Mining the Semantic Web: Requirements for Machine Learning

Fabio Ciravegna, Sam Chapman
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield, S14DP, United Kingdom,
{f.ciravegna, s.chapman} @dcs.shef.ac.uk

1 Introduction

In the current form of the Web, content is designed and published for human reading and it is not typically tractable by machines; the Semantic Web, SW, is expected to extend this by providing structured content via the addition of annotations. A prerequisite for the SW is the availability of structured knowledge, so methods need to be employed to generate it from existing unstructured content (document annotation).

A number of tools have been proposed for manual annotation of documents, e.g. (Staab et al., 2001); some of them use Information Extraction, IE, to reduce the burden on the user side (Ciravegna et al., 2002)(Vargas-Vera et al., 2002). Relying on a manual process presents some risks for the SW, because it creates a bottleneck: convincing millions of users to annotate documents requires a world-wide action of unlikely outcome. Moreover there are some serious concerns about the quality of manual annotation, due to user inability or to spamming (Dingli et al., 2003) (Dill et al., 2003). To produce a viable and maintainable SW, large scale automatic annotation services (Dill et al., 2003), similar to today's search engines, are needed. They must be: (1) easily defined for a specific ontological component or service; (2) able to constantly re-index documents (so to solve problem of obsolete/misaligned annotation).

Machine Learning, ML, and IE become then indispensable for developing SW tools able to extract and structure information: in this paper we focus on identifying requirements an challenges for future research in ML and IE applied to SW. When detailing the requirements and challenge we refer, as an example, to Armadillo (Dingli et al., 2003). Armadillo is a tool for extracting and integrating information from large repositories (e.g. the Web) developed at Sheffield. Armadillo is able to (1) learn to extract facts and entities in a largely unsupervised way; (2) cope with unstructured documents such as semi-structured and free documents as well. The learning algorythm currently integrated into Armadillo is $(LP)^2$, implemented in Amilcare (Ciravegna and Wilks, 2003). The requirements and challenges that we identify, however, are not related simply to Armadillo but can be shared by other SW tools with similar aims.

2 Large Scale IE

A first requirement is to be able to work on a large scale and across corpora/sources boundaries. Most of the current IE literature considers small corpora of the order of hundreds of documents (typically newspaper articles). In the SW, this scenario is not suitable, as documents and collections can be heterogeneous (see below) and the document space can have large and increasing dimensions. For this reason, a new research community has emerged at the intersection of the Semantic Web and Information Extraction fields, with the aim of producing large scale extraction tools for document annotation for the Semantic Web. SemTag (Dill et al., 2003) is a system able to annotate large document repositories (e.g. the Web) for retrieval purposes, using very large ontologies (Stanford's TAP ontology: 17,000 objects). The process is entirely automatic and the methodology is largely application independent, i.e. it does not require human intervention. The only task the system is able to cope with is entity extraction (i.e. not facts or events) and disambiguation (a first step in information integration). Dome (Leonard and Glaser, 2001), the harvester of the AKT triple store [1], is able to build large knowledge bases of facts for a specific application (not just entities). The aim is both large scale and deep ontology-based information extraction and integration. A large number of manually encoded wrappers are used to extract information from Web sites; therefore porting to a new application requires a great deal of manual programming. Maintenance is complex because when the web pages change their format, it is necessary to re-program the wrapper. The manual approach makes using very large ontologies impossible. Extraction is limited to highly regular and structured pages selected by the designer and it is not applicable to irregular pages or free text documents. In Armadillo the learning is currently limited to recognition of entities and implicit relation. The explicit modeling of relation is done in an indirect way by observing frequency of co-citations of entities with associated implicit relations. The methodology works well when it extracts facts that are repeated some times (redundancy of information), but it is unable to capture information that is not repeated in dif-

[1] http://triplestore.aktors.org/

ferent places. The challenge for the SW is to make accessible also information that is not necessarily of public domain, so not very frequently mentioned. New developments are therefore needed for enabling detailed information extraction of infrequent information.

