Frequency Prediction through Latent Dirichlet Allocation Model

Abstract

In this paper, we aim to predict the approximate frequency of metaphorical concepts in Persian language. As a first step, we apply LDA topic modeling on so-called Bijankhan corpus to extract topics. The extracted topics carry the words which share the most natural semantic proximity. Then, we develop a system for classifying natural and metaphorical sentences. Using the words of the topics, our system determines an overall topic for each sentence in the corpus. This system works on the assumption that if the overall topic of the sentence diverges from the topic of a word in the sentence, metaphoricity is detected. We have evaluated the system manually on 100 sentences and achieved the f-measure of 68.17%. Finally, we experiment and conclude that every at least two and at most four sentence seen in the corpus carries metaphoricity.

Keywords: Frequency, Persian language, LDA, Metaphoricity

1. Introduction

Human daily communication seldom happens in an invariant fashion and usually keeps pace with his creative thought. This creative thought which could often be a bridge between an abstraction and concreteness is built through metaphors. Metaphors help human readily understand one abstract idea in terms of, or in relation to another more concrete and physical one. The following sentence simply illustrates a rudimentary example of metaphor in Persian with its literal translation.

افت کرده است	روحیه‌ام
oft kærde æst	rʊhɪeæm
Dropped	My mood

Meaning: I am sad

In the above metaphorical expression, *my mood (rʊhɪeæm)* is considered something physical and, therefore, its change is associated with the act of dropping. These are: the orientational metaphor, the ontological metaphor, and the structural metaphor. Lakoff and Johnson (1980) extended the definition of metaphor to any symbolic type of expressions, like the concept of hate, the spatial direction "up", or the experience of inflation. According to them, three basic types of metaphor are: the orientational metaphor, the ontological metaphor and the structural metaphor. The metaphor in the abovementioned sentence exemplifies *orientational* or up-down *spatialization* metaphor, here SAD IS DOWN.

Recently, interest has grown tremendously in the studies of metaphor in Cognitive science. Cognitive studies of metaphor do recognize and understand metaphorical language comprehension by presenting subjects with linguistic stimuli and observing their responses. Unfortunately, however, less data amount and more time for recording data are the major obstacles for the cognitive researchers to achieve an acceptable output in a short period of time. To remove these obstacles, corpus linguistics could help provide a large amount of data for cognitive and psycholinguistic studies. Therefore, we aimed to use Persian corpus instead of Persian subjects in this research. Our hope is that cognitive science studies with unlabeled data and NLP techniques correspond to high-accuracy metaphor analysis in Persian language, even when our experiment is naïve for Persian language.

Our major goal, in this research, is to analyze conceptual complexity in Persian culture through predicting the metaphor frequency in Persian language. For this purpose, we intend to develop an automated system to classify the whole Persian corpus into natural and metaphorical expressions. Since Persian is a low-resource language and there is not any corpus specified with cognitive metaphors, we apply Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) which requires only an adequate amount of raw text. In our research, the task is one of recognition, and we use heuristic-based methods in an unsupervised approach to identify and predict the presence of metaphor in unlabeled textual data. To keep applying the results of it to psycholinguistic area too, the present study aims to produce a model which can automatically estimate how often a word is used metaphorically.

The remainder of the paper is as follows. In section 2, some prior works on manual and automatic metaphor estimation methods done in other languages but Persian are reviewed. In section 3, the data and methodology is descried. In Section 4, the experimental results are reviewed. The last section is devoted to make the conclusion and have further discussions.

2. Related works

Researchers have used different methods to estimate metaphor frequency in different languages. Pollio et al. (1990) analyzed a variety of texts manually and concluded that five metaphors exist in every text of about 100 words. Martin (1994) calculated the frequency of the types of metaphor on a sample of 600 sentences from the Wall Street Journal (WSJ), and concluded among other things that the most frequent type of WSJ metaphor was VALUE AS LOCATION. Martin (2006) in another paper noted that the probability of metaphoric concept was greatly increased in 2400 WSJ after a first metaphorical concept had already been observed.

Sardinha (2008) used a corpus of Portuguese conference calls and general Brazilian corpus to identify 432 terms that were used metaphorically. He found that on average these terms were used metaphorically 70% of the time.

What these researches have concluded may yield a small output to introduce a limited illustration of metaphor. However, for a general and every reliable analysis a large data set is needed. On the other hand, working with large data set and annotating them with either metaphorical or natural sentences is such an absolutely time consuming task. As a result, NLP machine learning techniques should be applied. One of the most reliable techniques is Latent Dirichlet Allocation which is introduced by Blei et al. in 2003.

