Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation

Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal

Wendy Hall, Paul H. Lewis, Nigel Shadbolt

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

ABSTRACT

The Web is probably the largest and richest information repository available today. Search engines are the common access routes to this valuable source. However, the role of these search engines is often limited to the retrieval of lists of potentially relevant documents. The burden of analysing the returned documents and identifying the knowledge of interest is therefore left to the user. The Artequakt system aims to deploy natural language tools to automatically extract and consolidate knowledge from web documents and instantiate a given ontology, which dictates the type and form of knowledge to extract. Artequakt focuses on the domain of artists, and uses the harvested knowledge to generate tailored biographies. This paper describes the latest developments of the system and discusses the problem of knowledge consolidation.

Categories and Subject Descriptors

I.2.6 Learning – *Knowledge acquisition*
I.2.7 Natural Language Processing – *Text analysis, Language parsing and understanding*

Keywords

Information Extraction, Ontology Instantiation, and Knowledge Consolidation.

INTRODUCTION

Web pages are the source of vast amounts of knowledge. This knowledge is often buried by layers of text and scattered over numerous sites. Associating web pages with annotations to identify their knowledge content is the ambition of the Semantic Web [3]. Much research is now focused on developing ontologies to manipulate this knowledge and provide a variety of knowledge services. Automatic instantiation of ontologies and building knowledge bases (KB) with knowledge extracted from the web corpus is therefore very beneficial. Artequakt is concerned with automating ontology instantiation with knowledge triples (subject - relation - object) about the life and work of artists, and providing this knowledge for biography generation services.

When analysing and extracting information from multi sourced documents, it is inevitable that duplicated and contradictory information will be extracted. Handling such information is challenging for automatic extraction and ontology instantiation approaches [18]. Artequakt applies a set of heuristics and reasoning methods in an attempt to distinguish conflicting information, to verify it, and to identify and merge duplicate assertions in the KB automatically.

This paper describes the main components of the Artequakt system, focusing on the latest development with respect to knowledge consolidation and ontology instantiation.

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 the area of text summarisation. A number of summarisation techniques have been described to help bring together important pieces of information from documents and present them to the user in a compact form.

Even though most summarisation systems deal with single documents, some have targeted multiple resources [12][23]. Statistical based summarisations tend to be domain independent, but lack the sophistication required for merging information from multiple documents [17]. On the other hand, Information Extraction (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.

Radev developed the SUMMONS system [17] to extract information and generate summaries of individual events from MUC (Message Understanding Conferences) text corpuses. The system compares information extracted from

K-CAP'03, October 23-25, 2003, Sanibel Island, FL, USA.

multiple resources, merges similar content and highlights contradictions. However, like most IE based systems; information merging is often based on linguistics and timeline comparison of single events [17][23] or multiple events [18].

Artequakt's knowledge consolidation is based on the comparison of individual knowledge fragments, rather than linguistic analyses or timeline comparison. Furthermore, Artequakt's consolidation is more fine-grained, focusing on the comparison and merging of individual entities (e.g. places, people, dates).

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 [23], sports matches [18]). Adaptive IE systems [4] can ease this problem by identifying new extraction rules induced from example annotations supplied by users. However, training such tools can be difficult and time consuming. Promising results are offered by more advanced adaptive IE tools, such as Armadillo [6], which discovers new linguistic and structural patterns automatically, thus requiring limited bootstrapping.

Using ontologies to back up IE is hoped to support information integration [2][18] and increase domain portability [10][11]. Poibeau [16] investigated increasing domain independency by using clustering methods on text corpuses to aid users construct primitive ontologies to represent the main corpus topics. Templates could then be generated from the ontology and guide the IE process. Ontologies produced by this approach are limited to the content of the corpus, rather than representing a specific domain. In some cases (such as in Artequakt) the corpus is very large and diverse (e.g. the Web). Creating ontologies from such corpus is infeasible. Furthermore, these ontologies are likely to be rough, shallow, and include undesired concepts that happen to be in the text corpus. Consequently, the cost of bringing such ontologies to shape might exceed the benefit.

