Entities, Topics and Events in Community Memories ⋆

Elena Demidova[1], Nicola Barbieri[2], Stefan Dietze[1], Adam Funk[3], Gerhard Gossen[1], Diana Maynard[3], Nikos Papailiou[4], Vassilis Plachouras[4], Wim Peters[3], Thomas Risse[1], Yannis Stavrakas[4], and Nina Tahmasebi[5]

[1]L3S Research Center, Hannover, Germany. [2]Yahoo Iberia SLU, Barcelona, Spain. [3]University of Sheffield, United Kingdom. [4]ATHENA RIC in Information Communication & Knowledge Technologies, Athens, Greece. [5]Computer Science & Engineering Department, Chalmers University of Technology, Goteborg, Sweden.

Abstract. This paper briefly describes the components of the ARCOMEM architecture concerned with the extraction, enrichment, consolidation and dynamics analysis of entities, topics and events, deploying text mining, NLP, and semantic data integration technologies. In particular, we focus on four main areas relevant to support the ARCOMEM requirements and use cases: (a) entity and event extraction from text; (b) entity and event enrichment and consolidation; (c) topic detection and dynamics; and (d) temporal aspects and dynamics detection in Web language and online social networks.

1 Introduction

Community memories largely revolve around events and the entities, topics and opinions related to these events. 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. In this context, the main logical concepts considered in ARCOMEM extraction and enrichment activities are Entities, Topics, Opinions and Events (ETOEs). To create incrementally enriched Web archives that allow access to all sorts of Web content in a structured and semantically meaningful way, extraction, enrichment and consolidation of ETOEs are of crucial importance [3]. To this extent, the main challenges we face are related to the extraction, detection and correlation of ETOEs and related information in a vast number of heterogeneous Web resources as well as analysis of their dynamics. These processes face issues arising from the diversity of the nature and quality of Web content, in particular when considering social media and user-generated content, where further issues are posed by informal use of language.

While entities and events resulting from the automatic extraction processes provide an initial classification and structure for the crawled Web documents, they can be heterogeneous, ambiguous and provide only very limited information. Therefore, data enrichment and consolidation in ARCOMEM follows three aims: (a) enrich extracted entities with related publicly available knowledge; (b) disambiguation, and (c) identify data correlations by aligning ARCOMEM entities with reference datasets. To achieve

⋆ This work was funded by the European Commission under grant agreement n. 270239 (ARCOMEM).

these aims we enrich ARCOMEM entities using Linked Open Data (LOD) sources such as DBpedia (www.dbpedia.org) and Freebase (www.freebase.com) and correlate the entities using their direct and indirect relationships within the LOD graph.

In addition to entity and event extraction, topic discovery and analysis techniques provide an effective way of analyzing and browsing large collections of textual data, such as the ones collected during the ARCOMEM's crawling campaigns, as they are able to uncover the hidden thematic, and semantic structure of the data. By exploiting the potentiality of this kind of analysis, the ARCOMEM framework provides state-of-the art tools for probabilistic topic detection and to analyze the trendiness and dynamical change of topics into time. These tools can provide a high level perspective of the textual data retrieved during a crawling campaign, and help the user to understand the semantic concepts behind each document. In addition, we investigate relationships between topics and social influence of the users and proposed topic-aware social influence framework that jointly learns topics and users' influence in those topics[1].

Since archiving has to consider evolution of content and metadata over time, temporal and dynamics-related aspects of entities and social networks are of special importance for ARCOMEM. Whereas language evolution is reflected in document archives, new challenges arise to automatically determine relevant information, even when it is expressed using forgotten terms. Because language changes over time, we run the risk of a gap between language known to the user and language stored in digital archives. Users cannot find information in the archives, when the names they know have changed over time. In ARCOMEM, we have developed the NEER system, the Named Entity Evolution Recognizer[8]. This tool is used to find and present different names for the same entity to the user and help the user uncover otherwise lost information.

