Preservation of Social Web Content based on Entity Extraction and Consolidation

Stefan Dietze[1], Diana Maynard[2], Elena Demidova[1], Thomas Risse[1], Wim Peters[2,]
Katerina Doka[3], Yannis Stavrakas[3]

[1] L3S Research Center, Leibniz University, Hannover, Germany
{dietze, nunes, demidova, risse}@l3s.de

[2] Department of Computer Science, University of Sheffield, Sheffield, UK
{diana,wim}@dcs.shef.ac.uk

[3] IMIS, RC ATHENA, Artemidos 6, Athens 15125, Greece
katerina@cslab.ece.ntua.gr; yannis@imis.athenainnovation.gr

Abstract. With the rapidly increasing pace at which Web content is evolving, particularly social media, preserving the Web and its evolution over time becomes an important challenge. Meaningful analysis of Web content lends itself to an entity-centric view to organise Web resources according to the information objects related to them. Therefore, the crucial challenge is to extract, detect and correlate entities from a vast number of heterogeneous Web resources where the nature and quality of the content may vary heavily. While a wealth of *information extraction* tools aid this process, we believe that, the *consolidation* of automatically extracted data has to be treated as an equally important step in order to ensure high quality and non-ambiguity of generated data. In this paper we present an approach which is based on an iterative cycle exploiting Web data for (1) targeted archiving/crawling of Web objects, (2) entity extraction, and detection, and (3) entity correlation. The long-term goal is to preserve Web content over time and allow its navigation and analysis based on well-formed structured RDF data about entities.

Keywords. Knowledge Extraction, Linked Data, Data Consolidation, Data Enrichment, Web Archiving, Entity Recognition

1 Introduction

Given the ever increasing pace at which Web content is constantly evolving, adequate Web archiving and preservation have become a cultural necessity. Along with "common" challenges of digital preservation, such as media decay, technological obsolescence, authenticity and integrity issues, Web preservation has to deal with the sheer size and ever-increasing growth rate of Web content. This in particular applies to user-generated content and social media, which is characterized by a high degree of *diversity*, heavily *varying quality* and *heterogeneity*. Instead of following a *collect-all* strategy, archival organizations are striving to build *focused archives* that revolve around a particular *topic* and reflect the diversity of information people are interested in. Thus, focused archives largely revolve around the *entities* which define a topic or area of interest, such as persons, organisations and locations. Hence, extraction of entities from archived Web content, in particular social media, is a crucial challenge in order to allow semantic search and navigation in Web archives and the relevance assessment of a given set of Web objects for a particular *focused crawl*.

However, while tools are available for information extraction from more formal text, social media affords particular challenges to knowledge acquisition, such as multilinguality (not only across but within documents), varying speech quality (e.g. poor grammar, spelling, capitalisation, use of colloquialisms etc), and greater heterogeneity of data. Due to these reasons, data extracted by automated means from social media often suffers from varying, non-optimal quality, noise, inaccuracies, redundancies as well as inconsistencies. In addition, it tends to lack sufficient descriptiveness, usually consisting of labeled and, at most, classified entities, which leads to ambiguities. This calls for a range of specific strategies and techniques to *consolidate*, *enrich*, *disambiguate* and *interlink* extracted data. This in particular benefits from taking advantage of existing knowledge, such as Linked Open Data [1], to compensate for, disambiguate and remedy degraded information. While data consolidation techniques traditionally exist independent from named entity recognition (NER) technologies, their coherent integration into unified workflows is of crucial importance to improve the wealth of automatically extracted data on the Web. This becomes even more crucial with the emergence of an increasing variety of publicly available and end-user friendly knowledge extraction and NER tools such as DBpedia Spotlight[1], GATE[2], Open Calais[3], Zemanta[4].

In this paper, we introduce an integrated approach to extracting and consolidating structured knowledge about entities from archived Web content. This knowledge will in the future be used to facilitate semantic search of Web archives and to further guide the crawl. In our approach, knowledge extraction and consolidation techniques are treated as equally important in order to gradually improve the quality – non-ambiguity, coherence and richness - of extracted information. This work was developed in the EC-funded Integrating Project ARCOMEM[5]. Note, while temporal aspects related to term and knowledge evolution are substantial to Web preservation, these are currently under investigation [26] but out of scope for this paper.

2 Related Work

Entity recognition is one of the major tasks within information extraction and may encompass both NER and term extraction. ER may involve rule-based systems [15] or machine learning techniques [16]. Term extraction involves the identification and filtering of term candidates for the purpose of identifying domain-relevant terms or entities. The main aim in automatic term recognition is to determine whether a word

[1] http://spotlight.dbpedia.org

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

[3] http://www.opencalais.com/

[4] http://www.zemanta.com/

[5] http://www.arcomem.eu

or a sequence of words is a term that characterises the target domain. Most term extraction methods use a combination of linguistic filtering (e.g. possible sequences of part of speech tags) and statistical measures (e.g. tf.idf) [17] and [18], to determine the salience of each term candidate for each document in the corpus [25].