Instantiating ontologies with assertions from textual documents can be a very laborious task. A number of tools have been developed that instantiate ontologies semi automatically with user driven annotations [20]. IE learning tools, such as Amilcare [4], can be used to automate part of the annotation process and speed up ontology instantiation [7][21].

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.

System Overview

Figure 1 illustrates Artequakt's architecture which 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.

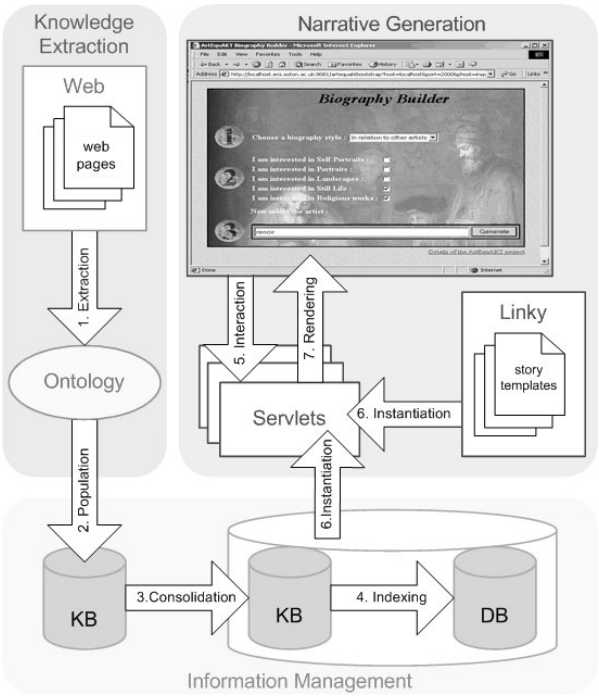

Figure 1. The Artequakt Architecture

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, addi-

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

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

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

tional tools could be added that take advantage of any semantically augmented pages passing the embedded knowledge through the KB.

As well as keeping open the interface between the KB and the extraction technology, a clear separation has been kept between the creation of a structured document from the knowledge base and the rendering of that document. In the current system, the information is rendered into an HTML page but alternative-rendering engines could be envisaged. For example, rather than presenting the biography as a linear textual document, the information might be rendered into a dynamic presentation system such as SMIL, converted into an audio stream using text to speech tools, or perhaps used to generate a dynamic hypertext with links referring back to queries to the KB on items such as artists names.

```
<kb:Person rdf:about="&kb;Person_1"
        kb:name="Pierre-Auguste Renoir"
        rdfs:label="Person_1">
        <kb:date_of_birth rdf:resource=
        "&kb;Date_1"/>
        <kb:place_of_birth rdf:resource=
        "&kb;Place_1"/>
        <kb:has_father rdf:resource=
        "&kb;Person_2"/>
        <kb:has_information_text rdf:resource=
        "&kb;Paragraph_1"/>
</kb:Person>
<kb:Date rdf:about="&kb;Date_1"
        kb:day="25"
        kb:month="2"
        kb:year="1841"
        rdfs:label="Date_1">
</kb:Date>
<kb:E53.Place rdf:about="&kb;Place_1"
        kb:name="Limoges"
        rdfs:label="Place_1"/>
<kb:Person rdf:about="&kb;Person_2"
        rdfs:label="Person_2">
        <kb:has_work_information rdf:resource=
        "&kb;Work_information_1"/>
</kb:Person>
<kb:Work_information rdf:about=
"&kb;Work_information_1"
        kb:job_title="tailor"
        rdfs:label="Work_information_1">
</kb:Work_information>
```

Figure 2. RDF representation of knowledge extracted from the paragraph: "Pierre-Auguste Renoir was born in Limoges on February 5, 1841. His father was a tailor."

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.

KNOWLEDGE EXTRACTION

The aim of our knowledge extraction tool is to identify and extract knowledge triples from text documents and to provide it as RDF files for entry into the KB [10]. Artequakt uses an ontology coupled with a general-purpose lexical database (WordNet) [14] and an entity-recogniser (GATE) [5] as guidance tools for identifying knowledge fragments.

Artequakt attempts to identify not just entities, but also their relationships following ontology relation declarations and lexical information.

Extraction Procedure

The extraction process is launched when the user requests a biography for a specific artist that is not in the KB. The query is passed to selected web search engines and the search results are analysed with respect to relevancy to the domain of artists.

