Labelling and Spatio-Temporal Grounding of News Events

Bea Alex
School of Informatics
University of Edinburgh, UK
balex@staffmail.ed.ac.uk

Claire Grover
School of Informatics
University of Edinburgh, UK
grover@inf.ed.ac.uk

Abstract

This paper describes work in progress on labelling and spatio-temporal grounding of news events as part of a news analysis system that is under development.

1 Introduction: News Event Analysis

The SYNC3 project[1] is developing a system that tracks news events and related blogs. A news event is defined like a TDT event as something that happened at a particular time and place (TDT, 2004). It constitutes a cluster of news items which all report on the same event. The system crawls news sources and clusters incoming news items. These clusters are then processed by a labelling and a relation extraction component. The former determines document and event-level labels and the later derives temporal, geographic and causal relations between events. Related blog posts are connected to news events and analysed for sentiment. In the user interface, users can search and select news events and related blogs, add comments and interact with other users. Users will also be able to visualise related news events in a map interface and timeline. In this paper, the focus is on the labelling of news events.

The input into the news event labeller is made up of news event clusters containing one or more news items from different sources. Each news item is fed through a linguistic processing pipeline, including named entity recognition, date and geo-resolution. Each cluster is then labelled with a LABEL (a title summarising the news event), a DESCRIPTION (the first sentence of a document), a LOCATION (the location where the event took place) and a DATE (the date of the event). We first compute this information for every news item as a document summary and then select the most representative document summary of the news event cluster.

[1] http://www.sync3.eu

1.1 News Event Label

News titles tend to be appropriate summaries of news items and events. They are coherent phrases or sentences that are understood by users. We therefore implemented variations of title labelling (Manning et al., 2008) made up of document-level title detection and cluster-level title selection. The first step is done by iterating through the sentences of a document and settling on a title if certain criteria are met (e.g. number of tokens is 3 or more, sentence does not match a set of filter strings etc.). Given all document titles, we select as the most representative LABEL:

1. the LABEL of the first published news item,
2. the LABEL of the news item closest to the cluster centroid or
3. the LABEL with the largest ratio of terms common to all titles divided by title length

The 1st method assumes that a news item which first reports an event is breaking news and most interesting to users. News items following it will provide the same or further information. The 2nd method assumes that the news item most representative of the cluster statistically summarises the news event best. The last method assumes that the most succinct title with the most common vocabulary in all titles is most informative about a news event.

1.2 News Event Location

We use the Edinburgh Geoparser (Tobin et al., 2010) to recognise location names and ground them to the GeoNames gazetteer.[2] Besides latitudes, longitudes and GeoNames IDs, we also assign population size and type of location (e.g. populated place, country etc.). Our Geoparser yields 81.2% accuracy when evaluating on SpatialML (Mani et al., 2008). It also compares favourably with Yahoo! Placemaker[3] in an end-to-end run.

[2] http://www.geonames.org

[3] http://developer.yahoo.com/geo/placemaker

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 27–28,
Los Angeles, California, June 2010.

We only consider locations grounded to lat/long values as potential news item locations, therefore restricting the set to more accurately recognised ones. We select the first location in the LABEL and DESCRIPTION or (if none can be found) either the first or most frequent location in the news item. The news item location associated with the most representative cluster LABEL is selected as the news event location. To allow consistency of the information, we treat all caps locations in the DESCRIPTION of each article as reporter locations and will investigate the percentage of cases in which this location is the same as, near or different from the news event location. We will also experiment with limiting the search space of locations to the excerpts of a news item that are evidence for it being part of its cluster.

1.3 News Event Date

We choose the publication date of the earliest published news item in the cluster as the news event date. Our linguistic processing recognises absolute, relative and under-specified temporal expressions (MUC-style TIMEX elements), normalises them and grounds them to a single number representation (the 1st of January 1 AD being `0`). This enables us to determine the day of the week, resolve relative dates and compute temporal precedence on a timeline. We are working towards evaluating the performance of the temporal expression recognition on the Timebank corpus (Pustejovsky et al., 2003).

2 Clustered News Data

We are developing our components using a static set of clusters containing 12,547 documents from 9 different news sources (AP: 16.7%, BBC: 12.9%, CNN: 5.2%, NYT: 9.2%, Reuters: 11.1%, Ria Novosti: 4.9%, USA TODAY: 12.3%, WP: 6.6% and Xinhua: 20.7%) which were crawled between May 20th and June 3rd 2009. The clustering of these documents changes in regular intervals. The current release contains 7,456 clusters with an average of 1.7 news items per cluster with up to 41 news items. 2,259 clusters (30.3%) contain 2 or more news items of which 1,091 (48.3%) contain news items from at least 2 sources. The duration of a news event is 4 days or less (<=1 day: 85.3%, 2 days: 12.4%, 3 days: 2.0%, 4 days: 0.3%).

The Geoparser extracts 188,932 locations assigned with lat/longs from this data. Using the 3rd labelling method, we currently detect a news event location in 7,325 of 7,456 news events (98.3%). If we only consider locations in news item LABELs and DESCRIPTIONs this figure drops to 83%. 117 clusters contain no location. An error analysis will show if this is due to false negatives or inexplicit locations.

3 Summary and Future Work

We have presented ongoing work on news event labelling, with a focus on title labelling and spatio-temporal grounding of news events, and have presented some initial statistics on development data.

We are in the process of creating gold standard data with which we can test the performance of the news event labelling. This will allow us to determine the appropriateness of the news event labels as well as the accuracy of news event locations and dates and enable us to fine-tune the labelling process. Our future work also includes identifying geographical, temporal and causal relations between news events for story detection.

Both the clustering of news into news events and their analysis are crucial for structuring and analysing the blogosphere accordingly, as one aim of SYNC3 is to extract news-event-related blog posts and identify their sentiment.

Acknowledgements

We would like to thank all project partners of the SYNC3 project (FP7-231854).

References

I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. 2008. SpatialML: Annotation scheme, corpora, and tools. In *Proceedings of LREC'08*.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

J. Pustejovsky, P. Hanks, R. Saurí, A. See, R.Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK corpus. *Corpus Linguistics*, pages 647–656.

TDT. 2004. TDT 2004: Annotation Manual Version 1.2. URL: `http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf`.

R. Tobin, C. Grover, K. Byrne, J. Reid, and J. Walsh. 2010. Evaluation of georeferencing. In *Proceedings of GIR'10*.