ACL 2007

Proceedings of the Workshop on Language Technology for Cultural Heritage Data

June 28, 2007
Prague, Czech Republic

The Association for Computational Linguistics

Production and Manufacturing by
Omnipress
2600 Anderson Street
Madison, WI 53704
USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Preface

Museums, archives, and libraries around the world maintain large collections of cultural heritage objects, such as archaeological artefacts, sound recordings, historical manuscripts, or preserved animal specimens. Large scale digitisation projects are currently underway to make these collections more accessible. The natural next step after digitisation is the development of powerful tools to search, link, enrich, and mine the digitised data. Language technology has an important role to play in this endeavour, even for collections which are primarily non-textual, since text is the pervasive medium used for metadata. At the same time, the cultural heritage domain poses special challenges for the NLP community, including the use of historical or non-standard language and orthography, the presence of OCR or transcription errors in the input data, and the necessity to deal with data from various media and languages. The cultural heritage domain is therefore also a challenging and interesting testbed for the robustness of existing language technology.

The ACL 2007 workshop on *Language Technology for Cultural Heritage Data* is to be seen in the context of a growing interest in the development of IT solutions for the cultural heritage domain, as witnessed by numerous national and international research initiatives, such as CATCH (Continuous Access to Cultural Heritage), DigiCULT (Digital Culture), MALACH (Multilingual Access to Large Spoken Archives), and MultiMatch (Multilingual/Multimedia Access To Cultural Heritage).

We solicited papers describing new and original work on all aspects of language technology for the cultural heritage domain. Out of the 22 submissions received, 11 were selected for inclusion in the workshop programme following a peer-review process. The list of papers reflects the current breadth of this exciting and expanding area, with topics covering improved access to cultural heritage data (combining digital libraries with treebanks, mono- and cross-lingual information retrieval, dealing with controlled vocabularies), methods for aligning hand-written documents with their transcripts, named entity recognition for historical texts, knowledge discovery in databases, and museum visitor path prediction. An invited talk by Douglas W. Oard on the MALACH project completes the workshop programme.

We would like to thank all authors who submitted papers for the hard work that went into their submissions. We are also extremely grateful to the members of the programme committee for their thorough reviews, and to the ACL 2007 organisers, especially the ACL 2007 Workshop Chair Simone Teufel, for their help with administrative matters. Special thanks to our invited speaker Doug Oard and to the MultiMatch project for their generous sponsorship of the workshop.

Antal van den Bosch
Claire Grover
Caroline Sporleder

Organizers

Chairs:

Caroline Sporleder, Saarland University
Antal van den Bosch, University of Tilburg
Claire Grover, University of Edinburgh

Program Committee:

Ion Androutsopoulos, Athens University of Economics and Business
Antal van den Bosch, Tilburg University
Kate Byrne, University of Edinburgh
Robert Dale, Macquarie University
Vania Dimitrova, University of Leeds
Mick O'Donnell, Universidad Autonoma de Madrid
Bassilis Gatos, NCSR Demokritos
Julio Gonzalo, Universidad Nacional de Educacion a Distancia
Claire Grover, University of Edinburgh
Jiyin He, University of Amsterdam
Marti Hearst, University of California Berkeley
Djoerd Hiemstra, University of Twente
Nancy Ide, Vassar College
Neil Ireson, University of Sheffield
Christer Johansson, University of Bergen
Franciska de Jong, University of Twente
Jaap Kamps, University of Amsterdam
Vangelis Karkaletsis, NCSR Demokritos
Piroska Lendvai, Tilburg University
Ruli Manurung, University of Indonesia
Maria Milosavljevic, University of Edinburgh
Marie-Francine Moens, Katholieke Universiteit Leuven
John Nerbonne, Rijksuniversiteit Groningen
Douglas Oard, University of Maryland
Hans Paijmans, Maastricht University
Martin Reynaert, Tilburg University
Maarten de Rijke, University of Amsterdam
Mark Sanderson, University of Sheffield
Caroline Sporleder, Saarland University
Efstathios Stamatatos, University of the Aegean
Erik Tjong Kim Sang, University of Amsterdam
Arjen de Vries, CWI, Amsterdam

Invited Speaker:

Douglas W. Oard, University of Maryland

Table of Contents

Conference Program

Thursday, June 28, 2007

9:00–9:05 Welcome

9:05–9:30 *Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature*
Lars Borin, Dimitrios Kokkinakis and Leif-Jöran Olsson

9:30–9:55 *Viterbi Based Alignment between Text Images and their Transcripts*
Alejandro H. Toselli, Verónica Romero and Enrique Vidal

9:55–10:20 *Retrieving Lost Information from Textual Databases: Rediscovering Expeditions from an Animal Specimen Database*
Marieke van Erp

10:20–10:45 *Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources*
Tandeep Sidhu, Judith Klavans and Jimmy Lin

10:45–11:15 Coffee Break and Poster Session

The Latin Dependency Treebank in a Cultural Heritage Digital Library
David Bamman and Gregory Crane

Cultural Heritage Digital Resources: From Extraction to Querying
Michel Généreux

Dynamic Path Prediction and Recommendation in a Museum Environment
Karl Grieser, Timothy Baldwin and Steven Bird

Anchoring Dutch Cultural Heritage Thesauri to WordNet: Two Case Studies
Véronique Malaisé, Antoine Isaac, Luit Gazendam and Hennie Brugman

Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation
Idan Szpektor, Ido Dagan, Alon Lavie, Danny Shacham and Shuly Wintner

11:15–11:40 *Deriving a Domain Specific Test Collection from a Query Log*
Avi Arampatzis, Jaap Kamps, Marijn Koolen and Nir Nussbaum

Thursday, June 28, 2007 (continued)

11:40–12:05 *Multilingual Search for Cultural Heritage Archives via Combining Multiple Translation Resources*
Gareth J. F. Jones, Ying Zhang, Eamonn Newman, Fabio Fantino and Franca Debole

12:05–12:55 Invited Talk: *Lessons from the MALACH Project: Applying New Technologies to Improve Intellectual Access to Large Oral History Collections*
Douglas W. Oard

12:55–13:00 Closing

Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature

Lars Borin, Dimitrios Kokkinakis, Leif-Jöran Olsson
Litteraturbanken and Språkdata/Språkbanken
Department of Swedish Language, Göteborg University
Sweden
{first.last}@svenska.gu.se

Abstract

This paper provides a description and evaluation of a generic named-entity recognition (NER) system for Swedish applied to electronic versions of Swedish literary classics from the 19th century. We discuss the challenges posed by these texts and the necessary adaptations introduced into the NER system in order to achieve accurate results, useful both for metadata generation, but also for the enhancement of the searching and browsing capabilities of *Litteraturbanken*, the Swedish Literature Bank, an ongoing cultural heritage project which aims to digitize significant works of Swedish literature.

1 Introduction

In this paper we investigate generic named entity recognition (NER) technology and the necessary adaptation required in order to automatically annotate electronic versions of a number of Swedish literary works of fiction from the 19th century. Both the genre and language variety are markedly different from the text types that our NER system was originally developed to annotate. This presents a challenge, posing both specific and more generic problems that need to be dealt with.

In section 2 we present briefly the background and motivation for the present work, and section 3 gives some information on related work. In section 4 we provide a description of the named entity recognition system used in this work, its entity taxonomy, including the animacy recognition component and the labeled consistency approach that is explored. Problems faced in the literary texts and the kinds of adaptations performed in the recognition system as well as evaluation and error analysis are given in section 5. Finally, section 6 summarizes the work and provides some thoughts for future work.

2 Background

Litteraturbanken <http://litteraturbanken.se/> (the Swedish Literature Bank) is a cultural heritage project financed by the Swedish Academy[1]. Litteraturbanken has as its aim to make available online the full text of significant works of Swedish literature, old and new, in critical editions suitable for literary research and for the teaching of literature. There is also abundant ancillary material on the website, such as author presentations, bibliographies, thematic essays about authorships, genres or periods, written by experts in each field.

Similarly to many other literature digitization initiatives, most of the works in Litteraturbanken are such for which copyright has expired (i.e., at least 70 years have passed since the death of the author); at present the bulk of the texts are from the 18th, 19th and early 20th century. However, there is also an agreement with the organizations representing authors' intellectual property rights, allowing the inclusion of modern works according to a uniform royalty payment scheme. At present, Litteraturbanken holds about 150 works – mainly novels – by about 50 different authors. The text collection is slated to grow by 80–100 novel-length works (appr. 4–6 million words) annually.

1 The present permanent version of Litteraturbanken was preceded by a two-year pilot project by the same name, funded by *the Bank of Sweden Tercentenary Foundation*.

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 1–8,
Prague, 28 June 2007.

Even at outset of the Litteraturbanken project, it was decided to design the technical solutions with language technology in mind. The rationale for this was that we saw these literary texts not only as representing Sweden's literary heritage, but also as high-grade empirical data for linguistic investigations, i.e. as corpus components. Hence, we wanted to build an infrastructure for Litteraturbanken which would allow this intended dual purpose of the material to be realized to the fullest.[2] However, we soon started to think about how the kinds of annotations that language technology could provide could be of use to others than linguists, e.g. literary scholars, historians and researchers in other fields in the humanities and social sciences.

Here, we will focus on one of these annotation types, namely NER and entity annotation. Combined with suitable interfaces for displaying, searching, selecting, correlating and browsing named entities, we believe that the recognition and annotation of named entities in Litteraturbanken will facilitate more advanced research on literature (particularly in the field of literary onomastics; see Dalen-Oskam and Zundert, 2004), but also, e.g., historians could find this facility useful, insofar as these fictional narratives also contain, e.g. descriptions of real locations, characterizations of real contemporary public figures, etc. Flanders et al. (1998: 285) argue that references to people in historical sources are of intrinsic interest since they may reveal "networks of friendship, enmity, and collaboration; familial relationships; and political alliances […] class position, intellectual affiliations, and literary bent of the author".

3 Related Work

The presented work is naturally related to research on NER, particularly as applied to diachronic/historical corpora. The technology itself has been applied to various domains and genres over the last couple of decades such as financial news and biomedicine, with performance rates difficult to compare since the task is usually tied to particular domains/genres and applications. For a concise overview of the technology see Borthwick, (1999). Even though this technology is widely used in a number of domains, studies dealing with historical corpora are mostly comparatively recent (see for instance the recent workshop on historical text mining; <http://ucrel.lancs.ac.uk/events/htm06/>).