Each selected document is then divided into paragraphs and sentences. Each sentence is analysed syntactically and semantically to identify any relevant knowledge to extract. Below is an example of an extracted paragraph:

> "Pierre-Auguste Renoir was born in Limoges on February 25, 1841. His father was a tailor and his mother a dressmaker. "

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. Relation extraction is determined by the categorisation result of the 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*' and '*Limoges*' respectively, this sentence generates the following knowledge triples about Renoir:

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

The second sentence generates knowledge triples related to Renoir's family:

`Pierre-Auguste Renoir` *`has_father`* `Person_2`

- `Person_2` *`job_title`* `Tailor`
- `Pierre-Auguste Renoir` *`has_mother`* `Person_3`
- `Person_3` *`job_title`* `Dressmaker`

Inaccurately extracted knowledge may reduce the quality of the system's output. For this reason, our extraction

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

rules were designed to be of low risk levels to ensure higher extraction precision. Advanced consistency checks can help 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!

The extraction process terminates by sending the extracted knowledge to the ontology server. Figure 2 is the RDF representation of the extracted knowledge. Artequakt's IE process is out of the scope of this paper, and is fully described in [2] and [10].

BIOGRAPHY GENERATION

Once the information has been extracted, stored and consolidated, the Artequakt system repurposes it by automatically generating biographies of the artists. Figure 3 shows a biography of Renoir.

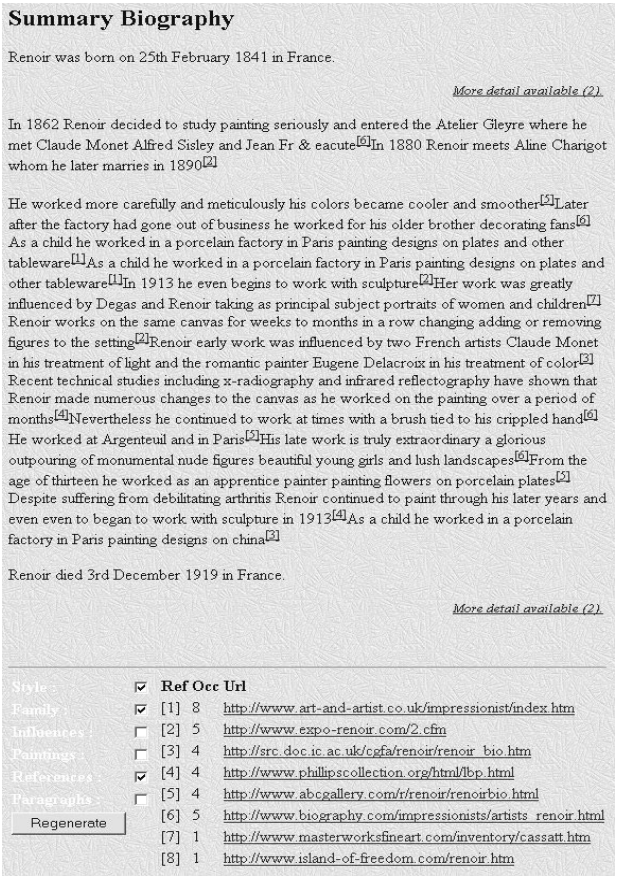

Figure 3. A Biography Generated Using Sentences.

The biographies are based on templates authored in the Fundamental Open Hypermedia Model (FOHM) and stored in the Auld Linky contextual structure server [13]. Each section of the template is instantiated with paragraphs or sentences generated from information in the KB. The KB informs the templates of the *theme* of the sentences and paragraphs (e.g. influences, family info, painting) and the generation tool select the relevant ones and structure them in the desired form and order. Very little text generation is used in the current implementation (e.g. Figure 3, 1st and last sentences), but this will be the focus of the next phase.

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.

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 this information.

More about Artequakt's biography generation is available at [14].