Data consolidation has to cover a variety of areas such as enrichment, entity/identity resolution for disambiguation as well as clustering and correlation to consolidate disparate data. In addition, link prediction and discovery is of crucial importance to enable clustering and correlation of enriched data sources. A variety of methods for entity resolution have been proposed, using relationships among entities [9], string similarity metrics [8], as well as transformations [11]. An overview of the most important works in this area can be found in [10]. As opposed to entity correlation techniques exploited in this paper, text clustering of documents exploits feature vectors, to represent documents according to contained terms [12][13][14]. Clustering algorithms measure the similarity across the documents and assign the documents to the appropriate clusters based on this similarity. Similarly, vector-based approaches have been used to map distinct ontologies and datasets [4][5].

As opposed to text clustering, entity correlation and clustering takes advantage of background knowledge from related datasets to correlate previously extracted entities. Therefore, link discovery is another crucial area to be considered. Graph summarization predicts links in annotated RDF graphs. A detailed survey of link predictions techniques in complex networks and social network are presented by [6] and [7], respectively.

3 Challenges and overall approach

ARCOMEM follows a use case-driven approach based on scenarios aimed at creating focused Web archives, particularly of social media, by adopting novel entity extraction and crawling mechanisms. ARCOMEM focused archives deploy (a) a document repository of crawled Web content and (b) a structured RDF knowledge base containing metadata about entities detected in the archived content. Archivists will be enabled to specify or modify crawl specifications (fundamentally consisting of selected sets of relevant entities and topics). The intelligent crawler will be able to learn about crawl intentions and to refine a crawling strategy on-the-fly. This is especially important for long running crawls with broader topics, such as the *financial crisis* or *elections*, where involved entities are changing more frequently compared to highly focused crawls, and hence, require regular adaptation of the crawl specification. End-user applications allow users to search and browse the archives by exploiting automatically extracted metadata about entities and topics.

Fundamental to both crawl strategy refinement and Web archive navigation is the efficient extraction of entities from archived Web content. In particular, social media poses a number of challenges for language analysis tools due to the degraded nature of the text, especially where tweets are concerned. In one study, the Stanford NER tagger dropped from 90.8% F1 to 45.88% when applied to a corpus of tweets [19]. [21] also demonstrate some of the difficulties in applying traditional POS tagging, chunking and NER techniques to tweets, while language identification tools typically also do not work well on short sentences. Problems are caused by incorrect spelling and grammar, made-up words (eg swear words, additional infixes), unusual but important tokens such as hashtags, @ signs and emoticons, unorthodox capitalisation, and spellings (e.g duplication of letters in words for emphasis, text speak). Since tokenisation, POS tagging and matching against pre-defined gazetteer lists are key to NER, it is important to resolve these problems: we adopt methods such as adapting tokenisers, using techniques from SMS normalisation, retraining language identifiers, use of case-insensitive matching in certain cases, using shallow techniques rather than full parsing, and using more flexible forms of matching.

In addition, to compensate for noise and lack of semantics of entities extracted automatically from heterogeneous social media, we include data consolidation and enrichment techniques into a coherent processing chain. The entity extraction and enrichment experiments described in this paper were applied to ARCOMEM specific datasets. These datasets consist of crawls which were provided as part of the use case applications with a particular focus on *financial crisis*-related content. Parts of the archived resources are available at *http://collections.europarchive.org/arcomem/*.

3.1 Processing chain overview

Entity extraction and enrichment is covered by a set of dedicated components which have been incorporated into a dedicated processing chain (Figure 1) which handles NER and consolidation (enrichment, clustering, disambiguation) as part of one coherent workflow. While the extraction and enrichment components (*Processing* layer in Figure 1) are detailed in the following sections, we would like to briefly introduce the ARCOMEM data model and the ARCOMEM Knowledge Base (*Storage* layer), which serves as central component for the extraction components according to the ARCOMEM data model. Note that in the following we focus in particular on entity recognition and consolidation from text (*GATE*, *Data Enrichment & Consolidation*), leaving aside the remaining components, such as *Event Recognition* ones.

