Requirements for Information Extraction for Knowledge Management

Philipp Cimiano*, Fabio Ciravegna∅, John Domingue⊕,
Siegfried Handschuh*, Alberto Lavelli+, Steffen Staab*, Mark Stevenson∅

*AIFB, University of Karlsruhe

∅ NLP Group, University of Sheffield

⊕ Open University, Milton-Keynes

+ ITC-irst, Trento

ABSTRACT

Knowledge Management (KM) systems inherently suffer from the *knowledge acquisition bottleneck* - the difficulty of modeling and formalizing knowledge relevant for specific domains. A potential solution to this problem is Information Extraction (IE) technology. However, IE was originally developed for database population and there is a mismatch between what is required to successfully perform KM and what current IE technology provides. In this paper we begin to address this issue by outlining requirements for IE based KM.

Keywords: Information Extraction, Knowledge Management, Ontologies, Annotation

1 INTRODUCTION

A large part of a company's knowledge is stored in textual documents available within intranets. However, this knowledge cannot be queried nor captured in a straightforward way, which reduces a company's efficiency. The challenge is to formally represent the knowledge contained in textual form such that it can be accessed and used by the workers in an enterprise through various knowledge-based services.

A similar scenario is encountered within the Semantic Web in which the central idea is to provide efficient access to heterogeneous and distributed web resources. This is only possible if the knowledge contained in the resources has been formalized so that it can be shared, understood and reused by other people or applications, such as crawlers, information brokering services and mediators. So the success clearly depends on the availability of machine-readable data, i.e. metadata.

Both scenarios are comparable to the extent mentioned above and in fact similar solutions have been proposed to overcome part of the problems associated with them. On the one hand, ontologies have been proposed as a formalism to externalize and share knowledge within KM [Staab et al. 02, Fensel 01, Mulholland et al. 01, Benjamins 98] as well as in the context of the Semantic Web [Berners-Lee et al. 01]. Ontologies are suitable for this purpose because they represent a formal, explicit specification of a shared conceptualization [Gruber 93]. A shared conceptualization in this sense has to be understood as an abstract model of some aspect or part of the world shared by a certain group of people with a common interest. Formal and explicit refer to the fact that such an ontology should also be readable for machines. On the other hand, semi-automatic or automatic methods have been proposed for KM as well as for the Semantic Web in order to reduce the cost of producing metadata [Ciravegna et al. 02], [Handschuh et al. 02] [Vargas-Vera et al. 02].

In this context, Information Extraction from text (IE) is a very promising technique for the Semantic Web as well as for KM [Ciravegna 01]. IE is an automatic method with the purpose of locating relevant entities and facts in electronic documents for further use and fits perfectly into the KM scenario described above. A first requirement derived from this potential use of IE within KM is the fact that the target knowledge structures produced by the IE system have to be compatible with the ontology used for formalizing externalized knowledge. Only then can the extracted knowledge be shared and further processed within a company's KM environment. This paper focuses on the way IE could be integrated into the existing KM technology as well as on the requirements that such integration poses

on the IE and KM technologies.

The remainder of this paper is organised as follows: in Section 2 we discuss the requirements for the integration of IE into Knowledge Management Systems. The requirements such integration poses on the IE technology itself are then covered in Section 3. The paper finishes with some conclusions and implications.

2 Knowledge Management Requirements

The most important requirement for a KM solution is its successful integration into the enterprise in question. The process concerned with the introduction of a KM system as well as its maintenance, evolution and refinement is commonly referred to as the knowledge meta process [Staab et al. 02]. The knowledge process on the other hand is concerned with issues related to the use of the introduced KM solution. In particular, it focuses on the cycle of information creation, capture, retrieval and use, for example to create new information and close the cycle (see [Staab et al. 02]) It is important that this cycle fits with existing (and emerging) work practices. Both processes are dependent on each other as the refinement of the KM solution can only take place by considering the working knowledge process, which in turn will be modified according to the introduced refinements. The information obtained in the retrieval/ access step of the knowledge management cycle is then typically included within a specific application and can also be used in the creation of new documents (see Figure 1).

