A Domain-Independent Approach to IE Rule Development

Kalliopi Zervanou [1], John McNaught [2]

[1] Department of Computation, UMIST
[2] Department of Computation, UMIST and UK National Text Mining Centre
P.O. Box 88, Manchester, M60 1QD, UK
{ K.Zervanou, J.McNaught }@co.umist.ac.uk

Abstract

A key element for the extraction of information in a natural language document is a set of shallow text analysis rules, which are typically based on pre-defined linguistic patterns. Current Information Extraction research aims at the automatic or semi-automatic acquisition of these rules. Within this research framework, we consider in this paper the potential for acquiring generic extraction patterns. Our research is based on the hypothesis that, terms (the linguistic representation of concepts in a specialised domain) and Named Entities (the names of persons, organisations and dates of importance in the text) can together be considered as the basic semantic entities of textual information and can therefore be used as a basis for the conceptual representation of domain specific texts and the definition of what constitutes an information extraction template in linguistic terms. The extraction patterns discovered by this approach involve significant associations of these semantic entities with verbs and they can subsequently be translated into the grammar formalism of choice.

Introduction

Information acquisition, dissemination and management have gained increasing importance in today's electronic age. The practical need to distinguish and retrieve information from large document collections, stored in electronic form in databases and the Internet, has motivated research in various areas of text processing. Research in Information Extraction focuses on a particular aspect of this problem, the recognition and extraction of facts and event relations and their respective characteristics. In order to achieve that level of textual analysis and understanding, Information Extraction systems rely on linguistic information from various NLP tools and knowledge resources. The subsequent recognition of salient pieces of information in text is performed through a set of linguistic rules. Typically, these rules involve syntactic relations between words or semantic classes of words and express in this way concepts and events of interest that are to be extracted from the text.

In order to respond to user specific information requirements, IE systems are innately constrained to user and domain-specific applications. An emerging issue in IE research is the facilitation of domain knowledge acquisition for the development of systems that can be easily adapted to new applications and there are currently numerous IE system approaches to the problem. These approaches typically involve manual or semi-automatic annotation of a training set of documents and/or the manual or semi-automatic construction of the lexico-semantic resources of the system for the creation of the extraction rule sets.

Within this research framework, we shall consider in this paper the potential for acquiring generic rule patterns by a less knowledge intensive approach and independently of user-specific requirements. Our research is based on the hypothesis that terms (the linguistic representation of concepts in a specialised domain) and Named Entities (NEs) (the names of persons, organisations and dates of importance in the text) together are most likely to convey the principal concepts of the document. We consider terms and NEs as the generic information elements to replace manual annotation and we investigate a methodology that combines recognition of these elements with other textual structure and syntactic cues as a basis for the conceptual representation of domain specific texts and the definition of what constitutes an IE template in linguistic terms. Our approach is inspired from automatic abstracting techniques and proposes an initially exhaustive identification of all possible linguistic patterns that express domain specific information, regardless of user-specific requirements.

In the first part of this paper, we briefly discuss the problem of knowledge acquisition and current approaches, and in the second we present our methodology and initial results of our research.

Knowledge acquisition for IE systems

Knowledge acquisition for IE systems can be viewed from a twofold perspective, the acquisition of lexico-semantic resources and the acquisition of linguistic rules. In both aspects, current solutions involve either reuse of existing resources and/or the automatic or semi-automatic development of domain specific resources.

Lexico-semantic resources, such as electronic dictionaries and thesauri, constitute key components of many NLP applications and an obvious reusable solution for IE systems. There have been approaches using general lexical resources such as the WordNet (Fellbaum, 1998) to incorporate semantic class information for rule acquisition (Califf & Mooney, 1999; Bagga et al. 1997). General semantic lexicons have been recognised as a valuable source for building IE system semantic lexica too.

