Automated Fact Repository Construction through Ontology Filtering and Natural Language Processing *

Akshay Java
University of Maryland
Baltimore County
1000 Hilltop Circle
Baltimore, MD, USA
aks1@umbc.edu

Jesse English
University of Maryland
Baltimore County
1000 Hilltop Circle
Baltimore, MD, USA
english1@umbc.edu

Sergei Nirenburg
University of Maryland
Baltimore County
1000 Hilltop Circle
Baltimore, MD, USA
sergei@cs.umbc.edu

ABSTRACT

In this paper, we present a novel approach to automatic fact repository construction from weblog data. The system we describe uses a combination of ontological filtering techniques and ontological-semantic natural language processing to collect a concise and thorough view of the data while processing only a fraction of the blog posts. Ontological filtering allows our system to weed through each weblog post and select only those relevant to the search topic on hand; the natural language processor extracts meaning representations from the selected texts. These meaning representations are aggregated to produce a collection of machine tractable data concerning the requested search.

1. INTRODUCTION

The blogosphere has become a new channel for communicating ideas, user reviews and thoughts. It reflects the daily world events and closely monitors the main stream media for new information. When new information arises there is very little delay before it is picked up by bloggers. Often there occurs a buzz around some events that triggers converstaions and debates on important political and social events.

Recently there has been interest in mining such information to generate reports of what the popular opinion is. These may be in the context of marketing a product or judging user sentiments about an event such as *'iraq war'* or *'presidential elections'*. Most of the available systems work by monitoring keyword occurences to find trends. While such systems provide a high level view of what is popular or what the current buzz is about, these are limited by their syntactic nature. In order to extract facts and relations about various entities in the blogosphere we need to be able to develop sophisticated Natural Language Processing tools that are capable of large-scale syntactic and semantic processing of text.

In this paper we present some of the preliminary results in adapting OntoSem, an ontologically based NLP system to process information from blogs and automatically construct a fact repository consisting of knowledge that was gleaned from the blogosphere. OntoSem goes beyond the capabilities of traditional Information Extraction (IE) tools, by not only identifying named entities but also exporting facts and relations between learned instances of ontological concepts.

Adapting OntoSem to automatically process text from blogs posed interesting callenges. Firstly, the language used in the blog domain tend to be quite informal and unstructured. There is also the presence of splogs, which are spam posts on the blogosphere, that lead to a lot of noise in the dataset. Secondly, due to the volume of information, it may be difficult to perform a complete and accurate analysis due, among other factors, to constraints on lexical coverage. OntoSem uses a graceful-degradation semantic analyzer that is capable of processing such unexpected inputs and tries to resolve the terms to the nearest ontological match by performing a constraint satisfaction search over the ontology. This allows us to be able to obtain significant knowledge even by processing a fraction of the blog posts.

In this paper we present preliminary results of using Natural Language Processing for analysis of blogs and automatic creation of fact repositories from significantly noisy datasets. The paper is organized as follows: Section 2 provides a basic overview of OntoSem NLP system. Section 3.1 describes the FACT system and show some of the results of automatic text processing in 4. Finally in 5 we describe the conclusions and work in progress.

2. RELATED WORK

Information extraction (IE) deals with identifying a set of known entity types from free text. Typical IE tools identify named entities such as person, location, organizations, dates and similar information. This was also one of the objectives of MUC [3]. Upcoming TREC Blog track [1] focuses on a related task of identifying events and extracting timelines from weblog datasets.

Recently there has been interest in automatically extracting opinions and sentiments from blog data [2]. Companies like Blogpulse [2] and Opinmind http://www.opinmind.com provide insightful details of what the latest opinions and buzz on the blogosphere is like.

SemNews [3] is an prototype application that demonstrates the feasibility of automatic language processing on news summaries from RSS sources. SemNews processes text using OntoSem and provides a semantic web representation of the meaning of the text [5, 4].

*This work is supported by

WWW2006, May 22–26, 2006, Edinburgh, UK.