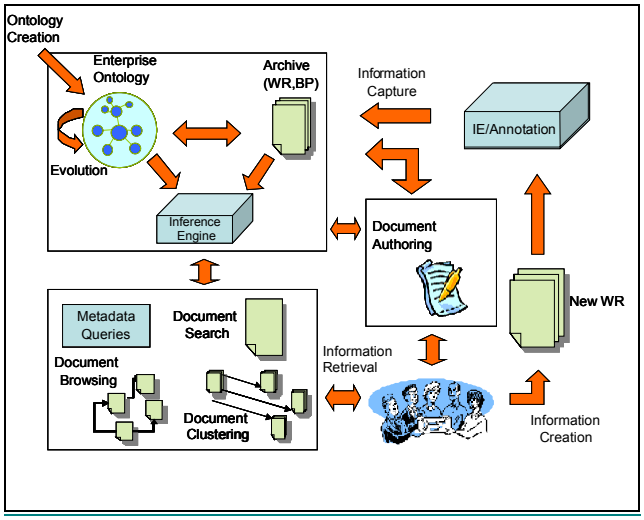

Figure 1

In order to close the knowledge process cycle, the information contained in the newly created documents has to be captured, i.e. the documents have to be annotated with regard to the ontology so that they can be fed back into the enterprise's archive for further use. This is where IE techniques come into play. As mentioned earlier, IE can be applied either in an automatic or semi-automatic way in order to produce annotations which are consistent with a given ontology. Thus, IE should directly exploit the underlying ontology in order to produce compatible knowledge structures. In particular, the mapping from knowledge structures produced by IE to ontological models represented in languages such as DAML+OIL, RDF(S) or OWL should be straightforward. An issue related to this requirement is the necessity to produce relational metadata, - instances of relations defined in the selected ontology.

One further important requirement is the need for some quality control of the output produced by IE before further processing it for KM purposes. In fact, IE is by definition an error-prone process. Consequently, the resulting knowledge structures cannot be directly used to populate an ontology without manual intervention. This quality control can for example take place directly in an annotation tool integrating IE as a plug-in. In this sense, the annotation framework would thus suggest annotations to the user, which have to be manually validated. We will make use of OntoMat Annotizer [Handschuh et al. 02] or MnM [Vargas-Vera et al. 02] for this purpose. However, it could also be thought of having an 'on the fly' validation of produced annotations in the sense that users may decide at some point during their work if a specific annotation is plausible or not and thus whether it can be kept or has to be rejected.

Documents are created in a context that is not captured in the text. It is thus important that annotations not only reflect the explicit content of a particular document but also knowledge related to its creation context, for example, reasons why particular items were omitted. Nevertheless, such an annotation should also be consistent with the underlying ontological model used within the enterprise so that this knowledge can be stored and reused as with 'conventional' document annotations.

Finally, it is important to mention that it cannot be expected that a reasonable and suitable ontology will be available right from the beginning. Moreover, we envision starting from a small seed ontology, which will be constantly extended, refined and modified. We intend to create such a seed ontology with the help of the text mining approach presented in [Cimiano et al. 03]. Thus the knowledge process and the knowledge meta process [Staab et al. 02] will be highly interleaved and dependent on each other. In this context it is important that knowledge about changes in the ontology is also made explicit and to have some ontology evolution support such as described in [Stojanovic et al. 02].

3 Information Extraction Requirements

Research in IE has been largely driven by the Message Understanding Conferences (MUC). These competitions focused on extracting information from newswire text. The participants were required to perform different tasks, from the identification of person, location and organization names (Named Entity recognition) to the identification of relations between entities (Template Relation) to the construction of complex templates (Scenario Template). The original aim of IE was to automatically fill database records from text and consequently systems have not, in general, been designed to carry out knowledge markup. In the remainder of this section we discuss the requirements for IE systems performing knowledge markup in the context of KM.

3.1 Relation Extraction

[Handschuh et al. 02] discuss the problems involved in using an IE system which carries out concept recognition (e.g. Amilcare [Ciravegna 03]) to produce relational metadata, i.e. instances of a certain ontological relation. For example, in the sentence "Mr. Jones was hired by Dot.Kom Ltd. last week" Amilcare can identify "Mr. Jones" as a person (and even as a "hiredPerson") and "Dot.Kom Ltd." as a company (or even "hiringCompany"). However, it cannot identify the relation between these two entities (i.e., that the specific person was hired by the specific company; this means that if different hiringCompany and hiredPerson exist it is not possible to connect them properly). [Handschuh et al..02] present a discourse analysis approach to map the entities tagged by Amilcare into graph structures such as those used in ontological formalisms as RDF, DAML+OIL or OWL. In order to use an IE system for KM purposes it is necessary that it produces relational metadata that can be used to directly populate an ontology. This means that some form of relation extraction is necessary (e.g. [Soderland 99, Yangarber et al. 00, Yangarber 03]). Such a component could be trained on relational annotation produced by a system like the OntoMat Annotizer [Handschuh et al. 02]. This type of approach could be supplemented by an ontology-based discourse analysis approach such as the one proposed in [Handschuh et al. 02].