Finally, evolution and changes of discussed topics in online social networks are of interest for the archive users. The users need to view and analyze data from online social networks, such as Twitter, for producing various kinds of results. To support this task, we proposed a model and a query language for defining in a flexible way data views on tweets. The key point of our approach is the representation of term associations and their evolution through time; queries combine complex conditions on time, terms, and their associations[7],[5].

In this paper, we provide an overview of the ARCOMEM offline processing chain and cross-crawl analysis that incorporate text processing components supporting automated extraction, enrichment and dynamics analysis for ETOEs from text.

2 The Offline Processing Chain and Cross Crawl Analysis

The overall ARCOMEM architecture includes several levels [6]. First, in the *crawler level* the system decides and fetches the relevant Web objects as initially defined by the archivists. In parallel to the crawler level, *online processing level* performs time-efficient analysis of the fetched resources to support prioritization of the crawler processing queue. Following that, in the *offline processing level*, fully-featured versions of the Entity, Topic, Opinion and Event (ETOE) analysis and the analysis of the social contents operate over the cleansed data from the crawl that are stored in the ARCOMEM Knowledge Base (ARCOMEM KB). These processing tools make use of linguistic, ma-

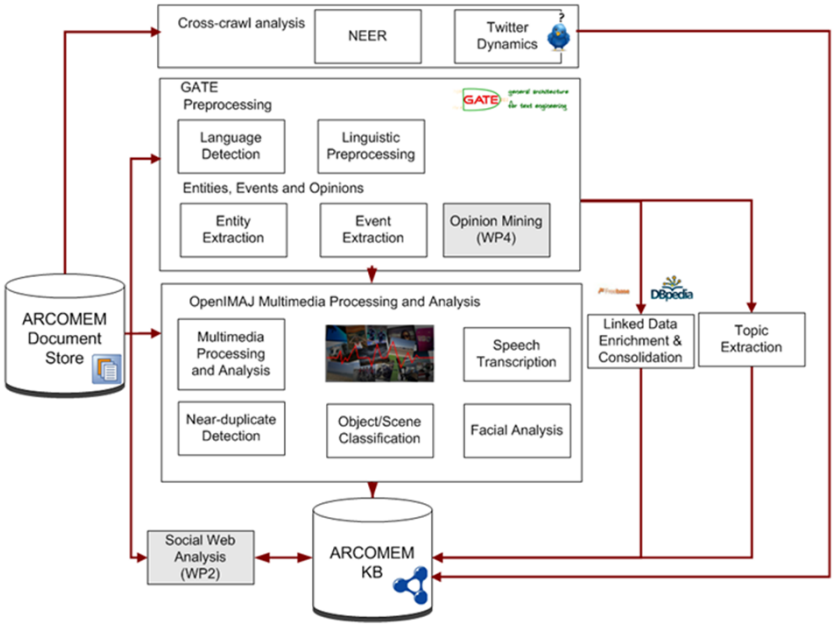

Fig. 1: Processing chain for ETOE extraction, analysis and enrichment

chine learning and NLP methods in order to provide a rich set of metadata annotations that are interlinked with the original data. The respective annotations are stored back in the ARCOMEM database and are available for further processing and information mining. After all the relevant processing has taken place, the Web pages to be archived and preserved are selected in a semi-automatic way. The selected original pages are transferred to the Web archive (in the form of WARC files). Finally, *cross crawl analysis level* operates on collections of Web objects that have been collected over time in order to register the evolution of various aspects identified by the ETOE and Web analysis components. As such, it produces aggregate results that pertain to a set of Web objects crawled at different points in time, rather than to a single crawl. Both the Web Archive and the ARCOMEM KB are accessible for the ARCOMEM applications (such as Crawler Cockpit and Search and Retrieval Application) to support their functionality.

ETOE processing chain within the offline processing and cross crawl analysis levels of the ARCOMEM architecture is presented in Fig.1. In the following, we briefly introduce components responsible for processing text resources.