Shoemaker (2005) reports on how the *Old Bailey Proceedings*, which contain accounts of trials that took place at the Old Bailey, the primary criminal court in London, between 1674 and 1834, was marked up for a number of semantic categories, including the crime date and location, the defendant's gender, the victim's name etc. Most of the work was done manually while support was provided for automatic person name[3] identification (*cf.* Bontcheva et al., 2002). The author mentions future plans to take advantage of the structured nature of the Proceedings and to use the lists of persons, locations and occupations that have already been compiled for annotating new texts.

Crane and Jones (2006) discuss the evaluation of the extraction of 10 named entity classes (personal names, locations, dates, products, organizations, streets, newspapers, ships, regiments and railroads) from a 19th century newspaper. The quality of their results vary for different entity types, from 99.3% precision for *Streets* to 57.5% precision for *Products*. The authors suggest the kinds of knowledge that digital libraries need to assemble as part of their machine readable reference collections in order to support entity identification as a core service, namely, the need for bigger authority lists, more refined rule sets and rich knowledge sources as training data.

At least least two projects are also relevant in the context of NER and historical text processing, namely NORA <http://www.noraproject.org/> and ARMADILLO <http://www.hrionline.ac.uk/armadillo/>. The goal of the first is to produce text mining software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries. The goal of the latter is to evaluate the benefits of automated mining techniques (including information extraction) on a set of online resources in eighteenth-century British social history.

2 This precluded the use of ready-made digital library or CMS solutions, as we wanted to be compatible with emerging standards for language resources and tools, e.g. TEI/(X)CES and ISO TC37/SC07, which to our knowledge has never been a consideration in the design of digital library or CM systems.

3 By using the General Architecture for Text Engineering (GATE) platform; <http://gate.ac.uk>.

4 Named Entity Recognition

Named entity recognition (NER) or entity identification/extraction, is an important supporting technology with numerous applications in a number of human language technologies. The system we use originates from the work conducted in the *Nomen Nescio* project; for details see Johannessen et al. (2005). In brief, the Swedish system is a multipurpose NER system, comprised by a number of modules applied in a pipeline fash-ion. Six major components can be distinguished, making a clear separation between lexical, gram-matical and processing resources. The six compo-nents are:

- lists of **multiword names**, taken from various Internet sites or extracted from various corpora, running directly over the tokenised text being processed;
- a rule-based, **shallow parsing** component that uses finite-state grammars, one grammar for each type of entity recognized;
- a module that uses **the annotations produced by the previous two components**, which have a high rate in precision, in order to make decisions regarding other unannotated entities. This module is further discussed in Section 4.2;
- lists of **single names** (approx. 100,000);
- **name similarity,** this module is further discussed in Section 4.3;
- a **theory revision and refinement** module, which makes a final control of an annotated document, in order to detect and resolve possible errors and assign new annotations based on existing ones, for instance by applying name similarity or by combining various annotation fragments.

4.1 Named-Entity Taxonomy

The nature and type of named entities vary depending on the task under investigation or the target application. In any case, *personal names*, *location* and *organization names* are considered "generic". Since semantic annotation is not as well understood as grammatical annotation, there is no consensus on a standard tagset and content to be generally applicable. Recently, however, there have been attempts to define and apply richer name hierarchies for various tasks, both specific (Fleischman and Hovy, 2002) and generic (Sekine, 2004). Our current system implements a rather fine-grained named entity taxonomy with 8 main named entitiy types as well as 57 subtypes. Details can be found in Johannessen et al., 2005, and Kokkinakis, 2004. The eight main categories are:

- **Person** (PRS): people names (forenames, surnames), groups of people, animal/pet names, mythological, theonyms;
- **Location** (LOC): functional locations, geographical, geo-political, astrological;
- **Organization** (ORG): political, athletic, media, military, etc.;
- **Artifact** (OBJ): food/wine products, prizes, communic. means (vehicles) etc.;
- **Work&Art** (WRK): printed material, names of films and novels, sculptures etc.;
- **Event** (EVN): religious, athletic, scientific, cultural etc.;
- **Measure/Numerical** (MSR): volume, age, index, dosage, web-related, speed etc.;
- **Temporal** (TME).

Time expressions are important since they allow temporal reasoning about complex events as well as time-line visualization of the story developed in a text. The temporal expressions recognized include both relative (*nästa vecka* 'next week') and absolute expressions (*klockan 8 på morgonen i dag* '8 o'clock in the morning today'), and sets or sequences of time points or stretches of time (*varje dag* 'every day').

4.2 Animacy Recognition

The rule-based component of the person-name recognition grammar is based on a large set of designator words and a group of phrases and verbal predicates that most probably require an animate subject (e.g. *berätta* 'to tell', *fundera* 'to think', *tröttna* 'to become tired'). These are used in conjunction with orthographic markers in the text, such as capitalization, for the recognition of personal names. In this work, we consider the first group (designators) as relevant knowledge to be extracted from the person name recognizer, which is explored for the annotation of animate instances

in the literary texts. The designators are implemented as a separate module in the current pipeline, and constitute a piece of information which is considered important for a wide range of tasks (*cf.* Orasan and Evans, 2001).