However in these approaches the problem of having a too general lexicon adapted to a specific domain has to be solved by reducing lexical ambiguity (Bagga et al. 1997; Cavaglia & Ciravegna, 1998; Harabagiu & Maiorano 2000). Another problem in using general lexica is the poor coverage of domain-specific terminology. In response, other approaches opt for the main use of terminological thesauri such as the UMLS or the EMP (Humphreys et al., 2000). Most of these approaches apply for the biomedical domain, but the availability of quality resources for other domains can be problematic. Research approaches to building specialised semantic lexica in a semi-automatic way include identifying key-concepts based on keyword extraction (Riloff, 1996), initial annotation of a training corpus (Soderland et al., 1995) or the expansion of an initial seed-domain lexicon or ontology (Brewster et al., 2002).

For the development of linguistic rules, systems are largely based on manual or semi-automatic annotation of texts. Earlier approaches required much manual effort in building the initial case-frame templates, either by constructing a seed dictionary of templates (Riloff, 1993) or by extensive manual annotation of example case-frames in a training corpus (Soderland et al., 1995). More recent approaches assist the annotation of such a training corpus in an adaptive, interactive way (Ciravegna et al., 2002; Vasilakopoulos et al., 2004). In approaches where annotation of a training corpus is not required, the effort is mostly shifted to the creation or the adaptation of the necessary lexico-semantic knowledge bases (Yangarber, 2000; Ciravegna et al., 2002).

To date, IE systems have made great advances in learning linguistic rules in an automated way. General resources such as the WordNet and the recently developed FrameNet (Fontenelle, 2003) are expanded to provide more information and find more applications. Coupling such resources with machine learning and statistical techniques resulted in much progress in the customisation and reusability of lexico-semantic resources. However, the problem of acquiring the information necessary for the learning of IE templates is faced by shifting the effort load onto either the annotation of training corpora or onto the development of the necessary lexico-semantic resources. Recent advances in the development and availability of lexico-semantic resources have promoted research on knowledge intensive approaches, although in cases where there is a lack of resources, the user or the developer is still faced with the problem of building them.

Methodology

In our approach we attempt to acquire generic extraction patterns without use of any annotation of target information and without use of any domain specific lexical resources, apart from a generic gazetteer used in NE recognition. Instead, we use existing technologies for the automatic recognition of generic semantic elements, namely NEs and terms, in combination with syntactic information, to identify significant extraction patterns.

Terms form an important feature, although not the only one, of any sublanguage domain and IE applications typically target domain specific information. Theoretically, terms are the embodiment of specialised concepts. Therefore, terms are more likely to convey the domain specific information in a document. The identification and the extraction of terms is, according to Boguraev & Kennedy, "one of the better understood and most robust natural language processing technologies within the current state of the art of language engineering" (Boguraev & Kennedy, 1997). Their research has shown that linguistic processing targeted at term identification can be applied for content characterisation in domain specific texts and can be extended to cover domain independent representations for automatic summarisation purposes. We can therefore take advantage of the research done in the area of term extraction for information extraction purposes and make use of existing tools.

Named Entities is another important element of the information that has to be extracted. Research in Information Extraction has shown that the NE recognition task has achieved the highest accuracy of all IE tasks, as defined and evaluated in the Message Understanding Conferences. In particular, NEs were extracted with reliability of F-Measure > 97%, in MUC-6, and F > 94%, in MUC-7 (DARPA, 1998). Moreover, NE recognition does not constitute a knowledge bottleneck, as relatively small gazetteers and a judicious use of internal and external evidence for NE recognition rules are sufficient for satisfactory results (Mikheev et al., 1999).

Therefore, term and NE extraction can provide a reliable linguistic basis for the identification of extraction templates when this information is combined with other syntactic and grammatical category information.

Our approach for rule acquisition consists of the following stages:

1. Corpus collection and pre-processing;
2. Morphological, POS and syntactic analysis;
3. Automatic term extraction;
4. NE recognition;
5. Automatic identification of extraction patterns;
6. Template rule creation.