AUTOMATIC INSTANTIATION

Storing knowledge extracted from text documents in KBs offers new possibilities for further analysis and reuse. *Ontology instantiation* refers to the insertion of information into the KB, as described by the ontology (sometimes referred to as *ontology population*). Instantiating ontologies with a high quantity and quality of knowledge is one of the main steps towards providing valuable and consistent ontology-based knowledge services. Manual ontology instantiation is very labour intensive and time consuming. Some semi-automatic approaches have investigated creating document annotations and storing the results as assertions [7][20][21]. [7] and [20] describe two frameworks for user-driven ontology-based annotations, enforced with the IE learning tool; Amilcare [3]. However, the two frameworks are manually driven and mainly focus on entity annotations. They lack the capability of identifying relationships reliably. In [20], relationships were added automatically between instances, *but* only if these instances already existed in the KB, otherwise user intervention is required.

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 RDF or XML files using tags mapped directly from names of classes and relationships in that ontology. When the ontology server receives a new RDF file, a *feeder* tool is activated to parse the file and adds its knowledge triples to the KB automatically. Once the feeding process terminates, the consolidation tool searches for and merges any duplication in the KB.

KNOWLEDGE BASE CONSOLIDATION

Automatically instantiating an ontology from diverse and distributed resources poses significant challenges. One persistent problem is that of the consolidation of duplicate information that arises when extracting similar or overlapping information from different sources. Tackling this problem is important to maintain the referential integrity and quality of results of any ontology-based knowledge service. [18] relied on manually assigned object identifiers to avoid duplication when extracting from different documents.

Little research has looked at the problem of information consolidation in the IE domain. This problem becomes more apparent when extracting from multiple documents. Comparing and merging extracted information is often based on domain dependent heuristics [17] [18] [23]. Our approach attempts to identify inconsistencies and consolidate duplications automatically using a set of heuristics and term expansion methods based on WordNet [22].

Duplicate Information

There exist two main type of duplication in our KB; duplicate instances (e.g. multiple instance representing the same artist), and duplicate attribute values (e.g. multiple dates of birth extracted for the same artists).

Artequakt's IE tool treats each recognised entity (e.g. Rembrandt, Paris) as a new instance. This may result in creating instances with overlapping information (e.g. two Person instances with the same name and date of birth). The role of consolidation in Artequakt includes analysing and comparing attribute values of the instances of each type of concept in the KB (e.g. Person, Date) to identify inconsistencies and duplications.

The amount of overlap between the attribute values of any pair of instances could indicate their duplication potential. However, this overlap is not always measurable. IE tools are sometimes only able to extract fragments of information about a given entity (e.g. an artist), especially if the source document or paragraph is small or difficult to analyse. This leads to the creation of new instances with only one or two facts associated with each. For example two artist instances with the name Rembrandt, where one instance has a location relationship to Holland, while the other has a date of birth of 1606. Comparing such *shallow* instances will not reveal their duplication potential. Furthermore, neither the source information nor the information extraction is always accurate. For example a Rembrandt instance can be extracted with the correct family attribute values, but with the wrong date of birth, in which case this instance will be mismatched with other Rembrandt instances in spite of referring to the same artist.

Unique Name Assumption

One basic heuristic applied in Artequakt is that artist names are unique; where artist instances with identical names are merged. According to this heuristic, all instances with the name Rembrandt are combined into one instance. This heuristic is obviously not fool proof, but it works well in the limited domain of artists.

Information Overlap

There are cases where the full name of an artist is not given in the source document or its extraction fails, in which case they will *not* be captured by the unique-name heuristic. For example, when we extracted information about Rembrandt and merged same-name artists, two instances remained for this artist; *Rembrandt* and *Rembrandt Harmenszoon van Rijn*. In such a case we compare certain attribute values, and merge the two instances if there is sufficient overlap. For the two Rembrandt instances, both had the same date and place of birth, and therefore were combined into one instance. The duplication would have not been caught if these attributes had different values.

Attribute Comparison

When the above heuristics are applied, merged instances might end up having multiple attribute values (e.g. multiple dates and places of birth), which in turn need to be analysed and consolidated. Note that some of these attributes might hold conflicting information that should be verified and held for future comparison and use.