Bethard et al. (2009) trained an SVM model with LDA-based features to recognize metaphorical sentences in large corpora. There the work is framed as a classification task, and supervised methods are used to label metaphorical and literal text.

Heintz et al. (2013) based a heuristic based model on LDA topic modeling, enabling metaphor recognition application to English and Spanish texts with no labeled data. He achieved an F-score of 59% for English.

Since Persian is a low-resource language and NLP combined with Cognitive analysis have not done on it yet, we base our model on the aforementioned LDA topic modeling and develop a classifier to predict the location of metaphoricity in Persian Corpus which represents a Persian Language.

3. Data

Persian or so-called Bijankhan corpus (2011) is a first and foremost corpus that is suitable for natural language processing research on the Persian (Farsi) language. This large corpus consists of daily news and common texts. We choose this rich corpus to serve as our data for exploring the frequency estimation of Persian metaphorical concepts.

Since the characters in Bijankhan corpus lack homogeneity and this problem disturbs the processing of our task and affects the accuracy substantially, we used Aminian[1] (2013) version of the corpus. Then, the whole corpus was normalized based on our convention so that we should yield acceptable results. Furthermore, we did a stemming task on all the words in the corpus to help topic modeling process not get trapped in lots of different forms of the same words. We stemmed all the words from different syntactic categories in a rule-based manner.

4. Latent Dirichlet Allocation

To operationalize the identification and prediction of the presence of Persian metaphor in unlabeled text, we employed a statistical generative topic model named Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA examines how word tokens, as a discrete data, co-occur in a corpus and identifies topics consisting of a group or mixture of statistically semantically similar words. For example, Table 1 shows a few topics from the Bijankhan corpus. These topics can be thought of as grouping words by their semantic domains. For example, we might think of topic 03 as the Animal (*he ɪ van*) domain.

[1] . In order to keep homogeneity until the end of the paper, we call Aminian's version of Bijankhan corpus also Bijankhan Corpus

The LDA algorithm is compared to a process someone might go through when writing a text. This generative process looks something like what Bethard (2009) brought in the following steps metaphorically:

1. Decide what topics you want to write about.
2. Pick one of those topics.
3. Think of words used to discuss that topic.
4. Pick one of those words.
5. To generate the next word, go back to 2.

T	Words
03	گربه(3%)، سگ(2%)، ببر(2%)، گوشت(2%) (2%)gʊʃt ،(2%)bæbr ،(2%)sæg ،(3%)gorbe (2%)meat ،(2%)tiger ،(2%)dog ،(3%)cat

Table 1: Topics and words

Formally, Bethard described the process above as:

1. For each document d select a topic distribution $\theta^{d} \sim Dir(\alpha)$
2. Select a topic $z \sim \theta^{d}$
3. For each topic select a word distribution $\varphi^{z} \sim Dir(\beta)$
4. Select a word $w \sim \varphi^{z}$

LDA learning algorithm is maximizing the likelihood of all the documents, where for one document we have the following equation.

$$(1) \qquad p(d|\alpha, \beta)=\prod_{i=1}^{N=1} p(w_i|\alpha,\beta)$$

In this research, Gibbs sampling (sampling from posterior distribution in case of joint distribution or full conditional distribution) is used to estimate the probabilities as it has been used by Bethard (2009) too and is available in the Mallet toolkit (McCallum, 2002).

Gibbs sampling begins by randomly assigning topics to all the words in the corpus. Then, the word-topic distributions and document-topic distributions are estimated using the following equations:

$$P((z_i|z_{i-}, w_i, d_i, w_{i-}, d_{i-}, \alpha, \beta)) = \frac{\varphi_{ij}\theta_{jd}}{\sum_{k=1}^{T} \varphi_{it}\theta_{td}} \tag{2}$$

$$\varphi_{ij} = \frac{C_{word_{ij}} + \beta}{\sum_{k=1}^{W} C_{word_{kj}} + W\beta} \qquad \theta_{jd} = \frac{C_{doc_{dj}} + \alpha}{\sum_{k=1}^{T} C_{doc_{dk}} + T\beta} \tag{3}$$

$C_{word_{ij}}$ is the number of times word i was assigned topic j, $C_{doc_{dj}}$ is the number of times topic j appears in document d, W is the total number of unique words in the corpus, and T is the number of topics requested. In fact, LDA counts the number of times that a word is assigned a topic and the number of times a topic appears in a document, and it uses these numbers to estimate word-topic.

We ran LDA over the documents in the Bijankhan corpus, extracting 50 topics after 2000 iterations of Gibbs sampling. We left the α and β parameters at their Mallet defaults of 0.1 and 0.01, respectively.