[1] http://trec.nist.gov/call06.html
[2] http://www.blogpulse.com
[3] http://semnews.umbc.edu

In a recent study, we have found that spam is a major issue in the blog domain [7]. We use a Support Vector Machine based splog identifier [6] to filter out the splogs. This is an important preprocessing step since the spurious nature of the information present in splogs would reduce the quality of the automatically constructed fact repositories.

3. ONTOSEM

Ontological Semantics (OntoSem) is a theory of meaning in natural language text [9]. The OntoSem environment is a rich and extensive tool for extracting and representing meaning in a language independent way. The OntoSem system is used for a number of applications such as machine translation, question answering, information extraction and language generation. It is supported by a *constructed world model* [10] encoded as a rich ontology. The Ontology is represented as a directed acyclic graph using IS-A relations. It contains about 8000 concepts that have on an average 16 properties per concept. At the topmost level the concepts are: OBJECT, EVENT and PROPERTY.

The OntoSem ontology is expressed in a frame-based representation and each of the frames corresponds to a concept. The concepts are named collections of property-value pairs. Properties include attributes and relations. Values are, in fact, graduated through a variety of specialized facets which aid in establishing preferences during semantic analysis. It is important to mention that concepts in this ontology are connected not only through subsumption links but through any of the ontological relations. The set of OntoSem properties, the metalanguage of the ontology, at present consists of about 350 elements.

The ontology is also supported by an *Onomasticon* [10], which is a lexicon of proper names. A more detailed description of OntoSem and its underlying theory, ontological semantics, is available in [10] and [1].

OntoSem text analyzer carries out a syntactic, semantic, and pragmatic analysis of the text, producing text meaning representations or TMRs. In addition to providing information about the lexical-semantic dependencies in the text, the TMR represents stylistic factors, discourse relations, speaker attitudes, and other pragmatic factors present in the input; in other words, the TMR captures both propositional and non-propositional components of textual meaning. OntoSems TMRs are represented in a custom frame-based representation language and can be exported into and imported from an OWL/RDF-oriented formalism.

The OntoSem environment takes as input unrestricted text and performs a variety of morphological, syntactic and semantic processing steps to convert it into a set of Text Meaning Representations (TMR). The basic steps in processing the sentence to extract the meaning representation are illustrated in figure 1. The preprocessor deals with identifying sentence and word boundaries, part of speech tagging, recognition of named entities and dates, etc. The syntactic analysis phase identifies the various clause level dependencies and grammatical constructs of the sentence. The TMRs are produced as a result of semantic analysis which uses knowledge sources such as lexicon, onomasticon and fact repository to resolve ambiguities and time references. Those of the concept instances derived from the text and encoded in TMRs that are relevant to a particular application are stored in the *fact repository* for that application. Thus, the fact repository essentially forms the knowledge

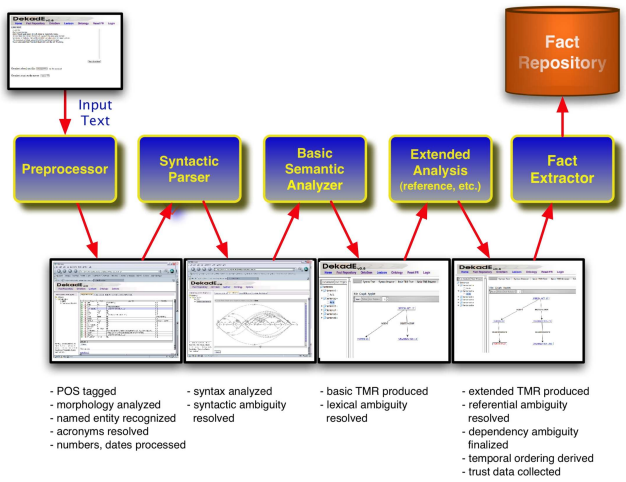

Figure 1: OntoSem goes through several stages in converting a sentence into a text meaning representation (TMR). The text is preprocessed for end of sentence tagging and part of speech analysis. Then the syntactic structure of the sentence is found. Following this, the semantics of the sentence are extracted. Any reference in the sentence is resolved in the fourth stage. Finally, the knowledge is extracted from the analysis and passed to the fact repository.

base of assertions in OntoSem. Note that the fact repository both helps to generate TMRs and is used as a storage of knowledge for a variety of other applications, including the one described in this paper.