Comparing the values of instance attributes is not always straightforward as these values are often extracted in different formats and specificity levels (e.g. synonymous place names, different date styles) making them harder to match. Artequakt applies a set of heuristics and expansion methods in an attempt to match these values. Consider the following sentences:

1. *Rembrandt was born in the 17th century in Leyden.*
2. *Rembrandt was born in 1606 in Leiden, the Netherlands.*
3. *Rembrandt was born on July 15 1606 in Holland.*

These sentences provide the same information about an artist, written in different formats and specificity levels. Storing this information in the KB in such different formats is confusing for the biography generator which can benefit from knowing which information is repetitive and which is contradictory. Matching the above sentences required enriching the original ontology with some temporal and geographical reasoning.

Geographical Consolidation

There has been much work on developing gazetteers of place names, such as the Thesaurus of Geographic Names (TGN) [8] and Alexandria Digital Library [9]. Ontologies can be integrated with such sources to provide the necessary knowledge about geographical hier-

archies, place name variations, and other spatial information [1]. Artequakt derives its geographical knowledge from WordNet [14]. WordNet contains information about geopolitical place names and their hierarchies, providing three useful relations for the context of Artequakt; synonym, holonym (part of), and part_meronym (sub part). The Artequakt ontology is extended to add this information for each new instance of place added to the KB.

Place Name Synonyms

The synonym relationship is used to identify equivalent place names. For example the three sentences above mention several place names were Rembrandt was born. Using the synonym relationship in WordNet, *Leyden* can be identified as a variant spelling for *Leiden*, and that *Holland* and *The Netherlands* are synonymous.

Place Specificity

The part-of and sub-part relationships in WordNet are used to find any hierarchical links between the given places. WordNet shows that *Leiden* is part of the *Netherlands*, indicating that *Leiden* is the more precise information about Rembrandt's place of birth.

Shared Place Names

It is common for places to share the same name. For example according to the TGN, there are 22 places worldwide named *London*. This problem is less apparent with WordNet due to its limited geographical coverage.

In Artequakt, disambiguation of place names is dependent on their specificity variations. For example after processing the three sentences about Rembrandt, it becomes apparent that he was born in a place named Leiden in the Netherlands. If the last two sentences were not available, it would have not been possible to tell for sure which Leiden is being referred to (assuming there is more than one). One possibility is to rely on other information, such as place of work, place of death, to make a disambiguation decision. However, this is likely to produce unreliable results.

Temporal Consolidation

Dates need to be analysed to identify any inconsistencies and locate precise dates to use in the biographies. Simple temporal reasoning and heuristics can be used to support this task.

Artequakt's IE tool can identify and extract dates in different formats, providing them as day, month, year, decade, etc. This requires consolidation with respect to precision and consistency. Going back to our previous example, to consolidate the first date (17th century), the process checks if the years of the other dates fall within the given century. If this is true, then the process tries to identify the more precise date. The date in the third sentence is favoured over the other two dates as they are all consistent, but the third date holds more information than the other two. Therefore, the third date is used for the instance of Rembrandt. If any of the given facts is inconsistent then it will be stored for future verification and use.

At the end of the consolidation process, the knowledge extracted from the three sentences above will be stored in the KB as the following two triples for the instance of Rembrandt:

- *Rembrandt date_of_birth 15 July 1606*
- *Rembrandt place_of_birth Leiden*

Inconsistent Information

Some of the extracted information can be inconsistent, for example an artist with different dates or places of birth or death, or inconsistent temporal information, such as a date of death that falls before the date of birth. The source of such inconsistency can be the original document itself, or an inaccurate extraction. Predicting which knowledge is more reliable is not trivial. Currently we rely on the frequency in which a piece of knowledge is extracted as an indicator of its accuracy; the more a particular piece of information is extracted, the more accurate it is considered to be. For example, for Renoir, two unique dates of births emerged; 25 Feb 1841 and 5 Feb 1841. The former date has been extracted from several web sites, while the latter was found in one site only, and therefore considered to be less reliable.

A more advanced approach can be based on assigning levels of trust for each extracted piece of knowledge, which can be derived from the reliability of the source document, or the confidence level of the extraction of that particular information. The knowledge consolidation process is not aimed at finding 'the right answers' however. The facts extracted are stored for future use, with references to the original material.

PORTABILITY TO OTHER DOMAINS

The use of an ontology to back up IE is meant 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 (name, date and place of birth, family relations, etc). However, some domain specific extraction rules, such as painting style, will eventually have to be retuned to fit the new domain.