3 Heterogeneity of formats

From a Natural Language/Information Extraction point of view, dealing with heterogeneous documents formats imposes requirements on the methodologies that can be used, in particular in terms of: (1) portability with limited or no user intervention (2) ability to cope with a variety of document types. When a paper citation is found in personal bibliographic pages, Armadillo learns to recognise regularities in the specific bibliographic style used: each person may use a different and somewhat rough style and the task is then discovering how the bibliographic page is organised. Here machine learning is essential: systems requiring manual rule development are not suitable because of the large number of pages to model. Regularities come both from the page layout (e.g. html lists when discovering new people names) or are language related (when discriminating among authors and editors in a bibliographic reference or in discovering departmental membership in sentences like "Dr J. Smith has been promoted to professorship"). This is a typical strategy needed when coping with Web documents, as they can be very rigidly structured (e.g. when produced by a database) or fairly rigid (as manually compiled bibliographic pages that are regular, but with many inconsistencies) or unstructured (e.g. free text). Sometimes those types are mixed within a single document, which contains sections with different styles. For example the main page of The Times (http://www.timesonline.co.uk/) contains free texts in the summaries of the articles, semi structured information in the headline lists (titles are mainly choppy sentences) and completely structured ones in the list of sections. The learning algorithm must adapt to the different document types smoothly without user intervention. Methods relying on deep linguistic analysis are bound to fail when confronted with rigid formatting (e.g. tables) and choppy sentences. Systems that rely on the use of formatting only (e.g. wrappers) are not able to cope with free texts and even choppy sentences. The $(LP)^2$ algorithm used by Armadillo uses a methodology where the level of linguistic analysis is one of the parameters the learner tunes during learning; such tuning can be different for different pieces of information located in different parts of the documents (so to cope with variously formatted information).

3.1 Heterogeneity of IE tasks

Another challenge for learning in Armadillo relies in the type of IE task to be performed. IE tasks can differ much accordingly to the kind of information that they have to extract and from the sources they have to analyse. One main limitation of most the current approaches to ML-based IE is the focus on entity extraction and/or on implicit relation recognition. This means that - for example - the system is able to identify speaker (a person who is giving the talk) and start-time (the time in which a talk starts) of a seminar, but not to relate elements of the relation directly. Therefore if two seminars are mentioned in a document, the system is not able to assign the speaker to the correct start-time (Ciravegna and Wilks, 2003). Only recently, the community has started exploring ML based methodologies for explicit relation recognition [Roth02], [Sudo03]. Despite the success of some tools performing only implicit relation recognition (e.g. Amilcare), relation extraction is of capital importance in IE for the Semantic Web. As an example, Armadillo deals with:

- extracting events from sentences;
- correlated information throughout a document;
- extracting dispersed information from multiple documents;
- extracting multi-document co-references.

Very effective default strategies for multi-document extraction are used that drastically simplify the task; for example in the CS task, we expect that all the potentially coreferential names (e.g. J Smith and John Smith) are always co-referring, unless there is strong evidence from the IE point of view that they are not (e.g. a dissimilarity in other personal details). Although these strategies are effective in many cases, more sophisticated strategies need to be developed. Moreover, Armadillo performs intra-document relation recognition as a task of information integration built on the top of implicit relation recognition provided by Amilcare. Learning based relation extraction is still largely in its infancy (Roth and tau Yih, 2002)(Sudo et al., 2003) and we believe that further research is needed in order to provide robust and effective methodologies that can be ported to new domain without expert intervention.

3.2 Bootstrapping Learning

In classic IE, training is performed by providing the system with a set of example texts where the information to be extracted is manually annotated (Vargas-Vera et al., 2002) (Handschuh et al., 2002), (Ciravegna et al., 2002). In environments with mixed text types and multiple sources (as in the generic SW scenario), exhaustive manual annotation is largely unfeasible, because it is impossible to annotate a considerable amount of documents for each type. The problem is increased when coping with explicit relation, because data sparsity becomes a problem: as a matter of fact, implicit relation recognition and entity recognition (as performed in most adaptive systems) only require the identification of one entity; therefore it is easy to find and annotate examples to train the system. When modelling explicit relations, it is necessary to identify triples (par-

ticipant1 relation participant2). Such triples tend to be less frequent and therefore the data sparsity problem arises. Annotating sparse examples requires a considerable effort on the user side in selecting the correct annotation corpus; also the annotation effort can be largely ineffective if the selection of the corpus is not adequate (Ciravegna and Petrelli, 2001). A different approach is used in Armadillo, where largely unsupervised methodologies are used and cannot be controlled by the user whose only role is to validate the knowledge extracted by the system at the end of the process. The shortcoming of the approach is that the user has no way to drive the system towards the most interesting documents and facts, therefore there is the concrete risk that - although the system is effective in extracting knowledge - the extracted knowledge is not the one that the user judges the most relevant one. We believe that the challenge for future research in IE is to study a way of bootstrapping learning on a large scale by focusing on a mixed initiative strategy where user and system collaborate to retrieve relevant documents and to annotate them. The approach will still retort to largely unsupervised learning, but the user needs to have a role in determining the direction the learning is taking. We believe that this approach is more suitable to Knowledge Management than both Armadillo's unsupervised approach and the completely supervised approach

3.3 Content cleaning and normalisation for information integration.

Noise is an intrinsic feature of the Web. Low quality annotation and spamming introduce noise, as well as the Web dynamicity: pages are often built or modified in a careless way. As an example the format of personal bibliographic pages tend to be quite regular, but not totally: they tend to be modified every now and then when one item is added and the format used can change over time: for example the title of the paper that initially was formatted in bold, can become italic and the title of the collection the other way round, making it more difficult to build rules relying on the format. ML methods able to cope with noisy data are necessary. For example, most of the wrapper induction methodologies require noise-free data. Amilcare is generally able to accept fairly noisy training data without degrading accuracy, but there is space for improvement in designing new algorithms.

