Automatic Extraction of Knowledge from Web Documents

Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal,
Paul H. Lewis, Wendy Hall, Nigel Shadbolt

I.A.M. Group, ECS Dept.
University of Southampton
Southampton, UK
{ha, sk, dem, mjw, phl, wh, nrs}@ecs.soton.ac.uk

Abstract. A large amount of digital information available is written as text documents in the form of web pages, reports, papers, emails, etc. Extracting the knowledge of interest from such documents from multiple sources in a timely fashion is therefore crucial. This paper provides an update on the Artequakt system which uses natural language tools to automatically extract knowledge about artists from multiple documents based on a predefined ontology. The ontology represents the type and form of knowledge to extract. This knowledge is then used to generate tailored biographies. The information extraction process of Artequakt is detailed and evaluated in this paper.

1 Introduction

Quick analysis and understanding of unstructured text is becoming increasingly important with the huge increase in the number of digital documents available. This has led to an increased use of various tools developed to help levy the problem of processing unstructured text documents through automatic classification, concept recognition, text summarisation, etc.

These tools are often based on traditional natural language techniques, statistical analysis, and machine learning, dealing mostly with single documents. The ability to extract certain types of knowledge from multiple documents and to maintain it in structured Knowledge Bases (KB) for further inference and report generation is a more complex process. This forms the aim of the Artequakt project.

1.1 Relation Extraction

There exist many information extraction (IE) systems that enable the recognition of entities within documents (e.g. ‘Renoir’ is a ‘Person’, ‘25 Feb 1841’ is a ‘Date’). However, such information is incomplete and sometimes insufficient for certain requirements without acquiring the relation between these entities (e.g. ‘Renoir’ was

born on ‘25 Feb 1841’). Extracting such relations automatically is difficult, but crucial to complete the acquisition of knowledge fragments and ontology population (building the KB). The MUC-7 systems [8] are example attempts for extracting a limited number of relations. Whereas the MUC participant systems used training examples to induce a set of rules for named-entity and relation extraction, Artequakt assumes a case where the number and type of relations to be extracted is non-static. Artequakt attempts to identify relations between the entities of interest within sentences, following ontology relation declarations and lexical information.

1.2 Ontology Population

Artequakt is also concerned with automating ontology population with knowledge triples, and providing this knowledge for a biography generation service.

When analysing documents and extracting information, it is inevitable that duplicated and contradictory information will be extracted. Handling such information is challenging for automatic extraction and ontology population approaches [16]. Artequakt applies a set of heuristics and reasoning methods in an attempt to distinguish conflicting information, verify it, and to identify and merge duplicate assertions in the KB automatically

1.3 Biography Generation

Storing information in a structured KB provides the needed infrastructure for a variety of knowledge services. One interesting service is to reconstruct the original source material in new ways, producing a dynamic presentation tailored to the users needs.

Previous work in this area has highlighted the difficulties of maintaining a rhetorical structure across a dynamically assembled sequence [14]. Where dynamic narrative is present it has been based around robust story-schema such as the format of a news programme (a sequence of atomic bulletins) [6].

It is our belief that by building a story-schema layer on top of an ontology we can create dynamic stories within a specific domain. In Artequakt we explore the generation of biographies of artists. Populating the ontology through automatic extraction tools might allow those biographies to be constructed from the vast wealth of information that exists on the World Wide Web, thus bringing together pieces of information from multiple sites into one single repository.

2 Artequakt

The Artequakt project has implemented a system that searches the Web and extracts knowledge about artists, based on an ontology describing that domain, and stores this knowledge in a KB to be used for automatically producing personalised biographies of artists. Artequakt draws from the expertise and experience of three separate

projects; *Sculpteur*[1], *Equator*[2], and *AKT*[3]. The main components of Artequakt are described in the following sections.

