Topic Models to Contextualize and Enhance Text-Based Discourses Using Ontologies

Dimitris Gkoumas
Future Internet Living Lab Budapest
Budapest, Hungary
e-mail: dgoumas@corvinno.com

Réka Vas
Department of Information Systems
Corvinus University of Budapest
Budapest, Hungary
e-mail: reka.vas@uni-corvinus.hu

***Abstract*—Public policy-making has a clear and unique purpose: achieve a desired goal that supports the best interest of all members of the society by providing guidance for addressing selected public concerns. Examples include clean air, healthcare, waste management etc. The identification of social targets and pathways – by which these targets could be reached – are at the core of policy-making. This paper is part of an ongoing research aiming at enhancing public policy-making in the field of waste management by contextualizing and enriching text-based, Web forum discourses on waste management. For that purpose, an ontology model describing the waste management domain has been created. In the next step, the actual forum discussions are connected to one or more subdomains of the ontology by determining what proportion of the sub-domain is covered by that discourse. Finally, applying text mining techniques semantically enriched domain concepts are assigned to the discourse. This paper also provides a critical discussion on two text mining approaches that could be applied for this purpose, also highlighting points that deserve further investigation.**

Keywords-Discourse contextualization; discourse enhancement; clustering topic model; probabilistic topic model; ontologies in NLP.

I. INTRODUCTION

Humans interact with the *real world* and they observe it from different perspectives trying to give an interpretation of it by creating mental concepts. Different people look at the world from different angles, paying attention to different things. Even the same person might also pay attention to different aspects of the world in different periods of time. This is called *reflected world* and is different from the real world because the perspective a person takes or has taken is often biased. The reflected world is mainly represented by speech or writing using a natural language, and in most cases the result is textual data.

Textual data plays a major role in conveying knowledge and information about the reflected world, which could be further used for problem solving or decision making as a result of public policy. However, the rapid increase in the amount of the textual data and its unstructured format make information extraction (IE) a challenging task. Additionally, acquiring knowledge from textual information is not always a straightforward process since textual data also derives properties of language. *Synonymy*, expressing a single concept in a number of ways (i.e., car and automobile) and *polysemy*, using the same term to refer to multiple concepts (i.e., jaguar which can mean a special car or an animal, as well), are two major obstacles in IE since in reality there is often no one-to-one correspondence between concepts and textual terms [1]. That word-sense ambiguity could utterly fool algorithms, which search terms only as a sequence of characters [2]. Lexical co-occurrence that is determined on the basis of statistical significances is an important indicator for term associations. According to this approach, two terms or a sequence of terms (*n-gram*) are associated when a presented term triggers the mental activation of another one. However, lexical co-occurrence cannot handle the above described ambiguity because it is not only invalid from a linguistic-semantical point of view but also prone to overestimate the semantic similarity [2][3]. On the other hand, incorporating knowledge in the form of an ontology bridging the conceptual and real world [2][4][5] can help to overcome challenges in text mining. Ontologies allow storing domain knowledge in a more sophisticated form, conceptualizing a domain [6]. By using ontologies, text terms could be indexed by ontology concepts, which reflect terms' meaning rather than words considered as lists with all the ambiguity they convey.

The main goal of this study is to contextualize semantically enriched text-based discourses to gain information from the scope of a specific domain eliminating the ambiguity of the discussion. After that, the next step is to enhance the discussion by supplying it with connected wiki pages. For this purpose, an ontology is used as a representation of the domain knowledge to match concepts with terms in the discussion. In the literature, the most common method applied in such cases is to map concepts on text. At the same time, this paper suggests two different approaches for tackling the above-described issues. Both of our approaches make use of topic models to discover the hidden semantic structure of the text-based discourse. The discourse may concern one or multiple topics in different proportions. After that, text mining methods are used to measure the similarity between the discourse and concepts belonging to the concerned topics.

 ISBN: 978-1-61208-552-4

Section II provides specific details about the case study. In Section III, we describe in detail the clustering topic model approach to contextualize and enhance the text-based discourse developing by that way a public policy. Section IV presents a probabilistic topic model approach to tackle the same issue. At the end, conclusions are drawn in Section V.

II. THE CASE

The textual data to be analyzed is collected from forum discussions, which collect conversations in the form of posted messages. On the investigated forum, people having ideas regarding eco-friendliness, and experts from the field of waste management can leave comments to provide help and enhance problem-solving. A domain ontology of waste management [7], holding knowledge about ten subdomains, is used to contextualize and enhance text-based discourses. Each concept in the domain ontology includes a label, which consists of one up to five terms, a set of synonyms, and a wiki page describing in detail the given concept. WordNet [8] – a language engineering tool – has been used to extract the set of synonyms for each concept label.