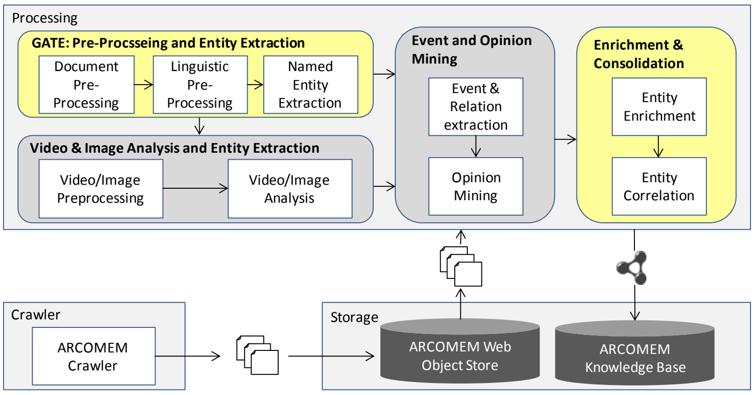


Fig. 1. Entity extraction and consolidation processing chain

3.2 Object store, knowledge base and data model

The ARCOMEM storage is a component that plays a central role in the platform. Its task is to provide storing, indexing, and retrieving mechanisms for all data produced and utilized by the rest of the architectural components. As such, it is expected to store, serve, and update different kinds of data: (a) binary data, in the form of Web objects, which represent the original content collected by the crawler; and (b) semi-structured data, in the form of RDF triples, which serve as Web object annotations and are primarily used by the ETOE and Social Web analysis, the dynamics analysis, as well as the applications. The sheer volume of information available on the Internet combined with the requirement of our system to capture multiple versions of Web objects over time creates enormous demands for storage as well as memory. Moreover, as some decisions are made at runtime (e.g., during the online processing), queries need to be resolved in near-real-time. Even for complex analytics tasks, high throughput is important since they may trigger a large number of queries and reduces performance. To cover the functional requirements, we have designed and implemented a storage module consisting of two components: the *object store* and the *knowledge base*. Both of them rely on distributed solutions that combine the MapReduce [9] paradigm and NoSQL databases. *MapReduce* is an ideal paradigm for harnessing scale-out architectures to build huge indices on distributed storage, while NoSQL databases, based on shared-nothing architectures, offer scalability and availability at low cost. Both, object store and knowledge base are realized based on HBase[6]. For the knowledge base, an enhanced version of the centralised Hexastore indexing scheme [28] over HBase was implemented.

The main logical *concepts* considered in extraction and enrichment activities are *entities*, *roles*, *relations*, *events, and topics*. We focus on entities, since these form the cornerstone for the extraction of other concepts. The *ARCOMEM data model* was created to reflect the informational needs for knowledge capturing, crawling, and preservation (see [22] for details). The central concepts in this configuration are *InformationObject*, *InformationRealization* and *CrawlingConcept*. InformationObject is the class that subsumes all information types: *Entities*, *Events*, *Opinions*, and *Topics*. Multilingual instances of this class are classified according to the language they belong to. InformationRealization captures the concrete instantiations of these information objects in the form of multimedia Web objects such as text documents or images. CrawlingConcept describes required aspects of the crawling workflow. The ARCOMEM data model[7] is represented in RDF[8]. Links to concepts from various established vocabularies ensure high interoperability. While entity enrichments and consolidation is an integral aspect of ARCOMEM, the data model contains dedicated *Enrichment*, *EnrichmentContext* and *Cluster* concepts. While the Enrichment concept is used to represent external concepts within the ARCOMEM knowledge base, the EnrichmentContext defines how a particular ARCOMEM entity relates to the particu-

[6] Apache Foundation; The Apache HBase Project: http://hbase.apache.org/

[7] http://www.gate.ac.uk/ns/ontologies/arcomem-datamodel.rdf

[8] http://www.w3.org/RDF/

lar Enrichment, to describe, for instance, the property which was enriched (e.g. *rdfs:label*) and the confidence of the enrichment.

3.3 Entity extraction, enrichment and consolidation

Within the ARCOMEM model, "entity" encompasses both traditional Named Entities and also single and multi-word terms: the recognition of both is done using GATE. While extracted data is already classified and labeled as a result of this process, it is nevertheless (i) heterogeneous, i.e. not well interlinked, (ii) ambiguous and (iii) provides only very limited information. This is due to data being extracted by different components and during independent processing cycles, since the tools in GATE have no possibility to perform co-reference on entities generated asynchronously across multiple documents. For instance, during one particular cycle, the text analysis component might detect an entity from the term “Ireland”, while during later cycles, entities based on the term “Republic of Ireland”" or the German term “Irland” might be extracted, together with, the entity “Dublin”. These would all be classified as entities of type *Location* and correctly stored in the ARCOMEM data store as disparate entities described according to the ARCOMEM RDF schema. Thus, *Enrichment and Consolidation* (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 illustrated above. This is achieved by mapping isolated ARCOMEM entities to concepts (nodes) within reference datasets (*enrichment*) and exploiting the corresponding graphs to discover correlations. Therefore, we exploit publicly available data from the Linked Open Data cloud[9] which offers a vast amount of data of both domain-specific and domain-independent nature (the current release consists of 31 billion distinct triples, i.e. RDF statements[10]).