Artequakt's architecture comprises of three key areas. The first concerns the knowledge extraction tools used to extract factual information from documents and pass it to the ontology server. The second key area is information management and storage. The information is stored by the ontology server and consolidated into a KB which can be queried via an inference engine. The final area is the narrative generation. The Artequakt server takes requests from a reader via a simple Web interface. The request will include an artist and the style of biography to be generated (chronology, summary, fact sheet, etc.). The server uses story templates to render a narrative from the information stored in the KB using a combination of original text fragments and natural language generation.

The architecture is designed to allow different approaches to information extraction to be incorporated with the ontology acting as a mediation layer between the IE and the KB. Currently we are using textual analysis tools to scrape web pages for knowledge, but with the increasing proliferation of the semantic web, additional tools could be added that take advantage of any semantically augmented pages passing the embedded knowledge through the KB.

2.1 The Artequakt Ontology

For Artequakt the requirement was to build an ontology to represent the domain of artists and artefacts. The main part of this ontology was constructed from selected sections in the CIDOC Conceptual Reference Model (CRM[4]) ontology. The CRM ontology is designed to represent artefacts, their production, ownership, location, etc. This ontology was modified for Artequakt and enriched with additional classes and relationships to represent a variety of information related to artists, their personal information, family relations, relations with other artists, details of their work, etc. The Artequakt ontology and KB are accessible via an ontology server.

3 Knowledge Extraction

The aim of the knowledge extraction tool of Artequakt is to identify and extract knowledge triplets (concept – relation – concept) from text documents and to provide it as XML files for entry into the KB [5]. Artequakt uses an ontology coupled with a general-purpose lexical database (WordNet) [11] and an entity-recogniser (GATE) [3] as supporting tools for identifying knowledge fragments.

---

[1] http://www.sculpteurweb.org/

[2] http://www.equator.ac.uk/

[3] http://www.aktors.org

[4] http://cidoc.ics.forth.gr/index.html

3.1 Document Retrieval

The extraction process is launched when the user requests a biography for a specific artist that is not in the KB. A script was developed to query the artist's name in general-purpose search engines, such as Google and Yahoo.

Documents returned by the search engines need to be filtered to remove irrelevant ones. Expanding queries with additional terms was not very effective for improving the web search. The approach followed in Artequakt is based on query-by-example. In order to pick up pages related to an artist, a short description of the artist from a well-known museum web site (e.g. WebMuseum[5]) is analysed and used as a similarity vector. Structural evidence, such as paragraph length or number of sentences within a paragraph, is used in order to identify and remove pages which mainly consist of links, tables, etc. If the similarity vector is unobtainable (e.g. a search for a relatively new or unknown artist whose entry is not available in the exemplar museum site) the ontology itself is used to create the vector. The quality of entity recognition has a direct effect on the accuracy of relation extraction.

Search for documents stops and the extraction process starts when the number of relevant documents found reaches a specified threshold.

3.2 Entity Recognition

Entity recognition is the first step towards extracting knowledge fragments. GATE is a syntactical pattern matching entity recogniser enriched with gazetteers. GATE's coverage can be expanded with additional extraction rules and gazetteers to enable the identification of further type of entities. However, the process of discovering and setting up new syntactic rules can be difficult and labour intensive. To this end, we deploy WordNet as a supplementary information source in order to identify additional entities not recognised by the default GATE. WordNet is also used to support relation extraction.

3.3 Extraction Procedure

Each selected document is divided into paragraphs and sentences. Each sentence is analysed syntactically and semantically to identify and extract relevant knowledge triples. Figure 1 shows the overall procedure of the extraction process as applied on the sentence:

"Pierre-Auguste Renoir was born in Limoges on February 25, 1841."

