The Topics they are a-Changing — Characterising Topics with Time-Stamped Semantic Graphs

A. Elizabeth Cano,[1] Yulan He,[2] and Harith Alani[1]

[1] Knowledge Media Institute, Open University, UK
ampaeli@gmail.com, h.alani@open.ac.uk
[2] School of Engineering and Applied Science, Aston University, UK
y.he@cantab.net

Abstract. DBpedia has become one of the major sources of structured knowledge extracted from Wikipedia. Such structures gradually re-shape the representation of Topics as new events relevant to such topics emerge. Such changes make evident the continuous evolution of topic representations and introduce new challenges to supervised topic classification tasks, since labelled data can rapidly become outdated. Here we analyse topic changes in DBpedia and propose the use of semantic features as a more stable representation of a topic. Our experiments show promising results in understanding how the relevance of features to a topic changes over time.

Keywords: social media, topic detection, DBpedia, concept drift, feature relevance decay

1 Introduction

Supervised topic classifiers which depend on labelled data can rapidly become outdated since new information regarding these topics emerge. This challenge becomes apparent when applying topic classifiers to streaming data like Twitter. The continuous change of vocabulary – in many cases event-dependent– makes the task of retraining such classifiers with fresh topic-label annotations a costly one. In event-dependent topics not only new lexical features re-characterise the topic but also existing features can potentially become irrelevant to the topic (e.g., Jan25 being relevant to violence in the Egyptian revolution is now less relevant to current representations of the topic violence). In dynamic environments the expectation that the progressive feature drifts of topics to be in the same feature space is not normally met.

The incorporation of new event-data to a topic representation leads to a linguistic evolution of a topic, but also to a change on its semantic structure. To the best of our knowledge, none of the existing approaches for topic classification using semantic features [4][2][5][7], has focused on the epoch-based transfer learning task. In this paper we aim to disseminate our work presented in [1] by summarising our proposed transfer learning approach for the epoch-based topic classification ot tweets. In [1] we investigate whether the use of semantic features as opposed to lexical features can provide a more stable representation of a topic. Here we extend our work by representing cross-epoch settings gain in F-measure for both lexical and semantic feature with infographics. This enables us to highlight the relevance of the studied semantic features over the lexical ones.

1.1 Evolving Topics

DBpedia is periodically updated to incorporate any additions and modification in Wikipedia. This enables us to track how specific resources evolve over time, by comparing these resources over subsequent DBpedia editions. For example, changes to the semantic graph for the concept Barack_Obama can be derived from snapshots of this resource's semantic graph from different DBpedia dumps [3]. E.g., in Figure 1, although some of the triples remain unchanged in consecutive dumps, new triples provide further information on the resource.

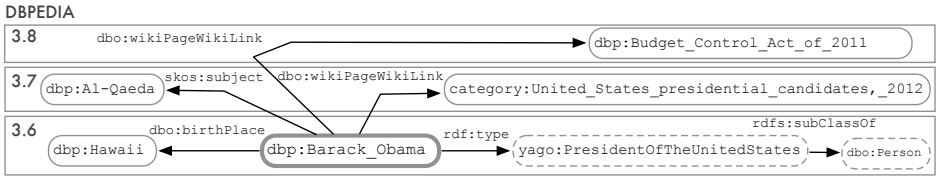

Fig. 1. Triples of the Barack_Obama resource extracted from different DBpedia dumps (3.6 to 3.8). Each DBpedia dump presents a snapshot in time of factual information of a resource.

Changes regarding a resource are exposed both through new semantic features (i.e triples) and new lexical features –appearing on changes in a resource's abstract–. In DBpedia a topic can be represented by the collection of resources belonging to both the main topic (e.g. `cat:War`) and resources (e.g `dbp:Combat_assessment`) belonging to subcategories (e.g. `cat:Military_operations`) of the main Topic. Therefore a topic's evolution can be easily tracked by tracking changes in existing and new resources belonging to it.

