IJICTR International Journal of Information & Communication Technology Research

Volume 7- Number 2- Spring 2015 (33-40)

Impact of Topic Modeling on Rule-Based Persian Metaphor Classification and its Frequency Estimation

Hadi Abdi Ghavidel
Language and Linguistics Center
Sharif University of Technology
Tehran, Iran
hadi_stlt@yahoo.com

Parvaneh Khosravizadeh
Language and Linguistics Center
Sharif University of Technology
Tehran, Iran
khosravizadeh@sharif.ir

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

Received: January 12, 2014-Accepted: December 29, 2014

Abstract—**The impact of several topic modeling techniques have been well established in many various aspects of Persian language processing. In this paper, we choose to investigate the influence of Latent Dirichlet Allocation technique in the metaphor processing aspect and show this technique helps measure metaphor frequency effectively. In the first step, we apply LDA on Persian or so-called Bijankhan corpus to extract classes containing the words which share the most natural semantic proximity. Then, we develop a rule-based classifier for identifying natural and metaphorical sentences. The underlying assumption is that the classifier allocates a topic for each word in a sentence. If the overall topic of the sentence diverges from the topic of one of the words in the sentence, metaphoricity is detected. We run the classifier on whole the corpus and observed that roughly at least two and at most four sentence in the corpus carries metaphoricity. This classifier with an f-measure of 68.17% in a randomly 100 selected sentences promises that a LDA-based metaphoricty analysis seems efficient for Persian language processing.**

Keywords-Impact, LDA, Persian language, Metaphoricity

I. INTRODUCTION

With the advent of new natural language processing algorithms, the problems of looking for a needle in a haystack have been largely alleviated in big textual data. One of these new algorithms is topic modeling. A topic modeling is a type of statistical modeling for discovering the hidden topics or topic-based patterns that exist in a set of heterogeneous data collections. Three famous topic models are: Latent Semantic Allocation (LSA), Probabilistic Latent Semantic Allocation (PLSA) and Latent Dirichlet Alocation (LDA). They all work on the fundamental assumption that any set of documents consists of a body of topics and each topic contains a bag of words with close semantic proximity. However, their mathematical framework is different in terms of linearity and probability. Each model has its own advantages and disadvantages. PLSA seems to enjoy an accurate distribution and a statistically formulated hypothesis

over LSA. Modular LDA is a Bayesian version of PLSA, which helps the algorithm avoid overfitting the data. Accordingly, LDA is growing in popularity and being applied quite often in the analysis of many aspects of natural language processing. Recently, it penetrated into the micro and macro levels of Persian linguistics. Macro linguistics is more challenging and newer in the field of computational linguistics (CL) in comparison with micro linguistics. In the current research, we dare to analyze LDA applicability in one of the hot topics of Persian macro language aspects if not hotter for other languages, metaphor.

Intuitively, 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 compact and memorable mode of expressing meaning that would be difficult to communicate with normal words is highly pacified through this most valuable tool. The following sentence simply illustrates a rudimentary example of metaphor in Persian with its literal translation.

Example 1

افت کرده است	روحیه‌ام
oft kærde æst	ɾʊhɪeæm
dropped	My mood

Meaning: I am sad

In the abovementioned metaphorical expression, *my mood* (ɾʊhɪeæm) is considered something physical and, therefore, its change is associated with the act of dropping.

Lakoff and Johnson [1] 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, SAD IS DOWN.

Traditional studies of metaphor, however, treated it as a deviation from normal way of integrating concepts into discourse. They classified metaphor into two types: dead metaphors and live metaphors. Dead metaphors or conventional metaphors are said to be used once by a speaker and then added to the lexicon of the speaker's language. Live metaphors, on the other hand, are new to the listener and thus potential to become dead just after being uttered for the frost time. Nevertheless, cognitive researchers like Lakoff and Johnson gave new shape to the definition of metaphor and viewed it as associating an idea out of its natural environment. Example 1 illustrates this association.