[5] http://www.ibiblio.org/wm/

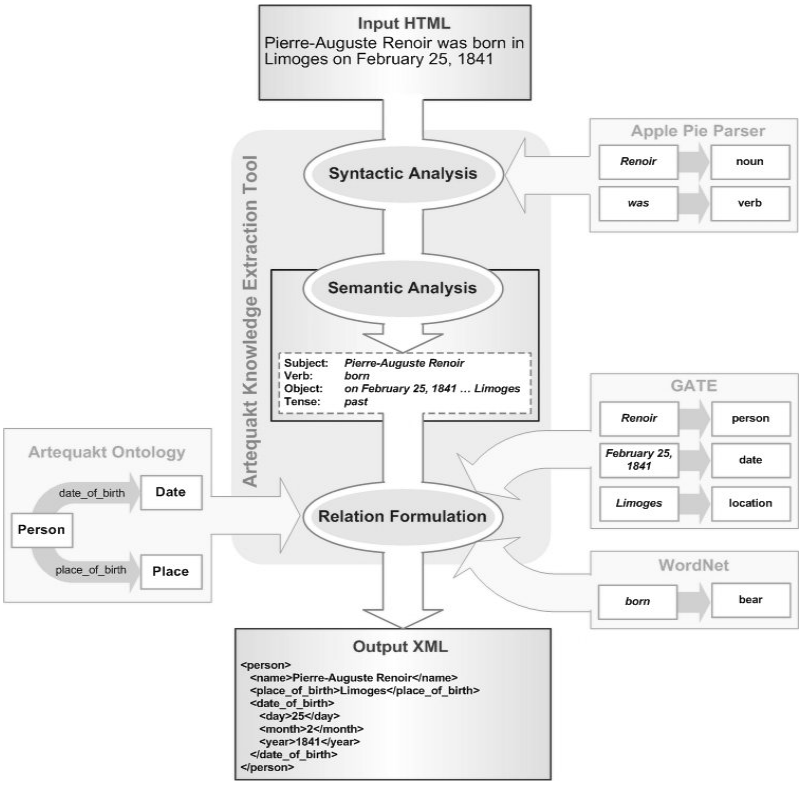


Fig. 1. Artequakt's IE Process

Syntactical Analysis. Syntactical parsing groups words into syntactic functions with no consideration to their semantic meaning. The Apple Pie Parser (APP) [15] is a bottom-up probabilistic chart parser derived from the syntactically tagged corpus; Penn Tree Bank (PTB). PTB contains a large number of example sentences; thus APP tends to have a broad-coverage performance with reasonable accuracy (over 70% for both precision and recall). The output of APP is structured according to the PTB bracketing.

Artequakt makes use of APP to gather syntactical annotations of sentences. For example in Figure 1, APP identified that 'Renoir' is a noun, and 'was' is a verb.

Semantic Analysis. Semantic examination then decomposes the sentence into simple sentences to locate the main components (i.e. subject, verb, object), and identifies named entities (e.g. "Renoir" is a Person, "Paris" is a Location). In the example sentence of Figure 1, "born" is tagged as the main verb in the "was born" verb phrase.

Annotations provided by GATE and WordNet highlight that "Pierre-Auguste Renoir" is a person's name, "February 25, 1841" is a date, and "Limoges" is a

location. GATE is also used to resolve anaphoric references of singular personal pronouns which is crucial for accurate relation extraction.

Term expansion tools are required if the terms identified by the named-entity recogniser differ from those in the ontology. For example, GATE annotates 'Museum of Art' as an Organisation while our ontology defines 'Legal Body' as a general concept for organisations. The system needs to map these two concepts to figure out that 'Museum of Art' is a Legal Body. Currently, we use WordNet for the mapping by looking up the lexical chains of the two terms in search of any overlap.

Relation Extraction. Artequakt is concerned with the extraction of relations between concepts within individual sentences. The aim is to extract relationships between any identified pair of entities within a given sentence. Knowledge about the domain specific semantic relations can be retrieved from the Artequakt ontology to find which relations are expected between the entities in hand.

Relations are extracted by matching the verb and entity pairs found in each sentence with an ontology relation and concept pairs respectively. Three lexical chains (synonyms, hypernyms, and hyponyms) from WordNet are used to expand entity names with related terms to reduce the problem of linguistic variations and increase the chance of matching with other semantically similar terms.