The ARCOMEM KB and data model: The ARCOMEM Knowledge Base (ARCOMEM KB) serves as a central integration point of the chain enabling its components to store and exchange processing results. The ARCOMEM KB is implemented as H2RDF[4] a fully distributed RDF store that combines the MapReduce processing framework with a NoSQL distributed data store. H2RDF is able to provide a fully distributed SPARQL query functionality on virtually unlimited number of RDF triples.

The information within the ARCOMEM KB is stored according to the ARCOMEM data model that describes concepts used in ARCOMEM and their relationships.

Entity extraction from text: The work performed in this task aims to advance the state of the art in the development and adaptation of language processing resources to new domains and languages, especially in the domain of social media, where such tools are not widespread and suffer from a variety of problems. Our main aim is to investigate new methodologies for language processing on social media and particularly on degraded texts (especially tweets).

The entity recognition components of ARCOMEM are all developed in GATE[1]. The application for entity recognition consists of a set of processing resources (PRs) executed sequentially over a corpus of documents. The application can be divided into the following components: (1) *Document pre-processing*, which separates the body of the content from the rest; (2) *Linguistic pre-processing* such as tokenisation, part of speech tagging, lemmatization and language identification for documents or even document parts in different languages; (3) *Named Entity Recognition* to detect crawl-related entities that occur in the document such as persons (e.g. artists, politicians, Web 2.0 users), organizations (e.g. companies, music bands, political parties), locations (e.g. cities, countries), dates and times. (4) *Term Extraction*; and (5) *RDF generation*.

In ARCOMEM, we have created a new branch of the GATE application which deals specifically with tweets, including specialized techniques to language identification, tokenization, normalization, and POS-tagging in tweets [2].

Event extraction from text: Along with entities, event recognition is one of the major tasks within Information Extraction, and has been successfully applied in research areas such as ontology generation, bioinformatics, news aggregation, business intelligence and text classification. Recognising events in these fields is generally carried out by means of pre-defined sets of relations, possibly structured into an ontology, which makes such tasks domain-dependent, but feasible. In ARCOMEM 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.

The event detection component consists of a combination of various 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 lower precision. The GATE application we have developed for event recognition is similar in structure to the one for entity recognition, and is designed to be run after the entity recognition application has first been run on the corpus.

ETOE enrichment and correlation: ETOEs extracted by the GATE component of the ARCOMEM offline analysis depicted in Fig. 1 provide an initial classification

[1] http://gate.ac.uk

and structure for the crawled Web documents, for instance, the association of terms with entity types defined in the ARCOMEM data model. However, as the content analysis extracts structured data from unstructured resources, such as text and images, the generated data is (i) heterogeneous, i.e. not well interlinked, (ii) ambiguous, and (iii) provides only very limited information. This is due to the data being generated by different components and during independent processing cycles.

Data enrichment and consolidation (achieved by the Linked Data Enrichment and Consolidation component in Fig.1) follows three aims: (a) enrich existing entities with related publicly available knowledge; (b) disambiguation, and (c) identify data correlations such as the ones above by aligning ARCOMEM entities with reference datasets. Both (a) and (b) exploit publicly available data from the Linked Open Data cloud[2], which offers a vast amount of data of both domain-specific and domain-independent nature (the current release of the LOD cloud consists of more than 31 billion distinct triples, i.e. RDF statements). The ETOE enrichment approach [3] first identifies correlating enrichments from reference datasets, which are associated with the respective entities and, secondly, uses these shared enrichments to identify correlating entities in the ARCOMEM KB. In particular, the current enrichment approach uses DBpedia and Freebase as reference datasets.

Topic detection, dynamics and topic-aware influence models: Exploring and retrieving meaningful information from large collections of textual documents is a challenging task. The hidden thematic structure in such collections can be discovered by applying recently proposed statistical analysis tools, such as Probabilistic Topic Models. At high level, the idea is to study the co-occurrence of words, assuming that words that tend to co-occur frequently, express, or belong to, the same semantic concept.