Metaphor definition is one side of the problem and its detection, interpretation and disambiguation is the far end of the line. Over the thousands of years of studies on metaphor, it's been believed that human mind once generated metaphor and the other time put intellectual creativity into detecting and disambiguating it on the basis of literary, philosophical or cognitive attitudes with a certain number of sentences. Recently, cognitive science has shown competing interest in the studies of metaphor. 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 Natural Language Processing (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 how LDA topic model has the desired effect in predicting the metaphor processing and frequency measurement in Persian language. It should be noted that Persian language[1] module of our classifier includes the one which is spoken specifically in Iran. To achieve our goal, we intend to develop an automated classifier to identify the natural and metaphorical expressions in the Persian corpus. Since Persian is a low-resource language and there is no corpus specified with any kind of metaphorical tags, LDA topic modeling (Blei et al., [2]) serves a helpful function on 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 show how effectively LDA helps us estimate the number of times a word is used metaphorically. In other words, the outcome of our analysis depicts the way LDA succeeds in highlighting clearly what density of a Persian mind's language module is made up of metaphorical concepts.

The remainder of the paper is as follows. In section 2, application of LDA topic model in several areas and also works on manual and automatic metaphor detection and estimation methods which have been done in other languages but Persian are reviewed comprehensively. In section 3, the topic construction is described. Persian metaphor frequency measurement is demonstrated and evaluated in section 4. In the last but one section of the paper, the experiments and results are illustrated in detail. Finally, the last section is devoted to making the conclusion and introducing the contributions.

[1] Persian is an Indo-European language which is spoken in Iran, Afghanistan and Tajikistan.

II. Related works

In recent years, researchers have used LDA to perform a variety of functions in the several areas of computational linguistics.

As soon as LDA was introduced by Blei [2] in 2003, Marlin [3] applied it in his *User Rating Profile* (URP) system. He proved LDA model is very useful for sifting out the fresh information through collaboration among different sources.

Rosen-Zvi et. al. [4] introduces an author-topic model using LDA. Rosen-Zvi constructed a probabilistic model to analyze the relationships between authors, documents, topics and words. The model was proved to yield better results in terms of perplexity compared to a more impoverished author model.

Purver et. al. [5] showed how Bayesian inference in the generative model can be used to simultaneously address the problems of topic segmentation and topic identification. The developed model segments multi-party meetings into topically coherent blocks. Therefore, LDA model leads to generate a well-established discourse.

Mei et. al. [6] discovered interesting spatiotemporal theme patterns, theme life cycles and theme snapshots effectively through LDA. They operated a robust sub-topic mining in weblogs and extracted themes from them. Finally, they generated a theme snapshot for any time period.

Bhattacharya and Getoor [7] extended LDA model for collective entity resolution in relational domains where there is a connection between each of them. Two real-world bibliographic datasets are evaluated for the applicability of the approach. Furthermore, their model calculates and estimates the number of entities from the references, which seems useful for detecting the underlying conditions.

Biro et. al. [8] applied a modification of LDA, the novel multi-corpus LDA technique for web spam classification. Their work is the first web retrieval application of LDA. They tested this new model on the UK2007-WEBSPAM corpus, and saw that the F-measure of the system increased by 11%.

Since LDA has been used as a model in the inference systems, it has recently used as a model to process metaphorical concepts in big data. However, manual identification of metaphors was conducted before the appearance of the generative models.

Smith et al. [9] analyzed metaphor density in a body of well-known works of American literature written between 1675 and 1975, a corpus of over 500,000 words from about 24 authors. Smith suggested the average number of metaphors is three among 500 words a page.

Arter [10] and Dixon et al. [11] investigated metaphoricity in the educational texts. They assigned level for the texts and generalized a view that there are two metaphors for every 120 words in a third-grade text. Also, there are five words carrying metaphoricity for every 500 words in an eleventh-grade text.

Pollio et al. [12] analyzed a variety of texts manually and concluded that five metaphors exist in every text of about 100 words. Martin [13] calculated the density 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 [14] in another paper noted that the probability of metaphoric concepts was greatly increased in 2400 WSJ after a first metaphorical concept had already been observed.

