Information Extraction as a Semantic Web Technology: Requirements and Promises

Mark Stevenson and Fabio Ciravegna
Department of Computer Science,
211 Regent Court, Portobello Street,
Sheffield S1 4DP
United Kingdom
{marks,fabio}@dcs.shef.ac.uk

Abstract

The Semantic Web will require services to annotate web pages with the necessary meta-data. Information Extraction (IE) may be a suitable technology for this purpose. This paper discuses the requirements generated for applications of IE to the Semantic Web and the degree to which current technology meets them.

1 Introduction

The Semantic Web (Berners-Lee et al., 2001) is an ongoing initiative to make the World Wide Web a more useful resource by standardizing the descriptions of available information and services. It is expected that this will be achieved by adding various forms of meta-data to web pages and, while this may be feasible for newly created pages, it is unlikely to be carried out for existing web content. The only reasonable approach is to annotate these pages automatically. Information Extraction (IE) may be a suitable technology for automating this process (Ciravegna, 2003).

Research in IE has been largely driven by the Message Understanding Conferences (MUC) (MUC, 1991 1993 1995 1997 1998). These exercises focused on identifying information from free text such as newswire stories. The participants were required to identify every item of a specific semantic type in the text, a process known as Named Entity (NE) recognition. For example in the sixth MUC the semantic types included `PERSON`, `ORGANIZATION` and `LOCATION`. Participants were also required to identify specific relations between these entities and combine them into templates where appropriate. The majority of IE systems carried out the stages of NE recognition and relation extraction as separate processes.

It is possible to define some basic requirements for an IE system to be useful for the Semantic Web; these pertain to issues portability, both to domain and extraction task. It is expected that the Semantic Web will be based on many small ontological components (Hendler, 2001) rather than large, complex ontologies like CYC (Lenat, 1995). These components will be continuously extended, merged or created, therefore the annotation services associated with them will have to be constantly adjusted or revised according to these changes. This poses a number of obvious constraints and requirements on the technology to support annotation in terms of usability, portability and maintainability that we will list in the next sections. If IE technology is used to support annotation then new applications will be required whenever a new ontological component is created. Machine learning offer techniques to adapt IE systems to new domains and extraction tasks and for this reason this paper focuses on these approaches.

Two types of resources must be dealt with in order to port an IE system: linguistic resources, such as tokenizers, part of speech taggers and parsers, and semantic resources like gazetteers and ontologies. Creators of Semantic Web ontologies cannot be expected to have expert knowledge of IE and so any annotation tools must be portable by a person with limited IE skills. This problem has already been discussed by Ciravegna (2001). Consequently, methodologies for learning linguistic rules will be needed. In principle the necessary semantic resources could be provided by the ontology. Unfortunately, the kind of ontology defined will satisfy the needs of the customer service, not of the IE component. This means that it will not be linguistically oriented and there is the possibility that it will not be useful for IE purposes. For example some distinction between two concepts could require deep reasoning on background knowledge which may be beyond the IE capabilities. Moreover some relations could be included that from a linguistic point of view require intermediate representation and reasoning (e.g. metonymic reasoning). This kind of information, if not provided to the IE system, could make the IE task quite complex if not infeasible. Some intermediate level of linguistically oriented ontology definition will be needed. Methodologies derived from the field of ontology learning (e.g. (Maedche and Staab, 2000), (Brewster et al., 2001)) could help in suggesting appropriate representations to non-linguistically-aware users.

The remainder of this paper discuses the suitability of current IE technology for requirements generated by the Semantic Web. Named entity identification and relation extraction technologies are discussed in Sections 2 and 3 respectively. Conclusions are presented in Section 4.

2 Named Entity Identification

Machine Learning techniques have proved to be very popular for named entity identification. Unfortunately, many systems require large amounts of training data to be ported to a new NE tasks. For example, BBN's SIFT system requires the annotation of a training corpus of 790,000 words in order to obtain 90% F-measure on the MUC7 task (Miller et al., 1998). Approaches such as Borthwick et al. (1998) and Mikheev et al. (1999) reduce the burden on the application developer by generalizing from the annotated text, seed rules or example names provided by the user. Riloff (1993) developed a system, AutoSlog, which learned from annotated text to generate semantic lexicons which could be used to identify NEs. Riloff and Shoen (1995) eliminated the need for annotations or rules in an extension of AutoSlog which only required the user to classify texts as relevant or irrelevant for the extraction task. Collins and Singer (1999) reduced the effort required further by using a bootstrapping algorithm which learned from just seven seed rules.

However, each of these approaches is very limited in terms of number of NE types and have often been restricted to those used in the MUC evaluations. These are generic and domain independent tags since the ontology used was both restricted and flat. It is expected that ontologies used in the Semantic Web will be significantly more complex, containing dozens of domain-specific concepts, instead of the seven used in the later MUC evaluations. These domain specific concepts will occur with less frequency than MUC-style ones which leads to the problem of data sparseness, making supervised learning approaches less feasible. Consequently this technology may not meet the Semantic Web requirements.