III. THE CLUSTERING TOPIC MODEL APPROACH

A. Methodology

In the current study, the first aforementioned approach tries to match concepts from an assigned subdomain to a text-based discourse. The process is broken down into three tasks. In the first step, a clustering topic model is applied on the domain ontology to verify that it is well-structured and there is no noise. The clustering algorithm automatically identifies subdomains in a group of concepts (Figure 1). Despite the fact, that the ontology structure will finally be used, the clustering algorithm is also used for extracting labels for each subdomain. Actually, the resulting centroids are being viewed as the resulting labels. In the second step, once a new text-based discourse comes out, it is assigned to one of the subdomains (Figure 2). In the last phase, the text-based discourse has to be associated with the concepts of the assigned subdomain. There are different methods to do that, however in this case we calculate the distance between the discussion and each concept in the assigned subdomain (Figure 3). Concepts with short distances are dominantly presented in the discussion while the ones with long distances are either weakly associated or not at all. After that, concepts with the shortest distances are chosen as predominant in the discussion, returning back a wikipage describing in detail the given ontology concept to enhance the discourse.

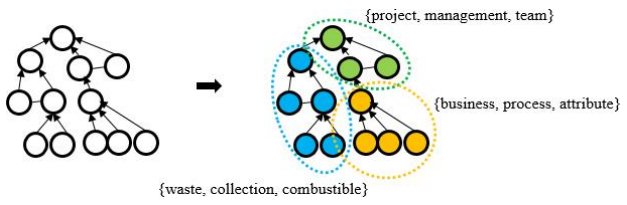

Figure 1. Automatic subdomain identification throughout the ontology and topic label extraction.

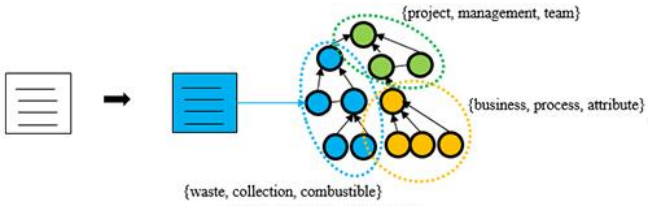

Figure 2. A text-based discourse is assigned to a subdomain.

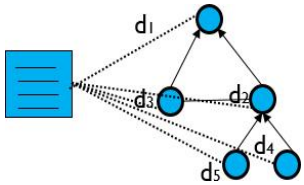

Figure 3. Distance calculation between text of discussion and concepts in the assigned subdomain.

B. Implementation

This section describes in detail the text preprocessing, the clustering topic model and the matching process to contextualize and enhance a text-based discourse.

1) Text preprocessing

At first, ontology concepts are extracted and saved into a text file. After the cleaning of the extracted concepts the most crucial part, the preprocessing of the extracted concepts, starts. In this phase, some basic techniques [9] are applied starting from *tokenization*. In this process, the text is split into a stream of words by removing all non-alphanumeric characters, such as punctuation and mathematical symbols, and then it is normalized to lowercase. This tokenized representation is then used for further processing, applying some filtering methods. Thus, words that bear little or no content information are removed. Initially, a stop-word filter removes high-frequency words, such as "the", "a", "or", with no content information. Then, a stemming or lemmatization filter is applied in order to reduce further the number of the words. At the end, one stemmed and only one tokenized vocabulary are created.

2) Clustering topic model in the domain ontology

Transformation is the next step after preprocessing. In the current approach, a vector space model is used to transform the textual data in a data structure before data mining techniques are applied. In order to assign a weight to each term or continuous sequence of n terms (n-grams) to a concept within a group of concepts, the *term frequency-inverse document frequency* (tf-idf) measure is applied looking at unigrams, bigrams, and trigrams. After that, the similarity between concepts in the domain ontology can be measured based on cosine similarity [10]. In reality, cosine similarity is a length-agnostic metric and measures the cosine of the angle between two text vector representations (formula (1)). Subtracting cosine similarity from one provides cosine distance, as it is seen in formula (2).

 ISBN: 978-1-61208-552-4

$$similarity = \cos(\theta) = A*B / ||A||*||B|| \quad (1)$$

$$d = 1 - \cos(\theta) \quad (2)$$

After having computed cosine distances between each concept and all the rest of the concepts in the domain ontology, a hierarchical clustering is performed on the concepts to find out the optimum number of clusters. In that case, we chose the agglomerative Ward clustering algorithm.

At the last step, the vector space model enclosing tf-idf measurements and the optimum number of clusters, that Ward clustering returned, are used as input to a k-mean clustering to assign each concept to a cluster with the nearest mean. Finally, the top n words that are nearest to the cluster centroids are sorted to be used as labels. The result is a set of important words for each cluster giving a sense of what a subdomain is about.

