A fully unsupervised Topic Modeling approach to metaphor identification*

Una aproximación no supervisada a la detección de metáforas basada en Topic Modeling

Borja Navarro-Colorado y David Tomás
Grupo de Procesamiento del Lenguaje Natural (GPLSI)
Universidad de Alicante
03690 San Vicente del Raspeig, Alicante (España)
{borja,dtomas}@dlsi.ua.es

Resumen: En este artículo se presenta una nueva aproximación no supervisada a la detección de metáforas, basada en el uso de LDA Topics Modeling. Se asume una correlación entre los tópicos extraídos mediante LDA Topic Modeling y los dominios conceptuales, de tal manera que se utilizan los tópicos de cada palabra para detectar la inconsistencia semántica entre una palabra y su contexto. El mı́odo que presentamos es totalmente no supervisado: no requiere de ningún recurso léxico ni corpus anotado. Se presentan varios experimentos con diferentes cantidades de tópicos, con el objetivo de definir el nivel de granularidad apropiado para esta tarea. Con una evaluación preliminar se obtiene una precisión de hasta el 70% y un nivel de F de 0,72 en los mejores casos.
Palabras clave: Deteccin de metforas, Topics Modeling, LDA

Abstract: This paper presents a new unsupervised approach to metaphor identification based on LDA topic modeling. Assuming a correlation between topic models and conceptual domains, the topics of each word are used to identify the semantic inconsistency between a word and its context. The system proposed is fully unsupervised, since it does not require any lexical resource nor manually annotated corpus. Some experiments with different topic granularities are used in order to define the best set of topics. The preliminary results obtained provide an accuracy level up to 70% and an F-Measure up to 0,72.
Keywords: metaphor detection, Topic Modeling, LDA

1 Introduction

Automatic metaphor analysis is usually divided into two tasks: metaphor recognition (or identification) and metaphor interpretation (Shutova, 2010). The goal of the first task is to distinguish between literal and metaphorical uses of words, expressions or collocations. The aim of the second task is to identify the appropriate and contextual meaning of the metaphorical expression. The work presented in this paper is focused on metaphor identification.

The application of Latent Dirichlet Allocation (LDA) topic modeling (Blei, Ng, and Jordan, 2003) to metaphor identification is not a new idea. Topic modeling is a family of algorithms that automatically discover topics from a collection of documents. Specifically, LDA assigns a topic to each word according to the topics of co-ocurrent words in the document and the topics assigned to this word in other documents. At the end, each topic is represented as a set relevant words.

Several papers have explored this approach in different ways (Bethard, Lai, and Martin, 2009; Heintz et al., 2013). The main idea of these approaches is to consider topic models as semantic domains (Bethard, Lai, and Martin, 2009). Therefore, following the framework of conceptual metaphor (Lakoff and Johnson, 1980), a metaphor is a word, expression or collocation usually related to a set of topics in the source domain, but used in a specific context to refer to a different set of topics in the target domain.

As we will show later, all of them present supervised or semi-supervised approaches. In this paper we want to go further with this approach and try to use it following a fully unsupervised

* We would like to thank the anonymous reviewers for their helpful suggestions and comments. Paper partially supported by the following projects: ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224), FIRST (FP7-287607) DIIM2.0 (PROMETEOII/2014/001)

approach.

Following Goatly's definition, metaphor occurs when words or expressions are used to refer unconventionally to a referent (real or conceptual), or colligate in an unconventional way (Goatly, 1997). By using topic models, metaphors can be seen as words, expressions or collocations related to a set of one or more topics used in an unconventional way in a specific context. Therefore, the key idea to metaphor detection is that of "unconventionality".

Our proposal is to calculate this unconventionality comparing the set of topics of a specific word (word domain) with the set of topic of its context (context domain). If both sets of topics are similar, the use of that word is considered as conventional. Therefore, there is a semantic coherence in the sentence and the word is used in a literal sense. On the other hand, if both sets are not similar, the use of that word is considered as unconventional: there are no semantic coherence between the word and its context, and therefore the word is used in a metaphorical sense.

The remainder of this paper is organized as follows: next section presents other papers that propose the use of topic modeling for metaphor identification; Section 3 describes the system developed whereas Section 4 describes its architecture; in Section 5 the experiments carried out and the results obtained are presented; finally, Section 6 shows the main conclusions and future work.

2 Related Work

There exist many different proposals to automatic metaphor identification (Shutova, Teufel, and Korhonen, 2012; Broadwell et al., 2013; Hovy, Srivastava, and Jauhar, 2013; Wilks et al., 2013; Shaikh et al., 2014; Schulder and Hovy, 2014) -among many others-. We will focuse only on that proposals that use LDA topic modeling as their core technique.

