Semiautomatic Creation of Semantic Networks

Lars Bröcker

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Lars.Broecker@iais.fraunhofer.de

1 Introduction

The vision of the Semantic Web ist one of extending the World Wide Web of today to one "[..] in which information is given well-defined meaning, better enabling computers and people to work in cooperation." (Tim Berners-Lee in an article for the Scientific American in 2001). This promises an exciting future for the WWW.

The advantages for users and machines alike are eminent, many of the building stones like RDF or OWL are in place already. But why has the Semantic Web not been adopted by more content creators, more web sites? The main technological reason for this lies in the complexity associated with the creation of ontologies. Ontologies are, following a definition of T. R. Gruber, a formal, explicit specification of a shared conceptualization of a given domain[1]. As such, they are an essential part of every semantic web application, since they define the language used to express the view on the world. But their creation is a time-consuming and expensive endeavor that is beyond many organizations or communities. Most therefore stay away from the Semantic Web altogether. This severely handicaps the efforts of bringing about the vision of the Semantic Web, by preventing the attainment of a critical mass of content available using it.

1.1 Research Problem

What is needed is a means to generate a meaningful description of the semantics of content collections in such a way that it necessitates as little manual interaction as possible. The results may not be as distinguished as a manually created ontology, but they at least provide a way to utilize the benefits of the semantic web. Two main problems need to be tackled: first, the extraction of the semantic network inherent in the collection, and second, the design of a surrounding system being both versatile and easy to expand to accommodate new features, data stores, or services.

The first problem is one of automating ontology engineering. The goal here is to extract the main entities and their relations from the corpus in order to gain an understanding of the topics the corpus contains. This boils down to three tasks: entity recognition, relation discovery, and creation of the semantic net from the results of the first two tasks. While there are good tools available for entity recognition, relation discovery as of now has to do without. Scientific approaches in this area typically consider binary relationships, higher order relations get almost no coverage. The task of network-creation is a translation step from the entities and their relations into a language of the Semantic Web framework.

The second problem addresses use-case necessities. Many interesting collections are not static, but are subject to many changes (e.g. wiki-webs). In order to accommodate this, a semantic representation needs to be able to continuously monitor the corpus and adapt itself accordingly. Other requirements may result in the necessity for integration of additional services into the system.

1.2 Contribution

This thesis concentrates on the task of relation discovery in order to generate meaningful connections for the network, since there already are numerous good tools for Named Entity Recognition (e.g. GATE[2]) available. Accordingly, the first contribution is an algorithm that gathers n-ary relations ($n \geq 2$) in a text corpus between entities from a set of previously agreed upon concept classes. The second contribution is an architecture containing the algorithm, as well as facilities for the monitoring of dynamic collections, paired with adaptation of the network where necessary.

1.3 Use Cases

The envisioned system provides a semantic representation of the content of a document repository without changing its data, i.e. it provides a semantic wrapper around the collection. The wrapper supplies a semantic view on the topics of the collection that can be used for further processing, data exchange, or provision of sophisticated search interfaces.

The first application of the approach is part of an ongoing research project financed by the German Ministry of Research and Education (BMBF) called WIKINGER[3], where it is used to bootstrap and subsequently monitor a wiki-web for the domain of Contemporary History.

In a similar manner, media providers like broadcasters, newspapers, or news agencies could use this approach to better organize and tap the contents of their digital archives.

2 Approach

For the sake of brevity, only the approach concerning the creation of the semantic network will be described in detail. It is a process divided into five separate steps. In the first step, a set of core concept classes is defined, followed by the annotation of examples of these classes. They are used to train a Named Entity Recognition tool. Next, the corpus is segmented into sentences. Those containing less than two entities are discarded. The remaining sentences serve as input for an algorithm computing association rules on the entity classes. The association rules express the degree of association between classes using two measures: the confidence that there is an association, and its coverage of the collection. This allows different ranking approaches depending on the strategy to be followed.

Given an ordering of the rules, the next step iteratively analyses the set of sentences belonging to a given rule. Since one rule describes an unknown amount of different relations between its constituents, the task is to find a clustering of the set such that each cluster describes one single relation. Since the amount of relations is not known beforehand, hierarchical clustering has to be employed.

The next step provides labels for the relation clusters. They are presented to the domain experts for review, who can change or remove labels, entities, or relation clusters.

The final step collects all entities and relations and creates the semantic web from them. While the entity translation is straightforward, special care has to be taken in expressing the relations between them, since not all relations will be binary. Preservation of the n-ary relations requires the introduction of proxy entities into the net, in order to conform to the triple schema of RDF.

2.1 Results so far

System architecture using a service-oriented architecture.

Internal data representations allows inclusion of external data sources given a suitable transformer.

Versioned repositories for the internal data, allow change montoring, detection, and adaptation.

2.2 Results still to be achieved

Clustering and labeling different distance measures and vector representations are evaluated.

Translation into RDFS algorithm needs to be designed and implemented

Change Management the service responsible needs to be implemented.