The designators are divided into four groups: designators that denote the nationality or the ethnic/racial group of a person (e.g. *tysken* 'the German [person]'); designators that denote a profession (e.g. *läkaren* 'the doctor'); those that denote family ties and relationships (e.g. *svärson* 'son in law'); and finally a group that indicates a human individual but cannot be unambiguously categorized into any of the three other groups (e.g. *patienten* 'the patient'). Apart from this grouping, inherent qualities, for at least a large group of the designators, (internal evidence/morphological cues) also indicate referent (natural) gender. In this way, the animacy annotation is further specified for male, female or unknown gender; unknown in this context means unresolved or ambiguous, such as *barn* 'child'.

Swedish is a compounding language and compound words are written as a single orthographic unit (i.e. solid compounds). This fact makes the recognition of animacy straightforward with minimal resources and feasible by the use of a set of suitable headwords, and by capturing modifiers by simple regular expressions. Approximately 25 patterns are enough to identify the vast majority of animate entities in a text; patterns such as "inna/innan/innor", "man/mannen/män/männen", "log/logen/loger", "ktör/ktören/ktörer" and "iker/ikern/ikerna". For instance, the pattern in (1) consists of a reliable suffix "inna" which is a typical designator for female individuals, preceded by a set of obligatory strings and an optional regular expression which captures a long list of compounds (2).

(1) `[a-zääö]*(kv|älskar|man|grev|…)inna`

(2) `taleskvinna, yrkeskvinna, idrottskvinna, ungkvinna, Stockholmskvinna, Dalakvinna, samboälskarinna, lyxälskarinna, ex-älskarinna, samlargrevinna, exälskarinna, markgrevinna, majgrevinna, änkegrevinna,…`

Examples of animacy annotations are given in (3). The attribute value *FAM* stands for *FAmily* relation and *Male*; *PRM* for *PRofession* and *Male*; *FAF* for *FAmily* relation and *Female* and finally *UNF* for *UNknown* and *Female*.

(3) `[…]<ENAMEX TYPE="FAM">riksgrefvinnans far</ENAMEX>, <ENAMEX TYPE="PRM">öfveramiralen</ENAMEX> […] hade till <ENAMEX TYPE="FAF">mor</ENAMEX> <ENAMEX TYPE="UNF">grefvinnan</ENAMEX> Beata Wrangel från […]`

Table (3) in Section 6.1 presents the results for the evaluation of this type of normative information. Note also, that in order to make the annotations more practical we have included the person name designators (e.g. 'herr' – 'Mr') in the markup as in (4); here *PRS* stands for *PeRSon*:

(4) `<ENAMEX TYPE="UNM">Herr</ENAMEX> <ENAMEX TYPE="PRS" SBT="HUM">Boman</ENAMEX>` becomes `<ENAMEX TYPE="PRS-UNM" SBT="HUM">Herr Boman</ENAMEX>`

4.3 Name Similarity

We can safely assume that the various system resources will not be able to identify all possible entities in the texts, particularly personal and location names. Although there is a large overlap between the names in the texts and the gazetteer lists, there were cases that could be considered as entity candidates but were left unmarked. This is because exhaustive lists of names even for limited domains are hard to obtain, and, in some domains even difficult to manage. Therefore, we also calculated the orthographic similarity between such words and the gazetteer content, according to the following criteria: a potential entity starts with a capital letter; it is ≥ 5 characters long; it is not part of any other annotation and it does not stand in the beginning of a sentence. We have empirically observed that the length of 5 characters is a reliable threshold, unlikely to exclude many NEs. As a matter of fact, only two such cases could be found in the evaluation sample, namely *ätten Puff* 'the family Puff' and "Yen-" in the context "Yen- kenberg"

As measure of orthographic similarity (or rather, difference) we used the Levenshtein distance (LD; also known as *edit distance*) between two strings. The LD is the number of deletions, insertions or substitutions required to transform a string into another string. The greater the distance, the more different the strings are. We chose to regard 1 and 2 as trustworthy values and disregarded the rest. We chose these two values since empirical observations suggest that contemporary Swedish and

19th century Swedish entities usually differ in one or two characters. In case of more than one match, we choose the most frequent alternative, as in the case of *Wenern* below. Table 1 illustrates various cases and the obtained results.

text word	#	gazeteer	LD	ann.	??
Dalarne	6	Dalarna	1	loc	yes
Asptomten	1	---	---	---	-
Härnevi*	1	Arnevi	2	prs	no
Sabbathsberg	1	Sabbatsberg	1	loc	yes
Wenern*	7	Werner,Waern Vänern	2 2	prs loc	no
Kaknäs	1	Valnäs,Ramnäs	2	loc	yes
Kallmar	1	Kalmar	1	loc	yes

Table 1. LD between potential NEs and the gazeteers; '*': both are locations;'??': correct annot.?

5 The Document Centered Approach

There is a known tradeoff between rule-based and statistical systems. Handcrafted grammar-based systems typically obtain better results, but at the cost of considerable manual effort by domain experts. Statistical NER systems typically require a large amount of manually annotated training data, but can be ported to other domains or genres more rapidly and require less manual work. Although the Swedish system is mainly rule-based, using a handcrafted grammar for each entity group, it can also be considered a hybrid system in the sense that it applies a document-centered approach (DCA) to entity annotation, which is a different paradigm compared to the local context approach, called *external evidence* by McDonald (1996). With DCA, information for the disambiguation of a name is derived from the entire document.