5. Persian Metaphor Frequency Prediction

5.1. Persian Metaphor Classifier

Our primary goal is to use the topics produced by LDA to help characterize words in terms of their metaphorical frequency. We develop a system for classifying natural and metaphorical sentences. Using the words in each topic, our classifier determines an overall or general topic for each sentence in the corpus. By a self-assumed hypothesis, we set a condition that if the overall topic of the sentence diverges from the topic of a word in the sentence, metaphoricity should be the result. In other words, the system checks all the words of a sentence and then names a sentence with one of the 50 topics extracted through LDA. The system further checks if there is any word which doesn’t belong to the overall topic. If yes, the sentence is marked as metaphor (MS). On the opposite side, the sentence is marked as natural sentence (NS).

Finally, we ran our system on the whole corpus and placed M before metaphorical and N before natural sentences. The following example makes this analysis clear:

داد	**نشان**	**پزشکی**	**تحقیقات**
dad	*ne∫an*	*peze∫k ɪ*	*tæhq ɪqat*
	Showed	*Medical*	*Researches*

(Medical researches showed)

Here the topic of the words *researches* and *medical* is summed up to the topic 23. However, the verb *show* belongs to the topic 12. This shows deviation from the overall or the most general topic. Therefore, a kind of metaphor could be observed here.

Another example makes the metaphor recognition even more clear:

دارد	جهانی	بازار	در	بالایی	قیمت	دلار
daræd	*jæhan ɪ*	*Bazar*	*Dær*	*bala ɪ*	*qe ɪmæt*	*dolar*
has	*World*	*market*	*In*	*high*	*Price*	*dollar*

(The dollar has a high price in the world market.)

In this example, the topic of the dollar, price, market and world are summed up to the topic 40. However, the word *high* is included in the topic 06. This shows deviation from the overall or the most general topic. Therefore, a kind of metaphor could have occurred here.

5.2. System Evaluation

In order to determine the quality of our classifier, we selected 100 sentences randomly from the corpus to analyze for metaphoricity. The number of words in these sentences is more than 4. Then, we gave these sentences to the system and analyzed them manually. Sixty-seven sentences out of them are correct and the rest are determined incorrect. For our classification task, we determined true positives, true negatives, false positives, and false negatives. Table 2 gives the numerical value information for each one of them. The terms positive (p) and negative (n) refer to our classifier's prediction (correct or incorrect), and the terms true and false refer to the states of metaphor and natural.

Number of Sentences	True (Metaphor)	False (Natural)	Number of Sentences
45	**tp**: correctly metaphor	**fp**: correctly natural	**22**
13	**tn**: incorrectly metaphor	**fn**: incorrectly natural	**20**

Table 2. Number of true positives, true negatives, false positives, and false negatives

Based on the information in Table 2 and the following formulas, we now calculate the accuracy, precision, recall and f-measure for our system.

$$(4) \quad \boldsymbol{accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$(5) \quad \boldsymbol{precison} = \frac{tp}{tp + fp}$$

$$(6) \quad \boldsymbol{recall} = \frac{tp}{tp + fn}$$

$$(7) \quad \boldsymbol{f - measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

According to the Figure 1, this system works well with the f-measure of 68.17. This shows a promising manner for our classifier in this very first step for analyzing metaphor in Persian language.

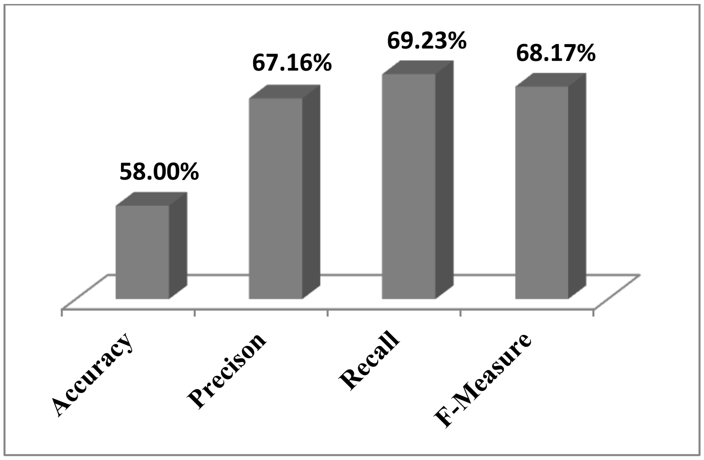

Figure 2. Persian Metaphor Classifier Evaluation

6. Experimental Results

We ran our system on the whole corpus to mark metaphorical and natural sentences. The number of sentences in the Bijankhan corpus is 381983 according to our tokenization algorithm and preprocessing (Aminian, 2013). After our first analysis, we concluded that there are 95453 sentences which carry metaphoricity. It means there is a sentence among every four sentences in the corpus that includes metaphorical concept.