Since a relation may have multiple matchings in WordNet (polysemous words), mapping between a term and an entry in WordNet should into account syntactic and semantic clues present in the given sentence. For example, according to WordNet, 'birth' has four noun senses and one verb sense. The first noun sense is selected since one of its hypernyms is 'time period' which has Date as a hyponym.

For the sentence used in Figure 1, the relation extraction is determined by the categorisation result of the main verb 'bear' which matches with two potential relations in the ontology; 'date_of_birth' and 'place_of_birth'. Since both relations are associated with "February 25, 1841" (Date) and "Limoges" (Place) respectively. After analysing the given sentence, Artequakt generates the following knowledge triples about Renoir:

- `Pierre-Auguste Renoir` *`date_of_birth`* `25/2/1841`
- `Pierre-Auguste Renoir` *`place_of_birth`* `Limoges`

The extraction process terminates by sending the extracted knowledge to the ontology server in XML.

4 Automatic Ontology Population

Storing knowledge extracted from text documents in KBs offers new possibilities for further analysis and reuse. Ontology population refers to the insertion of information into the KB. Populating ontologies with a high quantity and quality of instantiations is one of the main steps towards providing valuable and consistent ontology-based knowledge services. Manual ontology population is very labour intensive and time consuming. A number of semi-automatic approaches have investigated creating document annotations and storing the

results as ontology assertions. MnM [17] and S-CREAM [4] are two example frameworks for user-driven ontology-based annotations, enforced with the IE learning tool; Amilcare [2]. However, these frameworks lack the capability for identifying relationships reliably.

In Artequakt we investigate the possibility of moving towards a fully automatic approach of feeding the ontology with knowledge extracted from unstructured text. Information is extracted in Artequakt with respect to a given ontology and provided as XML files using tags mapped directly from names of classes and relationships in that ontology. When the ontology server receives a new XML file, a *feeder* tool is activated to parse the file and add its knowledge triples to the KB automatically. Once the feeding process terminates, the consolidation tool searches for and merges any duplication in the KB.

Tackling the problem of knowledge integration is important to maintain the referential integrity and quality of results of any ontology-based knowledge service. [16] relied on manually assigned object identifiers to avoid duplication when extracting from multiple documents. Artequakt's knowledge management component attempts to identify inconsistencies and consolidate duplications automatically using a set of heuristics and term expansion methods based on WordNet. Full description of the consolidation procedure is out of the scope of this paper.

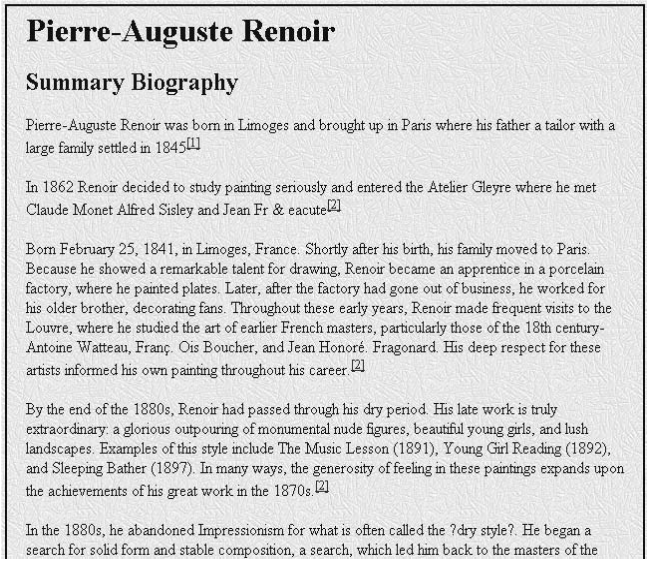
5 Biography Generation

Once the information has been extracted, stored and consolidated, the Artequakt system repurposes it by automatically generating biographies of the artists [1]. The biographies are based on templates authored in the Fundamental Open Hypermedia Model and stored in the Auld Linky contextual structure server [10]. Each section of the template is instantiated with paragraphs or sentences generated from information in the KB.

