Iterative Learning of Relation Patterns for Market Analysis with UIMA

Sebastian Blohm, Jürgen Umbrich, Philipp Cimiano, York Sure
Institute AIFB, Knowledge Management Research Group
University of Karlsruhe
D-76128 Karlsruhe, Germany
{blohm,juub,cimiano,sure}@aifb.uni-karlsruhe.de

Introduction

Monitoring news as well as corporate and community web sites is a important task in market analysis. Yet, it requires much manual processing and is inherently incomplete as only a small fraction of the data produced on the Web can be processed by human analysts. We present here our ongoing efforts to employ UIMA annotations and search on UIMA-annotated documents for supporting market analysis. Information is extracted based on relation patterns that serve as queries. The results are then visualized to allow detailed market analysis. The current status of the reported work is that we are already implemented most of the required infrastructure (UIMA, Pronto, Omnifind) and started some of the outlined experiments.

The **Unstructured Information Management Architecture (UIMA)** is a framework that enables a component-based analysis of unstructured documents(Ferrucci and Lally, 2004) providing a typed data structure for the derived structured information as well as means for defining and controlling process flows.

We have developed **Pronto**, a system for automatic learning of text patterns for extracting instances of semantic relations(Blohm and Cimiano, 2006). The system essentially uses the Web as a corpus accessing it via a Web search engine. Patterns are learned by abstraction over the textual occurrences of relation instances in the corpus. In the experiments we employ Pronto to produce UIMA annotations and additionally feed the UIMA annotation structure into the learning process to be able to learn structural patterns rather than plain text patterns. They will enable our system to annotated mentions of salient relations for market analysis. The output will allow retrieving facts about market events using structured queries.

```
<sentence>
  In <year>2005</year>
  <presented subj='Fiat' object='Punto'>
     Fiat presented the new Punto
  </presented>
  at the <exhibition>IAA</exhibition>
  in Frankfurt.
</sentence>
```

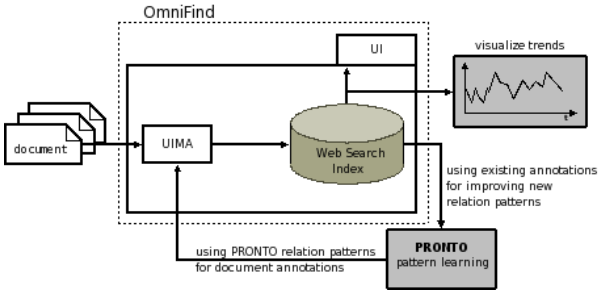

Figure 1: Architecture overview.

The Pronto System

The Pronto system allows automatic generation of appropriate text patterns serving as queries for Web search engines. Pronto uses an iterative pattern induction approach similar to that applied in the DIPRE (Brin, 1998) and Snowball (Agichtein and Gravano, 2000) systems. Pronto maintains a set of patterns for which confidence scores are computed using statistical models, frequency of co-occurrence or precision estimates based on previous extractions. The set is updated in each iteration which allows for flexibility and broader coverage. For each extracted relation, a confidence score is computed based on the number and quality of the patterns that retrieved that instance. Pronto takes as input a small number (e.g. 10) of example instances of a given relation and is able to produce around 1000 new instance per hour when querying Google as a remote search index. It has been applied to non-taxonomic relations from different domains and with different characteristics reaching between 45% and 85% precision.

Using index-based-search for information extraction has the advantages that large amounts of text become manageable and that new patterns can be introduced and tried-out without having to re-process the entire corpus.

System Architecture

The core of the architecture for our experiments consists of the integration of our web search based information extrac-

tion system Pronto with the UIMA analysis process as well as a search engine that allows retrieval on UIMA annotation structures. On top, **IBM OmniFind** will be used as a search engine. Prior to indexing, OmniFind runs a UIMA processing chain in which we intend to plug annotators that are based on relevant background knowledge and on annotation patterns produced by Pronto. The index is then used in three ways:

- To visualize the development of the market (e.g. by plotting extraction result counts for salient patterns over time)
- As a semantic search tool to further explore market developments.
- As interface to the Web corpus for the Pronto algorithm to improve its patterns. In particular by taking other annotations (both from the background knowledge and from the patterns previously produced by Pronto).

The UIMA API serves as the interface between the document crawler, the annotators and the indexer. The study will largely benefit from the Common Analysis Structure (CAS) which allows interaction between third party annotators we will use (for linguistic analysis and background knowledge) and the annotator that exploits the Pronto patterns. OmniFind's Search and Indexing API (SIAPI) will be employed as the interface to Pronto's extraction component as well as to the semantic search and trend visualization component. We will make use of OmniFind's the XML Fragments query language to query for text and UIMA annotations within a single search engine request.

Experiments

In this section we first describe the data sets on which we intend do do our experiments and then describe the course of experiments in further detail.

Datasets

We are using two fragments of the Web for our experiments.

- *Wikipedia:* which shows many characteristics of large corporate intranets: It shows higher quality, more structure and less spam as the Web average and structural information on the content is available (e.g. categories).
- *Automotive web sites:* a crawl from a number of salient web sites providing news about the automotive industry together with expert knowledge on a relevant semantic relations from market research.

Course of Experiments

Our experiments are executed in three phases of increasing integration of UIMA extractions into the pattern induction process. We are currently in the first phase.

In an initial, naive phase, extraction patterns previously derived on the web will be matched on documents prior to indexing to produce relation annotations. Patterns would look as follows:

```
* released the new * at
```

When integrated with the search API of OmniFind, patterns can be induced that use UIMA annotations to focus the search. Note that Pronto will be able to determine the relevant tags automatically by considering occurrences of seed instances and observing the UIMA annotations that are present. The relevant kinds of background knowledge will thus be determined automatically.

```
<carMaker/> released the new <carModel/> at
```

Finally, our aim is to increase Pronto's learning capabilities to learning structural patterns such as:

```
<article category='news'>
  <sentence>
   <carMaker/> released the new <carModel/> at
  </sentence>
</article>
```

Market visualization will then be possible by taking a fixed set of queries like:

```
<presented subj='Fiat' object='*'/>
<year>2007</year>
```

To monitor news on Fiat car releases. A substantial change in search result counts may indicate an interesting market event.

Conclusion

We presented the key components of our architecture (UIMA, Pronto, and Omnifind). Our aim is to integrate structural search on UIMA annotations into our information extraction system for market analysis. We expect two major benefits of UIMA to extraction quality, i.e. (i) increased precision of search results due to patterns that are sensitive to background knowledge annotations (e.g. noun categories) and (ii) increased coverage of patterns as text surface patterns which require high redundancy are replaced by more abstract structural patterns.

Acknowledgements

We are very grateful for the IBM UIMA Innovation Award 2006. This work was partially funded by the X-Media project (www.x-media-project.org) sponsored by the EC as part of the IST program under EC grant number IST-FP6-026978. The OmniFind search engine software is kindly provided by IBM through their Academic Initiative Program.

References

[Agichtein and Gravano, 2000] Agichtein, E. and Gravano, L. (2000). Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries (DL)*, pages 85–94.

[Blohm and Cimiano, 2006] Blohm, S. and Cimiano, P. (2006). Learning patterns from the web – evaluating the evaluation functions. In *OTT'06 - Ontologies in Text Technology: Approaches to Extract Semantic KnowledInformation*.

[Brin, 1998] Brin, S. (1998). Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*.

[Ferrucci and Lally, 2004] Ferrucci, D. and Lally, A. (2004). UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*.