3.1 Weblog Dataset

The dataset released by Intelliseek/Blogpulse [4] for this workshop consists posts from about roughly 1.5 million unique blogs. The data spans from about 20 days during the time period in which there were terrorist attacks in London. This time frame witnessed a significant activity in the blogosphere with a number of posts pertaining to this subject. We indexed about 8 million posts using Lucene, an open source search engine.

4. FACT

The FACT (Facts And Concepts from Text) system has been developed to encapsulate OntoSem in a complete throughput application of automating data acquisition. FACT facilitates flexible data set selection and automated fact repository construction. At this stage, FACT is essentially a batch processing shell surrounding the OntoSem analyzer. Using FACT, a user can select a dataset that has been catalogued and indexed, a concept from OntoSem's ontology, and a fact repository to work from. FACT will then extract all related texts in the dataset, send them to OntoSem for processing and store the resulting text meaning representations (TMRs) in the requested fact repository (see Figure 2).

Populating a FACT dataset involves splitting the data into sizes manageable to the native operating system, and then keyword indexing the entire data. Once this one-off process has been done, the user can select that dataset for

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

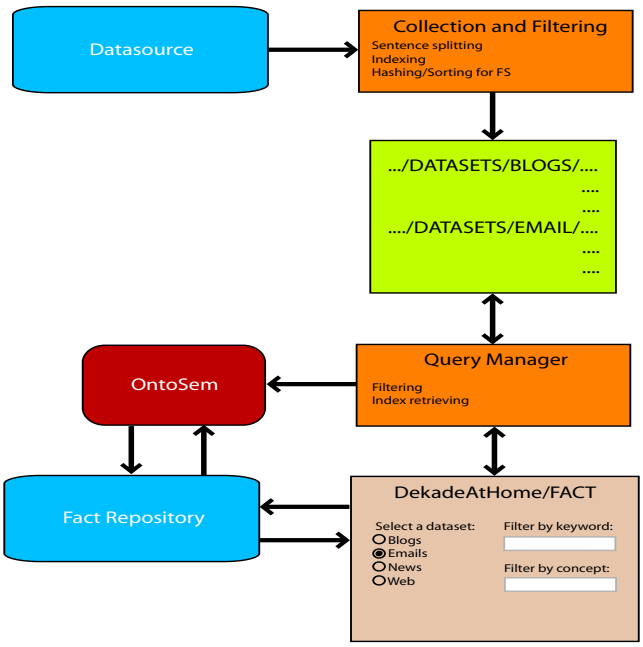

Figure 2: Various stages of processing in the FACT system are shown. The data is collected and filtered, and deposited into the local file system. The user interface can then formulate a query, which is gathers data, passes it to OntoSem, and generates the fact repository. This acquired data can then be viewed and edited directly through the interface.

processing in any number of ways. The key to efficient data extraction is not to naively process the entire dataset, generating enormous amounts of data, but rather to use the keyword filtering enhanced by ontological knowledge to cut the text set down to a more manageable size. A user can select an ontological concept from OntoSem's ontology; FACT will take this concept and extract all lexical mappings to it (the lexical mappings are obtained from OntoSem's semantic lexicon; currently, the English lexicon in OntoSem contains about 30,000 static entries; the semantics of the entries is explained in terms of the underlying ontology). As a result, an ontologically motivated set of keywords is produced that forms the search space over the indexed data. Having extracted the texts that are lexically matched to the users ontological query, FACT then splits the texts into sentences, and optionally disregards sentences not directly referring to the requested ontological concept. This is optional as it is done primarily for efficiency, but given the nature and complexity of entity reference resolution, it may be desirable to analyze an entire text, rather than individual sentences from the text. The desired sentences or texts are then passed to OntoSem, along with the user's requested fact repository to be queried and populated. The end result is the detailed analysis of the sentence in the form of a TMR, and the addition of facts to the fact repository. This data is then easily accessible, sortable, and searchable through the DekadeHome system (also developed at ILIT for use with OntoSem)