DCA as a term originates from the work by Mikheev (2000: 138), who claims that:

> important words are typically used in a document more than once and in different contexts. Some of these contexts create very ambiguous situations but some don't. Furthermore, ambiguous words and phrases are usually unambiguously introduced at least once in the text unless they are part of common knowledge presupposed to be known by the readers.

This implies a form of online learning from the document being processed where unambiguous usages are used for assigning annotations to ambiguous words, and information for disambiguation is derived from the entire document.

Similarly, label consistency, the preference of the same annotation for the same word sequence everywhere in a particular discourse, is a comparable approach for achieving qualitatively higher recall rates with minimal resource overhead (*cf.* Krishnan and Manning, 2006). Such an approach has been used, e.g., by Aramaki et al. (2006), for the identification of personal health information (age, id, date, phone, location and doctor´s and patient´s names).

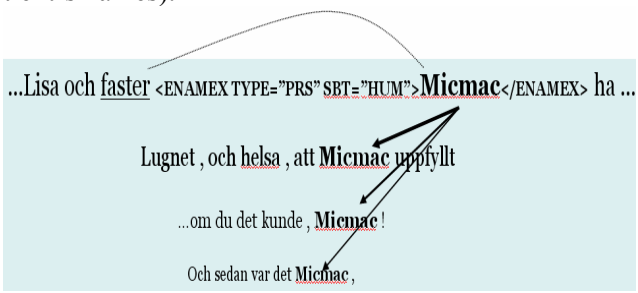


Figure 1. Example of label consistency

Figure 1 illustrates this approach with an example taken from *Almqvist's Collected Works, Vol. 30*. In this example, the first occurrence of the female person name *Micmac*, which is not in the gazetteer lists, is introduced by the author with the unambiguous designator *faster* 'aunt'. Many of the subsequent mentions of the same name are given without any reliable clue for appropriate labelling. However, as already discussed, there is strong evidence that subsequent mentions of the same name should be annotated with the same label, and since the same entity usually appears more than once in the same discourse, in our case a book, labelling consistency should guarantee better performance. There are exceptions for certain NE categories which may consist of words that are not proper nouns such as in the *Work&Art* category, and of course the temporal and measure groups which are blocked from this type of processing; *cf.* section 6.2.

6 Evaluation and Error Analysis

The system was evaluated twice, while no normalization or other preprocessing was applied to the original documents. Problems identified during the first evaluation round were taken under consideration and specific changes were suggested to the system by incorporating appropriate modifications.

During the first run, no adaptations or enhancements were made to the original NER system. After the first evaluation round, four major areas were identified in which the system either failed to produce an annotation or produced only partial or erroneous annotations. These failures were caused by:

- **Spelling variation**: particularly the use of <f/w/e/q> instead of <v/v/ä/k> as in modern Swedish. Most of the cases could be easily solved while other required different means such as calculating the LD between the name lists and possible name mentions in the texts (Section 4.3). One case that could be easily tackled was the addition of alternate spelling forms for a handful of keywords and designators, especially the preposition *av/af* common in temporal contexts, such as *i början af/av 1790-talet* 'in the beginning of the 1790s'; or words such *begge/bägge* 'both' and *qväll/kväll* 'evening';
- A number of **definite plural forms** of nouns, often designating a group of persons, with the suffix "erne" instead the "erna" as in modern Swedish, such as *Kineserne/Kineserna* 'the Chinese [people]' and *Svenskarne/Svenskarna* 'the Swedes';
- **Unknown names**: mentioned once with unreliable context;
- **Structure preservation**: the document structure of the texts in Litteraturbankens is designed to create a faithful rendering of the visual appearance of the original printed books. In extracting the texts from the XML format used in Litteraturbanken, we did not want to apply any kind of normalization or other processing. Such an approach would have altered the document structure. This implies that for a handful of the entities, for which the hyphenation in the original paper version has divided a name into two parts, as in (5), correct identification cannot be accomplished, while in some cases only a partial identification was possible, as in (6).

(5) `[…] Stock- holm`

(6) `<ENAMEX TYPE="PRS" SBT="HUM">Bertha von Lichten-</ENAMEX> ried`

6.1 Results

As a baseline for the evaluation we use the result of simple dictionary lookup in the single name gazetteer. This process is very accurate (w.r.t. precision). We could identify a number of cases with erroneous annotations, due to various circumstances: Names in the gazetteer lists may have multiple entity tags associated with them, and thus an entity may belong to more than one group that could not be disambiguated by the surrounding context, such as *Ekhammar* as a city and surname; many names are ambiguous with common nouns or verbs, such as *Stig* as a first name and as the verb 'step/walk'; the gazetteers contained a number of words that should not have been in the list in the first place, such as *Hvem* 'Who', *styrman* 'first mate' and *fänrik* 'lieutenant'. A probable cause of the latter problem is the fact that the name lists have been semi-automatically compiled from various sources including corpora and the Internet.