4 Implementation

For entity recognition, we use a modified version of ANNIE [20] to find mentions of *Person*, *Location*, *Organization*, *Date*, *Time*, *Money* and *Percent*. We included extra subtypes of *Organization* such as *Band* and *Political Party*, and have made various modifications to deal with the problems specific to social media such as incorrect English (see [23] for more details). The entity extraction framework can be divided into the following components (GATE component in Fig. 1) which are executed sequentially over a corpus of documents:

- Document Pre-processing (document format analysis, content detection)
- Linguistic Pre-processing (language detection, tokenisation, POS tagging etc)
- Named Entity Extraction: Term Extraction (generation of ranked list of terms and thresholding) & NER (gazetteers, rule-based grammars and co-reference)

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

[10] http://lod-cloud.net/state

For term extraction, we use an adapted version of TermRaider[11]. This considers all noun phrases (NPs) – as candidate terms (as determined by linguistic preprocessing), and ranks them in order of termhood according to 3 different scoring functions: (1) basic tf.idf (2) an augmented tf.idf which also takes into account the tf.idf score of any hyponyms of a candidate term, and (3) the Kyoto score based on [24] which takes into account the number of hyponyms of a candidate term occurring in the document. All are normalised to represent a value between 0 and 100. A candidate term is not considered an entity if it matches or is contained within an existing Named Entity, to avoid duplication. Also, we have set a threshold score above which we consider a candidate term to be valid. This threshold is a parameter which can be manually changed at any time – currently it is set to an augmented score of 45, i.e. only terms with a score of 45 or greater will be used by later processes.

The entity extraction generates RDF data describing NEs and terms as RDF/XML according to the ARCOMEM data model which is pushed to our knowledge base and directly digested by our *Enrichment & Consolidation* component (Fig. 1). The latter exploits (a) the *entity label* and (b) the *entity type* to expand, disambiguate and correlate extracted data. Note that an entity/event label might correspond directly to a label of one unique node in a structured dataset (as is likely for an entity of type person labelled "Angela Merkel"), but might also correspond to more than one node/concept, as is the case for most of the events in our dataset. For instance, the event labeled "Jean Claude Trichet gives keynote at ECB summit" will most likely be enriched with links to concepts representing the ECB as well as Jean Claude Trichet. Our approach is based on the following steps (reflected in Fig. 1):

S1. *Entity enrichment*
- S1.a. Translation: we determine the language of the entity label, and, if necessary, translate it into English using an online translation service.
- S1.b. Enrichment: extracted entities are co-referenced with related entities in reference datasets.

S2. *Entity correlation* and clustering

In order to obtain enrichments for these entities we perform queries on external knowledge bases. Our current enrichment approach uses DBpedia[12] and Freebase[13] as reference datasets, though it is envisaged to expand this approach with additional and more domain-specific datasets, e.g., event-specific ones. DBpedia and Freebase are particularly well-suited due to their vast size, the availability of disambiguation techniques which can utilise the variety of multilingual labels available in both datasets for individual data items and the level of inter-connectedness of both datasets, allowing the retrieval of a wealth of related information for particular items. In the case of DBpedia, we make use of the DBpedia Spotlight service which enables an approxi-

[11] http://gate.ac.uk/termraider.html

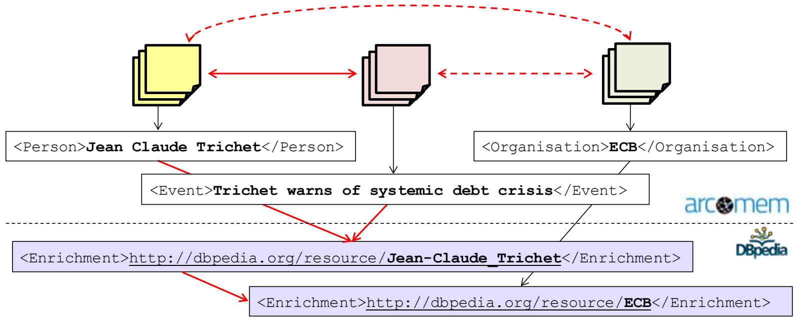
[12] http://dbpedia.org/

[13] http://www.freebase.com/

mate string matching with adjustable confidence level in the interval [0,1]. As part of our evaluation (Section 6), we experimentally selected a confidence level of 0.6 which provided the best balance of precision and recall. Note that Spotlight offers NER capabilities complementary to GATE. However, these were only utilised in cases where entities/events were not in a rather atomic form, as is often the case for events which mostly consists of free text descriptions such the one mentioned above.