3) The matching process

As it has been already mentioned, the matching process consists of two levels. In the first one, the text-based discourse is matched to a subdomain in the ontology to enable a general contextualization, while in the second step, it has to be calculated which subdomain concepts are the nearest to the discussion, not only to contextualize the discussion in a deeper semantic level but also to enhance it by returning a list of indexes, which point to the aforementioned wiki pages.

Once a new text-based discourse appears on the forum, we extract the textual data, filter the words based on the mentioned above vocabulary, and transform it to a space vector using tf-idf measure to assign weights to the terms. Then, the existing k-mean clustering model assigns the discussion to a subdomain. In the second phase, we define the nearest concepts of the discourse by running a k-nearest neighbor algorithm. The word tf-idf vectors are used to represent the concepts and the discussion, and cosine distance to measure distance.

IV. THE PROBABILISTIC TOPIC MODEL APPROACH

A. Methodology

In the previous approach, every cluster includes a prevalent topic and once a new text-based discourse comes out, it is assigned to a subdomain of the ontology. The question is what shall we do if a discussion covers more than one topic? In reality, a discussion can enclose many topics in different proportions. Even in the current case study, where only domain experts took part in the discourse about a quite specific topic, there is a high possibility that a variance of topics will come up in the discussion. In order to address this issue, the second approach makes use of latent Dirichlet allocation (LDA) model [11] - a probabilistic topic model - to be able to learn even about hidden topics in a discussion [12].

LDA is a probabilistic extension of latent semantic analysis (LSA) [13] assuming that each term is a mixture of topics and it is attributable to the LDA's topics. In general, a bag-of-words model – disregarding grammar and even word order – and the number of topics – given by an expert or applying a trial and error method – are used as input to the LDA model. After that, the model outputs a) topic vocabulary distributions b) topic assignments per term and c) topic proportion per text. Such a probabilistic approach not only has both favorable semantical and statistical quality [14] but also offers a dampening of synonymy [15].

In the current study, the concept labels are used as a domain corpus to train the LDA model to provide topic vocabulary distributions (TABLE I). In this case, once a text-based discourse appears, each term in the text is assigned to a topic. However, the goal of the mixed assignment is not only to associate the discourse with a collection of topics but also to calculate the relative proportion. The latter, besides a broader understanding of the text, could be leveraged to enhance a discussion and develop a public policy in a broader way. In order to calculate the topic proportion in the text, each assigned term is scored under the probabilistic topic vocabulary distributions (TABLE I). For instance, if the domain ontology includes three topics, the result will be a normal distribution over the prevalence of topics (π) in the discussion, as it is seen in formula (3).

$$\pi = [0.1*\pi1, 0.4*\pi2, 0.5*\pi3] \quad (3)$$

TABLE I. TOPIC VOCABILARY DISTRIBUTIONS

Topic 1		**Topic 2**		**Topic 3**	
Waste Management		***Business Process***		***Project Management***	
waste	0.1	business	0.18	project	0.15
collection	0.08	process	0.09	management	0.07
combustible	0.05	attribute	0.03	team	0.07
…	…	…	…	…	…

At this point, we have acquired a broader but quite general idea what the text-based discourse is about. In order to contextualize the discourse in a semantic level, it has to be associated with the concepts of the ontology. The main difference between the first and the second approach is that in the second case we match concepts from many subdomains that are presented in the discussion. However, subdomains with a low prevalence in the discussion are not taken into consideration. In order to match concepts to the discussion, we firstly compute the topic distribution for each concept, and then compute some sort of divergence between the discussion and concepts. As in the first approach, short distances between two topic distributions are dominantly presented in the discussion while the ones with long distances are either weakly associated or not at all.

 ISBN: 978-1-61208-552-4

B. Implementation

The same process is followed in the text preprocessing, that has been described in Section III./B. The main difference compared to the previous approach is that instead of a vector space model with tf-idf measures, we use a document-term (DT) matrix. LDA model is actually looking for repeating term patterns in the entire DT matrix. The optimum number of topics is equal to the number of the subdomains in the domain ontology while the number of terms composed in a single topic is chosen to a high number as we want to extract themes and concepts. Closing, we take 100 iterations to allow LDA algorithm for convergence.

The LDA model outputs topic- and weight terms. After that, we score all of the words in the text-based discourse under the above described probabilistic topic distributions to track the distribution of prevalent topics over the discussion. In the second phase, we calculate the distribution of topics for each concept belonging to prevalent subdomain. Finally, we compare the topic distributions between the text-based discourse and concepts using the Kullback–Leibler (KL) divergence measure [16].