Topic detection based on probabilistic topic models: The purpose of the Topic Detection module is to uncover the hidden thematic structure of a document collection related to an ARCOMEM campaign, and to identify semantic topics of interest for future data analysis and navigation. The detected topics provide a low-dimensional representation, in terms of abstract co-occurrence patterns, of the textual data analyzed. After detecting topics, the projection of each document into the topic space can be exploited to classify documents and named entities into semantic categories, allowing a more effective browsing of the content crawled during each campaign. To represent the topic space in a compact and intuitive way, the module also computes a semantic affinity score between topics, which can be used to summarize their relationships in a graph. Topics also constitute an important dimension for identifying experts and influential users. To this aim, we propose a machine learning framework which jointly models social influence and topics and is able to effectively detect users' authoritativeness and interests for each retrieved topic [1].

Dynamics on topics: The goal of this module is to analyze the temporal evolution of semantic meaningful POS which come from textual documents crawled during a campaign. This tool is particularly useful when dealing with long-term crawling campaigns, as it allows to monitor and detect the popularity and evolution of semantical concept, represented by a combination of named entities, bigrams or events.

[2] http://lod-cloud.net/

Dynamics in Web language (NEER): Language evolution is reflected in documents available on the Web or in document archives but is not sufficiently considered by current applications. Therefore, not knowing about different names referencing the same entity may severely compromise system effectiveness. We focus on named entity evolution, the detection of name changes, as it has a high impact, for example in information retrieval in archives, as well as linking of entities over long time spans. This research field is becoming increasingly important. However, most previous work depend on the availability of external knowledge sources or assume a static context around terms and expect the names to be the only changing factor. In ARCOMEM, we follow a statistical approach to eliminate the dependency on external resources and use a context based method that considers only periods with a high likelihood of name change, capturing evolving names with less computational effort [8].

Dynamics in online social networks: This module addresses temporal issues and dynamics of online social networks to enable end users and other modules of the ARCOMEM framework to query the evolution and changes of discussed topics in online social networks. In this module, we focus on identifying relations between terms in tweets (including hashtags) in the online social network of Twitter, and provide efficient ways to query the relations for given timespans [5].

3 Conclusion and Outlook

In this paper we briefly introduced components involved in the offline and cross-crawl levels of the ARCOMEM architecture. Most of the components were evaluated in isolation on recent ARCOMEM crawl data and the results confirmed good performance of these components. Due to space limitation we did not include evaluation results in this paper. As a next step, we plan to perform evaluation of the entire processing chain.

References

1. N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. *Knowledge and Information Systems*, pages 1–30, 2013.
2. L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In *Proc. of the 24th ACM Conference on Hypertext and Social Media*, HT '13, New York, NY, USA, 2013. ACM.
3. S. Dietze, D. Maynard, E. Demidova, T. Risse, W. Peters, K. Doka, and Y. Stavrakas. Preservation of social web content based on entity extraction and consolidation. In *Proc. of 2nd International Workshop on Semantic Digital Archives (SDA)*, 2012.
4. N. Papailiou. H2rdf: adaptive query processing on rdf data in the cloud. In *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 2012.
5. V. Plachouras and Y. Stavrakas. Querying term associations and their temporal evolution in social data. In *1st International Workshop on Online Social Systems (WOSS2012)*, 2012.
6. T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic web for guided web archiving. In *Proc. of TPDL'12*, 2012.
7. Y. Stavrakas and V. Plachouras. A platform for supporting data analytics on twitter: Challenges and objectives. In *KECSM2012, CEUR Workshop Proceedings Vol. 895*, 2012.
8. N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. Neer: An unsupervised method for named entity evolution recognition. In *Proc. of COLING 2012*, 2012.