Bethard et al. (2009) proposes the idea of representing metaphorical domains through LDA topics models following a supervised approach. First, LDA is applied on the British National Corpus in order to extract 100 topics. Then these topics are used as features to build a metaphor classifier, based on Support Vector Machines (SVM), to identify metaphorical uses. Training on a set of 400 sentences, the system achieves 61.3% accuracy on metaphor classification.

More recently, Heintz et al. (2013) presented a LDA topic model approach based on the Wikipedia corpus, aligning its topics to potential source and target concepts. These concepts are defined by small manually-created lists of seed words. The system first applies LDA inference on an input corpus to get topic probabilities for each document and sentence. Then it selects those sentences linked by LDA to both a source-aligned topic and a target-aligned topic. Finally, the system identifies the words in each selected sentence that are strongly related to each concept. With this information, a final score is determined. When the score of a word is above a certain threshold, it is labeled as metaphorical. Using 100 topics, this work reported an F-score of 59% for the English language.

These two systems require, at some point, a process of manual annotation. In this paper we propose going a step further and develop a fully unsupervised system to metaphor identification, using Wikipedia as a reference corpus and LDA topic modeling to represent semantic domains.

3 System Description

The system presented in this work detects if a word is used in a literal or metaphorical sense according to the set of topics that the word shares with its context. If the word is related with at least one topic of the context, the word is considered to be used in a literal sense. However, if the word has no topics in common with its context, then the word is considered as metaphorical. In our experiments, we considered the context of the word as the words that co-occur with it in the same sentence.

More formally, if T_w is the set of topics of a target word w in the whole corpus, T_i is the set topics of a word i in the context of w, and N is the number of words in the context of w:

$$T_S = \bigcup_{i \in N} T_i$$

represents the set of topics occurring in the context of w, that is considered as metaphorical if:

$$T_w \cap T_S = \emptyset.$$

In order to represent the semantics of each word, we run LDA topic modeling on Wikipedia, similar to (Heintz et al., 2013), following a bottom up approach. Formally, a topic is a distribution over a fixed vocabulary, and each word in the vocabulary has a probabilistic weight in each topic. Therefore, each topic is represented by a set of keywords: the most prominente words, the words with more probabilistic weight in the topic.

These topics extracted from Wikipedia are then used to represent the semantic domains of

each word in a new corpus, where we want to identify the metaphorical or literal use of different target words. Since Wikipedia is a general encyclopedia, for our system the word-topic relation extracted from it will be conventional relations.

Table 1 shows two examples of the most representative terms for topics extracted from Wikipedia. The first topic includes words related to family, such as mother ("madre"), daughter ("hija") and father ("padre"). The second topic relates to basketball, including terms such as season ("temporada"), basketball ("baloncesto") and rebounds ("rebotes").

mujer madre vida mujeres familia hija padre casa matrimonio relación amor hijos joven años hermana niños hijo padres esposa marido pareja hombre sexual
temporada equipo puntos liga partido jugador baloncesto año nba jugó temporadas universidad mejor rebotes profesional posición

Table 1: Examples of topics extracted from Wikipedia.

4 *System Architecture*

The system performs three main steps (see Figure 1):

1. LDA is run on Wikipedia, which is used as a reference corpus. We have used Mallet to perform topic modeling (McCallum, 2002). This step is carried out only once.
2. Given a new target corpus, the system extracts each sentence as the context. For each sentence, a vector of topics is created taking into account the topics previously associated to each word in the Wikipedia corpus. The same is done for the target word which we want to classify as literal or metaphorical.
3. Finally, the system compares the target word topics with the context (sentence) topics. If there is at least one topic in common, the system classifies the word as literal. Conversely, if there is no topic in common, the system classifies the word as metaphorical.

5 *Experiments and Evaluation*

The quality of the metaphor detection depends directly on two aspects:

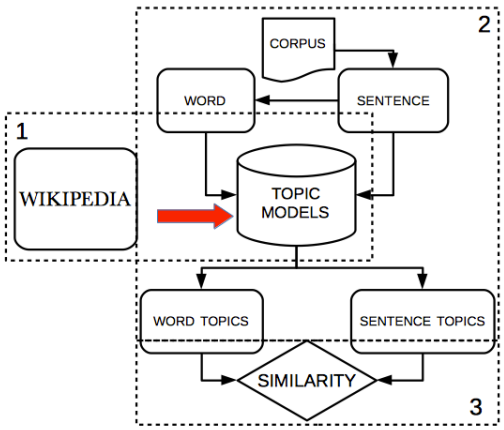

Figure 1: System architecture and steps carried out.