Similarly, the generation templates are currently manually set for 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.

Consolidation is often based on domain dependent heuristics. However, some of the heuristics used in Artequakt can be suitable for other domain. For example, Artequakt's approach for comparing and integrating place names using external gazetteers can be used in any domain. Similarly, heuristics concerning the comparison of specific facts to decide whether or not two instances of people are duplicates is also domain independent. Further work is planned to extend the scope of information integration

Building a cross-domain system is one of the aims of this project, and will be fully investigated in the next stage of development.

EVALUATION

We used the system to instantiate the KB with information on five artists, extracted from around 50 web pages.

Extraction Performance

Precision and recall were calculated for a set of 10 artist relations (about birth, death, places where they worked or studied, who influenced them, professions of their parents, etc). Results showed that precision scored higher than recall with average values of 85 and 42 respectively. The experiment is more detailed in [2].

Biography Evaluation

Although we have not conducted any formal evaluation of the biographies generated by the system, we are in the position to make a few observations. In general we found that the system is fairly successful in reproducing text for a given artist. We are currently looking at how best to perform a qualitative evaluation of the biographies, perhaps with a task-based user evaluation, comparing the Artequakt system with a traditional search engine.

Consolidation Rate

Table 1 shows the reduction rate in number of instances and relations after consolidating the KB. Applying the heuristics described earlier in the paper lead to the reduction in number of instances of the Person and Date classes by 90% and 64% respectively. Before consolidation, 283 instances representing Rembrandt were stored. The unique-name consolidation heuristic was the most effective with no identified mistakes.

When place instances are fed to the KB, they are expanded using WordNet and stored alongside their synonyms, holonyms (part of), and part_meronym (sub parts). The number of Place instances created in the KB has therefore increased significantly (94% rise). This gave the consolidation the power to identify and consolidate relationships to places as described in the geographical consolidation section. Some instances (mainly dates) were not consolidated due to slight syntactical differences, e.g. "25th/2/1841" versus "25/2/1841". This highlights the need for an additional syntactic-checking process that could eliminate such noise.

Table 1. Consolidation rates

Class	Before consld.	After consld.	Rate%
Person instance	1475	152	-90
Date instance	83	30	-64
Place instance	30	505	+94
Person relations	4240	1562	-63

CONCLUSIONS

This paper describes a system that automatically extracts knowledge, instantiates an ontology with knowledge triples, and reassembles the knowledge in the form of biographies. Problems related to this task, such as the identification and consolidation of duplicated knowledge and the verification of inconsistent knowledge, are highlighted. Artequakt's approaches to tackle these problems are described.

An initial experiment, using around 50 web pages and 5 artists, showed promising results, with nearly 3 thousand unique knowledge triples extracted (before consolidation). However, some of this knowledge was too sparse to be of any clear benefit. This indicates that more pages need to be processed, and further rules need to be constructed to cover additional ontology concepts and relations and expand the knowledge extraction scope.

The generated biographies were informative and brought together knowledge extracted from various sources. However, reusing original text to generate biographies highlighted several problems, including co-referencing and other textual deixis (such as 'Later', or 'Nevertheless'). This underlines the potential benefits of regenerating text directly from the extracted facts, which is part of our near future plans.

Our consolidation techniques significantly decreased the number of instances in the KB by up to 90% for certain classes and 63% for attributes related to instances of Person. Few instances remained undetected, mainly due to lack of information required for the knowledge comparison.

Future work on Artequakt will continue to develop its modular architecture and refine the information extraction and consolidation processes. In addition we are beginning to look at how we might leverage the full power of the underlying ontology to aid extracting information from multiple domains and produce different type of reports.

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., Jones, C., Tudhope, D.: Associative and Spatial Relationships in Thesaurus-Based Retrieval. Proc. 4th European Conf. on Digital Libraries, pages 45--58, Lisbon, Portugal, Sept. LNCS, 2000.

[2] Alani, H., Kim, S., Millard, D., Weal, M., Lewis, P., Hall, W., Shadbolt, N.: Automatic Extraction of Knowledge from Web Documents. Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd Int. Semantic Web Conf. Sanibel Island, Florida, USA, 2003.