We performed two evaluations, based on two different random samples consisting of 500 segments (roughly 30,000 tokens) each. A segment consists of an integral number of sentences (up to 10–20). The overall results for all tests are shown in table 2. Results for individual entities using the whole system during both runs are found in table 3. The samples were evaluated according to precision, recall and f-score using the formulas:

Precision = (Total Correct + Partially Corrtect) / All Produced
Recall = (Total Correct + Partially Correct) / All Possible
*F-score =2*P*R/P+R.*

	1st run – no adaptations			**2nd run** – with adaptations		
	P	**R**	**F**	**P**	**R**	**F**
Baseline (gazeteer lookup)	88,8%	69,8%	78,1%	93,1%	86,2%	89,5%
Rule-Based System (no time&animacy)	91,8%	75,4%	82,7%	96,9%	86,9%	91,6%
Rule-Based System (all categ.)	93,8%	69,4%	79,7%	96%	87,9%	91,7%
Rule-Based System & DCA	95,4%	83,4%	89%%	96,1%	88,8%	92,3%
Rule-Based System&DCA+ED	96%	84%	89,6%	96,6%	89,4%	92,8%

Table 2. Overall performance of the NER

NE Categories	1st run				2nd run			
	# of NEs corr prod./ possible	P	R	F	# of NEs corr.prod./ possible	P	R	F
PERSON	424/441	92,3%	96,1%	94,1%	410/419	96%	97,8%	96,9%
LOCATION	83/123	100%	67,4%	80,5%	74/95	97,3%	77,8%	86,4%
ORGANIZATION	7/10	70%	70%	70%	4/4	40%	100%	57,1%
ARTIFACT	0/4	---	---		0/6	---	---	
WORK/ART	3/9	75%	33,3%	46,1%	4/10	100%	40%	57,1%
EVENT	0/2	---	---		0/2	---	---	
TEMPORAL	102/114	99%	89,4%	97%	106/118	100%	89,8%	94,6%
MEASURE	1/4	100%	25%	40%	4/16	66,6%	25%	36,3%
ANIMACY	207/277	98,1%	74,7%	84,8%	292/339	96,3%	86,1%	90,9%
TOTAL	***827/984***	***96%***	***84%***	***89,6%***	***894/999***	***96,6%***	***89,4%***	***92,8%***

Table 3. Performance of the NER on the individual named entities including animacy

Partially correct means that an annotation gets partial credit. For instance, if the system produces an annotation for the functional location *Nya Elementarskolan* as in (7) instead of the correct (8), then such annotations are given half a point, instead of a perfect score.

(7) `Nya <ENAMEX TYPE="LOC" SBT="FNC">Elementarskolan</ENAMEX>`

(8) `<ENAMEX TYPE="LOC" SBT="FNC">Nya Elementarskolan</ENAMEX>`

If, on the other hand, the type is correct but the subtype is wrong, then the annotation is given a score of 0.75 points (e.g. a functional location instead of a geopolitical location).

6.2 Limitations of the Centering Approach

Labeling consistency and the DCA approach relies on the assumption that usage is consistent within the same document by the same author. However, we have observed that there are problems with entities composed of more than a single word, particularly within the group *Work&Art*, which can produce conflicting information, if we allow the individual words in such content (often nouns or adjectives) to be re-applied in the text.

For instance, the name of the novel *Syster och bror* occurred 32 times in one of the evaluation texts (Almqvist's Collected Works Volume 29). If we allow the individual words that constitute the title, *Syster*, *och* and *bror* to be re-applied in the text as individual words (2 common nouns and a conjunction), then we would have degraded the precision considerably since we would have allowed *Work&Art* annotations for irrelevant words. However, such cases can be resolved by simply letting the system ignore multiword *Work&Art* annotations during the DCA processing.

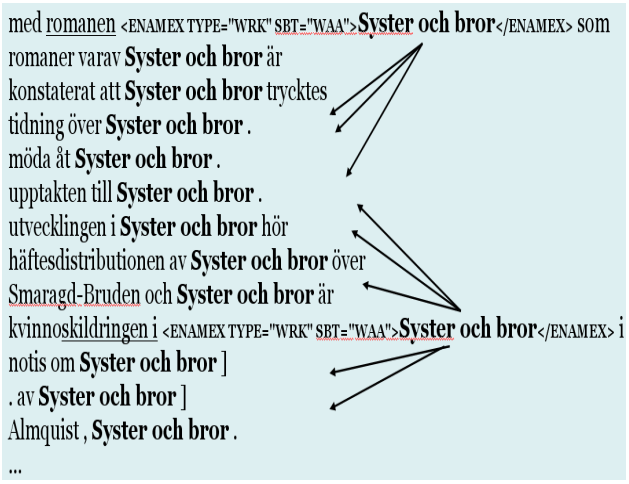


Figure 2. Occurrences of the multi-word entity *Syster och bror*; the rule-based system could reliably identify and annotate 2/32 occurrences.

Generally speaking, the experimental results have shown that any breaking of a multiword entity, except personal names, into its individual words often has a negative effect on performance. The best results are achieved when the DCA approach deals with single or bigram entities, particularly personal names.

7 Conclusions and Future Prospects

In this paper we have described the application of a generic Swedish named entity recognition system to a number of literary texts, novels from the 19th century, part of *Litteraturbanken*, the Swedish Literature Bank. We evaluated the results of the named entity recognition and identified a number of error sources which we tried to resolve and then introduce changes that would cover for such cases in the rule-based component of the system, in order to increase its performance (precision and recall) during a second evaluation round.

Entity annotations open up a whole new research spectrum for new kinds of qualitative and quantitative exploitations of literary and historical texts, allowing more semantically-oriented exploration of the textual content. In the near future, we will annotate and evaluate a larger sample and possibly integrate machine learning techniques in order to improve the results even more. We are also working to integrate the handling of named entity annotations into Litteraturbanken's search and browsing interfaces and hope to be able to conduct our first demonstrations and tests with users later this year.