Freebase [3] contains about 22 million entities and more than 350 millions facts in about 100 domains. Keyword queries over Freebase are particularly ambiguous due to the size and the structure of the dataset. In order to reduce query ambiguity, we used the Freebase API and restricted the types of the entities to be matched using a manually defined type mapping from ARCOMEM to Freebase entity types. For example, we mapped the ARCOMEM type “person” to the “people/person” type of Freebase, and the ARCOMEM type “location” to the Freebase types “location/continent”, “location/location” and “location/country”. The ARCOMEM entity types were determined previously in the entity extraction process. For instance, an ARCOMEM entity of type “Person” with the label “Angela Merkel” is mapped to the Freebase MQL query that retrieves one unique Freebase entity with the mid= "/m/0jl0g". With respect to data correlation, we distinct *direct* as well as *indirect* correlations. Please note, that a *correlation* does not describe any notion of equivalence (e.g. similar to *owl:sameAs*) but merely a meaningful level of relatedness.

Fig. 2 depicts both cases, direct as well as indirect correlations. Direct correlations are identified by means of equivalent and shared enrichments, i.e., any entities/events sharing the same enrichments are supposedly correlated and hence clustered. In Fig. 2, a direct correlation is visible between the entity of type person labeled “Jean Claude Trichet” and the event “Trichet warns of systemic debt crisis”. In addition, the retrieved enrichments associate the ARCOMEM entities and associated Web objects with the knowledge, i.e., data graph, available in associated reference datasets.


<Person>Jean Claude Trichet</Person>
<Organisation>ECB</Organisation>
<Event>Trichet warns of systemic debt crisis</Event>
arcomem
DBpedia
<Enrichment>http://dbpedia.org/resource/Jean-Claude_Trichet</Enrichment>
<Enrichment>http://dbpedia.org/resource/ECB</Enrichment>

Fig, 2..Enrichment and correlation example: ARCOMEM Web objects, entities/events, associated DBpedia enrichments and identified correlations

For instance, the DBpedia resource of the European Central Bank (*http://DBpedia.org/resource/ECB*) provides additional facts (e.g., a classification as organisation, its members, or previous presidents) in a structured, and therefore, machine-processable form. Exploiting the graphs of underlying reference datasets allows us to identify additional, *indirect correlations*. While linguistic/syntactic approaches would fail to detect a relationship between the two enrichments above (Trichet, ECB) and hence their corresponding entities and Web objects, by analysing the DBpedia graph we are able to uncover a close relationship between the two (Trichet being the former ECB president). Hence, computing the *relatedness* of enrichments would allow us to detect indirect correlations to create a relationship (dashed line) between highly releated entities/events, beyond mere equivalence.

Our current implementation is limited to detect direct correlations, while ongoing experiments based on graph analysis mechanisms aim to automatically measure *semantic relatedness* of entities in reference datasets to detect indirect relations. While in a large graph, all nodes are connected with each other in some way, a key research challenge is the investigation of appropriate graph navigation and analysis techniques to uncover indirect but semantically meaningful relationships between resources within reference datasets, and hence ARCOMEM entities and Web objects.

5 Results & evaluation

5.1 Evaluation datasets

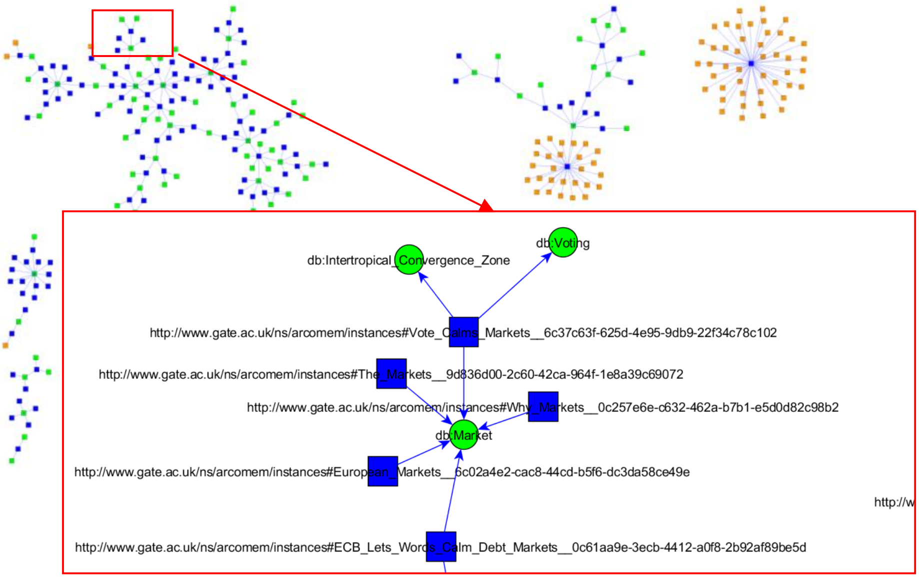
For our experiments, we used a dataset composed of English and German archived Web objects constituting a sample of crawls relating to the financial crsisis. The English content covered 32 Facebook posts, 41,000 tweets and 800 user comments from greekcrisis.net. The German content consisted of archived data from the Austrian Parliament[14] consisting of 326 documents (mostly PDF, some HTML).