[3] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, 2001.

[4] Ciravegna, F.: Adaptive Information Extraction from Text by Rule Induction and Generalisation. Proc.17th Int. Joint Conf. on Artificial Intelligence (IJCAI), pages 1251--1256, Seattle, USA, 2001.

[5] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: a framework and graphical development environment for robust NLP tools and applications. Proc. 40th Anniversary Meeting of the Association for Computational Linguistics, Phil, USA, 2002.

[6] Dingli, A., Ciravegna, F., Guthrie, D., Wilks, Y.: Mining Web Sites Using Unsupervised Adaptive Information Extraction. Proc. 10th Conf. of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 2003.

[7] Handschuh, S., Staab, S., Ciravegna, F.: S-CREAM – Semi Automatic Creation of Metadata. Semantic Authoring, Annotation and Markup Workshop, 15th European Conf. Artificial Intelligence, France, Lyon, 2002.

[8] Harpring, P.: Proper Words in Proper Places: The Thesaurus of Geographic Names. MDA Info. 2(3), 1997.

[9] Hill, L.L., Frew, J., Zheng, Q.: Geographic Names. The Implementation of a Gazetteer in a Georeferenced Digital Library. Digital Library Magazine, 5(1), 1999.

[10] Kim, S., Alani, H., Hall, W., Lewis, P.H., Millard, D.E., Shadbolt, N., Weal, M.J.: Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web. Workshop on Semantic Authoring, Annotation & Knowledge Markup, 15th Europ. Conf. on Artificial Intelligence, pp 1--6, France, 2002.

[11] Maedche, A., Neumann, G., Staab, S.: Bootstrapping an Ontology-based Information Extraction System. Intelligent Exploration of the Web. P. Szczepaniak, et al., Heidelberg, Springer 2002.

[12] McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B., Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. Proc. Human Language Technology Conf., San Diego, CA, USA. 2002.

[13] Michaelides, D.T., Millard, D.E., Weal, M.J., DeRoure, D.: Auld Leaky: A Contextual Open Hypermedia Link Server. Proc. 7th Hypermedia: Openness, Structural Awareness, and Adaptivity, pages 59--70, Springer Verlag, Heidelberg, 2001.

[14] Millard, D.E., Alani, H., Kim, S., Weal, M.J., Lewis, P., Hall, W., DeRoure, D., Shadbolt, N.: Generating Adaptive Hypertext Content from the Semantic Web. 1st International Workshop on Hypermedia and the Semantic Web, HyperText'03, Nottingham, UK. 2003.

[15] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to wordnet: An on-line lexical database. Int. J. Lexicography, 3(4):235--312, 1993.

[16] Poibeau, T.: Deriving a multi-domain information extraction system from a rough ontology. Proc. 17th Int. Conf. on Artificial Intelligence, Seattle. USA, 2001.

[17] Radev, D. R., McKeown. K. R.: Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24(3): 469—500, 1998.

[18] Reidsma, D., Kuper, J., Declerck, T., Saggion, H., Cunningham, H.: Cross document annotation for multimedia retrieval. EACL Workshop on Language Technology and the Semantic Web, Budapest, 2003.

[19] Staab, S., Maedche, A., Handschuh, S.: An Annotation Framework for the Semantic Web. Proc. 1st Int. Workshop on MultiMedia Annotation, Tokyo, 2001.

[20] Vargas-Vera, M., Motta, E., Domingue, J., Buckingham Shum, S., Lanzoni, M.: Knowledge Extraction by using an Ontology-based Annotation Tool. Proc. Workshop on Knowledge Markup & Semantic Annotation, 1st Int. Conf. on Knowledge Capture, pp 5--12, Victoria, B.C., Canada, 2001.

[21] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F.: MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. 13th Int. Conf. on Knowledge Engineering and Management (EKAW), Spain, 2002.

[22] Voorhees, E.M.: Using WordNet for Text Retrieval. Fellbaum (edt.) WordNet: An Electronic Lexical Database, pages 285--303, MIT Press**,** 1998.

[23] White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., Wagstaff, K.: Multidocument Summarization via Information Extraction. Proc. of Human Language Technology Conf. (HLT 2001), San Diego, CA, 2000.