Eventspotter

spotting entities that talk about events

Project Report

Amar Kadamalakunte , Meghana Sreekanta Murthy

Fall 2012 - Spring 2013

Supervisors:
Raphaël Troncy
Giuseppe Rizzo

EURECOM
Multimedia Department

Contents

Abstract

By definition, an event is an occurrence in the past, present or future. Events could be classified as musical events, sports events, news events, etc. The aim of our work is to develop an approach to spot the existence of musical event entities in free form text. The Eventspotter uses a large dataset called EventMedia to identify candidate event spots in the text, and disambiguates them by assigning a confidence score. The confidence score itself is based on the identification of correlated entities (event location or artists) in the surrounding text and cosine similarity measure between the surrounding text and existing description of event in the dataset. We used precision, recall and f-measure to evaluate our approach. We used event description from the Eventmedia dataset as input and found that the Eventspotter performed better than some exisiting approaches for event detection in the musical domain. We present our method and findings in this paper.

1 Introduction

2 Related Work

Many people have tried to tackle the problem of entity detection in the past.While some have relied purely on the linguistic properties of text others have used context and domain knowledge to accurately detect entities. [1] was one such approach that inspired us during the literature survey phase of the Eventspotter. In this case, the utilisation of context and domain knowledge along with the application of real world constraints proved to be beneficial in improving the accuracy of entity detection. The main goal was to establish semantic relationships between entities in social networking and content sharing sites like MySpace and Youtube. In this approach, the authors used an edit- distance based spotter to match the knowledge base with the text taken from online forums discussing popular music. They then checked for domain and context information and applied an SVM classifier. Further, a study of [2] introduced us to the approach of identifying entities based on knowledge from populated ontologies such as DBWorld and DBLP. They used clues such as entity names, text-proximity, text co-ocurrence, popularity of entities, semantic relationships between entities for disambiguation. This approach further vindicated our idea of using the large Eventmedia knowledge base to detect events in Eventspotter.

As part of another approach , DBpedia URIs were used for automatic annotation of texts. The DBPedia Spotlight [3] used an extensive ontology with 272 classes, called DBpedia, and also provided the flexibility of choosing domain of interest. 2 sources of textual references include : the lexicon created from the DBPedia graph (of labels, redirects, disambiguations) and wikilinks. The Eventspotter derives various facets of the DBpedia Spotlight approach including the 4 main stages : (1) Spot phrases that may indicate a mention of a DBpedia resource, (2) Select candidates by spotted phrase-to-resources mapping, (3) Disambiguate by using the context around spotted phrase, (4) Configure the annotation by customization from user with Resource Prominence, Topic Pertinence, Contextual Ambiguity, Disambiguation Confidence.We also studied two interesting methodologies that were applied to social media and wikipedia for detection of events. TWICAL [4] focused on linguistic properties and temporal expressions to detect events, classified them using the context of surrounding phrases and finally ranked

them according to strong associations with a unique date. The Wikipedia Live Monitor[5] tracked edits on various linguistic versions of a single document on wikipedia and relied on cross-language, full-text search across social networks in order to evaluate the plausability of the edits being events.

Although the above mentioned approaches are effective in entity spotting, there are a few limitations with each. In [1], the authors cite the absence of a training corpus and regions of text being commonly used words, as limiting factors.They also speak of the ambiguous nature of song titles being a major limitation. In [2] the non-exhaustive nature of any knowledge base proves to be the biggest hurdle.The same is the case with the DBpedia spotlight approach defined in [3], where the authors do not speak of using linguistic tools such as a Conditional Random Field classifier to detect entities that are non existent in the knowledge base.Both TWICAL [4] and Wikipedia Live monitor [5] are dependant on social networking sites to detect events. As a result, events that are not publicised in the social networking domain go undetected. The fact that each of these approaches have limitations of their own, goes to show that there is still scope for improvement in the field of event detection. The EventSpotter project is an attempt to meet this need for a more complete approach towards event detection. We first take a detailed look at the Eventspotter Approch and then in subsequent sections we speak at length about the golden data set creation, usage of 10-fold cross validation to evaluate the performance of the system and conclude with our thoughts on how we can potentially better the performance of the eventspotter with the help of machine learning techniques.

3 Approach

As mentioned earlier, the Eventspotter borrows the four stage approach towards event detection from DBpedia [3] .However, the Eventspotter differs in the definition of these stages. The four main stages of the Eventspotter approach are (1)Preprocessing (2)Candidate selection (3)Disambiguation and (4)Postprocessing. Let us look at each of these in greater detail.

3.1 Preprocessing

4 Evaluation

We tested the Eventspotter against Stanford Conditional Random Field Classifier and Opencalais. These tests were of a relatively small scale due to the need to manually create the the golden dataset. We felt it would be interesting to understand the performance of the flexible knowledge based approach of the Eventspotter against the purely grammar-based approaches of Stanford and Opencalais. Before we present the results of our experiments, it would be appropriate to understand the creation of the golden dataset.