The approach we present in this paper could certainly be complemented by use of domain specific terminological thesauri, ontologies or lexica. The intensive use of terminological thesauri for IE is currently mostly found in the biomedical domain. For our corpus, the use of such a specialised thesaurus would cover only a part of the information, it would not cover for example business and financial information. Moreover, the main objective of our research being to investigate an approach that is based on automatically acquiring data from a corpus rather than other resources, the use or the customisation of existing lexico-semantic resources has not been included in our methodology.

Implementation and Results

For our research we have collected a corpus of 840 newswire documents (approx. 452K words) related to biotechnology business information. These documents include information related to company research activities, such as drug discovery, trials, other biotechnology products such as new crops and herbicides and patent approval procedures. They also include business and financial information such as succession events, mergers and collaborations, stock market and other company financial information. These documents were selected on the basis of the variety of information types included that could possibly be of interest to different users. They are also representative of a text style used in newswire texts. The corpus has been collected from the internet, but apart from the standard HTML mark-up related to links, the main informational content of the document was in preformatted, free text format, without any other structural or semantic mark-up, such as author, date, header, table or paragraph tags that one may find in HTML documents.

After initial corpus pre-processing to remove all HTML annotation, URLs, tables and lists, the corpus was analysed by the ENGCG, Constraint Grammar Parser of English (Voutilainen, 1995). ENGCG adopts a linguistic approach to POS tagging. In order to resolve ambiguity, the analysis includes morphological analysis and a finite-state shallow syntactic parser that assigns surface syntactic function tags (subject, main verb, object, etc). The main advantages of this approach for our purposes, apart from its reported high accuracy in POS tagging (97%-99%) without need of training on a specific corpus, were the lemma and syntactic function information. Traditionally, IE systems opt for a shallow syntactic analysis, using either mere NP chunkers or both NP and VP chunkers. This is due to the fact that once the rules for the identification of information exist, full syntactic parsing is not necessary: it would merely slow down the analysis process. However, in our case, syntactic information is considered necessary for the rule acquisition stage, as it provides in an automated way information about the syntactic roles of NEs and terms and therefore an implicit way of discovering their semantic relations without manual annotation. Moreover, unlike approaches that opt for use of full syntactic parsing (Yakushiji, 2001) ENGCG provides surface and not canonical syntactic function information on the case frames, thus enabling the straightforward adaptation of the discovered patterns to a final IE system that does not include full parsing in the analysis pipeline.

For term recognition we used the C/NC value method (Frantzi & Ananiadou, 1999). This method is domain-independent and combines statistical and linguistic information for the extraction of multi-word and nested terms. The statistical part defining the termhood of the candidate strings outperforms the common statistical measure of frequency of occurrence used for term or mere keyword extraction, making it sensitive to nested terms. Evaluation of results from our corpus showed that the tool achieves 99% precision on NC values ranging from 697.39 to 103.97, where 88.46% were terms and 11.53% were NEs. For NC value range 99.20–10.05 total precision falls to 86% (73.12% terms and 12.92% NEs) and in the rest of the NC value distributions precision reaches similar rates (86%–77%). The results were manually assessed and term information was subsequently inserted in the corpus automatically whereas recognised NEs were classified and used to enrich the NE recognition tool's gazetteer.

For NE recognition we used the Basic Semantic Element Extraction (BSEE) component of the CONCERTO IE system (McNaught et al., 2000). In BSEE, NEs are identified and recognised by a combination of database look up and context-sensitive linguistic rules. These rules build up POS and semantic structural representations and identify instances of co-reference between names. The results of the analysis were manually inspected and a few minor modifications of the BSEE rules were made to improve the analysis. Subsequently, the results of NE, term and syntactic parser information were merged in a single corpus in XML format.