Our extraction and enrichment experiments resulted in an evaluation dataset[15] of 99,569 unique entities involving the types *Event*, *Location*, *Money*, *Organization*, *Person*, *Time*. Using the procedure described above, we obtained enrichments for 1,358 of the entities in our dataset using DBpedia (484 entities) and Freebase (975 entities). In total, we obtained 5,291 Freebase enrichments and 491 DBpedia enrichments. These enrichments built 5,801 entity-enrichment pairs, 5,039 with Freebase and 492 with DBpedia.

Our initial clustering technique uncovered a number of clusters, i.e. correlated entities. The following figure shows an excerpt of the graph generated by representing relations between ARCOMEM entities and enrichments. Blue nodes represent ARCOMEM entities, orange ones Freebase enrichments and green nodes DBpedia enrichments. As shown, several interesting entity clusters emerge, i.e., distinct blue nodes which are linked via jointly shared enrichments (orange/green nodes). In the figure, for instance, a number of such clusters appear in the upper left corner, centred around the DBpedia concept *http://dbpedia.org/resource/Market*.

[14] http://www.parliament.gv.at/

[15] The SPARQL endpoint of our dataset (extracted entities and enrichments) is available at http://arcomem.l3s.uni-hannover.de:9988/openrdf-sesame/repositories/arcomem-rdf?query.


db:Voting
db:Intertropical_Convergence_Zone
http://www.gate.ac.uk/ns/arcomem/instances#Vote_Claims_Markets__6c37c63f-625d-4e95-9db9-22f34c78c102
http://www.gate.ac.uk/ns/arcomem/instances#The_Markets__9d836d00-2c60-42ca-964f-1e8a39c69072
http://www.gate.ac.uk/ns/arcomem/instances#Why_Markets__0c257e6e-c632-462a-b7b1-e5d0d82c98b2
db:Market
http://www.gate.ac.uk/ns/arcomem/instances#European_Markets__6c02a4e2-cac8-44cd-b5f6-dc3da58ce49e
http://w
http://www.gate.ac.uk/ns/arcomem/instances#ECB_Lets_Words_Calm_Debt_Markets__0c61aa9e-3ecb-4412-a0f8-2b92af89be5d

Fig, 3. Generated ARCOMEM graph and clusters

5.2 Entity extraction evaluation

We have performed initial evaluations on the various text analysis components. We manually annotated a small corpus of 20 Facebook posts (in English) from the dataset described above with named entities to form a gold standard corpus against which we could compare the system annotations. The corpus contained 93 instances of Named Entities. For evaluating TermRaider, we took a larger set of 80 documents from the financial crisis dataset. From this, TermRaider produced 1003 term candidates (merged from the results of the three different scoring systems). Three human annotators selected valid terms from that list, and we produced a gold standard of 315, comprising each term candidate selected by at least two annotators (221 terms selected by exactly two annotators and 94 selected by all three). While inter-annotator agreement was thus quite low, this is normal for a term extraction task as it is extremely subjective; however, for future we will aim to tighten the annotation guidelines and provide further training to the annotators with the aim of reaching a better consensus.

For the NE recognition evaluation, we compared the system annotations with the gold standard, using the standard metrics of Precision, Recall and F-Measure. The system achieved a Precision of 80% and a Recall of 68% on the task of NE detection (i.e. detecting whether an entity was present or not, regardless of its type). On the task of type determination (getting the correct type of the entity (Person, Organization, Location etc.), the system performed with 98.8% Precision and 98.5% Recall. Overall (for the two tasks combined), this gives NE recognition scores of 79% Precision and 67% Recall. However, the results are slightly low because this actually includes Sentence detection also. Normally, Sentence detection is 100% accurate (or near enough), but in this case, it is subject to the language detection issue, because we only perform the entity detection on sentences deemed to be relevant (in the language of the task - in this case English, and which corresponds to the relevant part of the document - in this case, the actual text of the postings by the users). 26 of the missing system annotations in the document were outside the span of the sentences annotated, so could not have been annotated. Excluding these increase Recall from 68% to 83.9% for NE detection (shown in the table as "NE detection (adjusted)"), and from 67% to 73.5% for the complete NE recognition task (shown in the table as "Full NE recognition (adjusted)").