Sardinha [15] searched for metaphors in the general Brazilian corpus of conference calls. He found out that metaphorical meaning appears at a rate of one out of every 20 words. His generalization is based on 432 terms, which is very inadequate and incomprehensive.

What these researches have concluded is based on their own manual way of identifying metaphors and also counting them manually. This might yield a small output to introduce a limited illustration of manually labeled metaphors. To resolve these shortcomings, two measures were taken over the years. Firstly, metaphorical expressions should be processed in a highly automated manner and then counted within a few lines of code.

Automated metaphor processing is almost a new area of studies over the years of conducting NLP researches. It dates back to thirty years ago. Since then, there have been several studies conducted in this area on many languages like English, Russian, Japanese and etc. but a few in Persian. These studies could fall into a shallow categorization according to Abdi et. al. [16]:

- Data bank-based approaches: (see Barden [17])

- Ontology-based approaches: (see Gruber [18], Fass [19])

- Tagged corpus-based approaches: (see Gedigan [20])

- Linguistic sings-based approaches: (see Goalty [21])

- DOTL[2] fusion approaches: (see Krishnakumaran and Zhu [22])

It is axiomatic that 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 and costly task. As a result, NLP specialists decided to apply machine learning techniques in order to avoid further manipulation. They utilized several techniques, but they found out that LDA serves their purpose well in metaphoricity processing.

Bethard et al. [23] 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.

[2] . Abbreviation of Data bank-based, Ontology-based, Tagged corpus-based and Linguistic-based

Heintz et al. [24] 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.

Persian is a low-resource language, i.e., the number of well-structured recorded data is very low. On the other hand, carrying out cognitive analysis through data processing techniques have not been done on this language yet. As a result, 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. We show that LDA is the best available and most suitable model to process the Persian metaphor inference in the current Persian NLP status.

III. Topic Construction

A. Data Normalization

Persian or so-called Bijankhan corpus [25] is the 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. In this linguistic data set, all documents fall into different subject areas such as economic, sports, religious, politics and so on. We choose this rich corpus to serve as our data for exploring the LDA influence in the frequency measurement of Persian metaphorical concepts. The Bijankhan corpus contains about 2.6 million manually tagged words with a tag set that contains 40 Persian POS tags. After the data normalization phase, we ignore all these tags in order to build a fully-unsupervised classifier.

Since the characters in Bijankhan corpus lack homogeneity and this problem disturbs the processing phase of our task and affects the accuracy substantially, we used Aminian [26] version of the corpus. In this version, the whole corpus is converted into tokens. One of the distinguishing characteristics of this corpus is that it marks the verbs based on their arguments as a combined type or an incorporated type. Then, we further polished the whole corpus and normalized it based on our convention so that we should yield acceptable results. Characters like Persian semi-spaces, abbreviations, and Arabic and Persian letters are altered and homogenized without any deep semantic change.

Afterwards, we performed a shallow stemming task on all the words in the corpus to help topic modeling process not get trapped in lots of different forms of a same word. By stemming, we mean striping off the ending part of the words in the hope of removing the inflectional affixes. Our stemmer analyzed all the words within different syntactic categories such as nouns, adjectives, verbs, adverbs and even prepositions in a rule-based process. Since prepositions could carry inflectional morphemes in Persian language, they should not be neglected or supposed as unwanted tokens. Our rules determine measures to enable the classifier to trim any of the inflectional suffix morphemes from the end of the words and prefix morphemes from the beginning of the words (specifically verbs). Table 1 shows a list of these Persian inflectional morphemes. We consider sixteen suffixes and two prefixes to make up our miscellaneous collection. Since the negative forms of the verbs come before the prefixes, they are also trimmed from the very initial part of the verbs too. Finally, there are only the bare lexemes which will be fed into the training phase of the classifier.