Initial corpus analysis at sentence level showed that a relatively small number of sentences per document do not contain any NEs or terms. In documents of 2–26 sentences the number of these sentences ranges from 0–5, whereas in longer documents (45–55 sentences) sentences without any of these semantic entities rarely number more than 10. These results show that the existence alone of semantic entities in such documents of highly informational content is not enough for the identification of important information. However, in other measurements based on the frequency of semantic entities per sentence, the results show that sentences such as titles and sentences in the beginning and at the end of the document tend to be richer in terms and NEs. Early studies in automatic abstracting have come to similar results regarding the informational weight of sentences in the document using keyword (as opposed to term) frequencies (Luhn, 1958).

Regarding the syntactic roles of terms and NEs, our analysis showed that their position can also be used to distinguish highly informational content from the rest of the text, as they reach relatively low percentages against the totality of the respective recognised syntactic categories. For example, NEs and terms constitute 32.67% of recognised subjects, where 12.35% are organisations, 12.17% terms and 3.36% persons. Person names very rarely appear in other than subject role (less than 0.3%). Organisation names tend to dominate syntactic roles such as subject and pre-modifier in genitive or apposition. Amounts, percentages and dates tend to appear as subject complements. Dates dominate adverbial modifiers (4.39% out of 7.14%) whereas products and artefacts mostly appear in appositions. Semantic entities in apposition role reach the highest percentage of syntactic categories (49.59% out of all recognised appositions).

In this stage of our research we are investigating significant co-occurrent patterns of main verbs, terms and NEs. To do this, we have modified the linguistic rules of the C/NC value tool to expand noun phrase search into sentence level search for multi-word patterns, where the previously identified multi-word NEs and terms have been collapsed into a single word-element and where verbs are also taken into consideration. The C-value statistical measurement is used for the identification of significant co-occurences, i.e. not merely frequent co-occurrences.

We have identified in this way a set of 4412 extraction patterns, which have been subsequently classified by the pattern main verb into 395 categories. Analysis of surface syntactic function phenomena, such as verb usage in passive form or in a specific tense, reveals regularities not only in the verb used but also in the associated semantic entities.

For example, verbs such as *report, announce, say* mostly appear in our corpus extraction patterns in present tense, in indicative form and associated with specific NEs: *report* and *announce* have typically an Organisation subject and *say* a Person, the former are also usually associated with a Term object and a Date.

Initial evaluation of these extracted patterns has been performed by mere pattern matching techniques at this stage. The results were compared to a manually annotated subset of the corpus and yielded Recall in the range 0.45–0.70 Recall and Precision in the range 0.66–0.40. More extensive evaluation based on extracted patterns by rule application has to be performed.

Conclusion

We have implemented a domain-independent methodology for acquisition of information extraction rule patterns based on the identification of significant associations of verb and basic semantic elements (NEs and terms). Initial results show that this approach can provide developers with generic extraction patterns for the informative content of a document, regardless of user-specific requirements and without use of extensive knowledge resources. Although the result of such a process is not readily applicable for user-specific IE purposes, we believe that the subsequent exploitation of such patterns will not only facilitate knowledge acquisition in the development stage of IE systems, but it will also provide a reliable training basis for the automation of template rule acquisition. Such work is especially valuable in support of the development of multi-user, multi-domain, large scale IE services of the type associated with the JISC supported UK National Text Mining Centre, housed at UMIST. Moreover, an investigation of the linguistic features of what constitutes important information, based on basic semantic elements, can be beneficial for other areas of natural language processing.

References

Bagga, A., J. Y. Chai & A. Biermann (1997). The Role of WordNet in the Creation of a Trainable Message Understanding System. In Proceedings of the 14th National Conference on Artificial Intelligence and the 9th Conference on the Innovative Applications of Artificial Intelligence (AAAI/IAAI'97), (pp. 941–48), July 1997.

Boguraev, B. & Kennedy, C. (1997). Salience-Based Content Characterisation of Text Documents. In Proceedings of ACL/EACL'97 Workshop on Intelligent Scalable Text Summarisation, (pp. 2–9). Madrid, Spain.

Brewster, C., Ciravegna, F. & Wilks, Y. (2002). User-Centred Ontology Learning for Knowledge Management. In Proceedings of the 7th International Conference on Applications of Natural Language to Information Systems, (pp.203–207). Stockholm, June 27–28, 2002, Lecture Notes in Computer Science 2553, Springer Verlag.