Table 1. NER evaluation results

Task	Precision	Recall	F1
NE detection	80%	68%	74%
NE detection (adjusted)	80%	83.9%	81,9%
Type determination	98.8%	98.5%	98.6%
Full NE recognition	79%	67%	72.5%
Full NE recognition (adjusted)	79%	82.1%	80.5%

For term recognitions, we compared the TermRaider output for each scoring system with the gold standard set of terms, at different levels of the ranked list, as shown in Figure 4. For the terms above the threshold, we achieved Precision scores of 31% and Recall of 90% for tf.idf, 73% Precision and 50% Recall for augmented tf.idf and 63% Precision and 17% Recall for the Kyoto score. For any further processing, we only use the terms scored by the augmented tf.idf above the threshold.

5.3 Enrichment and correlation evaluation

For this evaluation we randomly selected a set of entity-enrichment pairs. Our evaluation was performed manually by 6 judges including graduate computer science students and researchers. The judges were asked to assign scores to each entity-enrichment pair, with “0” for *incorrect*, and “1” for *correct*. We judge an enrichment as correct if it partially defines a specific dimension of the entity/event, that is, an enrichment does not need to completely match an entity. For instance, enrichments referring to *http://dbpedia.org/resource/Doctor_(title)* and *http://dbpedia.org/page/Angela_Merkel* and enriching an entity of type person labelled “Dr Angela Merkel” were both equally ranked as correct. This is due to entities and events being potentially related to multiple enrichments, each enriching a particular facet of the source entity/event. Each entity/enrichment pair was shown to at least 3 judges and an average of their scores was built to alleviate bias. In case an entity label did not make sense to a judge, we assumed that there has been an error in the extraction phase. In this case we asked the judges to mark the corresponding entity as invalid and excluded it from the evaluation.

We computed the average scores of entity-enrichment pairs across judges and averaged the scores obtained for each entity type. Table 4 presents the average scores of the enrichment-entity pairs obtained using DBpedia and Freebase for different ARCOMEM entity types.

Table 2. Enrichment evaluation results

Entity Type	Avg. Score DBpedia	Avg. Score Freebase	Avg. Score Total
Location	0.94	0.94	0.94
Money	0.63	-	0.63
Organization	0.93	1	0.97
Person	0.72	0.89	0.8
Time	1	-	1
Total	**0.84**	**0.94**	**0.89**

Our initial clustering approach simply correlated entities/events which share equivalent enrichments. In total we generated 1013 clusters with 2.85 entities on average, with a minimum of 2 and a maximum of 112 entities. Ambiguous enrichments led to redundant clusters and require additional disambiguation. For instance, a location entity labelled "Berlin" might be (correctly) enriched with *http://rdf.freebase.com/ns/m/0xfhc* and *http://rdf.freebase.com/ns/m/047ckrl* (each referring to a different location "Berlin") requiring additional disambiguation to clean up the clusters. To this end, we exploit graph analysis methods to detect closeness of enrichments originating from the same object. For instance, measuring the relatedness of two location entities "Berlin" and "Angela Merkel" used to annotate the same Web object will allow us to disambiguate enrichments.

6 Discussion and future works

In this paper we have presented our current strategy for entity extraction and enrichment as realized within the ARCOMEM project, aimed at creating a large knowledge base of structured knowledge about archived heterogeneous Web content. Based on an integrated processing chain, we tackle entity consolidation and enrichment as implicit activity in the information extraction workflow.

The results of the entity extraction show respectable scores for this kind of social media data on which NLP techniques typically struggle. However, current work is focusing on better handling of degraded English (tokenisation, language recognition etc) and especially of tweets, which should improve the entity extraction further. The enrichment results indicate a comparably good quality of generated enrichments. The results obtained from DBpedia Spotlight provided a lower recall, but introduced less ambiguous enrichments due to Spotlights inherent disambiguation feature. On the other hand, partially matched keywords reduce the precision results. As future work, we foresee different directions to improve quality of the enrichment results. For example, one possibility is to use structured DBpedia queries to restrict entity types, similarly to the approach used for Freebase. On the other hand, we consider the introduction of sub-types of entities to further increase granularity of the types to be matched.

In addition, while preservation of Web content over time has to consider temporal aspects, evolution of entities and terms as well as time-dependent disambiguation are important research areas currently under investigation [26]. While our current data consolidation approach only detects direct relationships between entities sharing the same enrichments, our main efforts are dedicated to investigate graph analysis mechanisms. Thus, we aim to further take advantage of knowledge encoded in large reference graphs to automatically identify semantically meaningful relationships between disparate entities extracted during different processing cycles. Given the increasing use of both automated NER tools and reference datasets such as DBpedia, Wordnet or Freebase, there is an increasing need for consolidating automatically extracted information on the Web which we aim to facilitate with our work.