Some other approaches, specifically designed for use on the Web, use the regularity of the Web to learn entities contained in web pages. Brin (1998) uses a handful of user-defined examples to bootstrap learning for a task on finding book titles and authors. Ciravegna et al. (2003) employ multiple strategies to bootstrap learning on consistent repositories (e.g. Web sites). These approaches are promising since they remove much of the burden of manual annotation while still delivering good annotation services.

3 Relation Extraction

Simple recognition of entities is unlikely to generate complex enough meta-data for the Semantic Web and so identification of relations is considered to be important (Handschuh et al., 2002). From the IE point of view, relation extraction is a complex task which has not been studied in as much depth as NE recognition. The majority of MUC systems approached the relation extraction task using knowledge engineering approaches which relied on (para)linguistic rules manually created by an expert. The effort required to port these systems to a new domain or extraction task was often considerable, for example, the University of Massachusetts entered a system for the third MUC which required around 1,500 person-hours of expert labour to adapt the system for that extraction task (Lehnert et al., 1992). This overhead makes the knowledge engineering approach infeasible for the Semantic Web. A few systems have tackled this limitation using ML techniques.

WHISK (Soderland, 1999) is a system which learns extraction patterns directly from shallow parsed or unannotated text. WHISK assumes that the text has already been marked with named entities. Extraction patterns match directly to identify those which are part of a particular relation. The patterns can be applied to text which is either unannotated or partially parsed. This flexibility allowed the system to be applied to a wide range of text types including formal text and web pages. Evaluation was carried out using a simplified version of the management succession task used in the sixth MUC. WHISK was required only to identify relations which were described within a single sentence rather than across texts. The algorithm achieved an F-measure of 55.5. Chieu and Ng (2002) recast the relation extraction problem as a classification task. A score is computed for each pair of entities which occur in the same sentence to determine whether or not they represent a true relation. A maximum entropy learning algorithm was used to determine the score and pairs combined to form a template. Chieu and Ng reported an F-measure of 59.2 using Soderland's evaluation scheme. Yangarber et al. (2000) presented an unsupervised approach to relation learning. Text was pre-processed by examining the output from a parser to identify subject-verb-object tuples, for example `person-resigned-company` and `company-fired-person`. The user provides a set of seed patterns which are then generalized by substituting some elements with wildcards. Each of the patterns which occur in the corpus and match one of the generalized patterns are then evaluated and one chosen to be added to the pattern set. This approach was evaluated in terms of document relevance which makes comparison with other approaches difficult.

A crucial elements in applying ML to relation extraction is to find a way of generalizing patterns in a satisfactory way. Soderland (1999) and Yangarber et al. (2000) generalize patterns by removing restrictions on some elements of their patterns and then searching the corpus for instances which match the relaxed pattern. However, WordNet (Fellbaum, 1998) has been used to generalize patterns in a linguistically principled way (for example (Català et al., 2000)). One of the useful sources of information in WordNet is the lists of synonyms for terms and these could be used to generalize patterns. A potential problem is the fact that WordNet contains several senses for many of the words which might be of interest to an IE system. For example, there are nine senses for the verb "fire" and only one contains synonyms useful for generalizing patterns for the management succession task (e.g. "dismiss", "sack", "terminate"). Català et al. (2000) avoided this problem by requiring the user to identify the correct entry in WordNet when defining the extraction templates to be filled but this is a burden on the user. Chai and Biermann (1999) used word sense disambiguation to identify the correct WordNet entry. Performance of their IE system improved from an F-measure of 61.7 to 69.2 when disambiguation was used to guide the generalization process.

3.1 Suitability of technology

All the systems mentioned above require preliminary parsing. A parser processes the input so to produce formats more suitable for learning than unrestricted text and may resolve some of the ambiguity in language prior to learning. For example, the Connexor parser used by (Yangarber et al., 2000) analyses active and passive sentences to identify the semantic subject and object. So, for example, "The board fired Jones." and "Jones was fired by the board." produce the same triple (`board-fire-jones`).

Documents to be annotated for the Semantic Web will be of different types, from free texts (largely parsable) to very structured ones, were the information is largely carried by extralinguistic clues (e.g. HTML tags) and therefore largely unparsable. We have also noted how very often mixed types of documents can be found where some parts are highly structured and others consist of free text (Ciravegna, 2003). In this case generic linguistic methodologies (e.g. parsing) will not work properly and therefore the system will not be able to extract information. A system which can operate with or without the need for an intermediate representation is Soderland's WHISK (Soderland, 1999) which can learn patterns which match directly to unparsed text. However, it was found that this approach is more suitable for semi-structured text which is generally syntactically simpler and more regular than free text.

Moreover most of the technology mentioned above is not able to exploit the available ontological information, for example for generalizing rules and apply reasoning. With the potentially reach ontologies available for the Semantic Web, this could be a relevant limitation. New methodologies are needed for IE able to exploit such information, with the caveat mentioned above, though, that such ontologies could be far from linguistically oriented.