Different templates can be constructed for different types of biography. Two examples are the summary biography, which provides paragraphs about the artist arranged in a rough chronological order, and the fact sheet, which simply lists a number of facts about the artist, i.e. date of birth, place of study etc. The biographies also take advantage of the structure server's ability to filter the template based on a user's interest. If the reader is not interested in the family life of the artist the biography can be tailored to remove such information. An example of a biography generated by Artequakt can be seen in Figure 2.

By storing conflicting information rather than discarding it during the consolidation process, the opportunity exists to provide biographies that set out arguments as to the facts (with provenance, in the form of links to the original sources) by juxtaposing the conflicting information and allowing the reader to make up their own mind.

As well as searching the KB by name the user interface provides a search facility that allows users to select artists according to other extracted facts, for example the user can specify a range for date of birth and the system will search the appropriate fields with the correct constraints. This kind of query can not be easily formulated over the Web. Extracting the relevant knowledge and storing it in a KB made such queries more feasible.



Pierre-Auguste Renoir

Summary Biography

Pierre-Auguste Renoir was born in Limoges and brought up in Paris where his father a tailor with a large family settled in 1845[1]

In 1862 Renoir decided to study painting seriously and entered the Atelier Gleyre where he met Claude Monet Alfred Sisley and Jean Fr & eacute[2]

Born February 25, 1841, in Limoges, France. Shortly after his birth, his family moved to Paris. Because he showed a remarkable talent for drawing, Renoir became an apprentice in a porcelain factory, where he painted plates. Later, after the factory had gone out of business, he worked for his older brother, decorating fans. Throughout these early years, Renoir made frequent visits to the Louvre, where he studied the art of earlier French masters, particularly those of the 18th century-Antoine Watteau, Franç. Ois Boucher, and Jean Honoré. Fragonard. His deep respect for these artists informed his own painting throughout his career.[2]

By the end of the 1880s, Renoir had passed through his dry period. His late work is truly extraordinary: a glorious outpouring of monumental nude figures, beautiful young girls, and lush landscapes. Examples of this style include The Music Lesson (1891), Young Girl Reading (1892), and Sleeping Bather (1897). In many ways, the generosity of feeling in these paintings expands upon the achievements of his great work in the 1870s.[2]

In the 1880s, he abandoned Impressionism for what is often called the ?dry style?. He began a search for solid form and stable composition, a search, which led him back to the masters of the

Fig. 2. A biography generated using paragraphs.

6 Portability

The use of an ontology to back up IE is aimed to increase the system's portability to other domains. By swapping the current artist ontology with another domain specific one, the IE tool should still be able to function and extract some relevant knowledge, especially if it is concerned with domain independent relations expressed in the ontology, such as personal information. However, certain domain specific extraction rules, such as painting style, will eventually have to be altered to fit the new domain.

Similarly, the generation templates are currently manually tuned to fit biography construction. These templates may need to be modified if a different type of output is required. We aim to investigate developing templates that can be dynamically instructed and modified by the ontology. Building a cross-domain system is one of the ambitions of this project, and will be the focus of the next stage of development.

7 Knowledge Extraction Evaluation

We used the system to populate the KB with information about five artists, extracted from around 50 web pages. Precision and recall were calculated for a set of 10 artist relations (listed in Table 1). The experiment results given in Table 1 shows that precision scored higher that recall with average values of 85 and 42 respectively

Inaccurately extracted knowledge may reduce the quality of the system's output. For this reason, our extraction rules were designed to be of low risk levels to ensure higher extraction precision. Advanced consistency checks could help to identify some extraction inaccuracies; e.g. a date of marriage is before the date of birth, or two unrelated places of birth for the same person!