5. EXPERIMENTS

To evaluate the usefulness of the FACT system on the weblogs database, we selected a few ontological concepts that related to the presumed subject material of the data (based on knowledge of the world events at the time of the postings). To that end we selected the ontological concepts BOMB and TERRORIST, as well as the keyword "London". FACT then proceeded to search OntoSem's lexicon to derive all mapped entries to the selected concepts. This gave us three independent search spaces:

- BOMB
- grenade
    1. hand-grenade
    2. letter-bomb
    3. package-bomb
- TERRORIST
    1. terrorist
- LONDON

London is already a text string (it is an instance of the ontological concept CITY). So the string was simply passed in to the filter text query. For each of these spaces, we selected the first 100 matches from the indexed texts, and extracted any sentences from each text that keyword matched a value in the space. Each sentence was then individually analyzed by OntoSem, and the results were stored in an easily filterable fact repository. The results of the experiment showed a vast increase in gained structured knowledge in the form of fact repository entries, from the analyzed texts. We were able to sort out each instance of an ontological concept, and filter them by types: HUMAN, PLACE, EVENT, etc. We could easily see how often, for example, the city of London was mentioned in the text. But further, we could at a glance see all of the properties associated with the instance which were completely automatically acquired.

In Figure 3, we've constructed a small view into some of the automatically acquired fact repository entries, concerning a few instances of the concept PLACE that map to the real-world term "London". In the figure, it can be seen how the analyzer derived a variety of BOMB events, some with associated dates. The detailed relations between the HUMAN with the name "White", and the interaction concerning an ASSAULT ("attack"), which took place at an office located in London are also shown.

OntoSem produces all of this data automatically, but it could also be constructed by hand as proof of the expressive power of the meta language used in the TMRs. As the quality of OntoSem is constantly improving, the difference in automatically constructed and manually constructed TMRs is reduced. Shown below is the result of a interactively enhanced, or "gold-standard" [8], TMR, from the sentence *"A European terrorist organization launched an attack on London that has killed at least 37 and wounded at least 700."*

```
(ASSAULT-22
 (AGENT (VALUE ORGANIZATION-23))
 (THEME (VALUE CITY-25))
 (EFFECT (VALUE KILL-26 INJURY-27))
 (INSTANCE-OF (VALUE ASSAULT)))

(ORGANIZATION-23
```