Table 1. Persian Common Suffixes and Prefixes

ها	ان	ات	ون			Suffix
ha	an	æt	un			
تر	ترین					
tær	tærin					
م	ی	د	یم	ید	ند	
æm	i	æd	im	id	ænd	
م	ت	ش	مان	تان	شان	
æm	æt	æʃ	man	tan	ʃan	
می	ب					Prefix
mi	be					

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a robust generative model which clusters all the words of a body of documents into different topics. The model analyzes the statistically semantic proximity among the words and put them into the same groups which are called topics. In the NLP field, the underlying idea behind LDA is that language files are represented as random combinations over latent topics, where each topic is represented by a distribution over a number of words. The process of building topics and the inner distribution is conducted within a three-level hierarchical Bayesian model. Figure 1 shows a graphical plate of LDA. The bigger plate represents documents and the smaller one belongs to the topics and words.

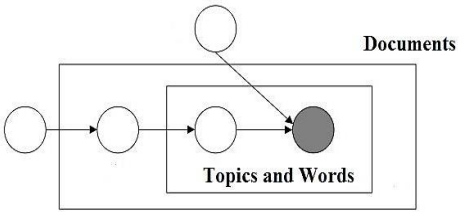

Fig. 1. LDA Plate

In a metaphorical sense, the LDA algorithm is compared to a process someone might go through when writing an essay or brainstorming and supporting an idea about a certain favorite topic. This generative process looks something like what Bethard [17] made clear in the following steps metaphorically. The steps are like hurdles which should be gotten over to go to the next one. It should be noted that all these steps need analytical thinking:

1. Determine a number of topics to write about.

2. Select one of the determined topics.

3. Fetch the appropriate words for the selected topic from the memory.

4. Select one of those fetched words.

5. To generate the next word, go back to 2.

In order to work like a human and think analytically, each step should be formalized and mathematized. Symbolically, the formation rule of the aforementioned process could be written in the following lines of code:

1. For each document d:

Sample θ^d ~Dir(α) (topic distribution)

2. Select a topic $z \sim \theta^d$

3. For each topic:

Sample φ^z ~Dir(β) (word distribution)

4. Select a word $w \sim \varphi^z$

In the LDA algorithm learning process, we have the following equation for each one of the documents in the bigger plate of Figure 1.

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

In this research, Gibbs sampling (sampling from posterior distribution in case of joint distribution or full conditional distribution) is used to estimate the probabilities. It's one of the most popular instances of a Markov Chain Monte Carlo technique, which provides a desired value by performing simulations which include probabilistic choices or decisions. Gibbs sampling is available in the MALLET toolkit (McCallum [27]). MALLET includes several methods for numerically optimizing functions, which alleviates a search for optimal parameters that maximize a log-likelihood function of our data.

Gibbs sampling [28] [29] assigns the conditioning variables, here topics, to all the words in our corpus through a recursive and random manner. Then, the word-topic distributions and document-topic distributions are estimated using the following equations:

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

$$(3) \qquad \varphi_{ij} = \frac{C_{word_{ij}} + \beta}{\sum_{k=1}^{W} C_{word_{kj}} + W\beta}$$

$$\theta_{jd} = \frac{C_{doc_{dj}} + \alpha}{\sum_{k=1}^{T} C_{doc_{dk}} + T\alpha}$$

$C_{word_{ij}}$ is the number of times word i was assigned topic j, $C_{word_{ij}}$ 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 α and β parameters at their Mallet defaults of 1 and 0.01, respectively. We optimized these parameters at ten optimize-interval iteration after a 200 iteration burn-in period. Cases in point for the topics or so-called classes which have been extracted through LDA could be observed in Table 2. These classes can be thought of as grouping words by their semantic domains. For example, we might think of topic 03 as the Animal (heɪvan) domain and topic 11 as the Municipality (ʃæhrdaɾɪ) domain.

Table 2. Topics and Words

T	Words
03	گربه(3%)، سگ(2%)، ببر(2%)، گوشت(2%) (3%) gorbe، (2%) sæg، (2%) bæbr، (2%) gʊʃt (3%) cat، (2%) dog، (2%) tiger، (2%) meat
11	تهران(3%)، شهرداری(2%)، شهر(4%)، شهرستان(3%) (3%) tehrɑn، (3%) ʃæhrdaɾɪ، (4%) ʃæhr، (3%) ʃæhrestan (3%) Tehran، (2%) municipality، (4%) city، (3%) town