3.2 Text types processed

The systems that participated in the MUC evaluations were required to extract information from well-formed newswire text. However, a KM system should be able to process a wider variety of texts since they will be expected to process web and intranet pages. IE systems have tended to extract information from a limited variety of text types, for example free and semi-structured text [Soderland 99] or tabular data [Hurst 00]. Initial attempt to cover all these types into single system has been done in Amilcare [Ciravegna 03]. This anyway still represents a challenge to the language processing community [Ciravegna 01].

3.3 Adaptivity and Usability

Traditional IE systems have tended to be difficult to port to new domains and extraction tasks. For example, [Lehnert et. al. 92] estimated that 1,500 person-hours of highly skilled labor were required to adapt their system for MUC-4. Clearly the applications will be limited for any tool that requires such an investment to be adapted to a new domain or extraction task.

It is therefore vital that IE systems can be adapted with the least possible effort and that this process can be carried out by non-experts. Machine learning techniques could be used for this (e.g [Soderland 99], [Yangarber et. al. 00, Yangarber 03]). Interaction with annotation tools requires little more than marking relevant concepts in text. However, the mode of interaction for marking relations in text is not as obvious as for marking concepts, which can be directly highlighted.

In general, the IE systems must be portable by non experts and users should be assisted in the whole application lifecycle. [Ciravegna 01] identifies the requirements in this respect, mentioning the need for tools for (1) scenario definition, (2) system adaptation and result validation and (3) application delivery. Scenario design is not an issue in ontology-based IE because the ontology will provide the scenario. Concerning system adaptation and result validation, experiences such as Melita [Ciravegna et al. 02] show that a great deal of control can be reached using simple HCI techniques. We are currently investigating in the direction of further improvement of usability through strong integration with the ontology as explained below.

3.4 Interaction with ontologies

It is crucial for the integration of IE into KM that its output can be directly used to populate ontologies or to enrich documents with ontology-based metadata. Thus, it is important that the output of IE systems can be mapped in a straightforward way to ontological models coded in languages such as RDF(S), DAML+OIL or OWL. Essentially this has four implications for IE:

1) *Detecting concepts over a hierarchy*: IE should directly interact with the ontological hierarchy and tag instances at different levels of hierarchical abstraction. From a practical point of view rules should be generalized semantically using the ontology.

2) *Exploiting conceptual markup as context:* It is possible to imagine that IE systems could operate in a bootstrapping-like fashion and make use of conceptual markup to extract the conceptual relation between two previously tagged entities.

3) *Exploiting lexical information:* It would be useful to include information about how certain conceptual relations are expressed linguistically in a text. This could for example allow the rule induction algorithm a more efficient exploration of the search space. Information about synonyms such as contained in linguistic ontologies as WordNet [Miller 90] could also turn out very useful in the context of the semantic generalization of extraction rules (e.g. [Chai et. al. 99] and [Harabagiu et al. 00]).

4) *Mapping between tags and concepts:* The mapping between the IE system and ontology should be one-to-one to allow the ontology to be exploited within the IE system and use the annotation produced by the IE system to populate the ontology.

Summarizing, the above mentioned requirements suggest some relevant directions for improving IE so that it can successfully satisfy KM needs. First of all, the importance of relation extraction will be addressed further investigating the approaches described in [Yangarber et al. 00, Yangarber 03]. Such unsupervised approaches take into consideration also the issue of adaptivity, crucial for reducing the cost of porting to new domains and applications. Adaptivity will be dealt with also experimenting with bootstrapping techniques, such as co-training.

4 Summary and Conclusion

In the context of KM, IE cannot be regarded as a stand-alone tool which can be applied quite independently of the KM technology used. In fact, it is important for the IE system to directly interact with the ontology to extract knowledge which is compatible with it and can thus be reused within the enterprise`s KM environment. Furthermore, the information extraction system should certainly be adaptive and applicable to a wide range of text types and genres. Concerning the knowledge cycle, it seems very important that the meta knowledge process and the knowledge process are highly interleaved and that the user is supported in the meta knowledge process by (semi-) automatic methods to produce a seed ontology which will be iteratively refined according to requirements derived from the working knowledge process.

In summary, the successful integration of IE into KM methodology presupposes a strong and direct interaction between the ontology, the IE system as well as the constantly changing information needs of the users.

ACKNOWLEDGMENTS

This work was carried out within the IST-Dot.Kom project (http://www.dot-kom.org), sponsored by the European Commission as part of the framework V, (grant IST-2001-34038). Dot.Kom involves the University of Sheffield (UK), ITC-Irst (I), Ontoprise (D), the Open University (UK), Quinary (I) and the University of Karlsruhe (D) . Its objectives are to develop Knowledge Management and Semantic Web methodologies based on Adaptive Information Extraction from Text.