1. Granularity of topics: the amount of topics extracted from the reference corpus (Wikipedia).
2. Granularity of keywords: the amount of representative words extracted for each topic.

On the one hand, if there is a lot of topics and keywords (high granularity), the system tends to classify words as literal: there are more probability to find a common topic between the word and its contexts. On the other hand, if there is only a few topics and keywords (low granularity), the system tends to classify words as metaphorical: there are low probability to find a common topic between the word and its contexts.

A set of experiments have been carried out on an evaluation corpus in order to define the appropriate values of granularity.

5.1 Evaluation corpus

The evaluation corpus comprises 100 sentences written in Spanish, grouped in two subcorpus according to the target word:

- 50 sentences with the word "desierto" (desert).
- 50 sentences with the word "oasis" (oasis).

These two sets are balanced, including 25 sentences with a metaphorical use of the target word, and another 25 representing a literal use of it.

Some examples of metaphorical usages of "desert" and "oasis" are

- Literal: "Las regiones montañosas de Marruecos, y los **oasis** saharauis..." (The mountain regions of Morocco, and saharauis **oasis**...).

- Metaphoric: “En su opinión, Pujol ha hecho de Cataluña un **oasis**...” (In his opinion, Pujol has turned Catalonia into an **oasis**...).
- Literal: “Un seísmo de 7,5 grados sacudió la localidad de Landers, en el **desierto** de California” (An earthquake of magnitude 7.5 shook the town of Landers, in the **desert** of California).
- Metaphoric: “Regresó al fútbol profesional después de atravesar el **desierto** de su adicción a la cocaín a” (He came back to profesional soccer after crossing the **dessert** of his cocaine addiction).

All these sentences have been manually annotated as “metaphorical” or “literal”. Due to the fact that the evaluation corpus is small, we will not achieve conclusive results. They are preliminary results that will be reevaluated with a bigger corpus in Future Work.

5.2 Experimental Setup

The baseline of our system follows a majority class approach, where the most common class in the corpus is assigned to every instance in the test set. Since the corpus is perfectly balanced (50% are metaphorical and another 50% are literal), our baseline is set to 50% accuracy.

A total of seven experiments have been carried out in order to study the effect of the granularity configuration in the performance of the system. The configuration of these experiments is shown in Table 2.

Run 1:	1000 topics with 20 keywords each one.
Run 2:	1000 topics with 50 keywords each one.
Run 3:	1000 topics with 100 keywords each one.
Run 4:	1000 topics with 200 keywords each one.
Run 5:	2500 topics with 20 keywords each one.
Run 6:	2500 topics with 50 keywords each one.
Run 7:	2500 topics with 100 keywords each one.
Run 8:	2500 topics with 200 keywords each one.

Table 2: Experiments carried out depending on granularity configuration.

5.3 Results and Comments

The goal of the system is to identify whether the sense of a target word in a sentence is metaphorical. Given the previous eight runs, we have calculated the following four measures to determine the performance of the system:

- Precision: the number of target words correctly identified as metaphorical divided by the total number of words classified by the system as metaphorical.
- Recall: the number of target words correctly identified as metaphorical divided by the total number of metaphorical words in the corpus.
- F-Measure: $F_1 = 2 * \frac{precision*recall}{precision+recall}$
- Accuracy: percentage of target words correctly classified as metaphorical or as literal.

Tables 3 and 4 show precision, recall, F-Measure and accuracy obtained for the runs previously defined.

Desierto	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Run 1:	0.61	**0.88**	**0.72**	**66**%
Run 2:	0.6	0.24	0.34	54%
Run 3:	0.85	0.24	0.375	60%
Run 4:	**1**	0.08	0.14	54%
Run 5:	0.55	**0.84**	0.67	50%
Run 6:	0.73	0.64	**0.68**	**70**%
Run 7:	**1**	0.2	0.33	60%
Run 8:	**1**	0.08	0.14	54%

Table 3: Results for the word “desierto” (desert)

Oasis	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Run 1:	0.53	**0.92**	**0.67**	**56**%
Run 2:	0.53	0.84	0.65	56%
Run 3:	0.45	0.2	0.28	48%
Run 4:	**0.8**	0.16	0.26	56%
Run 5:	0.53	**0.84**	0.65	56%
Run 6:	**0.61**	0.76	**0.68**	**64**%
Run 7:	0.6	0.72	0.65	62%
Run 8:	0.2	0.5	0.28	50

Table 4: Results word “oasis” (oasis)

The previous results show that, when keyword granularity increases (more keywords to define each topic), the accuracy of the system decreases. Precision increases with high granularity. However, recall shows a dramatic drop. For this reason, F-Measure indicates that it is better low granularity. The best precision with the best recall is obtained with 20 or 50 keywords.