IV. Persian Metaphor Frequency Measurement

A. Persian Metaphor Classifier

Our primary goal is to use the topics produced by LDA as classes to help classify sentences in terms of their metaphorical meaning. We develop an uncomplicated classifier for classifying natural and metaphorical sentences.On the basis of Selectional Preference (SP) [30] [31], the semantic content of the words are determined by their common shared properties. These shared properties could be found in what we have extracted out of the corpus i.e., the topics. Similar to the SP, LDA assigns to a specific topic only the words which include the most common sense between them. Incorporating these two points, namely LDA and SP, into our architecture leads us to think about a lexico-grammatical structure for Persian language. To clarify further, the semantic SP of the singing sense basically determines that the subject or theme must be physical object. From a deeper point of view, this theme cannot be a president in a normal sense. We believe this structure produces right conditions to operate metaphoricity analysis. Although the theory of SP focuses on predicates and arguments, we shall ignore the specific tags and attempt to build an unsupervised system. Accordingly, we devise a basic rule for our classifier which could be summed up in this way. 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 an immaculate condition that if the overall topic of the sentence diverges from the topic of a word in the sentence, metaphoricity should be the defining characteristics of the sentence. Figure 2 shows schematically the section division of the metaphor classifier system.

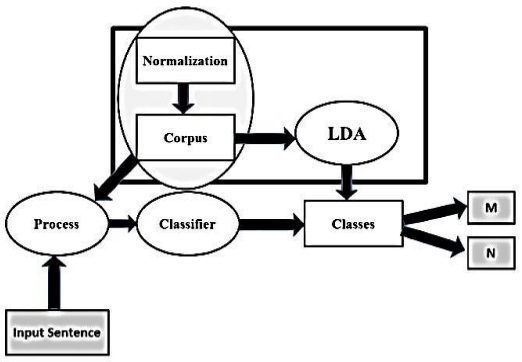

Fig. 2. Metaphor Classifier Diagram

As Figure 2 demonstrates, the classifier takes all the sentences from the normalized corpus which has not been segmented into the topics yet. Then, it checks all the words of a sentence for determining an overall topic from 50 topics extracted through LDA. The classifier further checks if there is any word which doesn't belong to the overall topic. When the topic of the word is recognized as deviant, the sentence is marked as metaphor (MS). On the opposite side, the sentence is marked as natural sentence (NS).

After running our classifier on the whole normalized corpus, we successfully built a metaphorically enriched corpus with an M tag before metaphorical and an N tag before natural sentences. The following example makes this analysis clear:

داد	نشان	پزشکی	تحقیقات
dad	*neʃan*	*pezeʃkı*	*tæhqıqat*
	showed	*Medical*	*researches*

(Medical researches showed)

In this example, the topic of the words *researches* and *medical* is summed up to occur in the topic 23. However, the verb *showed* belongs to the topic 12. This shows a form of deviation from the overall or the most general topic. Therefore, a kind of metaphor could be observed here.

Another example makes the metaphor recognition process even more clear:

دارد	جهانی	بازار	در	بالایی	قیمت	دلار
daræd	*jæhanı*	*bazar*	*dær*	*balaı*	*qeımæt*	*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 exist in the topic 40. However, the word *high* is included in the topic 06. This inclusion demonstrates an obvious deviation from the overall or the most general topic. Therefore, metaphorical concept could exist in this sentence.

B. Proposed Method 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 classifier and analyzed them manually. The number of correctly classified sentences is 76 and the rest of them are determined incorrect ones. For our classification task, we decided to set true positives, true negatives, false positives, and false negatives. Table 3 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 being metaphor and natural.

Table 3 shows that 45 out of 100 randomly selected sentences are identified as metaphorical, while 22 of all the sentences as natural. There's a low proportion of the wrong analysis, with thirteen sentences for incorrect metaphors and twenty sentences for incorrect naturals.

Table 3. Evaluation parameters and their values

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**

Based on the information in Table 3 and the following formulas, we now measure the classifier effectiveness by calculating the accuracy, precision, recall and f-measure.