Table 1. Precision/Recall of extracted relations from around 50 documents for 5 artists

Artist (P/R) / Relation	Rembrandt (P/R)	Renoir (P/R)	Cassatt (P/R)	Goya (P/R)	Courbet (P/R)	Average per relation
Date of birth	75/43	100/50	100/67	80/40	100/100	91/60
Place of birth	100/63	100/14	100/50	100/40	100/63	100/46
Date of death	100/63	100/67	100/50	N/A /0	100/50	100/46
Place of death	100/100	100/43	N/A /0	100/20	100/33	100/49
Place of work	100/50	67/33	33/100	N/A /0	0/0	40/37
Place of study	100/20	100/14	100/75	100/20	100/29	60/32
Date of marriage	100/50	100/33	N/A[1]	100/100	N/A /0	60/46
Name of spouse	100/38	N/A /0	N/A	N/A /0	N/A /0	100/10
Parent profession	100/57	50/67	0/0	67/100	100/100	63/65
Inspired by	100/43	50/60	0/0	100/17	100/33	83/31
Averages	98/53	85/38	61/43	92/34	88/41	**85/42**

The preference of precision versus recall could be dependent on the relation in question. If a relation is of single cardinality, such as a place of birth, then recall could be regarded as less significant as there can only be one value for each occurrence of this relation. A single accurate capture of the value of such a relation could therefore be sufficient for most purposes. However, multiple cardinality relations, such as places where a person worked, can have several values. Higher recall in such cases could be more desirable to ensure capturing multiple values. One possible approach is to automatically adjust the risk level of extraction rules with respect to cardinality, easing the rules if cardinality is high while restricting them further when the cardinality is low.

In Table 1, Goya is an example where few, short documents where found. The amount of knowledge extracted per artist could be used as an automatic trigger to start gathering and analyzing more documents.

8 Related Work

Extracting information from web pages to generate various reports is becoming the focus of much research. The closest work we found to Artequakt is in the area of text

summarisation. A number of summarisation techniques have been developed to help bring together important pieces of information from documents and present them to the user in a compact form. Artequakt differs from such systems in that it aims to extract specific facts and populate a knowledge base with these facts to be used in the generation of personalised reports (e.g. biographies).

Even though most summarisation systems deal with single documents, some have targeted multiple resources [9][18]. Statistical based summarisations tend to be domain independent, but lack the sophistication required for merging information from multiple documents [12]. On the other hand, IE based summarisations are more capable of extracting and merging information from various resources, but due to the use of IE, they are often domain dependent.

Merging information extracted from single or multiple sources is a necessary step towards maintaining the integrity of the extracted knowledge. In many existing IE based systems, information integration is based on linguistics and timeline comparison of single events [12][18] or multiple events [13]. Artequakt's knowledge consolidation is based on the comparison and merging of not just events, but also individual knowledge fragments (e.g. person, place).

Most traditional IE systems are domain dependent due to the use of linguistic rules designed to extract information of specific content, e.g. bombing events (MUC systems), earthquake news [18], sports matches [13]. Adaptive IE systems [2] caould ease this problem by identifying new extraction rules induced from example annotations supplied by users. Using ontologies to back up IE is hoped to support information integration [1][13] and increase domain portability [5][7].

9 Conclusions

This paper describes a system that extracts knowledge automatically, populates an ontology with knowledge triples, and reassembles the knowledge in the form of biographies. Initial experiment using around 50 web pages and 5 artists showed promising results, with nearly 3 thousand unique knowledge triples were extracted, with an average of 85% precision and 42% recall. Preference of precision over recall is subjective and should be associated with relations' cardinality. High precision could be more important for single cardinality relations (e.g. date of birth), while high recall could be preferred for multiple cardinality relations (e.g. places visited).

Future work on Artequakt will continue to develop its modular architecture and refine its information extraction and consolidation processes. In addition we are beginning to look at how we might leverage the full power of the underlying KB and produce biographies that use inference and a dynamic choice of templates to answer a variety of user queries with textual documents. We also intend to investigate the system's portability to other domains.

