An Entity-Centric Approach to Storing Community Memories

Diana Maynard, Wim Peters, Jonathon Hare, Adam Funk
University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
initial.surname@dcs.shef.ac.uk

21st October 2011

1 Introduction

With the rapidly growing volume of resources on the Web, Web archiving becomes an important challenge. In addition, the notion of community memories extends traditional Web archives with related data from a variety of sources on the Social Web. Community memories take an entity-centric view to organise Web content according to the events, entities and opinions related to them. To this end, the main challenge is to extract, detect and correlate such information from a vast number of heterogeneous Web resources, including multimedia. The ARCOMEM project aims to perform this task based on an iterative cycle consisting of (1) targeted archiving/crawling of Web objects, (2) entity, topic, opinion and event (ETOE) extraction and and (3) refinement of crawling strategy. Within this paper, we describe the ETOE extraction component.

2 System Architecture

The ETOE Detection architecture is shown in Figure 1. It passes over a single crawled web object in a simplified form (containing structural information, links, information about embedded multimedia objects and the text itself. The first step in the ETOE detection process is to determine the language of the text for all subsequent linguistic processing, using either metadata from the text or a language identification component. The Linguistic Preprocessing provides linguistic information that will be used in later processing, such as tokenisation, part of speech tagging, lemmatization and so on. Next, the Entity Detection module tries to detect crawl-related entities that occur in the document, followed by relation detection, which is used together with the entity information for the detection of events.

Beside text documents, the ETOE component also analyses images and videos to predict entities within images, as well as topics and events depicted by the image as a whole. Images can potentially be very useful in aiding the disambiguation of entities extracted from the textual content of the document. The training and data collection for these classifiers is an offline process that is

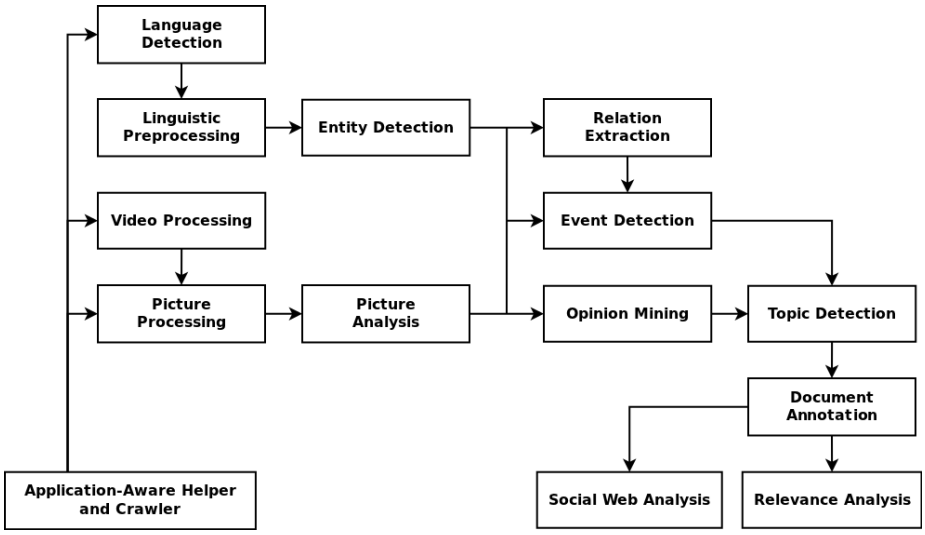

Figure 1: Architecture of the ETOE component

itself guided by the crawl specification. For the processing of videos, images are first extracted from the video and then picture analysis methods are applied.

The extracted meta-information and the linguistically tagged document are also used for detecting the opinion or sentiment of the document. Then, topic detection builds upon the results of the previous processing to identify the topic. Finally, the document will be annotated with all extracted information and used as a starting point for the dynamics analysis, social web analysis and relevance analysis modules. In the following sections, we describe the main individual ETOE components in more detail. Full description of these (and their evaluation) will be included in the final paper.

3 Entity extraction

The entity extraction component of the system makes use of the named entity recognition tools in GATE, which have been modified and extended for the purposes of this task. The set of entities has been extended to include:

- Persons (e.g. artists, politicans, web 2.0 users);
- Organizations (e.g. companies, music bands, political parties);
- Locations (e.g. location names, geo codes);
- Events (e.g. concerts, accidents, crimes);
- Brands and products (e.g. swr3, Tatort, iPhone);
- Dates and times (of events and content publication).

Entity extraction also includes the recognition of technical terms, for which we have adapted and extended the TermRaider plugin for GATE. Both named

entity and term recognition tools have also been adapted to work on German as well as English texts. The application can deal with a multilingual corpus, making use of a language identification plugin to first determine whether a piece of text is in English or German, and then calling the appropriate language processing components (tokeniser, POS tagger, named entity recognition components, etc.) for that language. In this way, the selection of appropriate resources for the language is entirely automatic and requires no human intervention.