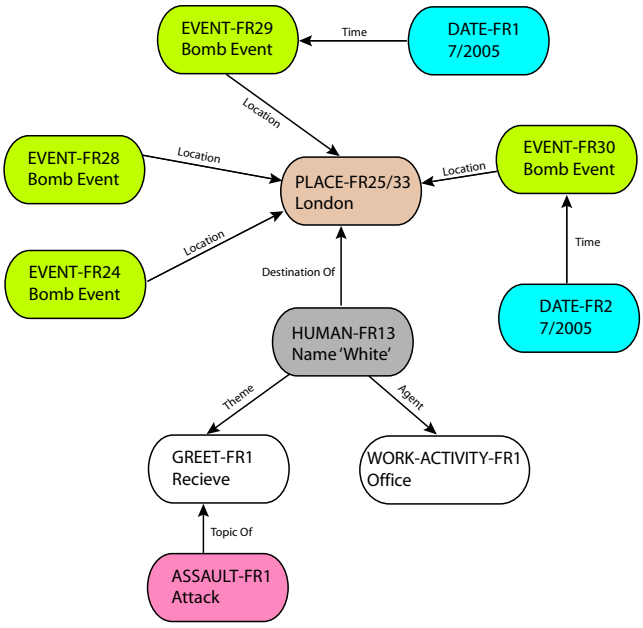

Figure 3: Partial view of the automatically acquired fact repository about London. This view covers some of the data acquired from multiple analyzed sentences, and shows four bomb events focused in London, two of which were flagged with date information. In addition, a human named "White" can be seen traveling to an office in London, where a conversation concerning the assaults occurred.

```
 (LOCATION (VALUE EUROPE))
 (RELATION (VALUE TERRORIST-24))
 (AGENT-OF (ASSAULT-22))
 (INSTANCE-OF (VALUE ORGANIZATION)))

(TERRORIST-24
 (INSTANCE-OF (VALUE TERRORIST)))

(CITY-25
 (HAS-NAME (VALUE LONDON))
 (THEME-OF (VALUE ASSAULT-22))
 (INSTANCE-OF (VALUE CITY)))

(KILL-26
 (CAUSED-BY (VALUE ASSAULT-22))
 (THEME (VALUE SET-28))
 (INSTANCE-OF (VALUE KILL)))

(INJURY-27
 (CAUSED-BY (VALUE ASSAULT-22))
 (EXPERIENCER (VALUE SET-29))
 (INSTANCE-OF (VALUE INJURY)))

(SET-28
 (MEMBER-TYPE (VALUE HUMAN))
 (CARDINALITY (>=  37))
 (THEME-OF (VALUE KILL-26))
 (INSTANCE-OF (VALUE SET)))

(SET-29
 (MEMBER-TYPE (VALUE HUMAN))
 (CARDINALITY (>= 700))
 (EXPERIENCER-OF (VALUE INJURY-27))
 (INSTANCE-OF (VALUE SET)))
```

6. CONCLUSIONS AND FUTURE WORK

Blogs have a wealth of information in them and extracting stuctured knowledge from them requires robust and scalable Natural Language Processing tools. Unlike web pages and mainstream media articles or newswires, the language in blogs can be much more informal and spontaneous. There is also the problem of splogs and presence of noisy data on the blogosphere. Our experience with automatic processing of weblogs using OntoSem has shown that using a combination of ontological filtering and with the aid of a robust semantic analyzerm, it is possible to extract useful information and build large fact repositories.

A number of directions of future work suggest themselves. First and foremost, they relate to improvements in OntoSem – work is under way on enhancing the entity reference resolution module, improving the capabilities of semantic disambiguation as well as treatment of ungrammatical inputs. Another line of improvement is related to introducing relevance filters for inclusion of TMRs into the fact repository.

7. ACKNOWLEDGEMENTS

We would like to thank Dr. Stephen Beale for his help with gold standard TMR construction and Dr. Timothy Finin for advice and guidance with this paper.

8. REFERENCES

[1] Institute for language and information technologies. `http://ilit.umbc.edu/`.

[2] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD*, pages 419–428, 2005.

[3] R. Grishman and B. Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 466–471, 1996.

[4] A. Java, T. Finin, and S. Nirenburg. Integrating language understanding agents into the semantic web. In T. Payne and V. Tamma, editors, *Proceedings of the AAAI Fall Symposium on Agents and the Semantic Web*, November 2005.

[5] A. Java, T. Finin, and S. Nirenburg. Text understanding agents and the Semantic Web. In *Proceedings of the 39th Hawaii International Conference on System Sciences*, Kauai HI, January 2006.

[6] P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006. to appear.

[7] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. 2006. submitted to AAAI 2006 - AI on the Web.

[8] S. Nirenburg, S. Beale, and M. McShane. Evaluating the performance of the ontosem semantic analyzer. In *Proceedings of the ACL Workshop on Text Meaning Representation*, 2004.

[9] S. Nirenburg and V. Raskin. Ontological semantics, formal ontology, and ambiguity. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 151–161, New York, NY, USA, 2001. ACM Press.

[10] S. Nirenburg and V. Raskin. *Ontological semantics*. MIT Press, 2005.