References

Eiji Aramaki, Takeshi Imai, Kengo Miyo and Kazuhiko Ohe. 2006. Automatic Deidentification by using Sentence Features and Label Consistency. *Challenges in NLP for Clinical Data Workshop*. Washington DC.

Kalina Bontcheva, Diana Maynard, Hamish Cunningham and Horacio Saggion. 2002. Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. PhD Thesis. New York University.

Gregory Crane and Alison Jones. 2006. The Challenge of Virginia Banks: an Evaluation of Named Entity Analysis in a 19th-century Newspaper Collection. *ACM/IEEE Joint Conference on Digital Libraries*, JCDL. Chapel Hill, NC, USA. 31–40.

Karina van Dalen-Oskam and Joris van Zundert. 2004. Modelling Features of Characters: Some Digital Ways to Look at Names in Literary Texts. *Literary and Linguistic Computing* 19(3): 289–301.

Julia Flanders, Syd Bauman, Paul Caton and Mavis Cournane. 1998. Names Proper and Improper: Applying the TEI to the Classification of Proper Nouns. *Computers and the Humanities* 31(4): 285–300.

Michael Fleischman and Eduard Hovy. 2002. Fine Grained Classification of Named Entities. *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan. 1–7.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitrios Kokkinakis, Paul Meurer, Eckhard Bick and Dorte Haltrup. 2005. Named Entity Recognition for the Mainland Scandinavian Languages. *Literary and Linguistic Computing*. 20(1): 91–102.

Dimitrios Kokkinakis. 2004. Reducing the Effect of Name Explosion. *Proceedings of the LREC-Workshop: Beyond Named Entity Recognition - Semantic Labeling for NLP*. Lisbon, Portugal.

Vijay Krishnan and Christopher D. Manning. 2006. An Efficient Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. *Proceedings of COLING/ ACL 2006*. Sydney, Australia. 1121–1128.

David D. McDonald. 1996. Internal and External Evidence in the Identification and Semantic Categorisation of Proper Nouns. *Corpus-Processing for Lexical Acquisition*. James Pustejovsky and Bran Boguraev (eds). MIT Press. 21–39.

Andrei Mikheev. 2000. Document Centered Approach to Text Normalization. *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens, Greece. 136–143.

Satoshi Sekine. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Lisbon, Portugal.

Constantin Orasan and Roger Evans. 2001. Learning to Identify Animate References. *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001)*. ACL-2001. Toulouse, France.

Robert Shoemaker. 2005. Digital London. Creating a Searchable Web of Interlinked Sources on Eighteenth Century London. *Program: Electronic Library & Information Systems* 39(4): 297–311.

Viterbi Based Alignment between Text Images and their Transcripts*

Alejandro H. Toselli, Verónica Romero and **Enrique Vidal**
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera s/n
46071 - València, Spain
[ahector,vromero,evidal]@iti.upv.es

Abstract

An alignment method based on the Viterbi algorithm is proposed to find mappings between word images of a given handwritten document and their respective (ASCII) words on its transcription. The approach takes advantage of the underlying segmentation made by Viterbi decoding in handwritten text recognition based on Hidden Markov Models (HMMs). Two HMMs modelling schemes are evaluated: one using 78-HMMs (one HMM per character class) and other using a unique HMM to model all the characters and another to model blank spaces. According to various metrics used to measure the quality of the alignments, encouraging results are obtained.

1 Introduction

Recently, many on-line digital libraries have been publishing large quantities of digitized ancient handwritten documents, which allows the general public to access this kind of cultural heritage resources. This is a new, comfortable way of consulting and querying this material. The *Biblioteca Valenciana Digital* (BiValDi)[1] is an example of one such digital library, which provides an interesting collection of handwritten documents.

Several of these handwritten documents include both, the handwritten material and its proper transcription (in ASCII format). This fact has motivated the development of methodologies to align these documents and their transcripts; i.e. to generate a mapping between each word image on a document page with its respective ASCII word on its transcript. This word by word alignment would allow users to easily find the place of a word in the manuscript when reading the corresponding transcript. For example, one could display both the handwritten page and the transcript and whenever the mouse is held over a word in the transcript, the corresponding word in the handwritten image would be outlined using a box. In a similar way, whenever the mouse is held over a word in the handwritten image, the corresponding word in the transcript would be highlighted (see figure 1). This kind of alignment can help paleography experts to quickly locate image text while reading a transcript, with useful applications to editing, indexing, etc. In the opposite direction, the alignment can also be useful for people trying to read the image text directly, when arriving to complex or damaged parts of the document.

Creating such alignments is challenging since the transcript is an ASCII text file while the manuscript page is an image. Some recent works address this problem by relying on a previous explicit image-processing based word pre-segmentation of the page image, before attempting the transcription alignments. For example, in (Kornfield et al., 2004), the set of previously segmented word images and their corresponding transcriptions are transformed into two different times series, which are aligned

*This work has been supported by the EC (FEDER), the Spanish MEC under grant TIN2006-15694-C02-01, and by the *Consellería d'Empresa, Universitat i Ciència - Generalitat Valenciana* under contract GV06/252.