All in all, the systems mentioned above are (very similar to) standard MUC systems where just the step of extracting relations from sentences are performed automatically. These systems still require IE experts for porting the other modules, so - we believe - are not suitable for use in the Semantic Web.

4 Conclusions

This paper has discussed some requirements for the suitability of IE as annotation support for the Semantic Web. The core requirements are portability by non-experts, ability to cope with different text types, ability to use ontological information and ability to train with limited annotated documents. We have discussed the extent to which current IE technology meets these requirements. It was found that current technology is limited in the following respects:

1. the ability to train with a limited amount of material.
2. the ability to learn relations without relying on deep linguistic annotation; the system should be able to exploit linguistic information when existent and reliable but rely on shallower methods when necessary.
3. the ability to use ontological information when available.

Some of these limitations have been addresses by various approaches to IE. For example, Yangarber et al. (2000) and Collins and Singer (1999) use unsupervised learning algorithms to reduce the input required from the user to a few seed patterns and this is an attempt to overcome limitation 1. Limitation 2 is addressed by the WHISK system (Soderland, 1999) which has the ability to learn patterns from both parsed and unparsed text. The final limitation has not really been addressed; while some approaches have made use of linguistic ontologies (see Section 3) these are different from the ontologies which are expected to be used for the Semantic Web.

In conclusion we believe that IE is a very promising technology for annotation for the Semantic Web. Potentially IE systems could become in the future Semantic Web as important as indexing systems are for search engines in the current for of the Web. In order to get this opportunity, some very focused research effort is needed that goes beyond the usual definitions and limitations of IE as derived from the MUC conferences.

Acknowledgments

This work was carried out within the European Commission framework V project dot.kom (IST-2001-34038). Further information is available from `http://www.dot-kom.org`

The authors are grateful for a number of interesting discussions on the topics in this paper with the dot.kom partners.

References

T. Berners-Lee, J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American*, 28(5):34–43.

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 152–160, Montreal, Canada.

C. Brewster, F. Ciravegna, and Y. Wilks. 2001. User-centred ontology learning for knowledge management. In *Proceedings of the 7th International Conference on Applications of Natural Language to Information Systems*, pages 203–207, Stockholm, Sweden.

S. Brin. 1998. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183, Valencia, Spain.

N. Català, N. Castell, and M. Martin. 2000. ESSENCE: A Portable Methodology for Acquiring Information Extraction Patterns. In *Proceedings of the 14th European Conference on Artifical Intelligence*, pages 411–415, Berlin, Germany.

J. Chai and A. Biermann. 1999. The use of word sense disambiguation in an information extraction system. In *Proceedings of the Eleventh Annual Conference on Innovative Applications of Artificial Intelligence*, pages 850–855, Portland, OR.

H. Chieu and H. Ng. 2002. A maximum entroy approach to information extraction from semi-structured and free text. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (AAAI-02)*, pages 768–791, Edmonton, Canada.

F. Ciravegna, A. Dingli, D. Guthrie, and Y. Wilks. 2003. Integrating information to bootstrap information extraction from web sites. In *Proceedings of the IJCAI 2003 Workshop on Information Integration on the Web.* Acapulco, Mexico.

F. Ciravegna. 2001. Challenges in information extraction from text for knowledge management. *IEEE Intelligent Systems and Their Applications*, 27:97–111.

F. Ciravegna. 2003. Designing adaptive information extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web.*

M. Collins and Y. Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110, College Park, MA.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and some of its Applications.* MIT Press, Cambridge, MA.

S. Handschuh, S. Staab, and F. Ciravegna. 2002. S-CREAM - Semi-automatic CREAtion of Metadata. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW-02)*, Sigüenza, Spain.

J. Hendler. 2001. Agents and the semantic web. *IEEE Intelligent Systems Journal*, 16(2):30–37.

W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland. 1992. University of massachusetts: Description of the CIRCUS system used for MUC-4. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pages 282–288, San Francisco, CA.

D. Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

A. Maedche and S. Staab. 2000. Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence*, pages 321–325. Berlin, Germany.

A. Mikheev, M. Moens, and C. Grover. 1999. Named entity recognition without gazeteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, Bergen, Norway.

S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, and R. Weischedel. 1998. BBN: Description of the SIFT system as used for MUC7. In *Proceedings of the 7th Message Understanding Conference*, Fairfax, VA.

1991, 1993, 1995, 1997, 1998. *Proceedings of the Third, Fourth, Fifth, Sixth and Seventh Message Understanding Conferences.* Morgan Kaufmann.

E. Riloff and J. Shoen. 1995. Automatically acquiring conceptual patterns without an annotated corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 148–161, Somerset, NJ.

E. Riloff. 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816, Washington, DC.

S. Soderland. 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 31(1-3):233–272.

R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen. 2000. Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of the Applied Natural Language Processing Conference (ANLP 2000)*, pages 282–289, Seattle, WA.