All these data mean that it is better to set up a low keyword granularity. The results are similar for both corpus: for 1000 topics, the best results are obtained with 20 keywords, for 2500 topics, the best results are obtained with 50 keywords.

If the system annotates all target words as metaphorical (baseline), the accuracy is 50%. All runs, except run 3, show an accuracy increment from 4 to 20%. However, the F-Measure of the

baseline is 0.6. In this case, only runs with low keyword granularity are over the baseline.

As main drawback, we have observed that some words do not appear related to any topic, because LDA does not give enough weight to low frequency words. Consequently, some sentences are semantically under-represented and the system has not enough data to classify the target word.

6 Conclusions

Despite the apparent simplicity of the proposal, based on the results achieved we can conclude that the comparison between word topics and context topics is a promising approach to metaphor identification. The preliminary results obtained improve the baseline proposed. Compared with current state of the art systems, our proposal has the clear advantage of being completely unsupervised: it does not require any manual annotation of a corpus and thus can be easily adapted to other languages.

As a Future Work we plan, first of all, to improve the evaluation process with a bigger corpus. The manual annotation of metaphors is a difficult task: there are few corpora annotated with metaphors, and they use to be annotated for English. In any case, we plan to extend our manually annotated corpus and reevaluate the system. Besides, we will apply our system to the evaluation corpora used by other systems, in order to compare the results.

On other hand, currently in our system the similarity of words and contexts is computed by means of the intersection of the set of topics that defines them. As a second future work, we plan to improve this comparison taking into account the weight assigned to each topic, and not just its presence or absence, by building a weighted vector for the topics identified in the word and in the context, and comparing them following the vector space model representation (Salton, Wong, and Yang, 1975; Turney and Pantel, 2010).

Finally, we also plan to extend this model to other types of metaphors and finally to texts following an "all word" approach: identify the literal or metaphorical sense of all the words in a corpus, and not only on a set of previously selected words. Beside, we will apply the model to other languages.

References

Bethard, S., V. T. Lai, and J. H. Martin. 2009. Topic Model Analysis of Metaphor Frequency for Psycholinguistic Stimuli. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*, pages 9–16, Boulder, Colorado.

Blei, D., A. Ng, and M. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Broadwell, G., U. Boz, I. Cases, T. Strzalkowski, L. Feldman, S. Taylor, S. Shaikh, T. Liu, K. Cho, N. Webb, and S. Taylor. 2013. Using Imageability and Topic Chaining to Locate Metaphors in Linguistic Corpora. In A.M. Greenberg, W.G. Kennedy, and N.D. Bos, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer-Verlag, Berlin Heidelberg, pages 102–110.

Goatly, A. 1997. *The Language of Metaphors*. Routledge, New York.

Heintz, I., R. Gabbard, M. Srinivasan, D. Barner, D. Black, M. Freedman, and R. Weischedel. 2013. Automatic Extraction of Linguistic Metaphor with LDA Topic Modeling. In *First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia.

Hovy, D., S. Srivastava, and S. Jauhar. 2013. Identifying Metaphorical Word Use with Tree Kernels. In *Workshop on Metaphor in NLP*.

Lakoff, G. and M. Johnson. 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.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of ACM*, 18(11):613–620.

Schulder, M. and E. Hovy. 2014. Metaphor Detection through Term Relevance. In *Proceedings of the Second Workshop on Metaphor in NLP*.

Shaikh, S., T. Strzalkowski, K. Cho, T. Liu, G. Broadwell, L. Feldman, S. Taylor, B. Yamrom, C. Lin, N. Sa, I. Cases, Y. Peshkova, and K. Elliot. 2014. Discovering Conceptual Metaphors Using Source Domain Spaces. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon*, pages 210–220, Dublin, Ireland.

Shutova, E. 2010. Models of Metaphor in NLP. In *Proceedings of the 48th Annual Meeting*

of the Association for Computational Linguistics, number July, pages 688–697. Association for Computational Linguistics.

Shutova, E., S. Teufel, and A. Korhonen. 2012. Statistical Metaphor Processing. *Computational Linguistics*, (July 2011):1–92.

Turney, P. D. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Wilks, Y., L. Galescu, J. Allen, and A. Dalton. 2013. Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction. In *Workshop on Metaphor in NLP*.