2 Topic Classification with Time-Stamped Semantic Graphs

In [1], we propose a novel transfer learning [6][3] approach to address the classification task of new data when the only available labelled data belongs to a previous epoch. This approach relies on the incorporation of knowledge from DBpedia graphs. This approach is summarised in Figure 2 and consists of the following stages: 1) Extraction of lexical and semantic features from tweets; 2) Time-dependent content modelling; 3) Strategy for weighting topic-relevant features with DBpedia; and 4) Construction of time-dependent topic classifiers based on lexical, semantic and joint features.

Our analysis involves the use of two main feature types: lexical and semantic features. The semantic features consist on Class, Property, Category, and Resource. The semantic feature representation of a document therefore is build upon the collection of such features derived from the document's entities mapped to a DBpedia resource. The mapping targets the available DBpedia dump when the document was generated. In [1], we proposed different weightening strategies some of which made use of graph properties of a Topic in a DBpedia graph. Such strategies incorporated statistics of the topic graph representation considering a DBpedia graph at time t.

2.1 Construction of Time-Dependent Topic Classifiers

We focus on the binary topic classification in epoch-based scenarios, where the classifier that we train on a corpus from epoch $t-1$, is tested on a corpus on epoch t. Our

[3] The DBpedia dumps correspond to Wikipedia articles at different time periods as follows: DBp3.6 generated on 2010-10-11; DBpedia 3.7 on 2011-07-22, DBp3.8 on 2012-06-01, DBp3.9 on late April. DBpedia have them available to download at DBpedia http://wiki.dbpedia.org/Downloads39

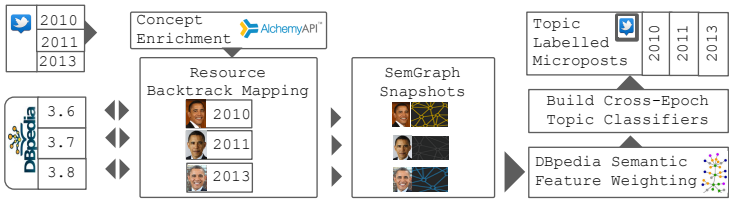

Fig. 2. Architecture for backtrack mapping of resources to DBpedia dumps and deriving topic-relevance based features for epoch-dependent topic classification.

analysis targeted our hypothesis that, as opposed to lexical features which are situation-dependent and can change progressively in time, semantic structures – including ontological classes and properties – can provide a more stable representation of a Topic.

Following the proposed weighting strategies the semantic feature representations of the $t-1$ corpus and the t corpus, are both generated from the DBpedia graph available at $t-1$. For example when applying a classifier trained on data from 2010, the feature space of a target test set from 2011 is computed based on the DBpedia version used for training the 2010-based classifier. This is in order to simulate the availability of resources in a DBpedia graph at a given time. The semantic feature f in a document x is weighted based on the frequency of a semantic feature f in a document x with Laplace smoothing and the topic-relevance of the feature in the $\mathcal{DB}\_t$ graph:

$$W_x(f)_{\mathcal{DB}\_t} = [\frac{[N_x(f)_{\mathcal{DB}\_t} + 1}{|F| + \sum_{f' \in F} N_x(f')_{\mathcal{DB}\_t}}] * (W_{\mathcal{DB}\_t}(f))^{1/2} \quad (1)$$

where $N_x(f)$ is the number of times feature f appears in all the semantic meta-graphs associated with document x derived from the $\mathcal{DB}\_t$ graph ; F is the semantic features' vocabulary of the semantic feature type and $W_{\mathcal{DB}\_t}(f)$ is the weighting function corresponding to the semantic feature type computed based on the $\mathcal{DB}\_t$ graph. This weighting function captures the relative importance of a document's semantic features against the rest of the corpus and incorporates the topic-relative importance of these features in the $\mathcal{DB}\_t$ graph.