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

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

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

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

Recall depicts the classifier's sensitivity and precision estimates the classifier's prediction. The harmonic mean of precision and recall is calculated through f-measure and the weighted arithmetic mean of precision and inverse precision as well as a recall and inverse recall calculated through accuracy. According to the Figure 3, this classifier works well (of course without being tuned) with the f-measure of 68.17. This shows a promising functionality for our classifier in this very first step of analyzing metaphor in Persian language through training data with LDA.

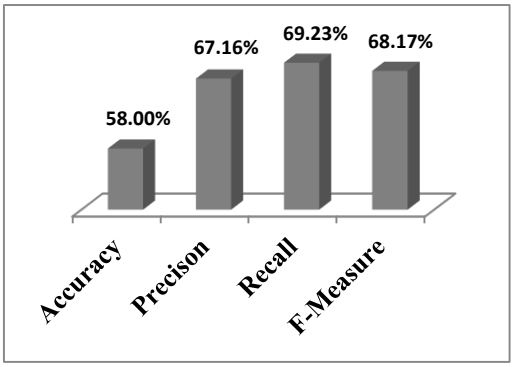

Fig. 3. Persian metaphor classifier evaluation based on the measurements percentage

V. Experiments and Results

We ran our classifier on the whole corpus to mark metaphorical and natural sentences. The number of sentences in the Bijankhan corpus [16] is 381983 according to our tokenization algorithm and preprocessing (Aminian [17]). After conducting 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 selection. 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. Eventually, the metaphor density of Bijankhan corpus ranges from 25% to 50%. An overview of our result could be seen in Figure 4.

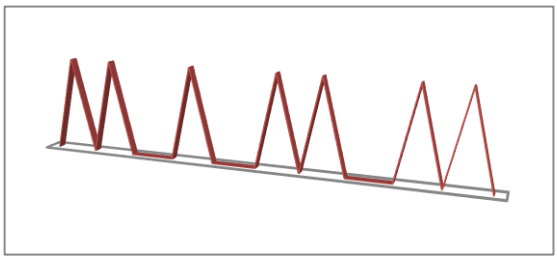

Fig. 4. Schematic panorama of metaphor existence in Persian according to Bijankhan corpus, where the x—axis represents metaphor and the y-axis represents the existence of metaphor

VI. Conclusions

We presented a classifier which identifies metaphorical characteristics of the sentences. This presentation is very novel for Persian language on the basis of extending the application of topic models. It could be directly transferable to a large number of Persian language processing applications that can benefit from processing the meaning and understanding the gist of natural language data.

We tested running LDA topic modeling technique efficacy 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 classifier calculates an overall topic for each sentence in the corpus. We determined 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.

We investigated the impact of LDA features on finding Persian metaphors. Overall, LDA was found to perform properly in Bijankhan corpus. Some general conclusions can be made: high F-measure obtained indicates topic models could be very effective as features to estimate Persian metaphor frequency. Accuracy over 50% also shows that topic model features outperform strong traditional and manual techniques.

Since this classifier works on unlabeled data, it may undergo some deficiencies like the lack of theta-roles (Fillmore [32]) in the corpus. A corpus enriched with theta tags catalyzes processing complex forms of metaphors for which we chose our system to manage sentences in a binary form and simply tags them metaphorical or natural. Another issue which must be taken into consideration in our future works is to recognize the exact type of metaphor according to Lakoff and Johnson [1]. The last but not the least is the feature of our rules. They could be validated more systematically according to the fuzzy logic because they are close to the exact reasoning and have been fixed partially yet. We have stepped in this Persian journey and will try to improve these deficiencies in our next steps. We hope this research could pave the way for conducting lots of automatic text understanding researches through NLP and CL techniques.

Acknowledgments

We appreciate Dr Kambiz Badie for his invaluable assistance which tuned us to finalize this research.

References

[1] G. Lakoff, and M. Johnson, Metaphors We Live By. University of Chicago Press, Chicago, , 1980.

[2] D. M. Blei, Y. Ng. Andrew and M. I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, 3:993–1022, 2003.

[3] B. Marlin. Modeling user rating profiles for collaborative filtering. NIPS, 2003.

[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and document. In UAI, 2004.