Califf, M. E. & R. J. Mooney (1999). Relational Learning of Pattern-Match Rules for Information Extraction. In Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99), (pp.328–334). Orlando, FL, July 1999.

Cavaglia, G. & Ciravegna, F. (1998) Combining Wordnet and Dewey Decimal Classification for Building Lexical Resources for Information Extraction from Text. In: Atti dell'Incontro del Gruppo di Lavoro sulla Rappresentazione della Conoscenza e Ragionamento Automatico su "Strumenti di Organizzazione e Accesso Intelligente per Informazioni Eterogenee", Padova, September 1998.

Ciravegna, F., Dingli, A., Petrelli, D. & Wilks, Y. (2002). User-System Cooperation in Document Annotation based on Information Extraction. In Proceedings of the Thirteenth International Conference on Knowledge Engineering and Knowledge Management (EKAW02), (pp.122–137). Siguenza, Spain, 1–4 October 2002.

DARPA (1998). Proceedings of the Seventh Message Understanding Conference (MUC-7). Defense Advanced Research Projects Agency, available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fontenelle, T. (ed.) (2003). FrameNet and Frame Semantics. Special issue of International Journal of Lexicography, 16(3).

Frantzi, K.T. & Ananiadou, S. (1999). The C-Value/NC-Value Domain Independent Method for Multi-Word Term Extraction. Journal of Natural Language Processing, 6(3): 145–180.

Harabagiu, S. & Maiorano, S. (2000). Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece, 31 May–2 June 2000.

Humphreys K., Demetriou, G. & Gaizauskas, R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. In Proceedings of the 5th Pacific Symposium on Biocomputing (PSB-2000), (pp.502–513). Honolulu, Hawaii, USA.

Luhn, H. P. (1958). The automatic creation of Literature abstracts. IBM Journal of Research and Development, 2(2): 159-165.

McNaught, J., Black, W. J., Rinaldi, F., Bertino, E., Brasher, A., Deavin, D., Catania, B., Silvestri, D., Armani, B., Leo, P., Persidis, A., Semeraro, G., Esposito, F., Zarri, G. P. & Gilardoni, L. (2000). Integrated Document and Knowledge Management for the Knowledge-based Enterprise. In Proceedings of the 3rd International Conference on the Practical Application of Knowledge Management (PAKeM2000) (pp. 89–108). Manchester, 12–14 April 2000.

Mikheev, A., Moens, M. & Grover, C. (1999). Named Entity Recognition without Gazetteers. In Proceedings of the 9th International Conference of the European Chapter of the Association for Computational Linguistics (EACL'99), (pp.1–8). Bergen, Norway, 8–12 June 1999.

Riloff, E. (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. In Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93), (pp.811–816). AAAI/MIT Press.

Riloff, E. (1996). An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. AI Journal 85(1–2):101–134, August 1996.

Soderland, S., Fisher, D., Aseltine, J. & Lehnert, W. (1995). CRYSTAL: Inducing a Conceptual Dictionary. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95), (pp.1314–1319). San Francisco: Morgan Kaufmann Publishers

Vasilakopoulos, A., Bersani, M. & Black, W.J. (2004). A Suite of Tools for Marking Up Textual Data for Temporal Text Mining Scenarios. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, 24th-30th May 2004. (to appear)

Voutilainen, A. (1995). A syntax-based part of speech analyser. In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, (EACL'95), (pp.157–164). Dublin, 27–31 March 1995.

Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. In Proceedings of the 6th Pacific Symposium on Biocomputing (PSB 2001), (pp. 408–419). Hawaii, U.S.A.

Yangarber, R., Grishman, R., Tapanainen, P. & Huttunen, S. (2000). Automatic Acquisition of Domain Knowledge for Information Extraction. In Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000), (pp. 940–946). Saarbrucken, Germany.