4 Event detection

In this work, we refer to an event as a situation within the domain (states, actions, processes, properties) expressed by one or more relations. These may be unique events such as the first landing on the moon or a natural disaster, or regularly occurring events such as elections or TV serials.

Events can be expressed by text elements such as:

- verbal predicates and their arguments (e.g. “The committee dismissed the proposal”);
- noun phrases headed by nominalizations (e.g. “economic growth”);
- adjective-noun combinations (e.g. “governmental measure”; “public money”);
- event-referring nouns (e.g. “crisis”, “cash injection”).

Events can denote different levels of semantic granularity, i.e. general events can contain more specific sub-events. For instance, the performances of various bands form sub-events of a wider music event, while a general event like “Turkey’s EU accession” has sub-events such as the European Parliament approving Turkey’s Progress Report.

The event detection component consists of a combination of two approaches. The top-down approach involves a form of template filling, by selecting a number of known events in advance, and then identifying relevant verbs and their subjects and objects to match the slots. For example, a "performance" event might consist of a band name, a verb denoting some kind of "performing" verb, and optionally a date and location. This kind of approach tends to produce high precision but relatively low recall. We therefore supplement this with a bottom-up approach which consists of identifying verbal relations in the text, and classifying them into semantic categories, from which new events can be suggested. This kind of approach produces higher recall, but precision can be low due to attachment ambiguities if full parsing is not used. We tend to avoid full parsing because, in addition to being slow, it is often highly inaccurate on social media where sentences are not grammatical.

5 Opinion Mining

The opinion mining component finds opinions and sentiments relating to the previously identified entities and events. The first stage is to perform a basic sentiment analysis, i.e., to associate a positive, negative or neutral sentiment with each relevant entity or event. We extend the work of [3] which looked

at identifying political opinions in tweets, and associating with each relevant tweet a triple denoting three kinds of entity: Person, Opinion and Political Party. Here, the opinion is extended to all kinds of entity and event, not just to political parties.

The detection of the actual opinion (sentiment) is performed via a number of different phases: detecting positive, negative and neutral words, identifying factual or opinionated versus questions or doubtful statements, identifying negatives, sarcasm and irony, and detecting extra-linguistic clues such as smileys. These processing resources are developed for both German and English, and make use of SentiWS [4] for German and SentiWordNet [1] for English as seed lexicons for opinionated words. The initial system uses a rule-based approach, but integration of additional machine learning components is planned.

6 Multimedia Processing

Many of the processes that can be applied to textual information have analogues in multimedia data. For example, entity extraction is analogous to object recognition (or face recognition for *person* entities) and topic detection is analogous to multimedia summarisation. Recent work has even investigated how sentiment analysis can be applied to mine coarse notions of opinion from images [5].

Unconstrained object and face recognition in images and video is probably impossible. However, by adding constraints based on the entities in the text surrounding the multimedia object, the problem is much more tractable. An interesting side effect of combining the textual and multimedia information is that the multimedia data can be used to enrich and disambiguate the textual information and vice versa. As an example, consider a news article about "George Bush" that includes a picture. The text of the article may not explicitly make it clear which George Bush (junior or senior) is being discussed, however, by applying face recognition/verification the it becomes possible to disambiguate the entity extracted from the text. We are actively extending OpenIMAJ to classify more complex objects and scene types using combined visual and textual features, in particular we are investigating new techniques for dynamically building object classifiers for previously unseen object types.

Another potential use for the multimedia data is to create links between documents that reuse the same, or a similar, multimedia object. Near-duplicate detection and matching of multimedia objects is a relatively advanced area of research [2, 6]. In terms of archiving, this is useful where articles occur in languages that are not supported by the NLP tools; if an image that occurs in an English article that has been deemed highly relevant based on features extracted through NLP, and the same image occurs in a different article, written in an unsupported language, then it is highly likely that the second article is also relevant to the archive.

7 Conclusions

This abstract has sketched the main extraction components for an entity-centric approach to community archiving. The system extends a core IE system to handle complex events, opinion mining and multimedia components, all of

which provide further information which is relevant for better understanding of the text, and which feeds back into the crawler and storage mechanisms for the archiving tool. Preliminary evaluations of the individual components are promising; further details will be provided in the full paper.

References

[1] A. Esuli and F. Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006*, 2006.

[2] David Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, January 2004.

[3] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In *Proceedings of MSM 2011: Making Sense of Microposts Workshop at 8th Extended Semantic Web Conference*, Heraklion, Greece, May 2011.

[4] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. Sentiws - a publicly available german-language resource for sentiment analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

[5] Stefan Siersdorfer, Jonathon Hare, Enrico Minack, and Fan Deng. Analyzing and predicting sentiment of images on the social web. In *ACM Multimedia 2010*, pages 715–718. ACM, October 2010. URL `http://eprints.ecs.soton.ac.uk/21670/`.

[6] Wengang Zhou, Yijuan Lu, Houqiang Li, Yibing Song, and Qi Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM Multimedia*, pages 511–520, 2010.