[5] M. Purver, K. Kording, T. Griffiths, and J. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In Proc. of COLING-ACL, 2006.

[6] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. Proc. 15th Int. World Wide Web Conference (WWW'06), 2006.

[7] I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. SIAM International Conference on Data Mining, 2006.

[8] I. Biro, D. Siklosi, J. Szabo, and A. A. Benczur. Linked latent dirichlet allocation in web spam filtering. Proc. 5th Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2009.

[9] M. K. Smith, H. R. Pollio, and M. K. Pitts, "Metaphor as intellectual history: Conceptual categories underlying figurative usage in American English from 1675-1975", Linguistics, 19, 911-935, 1982.

[10] J. L. Arter, "The effects of metaphor on reading comprehension". Doctoral dissertation, University of Illinois, Urbana, 1976.

[11] K. Dixon, A, Ortony, and D. Pearson, "Some reflections on the use of figurative language in children's books", In the proceedings of the 13th Annual Meeting of the National Reading Conference, New Orleans, April 1980.

[12] H R. Pollio, M. K. Smith and M. R. Pollio, "Figurative language and cognitive psychology", Language and Cognitive Processes, 5:141–167, 1990.

[13] J. H. Martin, "A Computational Model of Metaphor Interpretation", Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[14] J. H. Martin, "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, 2006.

[15] T. B. Sardinha, "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, 2008.

[16] H. Abdi Ghavidel, M. Bahrani and B. Vazirnezhad, "An investigation into the methods of detecting and disambiguating of metaphorical and metonymic concepts", In proc. of the 2nd conference of Computational Linguistics, Tehran, Iran, 2012. (in Persian)

[17] J. A. Barnden and G. M. Lee, "An artificial intel-ligence approach to metaphor understanding". Theoria et Historia. Scientiarum, 6(1):399–412, 2002.

[18] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2), 199–220, 1993.

[19] D. Fass, "Met*: A Method for Discriminating Metonymy and Metaphor by Computer", Computational Linguistics, 17(1), 49–90, 1991.

[20] M. Gedigian, J. Bryant, S. Narayanan and B. Ciric, "Catching Metaphors", In Proceedings of the 3rd Workshop on Scalable Natural Language Understanding, 41–48, New York, NY, 2006.

[21] A. Goatly, The Language of Metaphors, London: Routledge, 1997.

[22] S. Krishnakumaran and X. Zhu. "Hunting Eelusive Metaphors Using Lexical Resources", In Proceedings of the Workshop on Computational Approaches to Figurative Language, 13–20, Rochester, NY, 2007.

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

[24] I. Heintz, R. Gabbard, M. Srivastava, D. Barner, D. Black, M. Friedman and R. Weischede, "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, 2013.

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

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

[27] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu, 2002.

[28] S. Geman, and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 12: 609-628, 1984.

[29] M. A. Tanner, and W. H. Wong, "The calculation of posterior distributions by data augmentation", Journal of the American Statistical Association, 82: 528-549, 1987.

[30] N. Chomsky, Aspects of the Theory of Syntax. MIT Press, Cambridge, Massachusetts, 1965.

[31] R. Jackendo, Semantic Structures. MIT Press, Cambridge, MA and London, 1990.

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

Hadi Abdi Ghavidel received his M.Sc. degree in computational linguistics from Sharif University of Technology. His research interests include automatic metaphor processing, computational cognitive science, machine translation and computational discourse analysis.

Parvaneh Khosravizadeh received her B.A. in English Translation from Allameh Tabataba'i University,and her M.A. in General Linguistics from Central Tehran Branch, Islamic Azad University. She received her Ph.D. in General Linguistics at University of Tehran. She is now an assistant professor and faculty member at Sharif University of Technology in the field of computational linguistics. Her research interests include natural language processing, psycholinguistics and machine translation.

Afshin Rahimi received his B.Sc. in Computer Engineering and his M.Sc. in Computational Linguistics from Sharif University of Technology. He is currently a Ph.D. student in computer science at the University of Melbourne. His research interests include Natural Language Processing, Machine Learning for NLP and Information Retrieval.