Some noise can be introduced also by the unsupervised method itself during Armadillo's Extraction of potential knowledge. In our experiments in the CS task, three initial people seeds turned out to be wrong, i.e. they were people not working at the department. Keeping under control the effects of this kind of noise is a challenge especially in systems that build upon existing results. We currently use multiple weak evidence based validation before accepting a seed at any stage in learning. It is a quite an effective strategy, and it filters most noise, but with some domains and spamming this could become more problematic.

In general, we believe that the requirements for the future research in information integration for large scale IE are:

- Type of task: complex event/facts/entities recognition and integration across documents and archives;
- Adaptivity: such strategies should be adaptive and not require any expert intervention such as manual resources development;
- Reasoning: integration of facts and events requires some degree of reasoning, otherwise it is unfeasible. One type of reasoning that is needed is some degree of temporal and causal reasoning, in order to capture evolution of situations.
- Robustness: information in large repositories is by definition inconsistent and therefore the integration process must allow some degrees of uncertainty and inconsistency. Inconsistency can be preserved or removed according to specific cases and tasks.

3.4 Sparse annotation

When analysing a document, most algorithms used for IE consider the set of annotations as positive examples and the rest of the document as negative example. When seeding and learning in a web page, though, this is not a suitable methodology, because many of the unannotated examples are actually information to be extracted. It is like having to learn from a corpus that is at the same time training and testing corpus. The Amilcare strategy that we currently use is to limit the visibility window around annotated examples for each tag. Everything is in the window (e.g. +/- 5 words) is considered counterexample if it is not annotated. The idea is that the probability that an unannotated example is included in such a limited window is very low. From an experimental point of view it works quite well, but more sophisticated methodologies taking into account the previously ratings of found knowledge are needed.

4 Conclusion

In this paper we have outlined the challenges for Machine Learning models for bootstrapping the Semantic Web via mining the Web. We have focused on the Armadillo methodology, but the requirements are actually relevant to a range of methodologies and tools with similar aims.

5 Acknowledgements

This work was carried out within the AKT project (http://www.aktors.org), sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/N15764/01), and the Dot.Kom project, sponsored by the EU IST asp part of Framework V (grant IST-2001-34038).

References

Fabio Ciravegna and Daniela Petrelli. 2001. User involvement in adaptive information extraction: Position paper. In *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining held in conjunction with the 17th International Joint Conference on Artificial Intelligence*. Seattle, http://www.smi.ucd.ie/ATEM2001/.

Fabio Ciravegna and Yorick Wilks. 2003. Designing adaptive information extraction for the semantic web in amilcare. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web*, Frontiers in Artificial Intelligence and Applications. IOS Press.

Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli, and Yorick Wilks. 2002. User-system cooperation in document annotation based on information extraction. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*. Springer Verlag.

S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. 2003. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the World Wide Web Conference 2003*.

Alexiei Dingli, Fabio Ciravegna, and Yorick Wilks. 2003. Automatic semantic annotation using unsupervised information extraction and integration. In *Proc. of the K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*.

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, EKAW02*. Springer Verlag.

Thomas Leonard and Hugh Glaser. 2001. Large scale acquisition and maintenance from the web without source access. In Siegfried Handschuh, Rose Dieng-Kuntz, and Steffen Staab, editors, *Proceedings Workshop 4, Knowledge Markup and Semantic Annotation, K-CAP 2001*.

Dan Roth and Wen tau Yih. 2002. Probabilistic reasoning for entity & relation recognition. In *COLING*.

S. Staab, A. Maedche, and S. Handschuh. 2001. An annotation framework for the Semantic Web. In *Proc. of the First Workshop on Multimedia Annotation, Tokyo*.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003. An improved extraction pattern representation model for automatic ie pattern acquisition. In *ACL*, pages 224–231.

M. Vargas-Vera, Enrico Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. 2002. MnM: Ontology driven semi-automatic or automatic support for semantic markup. In *Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*. Springer Verlag.