2.3 Evaluation

Evaluation of the approach will be performed in the project WIKINGER. Domain experts will be on site to handcraft relation clusters. These will serve as ground truth for the automatically proposed relation clusters. Quality in a dynamic environment will be evaluated via periodical surveys when the system goes live in August of this year. In parallel, a similar setup for the domain of newspaper archives will be tested with the help of archive personnel from a newspaper company.

3 State of the Art

The approach presented of the thesis touches two areas of research: ontology learning and relation finding. This section highlights the approaches most relevant for this work.

3.1 Ontology Learning

Alexander Maedche from the AIFB in Karlsruhe describes a system called *Text-To-Onto*[4] that is used to aid ontology engineers in their work. Its objective is to find new concepts for the target ontology from domain taxonomies providing is-a relations, and hyponym relations gathered from texts using text mining methods. The candidate concepts are added manually to the ontology. An additional module deals with the discovery of non-taxonomic relations. It deducts possible relations from association rules. The module stops at this step and only considers concept-pairs.

Philipp Cimiano and Johanna Völker, also from the AIFB, present with *Text-2-Onto* an advanced system for the task of ontology learning from text. It holds the ontology in a so-called probabilistic ontology model (POM) that contains modelling primitives along with a confidence measure stating the probability of them being part of the ontology. A GUI allows manual changes to the ontology after the learned phase. The system reacts on changes in the corpus by only recalculating the parts of the ontology that are affected by the changes. Named entity recognition using GATE is performed on the collection, but only hyponym relations (kind-of) are extracted automatically from the texts.

3.2 Relation Learning

Takaaki Hasegawa et al. describe an algorithm for the discovery of relations in natural language texts[6], using named entity recognition with a small set of concept classes. This is followed by a per-sentence analysis of the corpus. All sentences containing two instances having a maximum distance of five words are considered for further processing. Finally, a cluster-analysis is performed on every class of pairs, resulting in clusters containing the different types of relation between pairs. Evaluation is done using a years worth of newspaper articles, and matching automatic performance against hand-picked relations. The best results (34 of 38 existing relations found) attain an F-measure of 60%.

Aron Culotta and Jeffrey Sorensen present an approach to relation extraction from texts using kernel methods [7]. The task is to extract previously learned binary relations from the corpus. This is achieved by first performing shallow parsing of a sentence and then using a kernel method on the smallest dependency tree containing both entities. This reduces the amount of words considered in the calculation of the kernel, thus reducing the amount of noise in the result. They reach 70% precision with 26% recall. Bunescu et al.[8] propose a variation of this approach: their kernels consider only the words on the shortest path between the two entities. Their evaluation is performed on the same data where they reach 71% precision with 39% recall.

3.3 Discussion

Text-To-Onto was developed as a tool for knowledge engineers, who are supposed to do the real modelling, and it shows. All additions to the ontology are performed manually, and while it contains a module for relation learning using association rules, it refrains from discovering the actual relations. Text-2-Onto uses an interesting storage model for the ontology, but is restricted to hyponym relations, thereby falling behind its predecessor with regard to relation discovery. The system described in this paper goes a step beyond these systems in two ways: it does not depend on the availability of ontology engineers, and it aims to discover all relevant relations contained in the text.

Hasegawa et al. use a clustering approach to find hitherto unknown relations but restrict themselves to pairs of entities, thus tearing apart relations of higher order that might have been present in the data. Their algorithm does not include a means to rank the pairs of prior to the clustering. The approaches by Culotta and Bunescu offer interesting possibilities for subsequent classification of relations, but cannot be used to discover them in the first place.

4 Conclusion

This paper summarizes the main topics of my PhD thesis. The approach promises to be a feasible way to bring the benefits of the Semantic Web to a larger audience, especially in those domains where creation of a specialized ontology is not feasible in the foreseeable future. The architecture has been designed such that it lends itself well for expansion in different ways. Inclusion of video or audio transcripts is an interesting option, since more and more such content finds its way onto the web. The inclusion of an easy interface allowing for the definition of new relations is another interesting expansion of the system, perhaps by graphical means using SVG or by an extended wiki-syntax as found in semantic wiki systems.

References

1. Gruber, T.R.: A translation approach to portable ontology specifications. In *Knowledge Acquisition*(5), 1993, pp. 199–220
2. Cunningham, H.:GATE, a General Architecture for Text Engineering. In *Computers and the Humanities*, vol 36, 2002, pp223 – 254
3. Bröcker, L.: WIKINGER – Semantically enhanced Knowledge Repositories for Scientific Communities. In: *ERCIM-News*, vol. 66, 2006, pp. 50–51.
4. Maedche, A.: The Text-To-Onto Environment. Chapter 7 in: *Maedche, A.: Ontology Learning for the Semantic Web.* Kluwer Academic Publishers, 2002.
5. Cimiano, P., Völker, J.: Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In *Proceedings of NLDB*, 2005.
6. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: *Proceedings of the 42nd Conf. of the ACL*, 2004. pp. 15–42
7. Culotta, A., Sorensen, J.: Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42nd Conf. of the ACL*, 2004. pp. 423–429.
8. Bunescu, R.C., Mooney, R.J.: A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of EMNLP 2005*, pp 724–731