Acknowledgments

This work is partly funded by the European Union under FP7 grant agreement n° 270239 (ARCOMEM).

References

[1] Bizer, C., T. Heath, Berners-Lee, T. (2009). Linked data - The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems.

[2] Rizzo, G., Troncy, L., Hellmann, S., Brümmer, M., NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud, Linked Data on the Web 2012 (LDOW2012), Lyon, France, 2012.

[3] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., Freebase: a collaboratively created graph database for structuring human knowledge. In Proc. of the SIGMOD 2008, pages 1247–1250, New York, NY,USA, 2008. ACM.

[4] Dietze, S., and Domingue, J. (2008) Exploiting Conceptual Spaces for Ontology Integration, Workshop: Data Integration through Semantic Technology (DIST2008) Workshop at 3rd Asian Semantic Web Conference (ASWC) 2008, Bangkok, Thailand.

[5] Dietze, S., Gugliotta, A., and Domingue, J. (2009) Exploiting Metrics for Similarity-based Semantic Web Service Discovery, IEEE 7th International Conference on Web Services (ICWS 2009), Los Angeles, CA, USA.

[6] Lü, L., Zhou, T.: Link prediction in complex networks: a survey, Physica A 390 (2011), 1150–1170.

[7] Hasan, M. A., Zaki, M. J.: A survey of link prediction in social networks. In C. Aggarwal, editor, Social Network Data Analytics, pages 243–276. Springer,

[8] Cohen, W. W., Ravikumar, P. D., Fienberg, S. E.. A comparison of string distance metrics for name-matching tasks. In IIWeb, 2003.

[9] Dong, X., Halevy, A., Madhavan, J., Reference reconciliation in complex information spaces. In SIGMOD, 2005.

[10] Elmagarmid, A. K., Ipeirotis, P. G., Verykios, V. S., Duplicate record detection: A survey. TKDE, 19(1), 2007.

[11] Tejada, S., Knoblock, C. A., Minton, S., Learning domain-independent string transformation weights for high accuracy object identification. In KDD, 2002.

[12] Boley, D., Principal Direction Divisive Partitioning. Data Mining and Knowledge Discovery, 2(4), 1998.

[13] Broder, A., Glassman, S., Manasse, M., Zweig, G., Syntactic Clustering of the Web. In Proceedings of the 6th International World Wide Web Conference, pages 1997.

[14] Hotho, A., Maedche, A., Staab, S., Ontology-based Text Clustering. In Proceedings of the IJCAI Workshop on \Text Learning: Beyond Supervision", 2001.

[15] Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y., Named Entity Recognition from Diverse Text Types. Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark, Bulgaria, 2001

[16] Li, Y., Bontcheva, K., Cunningham, H., Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, 15(02), 241-271, 2009.

[17] Buckley, C., G. Salton, G., Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.

[18] Maynard, D., Li, Y., Peters, W., NLP techniques for term extraction and ontology population. In: Buitelaar, P. and Cimiano, P. (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 171-199, IOS Press, Amsterdam (2008)

[19] Lui, M., Baldwin, T., 2011. Cross-domain feature selection for language identification. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 553–561, November.

[20] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).

[21] Ritter, A., Clark, S., Mausam, Etzioni, O., 2011. Named entity recognition in tweets: An experimental study. In Proc. of Empirical Methods for Natural Language Processing (EMNLP), Edinburgh, UK

[22] Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., Senellart, P., Exploiting the Social and Semantic Web for guided Web Archiving, The International Conference on Theory and Practice of Digital Libraries 2012 (TPDL2012), Cyprus, September 2012.

[23] Maynard, D., Bontcheva, K., Rout, D., Challenges in developing opinion mining tools for social media. In Proceedings of @NLP can u tag #user*generated*content?! Workshop at LREC 2012, May 2012, Istanbul, Turkey.

[24] Bosma, W., Vossen, P., 2010. Bootstrapping languageneutral term extraction. In 7th Language Resources and Evaluation Conference (LREC), Valletta, Malta.

[25] Deane, P. A nonparametric method for extraction of candidate phrasal terms, In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005.

[26] Tahmasebi, N., Risse, T., Dietze, S. (2011) Towards Automatic Language Evolution Tracking: A Study on Word Sense Tracking, Joint Workshop on Knowledge Evolution and Ontology Dynamics 2011 (EvoDyn2011), at the 10th International Semantic Web Conference (ISWC2011), Bonn, Germany.

[27] K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The Hadoop distributed file system. In Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, 2010.

[28] C. Weiss, P. Karras, and A. Bernstein. Hexastore: sextuple indexing for semantic web data management. Proceedings of the VLDB Endowment, 1(1):1008–1019, 2008.