REFERENCES

[Berners-Lee et al. 01] Berners-Lee, T., Hendler, J. and Lassila, O. "The Semantic Web". *Scientific American*, May 2001 Issue.

[Benjamins et al. 98] Benjamins, V., Fensel, D., Gómez Pérez, A. "Knowledge Management through Ontologies". In *Proceedings of the Conference on Practical Aspects of Knowledge Management (PAKM)*, 1998.

[Chai et. al. 99] Chai, J., Biermann, A. and Guinn, C. "Two Dimensional Generalization in Information Extraction" In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)* 1999.

[Cimiano et al. 03] Cimiano, P., Staab, S. and Tane, J. "Automatic Acquisition of Taxonomies: FCA meets NLP". In *Proceedings of the ECML/PKKD Workshop on Adaptive Text Extraction and Mining*, 2003.

[Ciravegna 01] Ciravegna, F. "Challenges in Information Extraction from Text for Knowledge Management", In *IEEE Intelligent Systems and Their Applications (Trends and Controversies)*, 2001.

[Ciravegna et al. 02] Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli and Yorick Wilks: User-System Cooperation in Document Annotation based on Information Extraction" in *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002)*, 1-4 October 2002 - Sigüenza (Spain).

[Ciravegna 03] Ciravegna, F. "Designing Adaptive Information Extraction for the Semantic Web in Amilcare" in S. Handschuh and S. Staab (eds), *Annotation for the Semantic Web* IOS Press, Amsterdam, 2003.

[Fensel 01] Fensel, D. *Ontologies: A Silver Bullet for Knowledge Management*, Springer Verlag, 2001.

[Gruber 93] Gruber, T. "A translation approach to portable ontology specifications". *Knowledge Acquisition*, 5:199-220, 1993.

[Handschuh et al. 02] Handschuh, S., Staab, S. and Ciravegna, F. "S-CREAM - Semi-automatic CREAtion of Metadata", In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, 2002.

[Harabagiu et al. 00] Harabagiu, S. and Maiorano, S. “Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction” In *Proceedings of the Language Resources and Evaluation Conference (LREC-2000)*, Athens, Greece, 2000.

[Hirschman 1998] Hirschman, L. "The Evolution of evaluation: Lessons from the Message Understanding Conferences", *Computer Speech and Language*, 12, pp. 281-305, 1998.

[Hurst 2000] Hurst, M. “Processing Tables for Information Extraction” Ph.D. thesis, University of Edinburgh.

[Lehnert et. al. 1992] Lehnert, W. and Cardie, C. and Fisher, D. and McCarthy, J. and Riloff, E. and Soderland, S. “University of Massachusetts: Description of the CIRCUS System used for MUC-4” *Proceedings of the Fourth Message Understanding Conference (MUC-4)*

[Miller 90] Miller, G. “WordNet: An On-line Lexical Database” *International Journal of Lexicography* 3(4):235-312

[Mulholland et al. 01] Mulholland, P., Zdrahal, Z.,Domingue, J., Hatala, M. and Bernardi, A. A Methodological Approach to Supporting Organisational Learning. *International Journal of Human Computer Studies.* Vol. 55, No. 3, September 1, 2001, pp. 337-367

[Soderland 99] .Soderland, S. “Learning Information Extraction Rules for Semi-Structured and Free Text”. *Machine Learning 34(1-3)*, 1999.

[Staab et al. 02] Staab, S., Studer, R. and Sure, Y. “Knowledge Processes and Meta Processes in Ontology-based Knowledge Management”. In *Handbook of Knowledge Management,* Springer Verlag, 2002.

[Stanoevska et al. 98] Stanoevska, K., Hombrecher, A., Handschuh, S. and Schmid, B. “Efficient Information Retrieval: Tools for Knowledge Management”. In *Proceedings of the Conference on Practical Aspects of Knowledge Management*, 1998.

[Stojanovic et al. 02] Stojanovic, L., Stojanovic, N. and Maedche, A.. “Change discovery in ontology-based knowledge management systems”. In *Proceedings of the 2nd International Workshop on Evolution and Change in Data Management (ECDM)*, 2002.

[Yangarber et al. 00] Yangarber, R., Grishman, R., Tapanainen, P. and Huttunen, S. “Automatic Acquisition of Domain Knowledge for Information Extraction”. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING),* 2000.

[Yangarber 03] Yangarber, R. “Counter-Training in Discovery of Semantic Patterns”. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003),* 2003.

[Vargas-Vera et al. 02] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A. and Ciravegna, F. “MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup”. In *Proceedings of the 13th International Conference on Knowledge Engineering and Management (EKAW 2002)*, ed. Gomez-Perez, A., Springer Verlag, 2002.