3 Experiments

We evaluated our approach using two collections: DBpedia and Twitter datasets. The DBpedia collection comprises four DBpedia dumps (3.6 to 3.9)[4]. The Twitter datasets consist of a collection of Violence-related topics: Disaster_Accident, Law_Crime and War_Conflict. Each of these datasets comprises three epoch-based collections of tweets, corresponding to 2010, 2011, and 2013. The Twitter dataset contained 12,000 annotated tweets [5]. To compare the overall benefit of the use of the proposed weighting strategies against the baselines on this three topics, we averaged the P, R and F-measure of these three cross-epoch settings for each topic. Table 1 presents a summarised version of our results in [1], showing only the best performing features. We can see that in average the Class-based semantic features improve upon the bag of words (BoW) features in F measures. This reveals that the use of ontological classes is a more stable option for the representation of a topic. In order to analyse the differences in gain in F measure for each topic in each of the examined features we used the radar plots in Figure 3. In

[4] General statistics of these dumps are available at http://wiki.dbpedia.org/Downloads39

[5] Further information about this dataset is available at [1]

this figure a positive value indicates an improvement on the classifier. While semantic features improve upon lexical feature in the three topics, the weighted features for resource, class and category exhibit a positive improvement on these scenarios. Moreover the class based features consistently outperform the BoW in all three topics.

	BoW	Cat_{sff}	Cat_{sfg}	Cat_j	Res_{sff}	Res_{sfg}	Res_j	Cls_{sff}	Cls_{sfg}	Cls_j	Sem_{sff}	Sem_{sfg}	Sem_j
P	0.808	0.719	0.784	0.775	0.764	0.775	0.777	0.692	0.691	0.705	0.708	0.751	0.75
R	0.429	0.433	0.434	0.383	0.438	0.426	0.408	0.649	0.638	0.640	0.438	0.373	0.404
F	0.536	0.524	0.550	0.501	0.544	0.529	0.517	0.660	0.658	0.665	0.525	0.490	0.518

Table 1. Average results for the cross-epoch scenarios for all three topics.

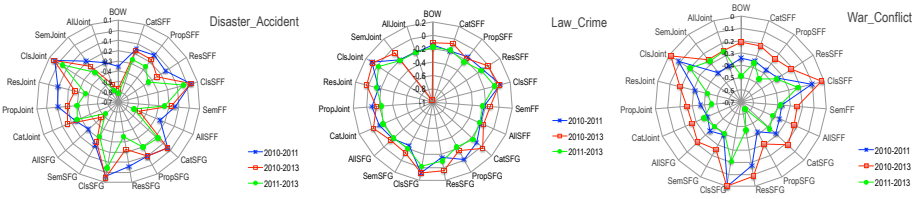

Fig. 3. Summary of performance decays for each feature for each Topic on the three cross-epoch scenarios.

4 Conclusions

Our results showed that Class-based semantic features are much slower to decay than other features, and that they can improve performance upon traditional BoW-based classifiers in cross-epoch scenarios. These results demonstrate the feasibility of the use of semantic features in epoch-based transfer learning tasks. This opens new possibilities for the research of concept drift tracking for transfer learning based on existing Linked Data sources.

References

1. A. E. Cano, Y. He, and H. Alani. Streching the life of twitter classifiers with time-stamped semantic graphs. In *ISWC 2014, Riva del Garda, Trentino, Italy, Oct 19-23, 2014. Proceedings*, Lecture Notes in Computer Science. Springer, 2014.
2. A. E. Cano, A. Varga, M. Rowe, F. Ciravegna, and Y. He. Harnessing linked knowledge source for topic classification in social media. In *Proc. 24th ACM Conf. on Hypertext and Social Media (Hypertext)*, Paris, France, 2013.
3. R. Caruana. Multitask learning. 28(1):41–75, 1997.
4. Y. Genc, Y. Sakamoto, and J. V. Nickerson. Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*, FAC'11, pages 484–492, Berlin, Heidelberg, 2011. Springer-Verlag.
5. Y. He. Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. *ACM Transactions on Asian Language Information Processing*, 11(2):4:1–4:19, June 2012.
6. S. Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*, pages 640–646. The MIT Press, 1996.
7. A. Varga, A. Cano, M. Rowe, F. Ciravegna, and Y. He. Linked knowledge sources for topic classification of microposts: A semantic graph-based approach. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web (JWS)*, 2014.