4.1 Golden Dataset Creation

We used 60 event descriptions for the creation of the golden dataset. 30 documents were known to include event titles as well as artist names while the remaining were chosen at random. We performed cross validation with factor k =10. We followed a set of rules in order to ensure consistent annotation for the golden dataset creation. For the purpose of annotation we defined a musical event title as : "a proper noun that refers to a musical concert or festival". In all cases, an agreement between 4 human annotators was necessary. Any disagreement was resolved after deliberation and discussion. A candidate spot, was tagged with <EVENT >and </EVENT>. For instance, if 'scala' is an event in the phrase 'scala:)', we annotate it as '<EVENT>scala </EVENT>:)'. For experimental purposes, we followed 2 types of annotation : synthetic annotation and manual annotation .
In the case of synthetic annotation, we carried out a straightforward string comparison between the input text and all event titles in the Eventmedia dataset. If there was a match, the spot was annotated as an event. There were several reasons for doing so. First, we were able to initially establish a base for assessing the performance of the Eventspotter. Second, the synthetic annotation was made available for training the Stanford Conditional Random Field Classifier. We will see during the presentation of the results that the Stanford Conditional Random Field performed quite differently with synthetically annotated training data when compared to manually annotated training data. Third, we realised by observation, that some aspect

of the synthetic annotation approach could be applied during the manual annotation process.
As we performed manual annotation, we realised that there was an inherent ambiguity in event titles. We found it extremely difficult, at times, to distinguish between event titles and artist names. Mainly because events were often titled as a mash up of the artist names, venue and sometimes the date of occurrence.Due to this ambiguity many event titles were left untagged despite strict adherence to the annotation rules. Thus, it made sense to not rely solely on the annotators‘ knowledge, but infact, partially adopt the synthetic annotation approach in order to generate a more robust golden dataset. The manual annotation was broken down into two passes. In the first pass, an unbiased, rule based manual annotation was performed. In the second pass, the human annotators were given the list of event titles that were being described in the documents. With this posteriori knowledge the human annotators were able to annotate event occurrences which would have otherwise gone untagged. As human annotators, we leveraged the context of the spot to perform word sense disambiguation. We refrained from annotating those spots where in the event title string was not used to directly refer to the event itself. For instance in the phrase “Carrie Underwood announces her North American tour Blown Away”, ‘Carrie Underwood’ refered to the artist and hence was not annotated. But in another instance “Carrie Underwood comes to town this monday.” since the annotator knew that Carrie Underwood was the event title and since this string was being used to refer to an event in this context, the spot was annotated as an event. We observed that this hybrid approach towards manual annotation was effective in improving the golden dataset and this was corroborated by the improvement in performance results as presented in the following section.

4.2 Results

The Eventspotter performed better than the Stanford CRF Classifier in terms of recall and F-measure. It obtained a score of 89.26% for recall and 72.73% for F-measure, over the scores of 4.27% and 8% respectively of the Stanford CRF Classifier trained with synthetic annotation and 54.70% and 60.95% respectively when trained with manual annotation. To highlight once again, these set of tests on the synCarrieCarriethetically annotated golden dataset were carried out merely to give ourselves an idea about the Eventspotter's performance. Opencalais did not identify any musical event entities at all.

The second set of tests run were run on the manually annotated golden dataset. Again the Eventspotter's performance outshined that of the Stanford CRF Classifier in terms of recall and F-measure. It obtained a score of 68.14% for recall and 43.14% for F-measure, over the scores of 54.87%

Table 1: Tests Performed with Synthetically Annotated Golden Dataset

Approach	Precision	Recall	F-Measure
Eventspotter	61.36%	**89.26%**	**72.73%**
Stanford trained on synthetic data	62.5%	4.27%	8%
Stanford trained on manual data	68.82%	54.70%	60.95%
Opencalais	no events	no events	no events

and 56.88% respectively of the Stanford CRF Classifier trained with synthetic annotation and 11.5% 18.71% respectively when trained with manual annotation. Again, Opencalais did not identify any musical event entities at all. We also tested a linear combination of Eventspotter and Stanford CRF Classifier trained on synthetically annotated data, with Stanford CRF Classifier acting as a gazateer to Eventspotter. Though there wasnt stark improvement, there was a slight increase in the recall score to 70.8%.

Table 2: Tests Performed with Manually Annotated Golden Dataset

Approach	Precision	Recall	F-Measure
Eventspotter	31.56%	**68.14%**	43.14%
Stanford trained on synthetic data	59.05%	54.87%	56.88%
Stanford trained on manual data	50%	11.5%	18.71%
Opencalais	no events	no events	no events
Gazeteer	31.25%	**70.8%**	43.36%

5 Conclusion and Future work

In this paper, we presented Eventspotter, a tool for detecting musical event entities in unstructured text. We brought to light the various existent ideas that inspired our approach for the Eventspotter. We then detailed the methodologies implemented followed by the evaluation results. We compared the Eventspotter with other open source services and obtained encouraging results that asserted our belief that there is scope for further research in the direction adopted by Eventspotter.

In the future, we plan to incorporate into the Eventspotter logic, Support Vector Machine classifier and machine learning techniques using WAEKA. Parts of speech tagging, domain specific terms, sentimental expressions and results from the Stanford CRF can be used as features to train the SVM classifier. Also, we can maintain a 'white-list' of verbs which are known occur in the proximity of event title in text. This list can be also used as feature to the SVM classifier. We plan to further extend our knowledge base to include other popular data sources such as DBpedia, DBLP, DBworld etc. We would like to explore the idea of generating.We hope that improvements such as these would be a stepping stone for the evolution of a high performance Eventspotter.

Bibliography

[1] D. Gruhl, M. Nagarajan, J. Pieper, and C. Robson, "Context and domain knowledge enhanced entity spotting in informal text."

[2] J. Hassell, B. Aleman-meza, and I. B. Arpinar, "Ontology-driven automatic entity disambiguation in unstructured text," in *In International Semantic Web Conference*, pp. 44–57, 2006.

[3] P. N. Mendes, M. Jakob, A. GarcÃa-silva, and C. Bizer, "Dbpedia spotlight: Shedding light on the web of documents," in *In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics*, 2011.

[4] A. Ritter, S. Clark, and O. Etzioni, "Open domain event extraction from twitter."

[5] T. Steiner, S. van Hooland, and E. Summers, "Mj no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection," *CoRR*, vol. abs/1303.4702, 2013.