V. CONCLUSION

Word-sense disambiguation (WSD) is an important and challenging process of determining which sense of a word is used in a given context. There are hundreds of WSD algorithms for bespoke applications. However, in this paper we follow another way. A domain ontology is used as a dictionary to specify the senses which are to be disambiguated and text-based discourses to be disambiguated. Actually, we propose two different topic models – a clustering and a probabilistic one – to contextualize text-based discourses. In the first case, cosine similarity is used to measure the similarity between the text-based discourse and concepts, while in the second one, KL-divergence measure is used to compare topic distributions between the discussion and concepts belonging to prevalent topics. On one hand, since cosine similarity is a length-agnostic metric, it lets us compare word distributions between texts of varying lengths. Thus, it seems to be a good metric to compare discussions with concepts consisting of one up to five words. On the other hand, measuring distances directly using vector representations may not be reliable because, in very high dimensions, a distance between any two points starts to look the same. An LSA faces efficiently the issue since it reduces the data dimensionality.

Even these methods have already been implemented, they have not been evaluated by a domain expert as they are part of an ongoing research project. However, we expect for the clustering approach to perform better in cases when discourses are strictly domain specific, framed in well-defined borders, and they have a low deviation from the discussed topic. The latter situation seems to be ideal or at least scarce. In reality, there is always a topic deviation in a discussion, as humans tend to integrate concepts from different domains when they are critically thinking. For this purpose, a probabilistic model seems to tackle the contextualization issue better.

There are also two important points, which deserve further investigation. It has been mentioned that the discourse is enhanced by adding wikitext describing an ontology concept in detail. However, what shall we do if people omit important topics and the quality of the discussion is not desirable? In that case, the ontology structure could be used to identify any of the important and omitted concepts. In the opposite case, it should be investigated what to do if people mention concepts that do not exist in the ontology? It is obvious there is a need for a two-way interaction. In that case, new concepts from the discussion should be extracted to enrich the domain ontology. Thus, the latter could process similar future discussions more efficiently.

Nevertheless, the two approaches look quite promising. These approaches just need to be experimented under different circumstances followed by an evaluation process to confirm or reject the aforementioned assumptions.

REFERENCES

[1] I. Spasic, S. Ananiadou, J. Mcnaught, and A. Kumar, “Text mining and ontologies in biomedicine: making sense of raw text,” Briefings in Bioinformatics, vol. 6, no. 3, Sep. 2005, pp. 239–251.

[2] G. Nagypál, “Improving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies,” in On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops, R. Meersman, Z. Tari, and P. Herrero, Eds. Springer Berlin Heidelberg, 2005, pp. 780–789.

[3] B. Lemaire and G. Denhière, “Effects of High-Order Co-occurrences on Word Semantic Similarities,” ArXiv08040143 Cs, Apr. 2008.

[4] N. Aussenac-Gilles and J. Mothe, “Ontologies As Background Knowledge to Explore Document Collections,” in Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, Paris, France, 2004, pp. 129–142.

[5] G. Solskinnsbakk and J. A. Gulla, “Ontological Profiles As Semantic Domain Representations,” in Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, Berlin, Heidelberg, 2008, pp. 67–78.

[6] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, “What Are Ontologies, and Why Do We Need Them?,” IEEE Intell. Syst., vol. 14, no. 1, pp. 20–26, Jan. 1999.

[7] R. Vas, “STUDIO – Ontology-Centric Knowledge-Based System,” in Corporate Knowledge Discovery and Organizational Learning, A. Gabor and A. Ko, Eds. Springer International Publishing, 2016, pp.33-58.

[8] C. Fellbaum, “WordNet: An Electronic Lexical Database,” Cambridge, MA: MIT Press, 1998

[9] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining,” LDV Forum, vol. 20, no. 1, May 2005, pp. 19–62.

[10] A. Huang, "Similarity measures for text document clustering," in Proceedings of the sixth new Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, Apr. 2008, pp.49-56.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, Jan. 2003, pp.993-1022.

[12] D. M. Blei, “Probabilistic Topic Models,” Communications of the ACM, vol. 55, no. 4, Apr. 2012, pp. 77–84.

[13] S.T. Dumais, “Latent semantic analysis,” Annual review of information science and technology, vol. 38, no.1, Jan. 2004, pp.188-230.

[14] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," Expert Systems with Applications, vol. 38, no. 3, Mar. 2011, pp. 2758–2765.

[15] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, vol. 41, no. 6, Sep. 1990, pp. 391–407.

[16] S. Kullback, and R.A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics, vol. 22, no. 1, Mar. 1951, pp. 79–86.

 ISBN: 978-1-61208-552-4