After doing the first phase, we also checked them manually in a random. We saw that some of the sentences are 50% metaphorical and 50% natural. We chose to suppose them as metaphorical to achieve a periodical result.

According to the number of metaphorical sentences in the first phase and in the second phase, we came to conclusion that every at least two and at most four sentence seen in the corpus carries metaphoricity. An overview of our result could be seen in Figure 2.

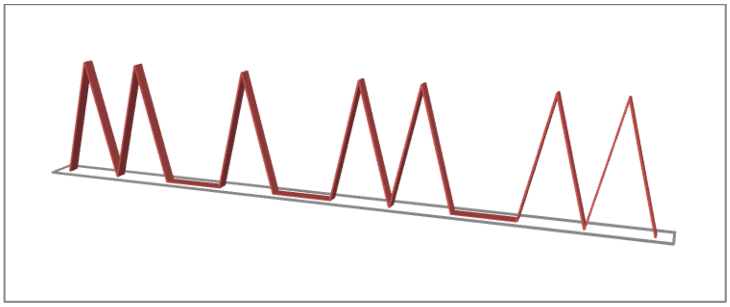

Figure 2. Schematic Panorama of Metaphor Existence in Persian Speech

7. Conclusions

We presented a system which identifies metaphorical sentences. This presentation is very novel for Persian language on the basis of cognitive studies. It could be directly transferable to a large number of Persian language processing applications that can benefit from Psycholinguistic studies on Persian subjects.

We tested running LDA topic modeling technique for metaphor discovery in Persian language. Our approach of looking for overlapping semantic concepts allows us to find metaphors of any syntactic structure. Using the topics extracted through LDA, our system calculates an overall topic for each sentence in the corpus. We showed that if the overall topic of the sentence diverges from the topic of a word in the sentence, Persian metaphoricity is detected. We concluded that every at least two and at most four sentences seen in the corpus carries metaphoricity.

Since this system works on unlabeled data, it may undergo some deficiencies like the lack of theta-roles (Fillmore, 1971) in the corpus or the exact type of metaphor according to Lakoff and Johsnon (1980). We have stepped in this Persian journey and try to improve these deficiencies in our next steps. We hope this research could pave the way for conducting lots of cognitive researches through NLP and CL techniques.

References

Aminian, M., Rasooli, M. S., and Sameti, H. (2013). *Unsupervised Induction of Persian Semantic* Verb Classes Based on Syntactic Information, Language Processing and Intelligent Information Systems. Springer Berlin Heidelberg, 112-124.

Bethard, S., Lai, V. T., Martin , J. H. (2009). *Topic Model Analysis of Metaphor Frequency for Psycholinguistic Stimuli*. In Proceedings of the North American Chapter of the Association for Computational Linguistics.

Bijankhan, M., Seikhzadeghan, J., Bahrani, M. and Ghayoomi, M. (2011). *Lessons from Creation of a Persian Written Corpus: Peykare, Language Resources and Evaluation Journal*, 45(2):143-164.

Blei, D. M., Andrew, Y. Ng, and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Fillmore, C. J. (1971). *Types of lexical information*, in Steinberg, D.; Jacobovitz, L., Semantics: An interdisciplinary reader in philosophy, linguistics and psychology, Cambridge University Press.

Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M. and Weischede, R. (2013). *Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling*. In Proceedings of The 1st Workshop on Metaphor in NLP (co-located with NAACL-HLT 2013), Atlanta, Georgia, USA.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press, Chicago.

McCallum, A. K. (2002) *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu

Martin, J. H. (1990). *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA, USA.

Martin, J. H. (2006). *A rational analysis of the context effect on metaphor processing*. In Stefan Th. Gries and Anatol Stefanowitsch, editors, Corpus-Based Approaches to Metaphor and Metonymy. Mouton de Gruyter.

Pollio, H R., Smith, M. K. and Pollio, M. R. (1990). *Figurative language and cognitive psychology*. Language and Cognitive Processes, 5:141–167.

Sardinha, T. B. (2008). *Metaphor probabilities in corpora*. In Mara Sofia Zanotto, Lynne Cameron, and Marilda do Couto Cavalcanti, editors, Confronting Metaphor in Use, pages 127–147. John Benjamins.