Acknowledgements

This research is funded in part by EU Framework 5 IST project "Scultpeur" IST-2001-35372, EPSRC IRC project "Equator" GR/N15986/01 and EPSRC IRC project "AKT" GR/N15764/01

References

1. Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., and Shadbolt, N. "Automatic Ontology-based Knowledge Extraction from Web Documents". *IEEE Intelligent Systems*, 18(1), pages 14-21, 2003.
2. Ciravegna, F. "Adaptive Information Extraction from Text by Rule Induction and Generalisation". *Proc.17th Int. Joint Con. on AI (IJCAI)*, pp 1251--1256, Seattle, USA, 2001.
3. Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. "GATE: a framework and graphical development environment for robust NLP tools and applications". *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Phil.,USA, 2002.
4. Handschuh, S., Staab, S., and Ciravegna, F. "S-CREAM – Semi Automatic Creation of Metadata". *Semantic Authoring, Annotation and Markup Workshop, 15th European Conf. on Artificial Intelligence*, pages 27--33, Lyon, France, 2002.
5. Kim, S., Alani, H., Hall, W., Lewis, P.H., Millard, D.E., Shadbolt, N., and Weal, M.J. "Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web". *Workshop on Semantic Authoring, Annotation & Knowledge Markup, 15th European Conf. on Artificial Intelligence (ECAI)*, pages 1--6, Lyon, France, July 2002.
6. Lee, K., D. Luparello, and J. Roudaire, "Automatic Construction of Personalised TV News Programs," *Proc. 7th ACM Conf. on Multimedia*, Orlando, Florida, 1999, pp. 323-332.
7. Maedche, A., G. Neumann and S. Staab. Bootstrapping an Ontology-based Information Extraction System. Intelligent Exploration of the Web. Springer 2002.
8. Marsh, E. & D. Perzanowski (NRL), MUC-7 Evaluation of IE Technology: Overview of Results, available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html
9. McKeown, K. R., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman and S. Sigelman. "Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster". Proc. Human Language Technology Conf., CA, USA. 2002.
10. Michaelides, D.T., Millard, D.E., Weal, M.J., and DeRoure, D. "Auld Leaky: A Contextual Open Hypermedia Link Server". *Proc. of the 7th Hypermedia: Openness, Structural Awareness, and Adaptivity*, pages 59--70, Springer Verlag, Heidelberg, 2001, LNCS.
11. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. "Introduction to wordnet: An on-line lexical database". *Int. Journal of Lexicography*, 3(4):235--312, 1993.
12. Radev, D. R. and K. R. McKeown. "Generating natural language summaries from multiple on-line sources." Compuutational Linguistics 24(3): 469—500, 1998.
13. Reidsma, D., J. Kuper, T. Declerck, H. Saggion and H. Cunningham. Cross document annotation for multimedia retrieval. EACL Workshop on Language Technology and the Semantic Web (NLPXML), Budapest, Hungary, 2003.
14. Rutledge, L., B. Bailey, J.V. Ossenbruggen, L. Hardman, and J. Geurts, "Generating Presentation Constraints from Rhetorical Structure," *Proc. 11th ACM Conf. on Hypertext and Hypermedia*, San Antonio, Texas, USA, 2000, pp. 19-28.
15. Sekine, S. and Grishman R., "A corpus-based probabilistic grammar with only two non-terminals", Proc. of the 1st Int. Workshop on Multimedia annotation, Japan, 2001.
16. Staab, S., Maedche, A., and Handschuh, S. "An Annotation Framework for the Semantic Web". Proc. 1st Int. Workshop on MultiMedia Annotation, Tokoyo, Japan, January 2001.
17. Vargas-Vera, M., E. Motta, J. Domingue, M. Lanzoni, A. Stutt and F. Ciravegna. "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup". 13th Int. Conf on Knowledge Engineering and Management (EKAW 02), Spain, 2002.
18. White, M., T. Korelsky, C. Cardie, V. Ng, D. Pierce and K. Wagstaff. Multidocument Summarization via Information Extraction. Proc. of Human Language Technology Conf. (HLT 2001), San Diego, CA, 2001.