[1] http://bv2.gva.es

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 9–16,
Prague, 28 June 2007.

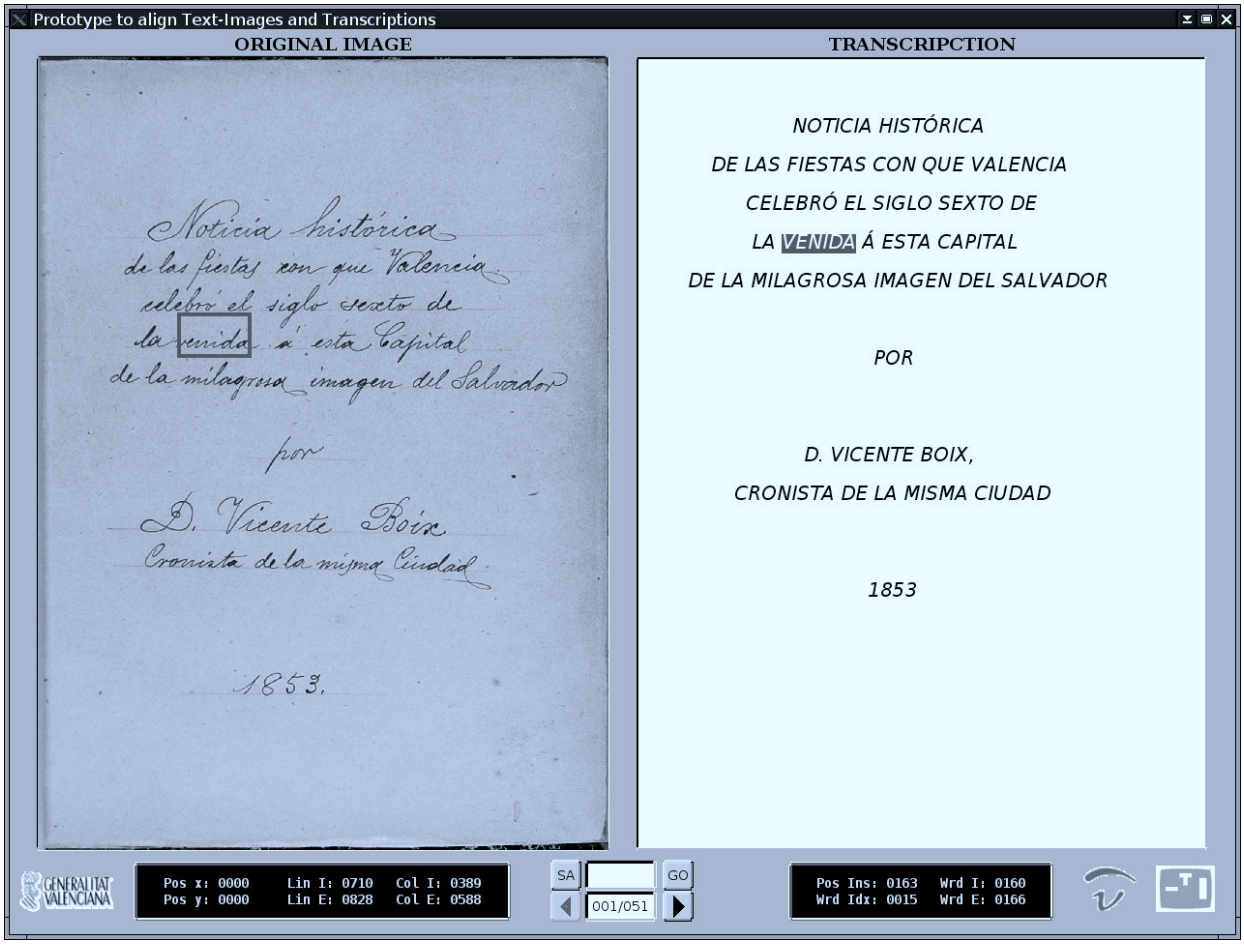


Figure 1: Screen-shot of the alignment prototype interface displaying an outlined word (using a box) in the manuscript (left) and the corresponding highlighted word in the transcript (right).

using *dynamic time warping* (DTW). In this same direction, (Huang and Srihari, 2006), in addition to the word pre-segmentation, attempt a (rough) recognition of the word images. The resulting word string is then aligned with the transcription using dynamic programming.

The alignment method presented here (henceforward called Viterbi alignment), relies on the Viterbi decoding approach to handwritten text recognition (HTR) based on Hidden Markov Models (HMMs) (Bazzi et al., 1999; Toselli et al., 2004). These techniques are based on methods originally introduced for speech recognition (Jelinek, 1998). In such HTR systems, the alignment is actually a byproduct of the proper recognition process, i.e. an implicit segmentation of each text image line is obtained where each segment successively corresponds to one recognized word. In our case, word recognition is not actually needed, as we do already have the correct transcription. Therefore, to obtain the segmentations for the *given* word sequences, the so-called "forced-recognition" approach is employed (see section 2.2). This idea has been previously explored in (Zimmermann and Bunke, 2002).

Alignments can be computed line by line in cases where the beginning and end positions of lines are known or, in a more general case, for whole pages. We show line-by-line results on a set of 53 pages from the *"Cristo-Salvador"* handwritten document (see section 5.2). To evaluate the quality of the obtained alignments, two metrics were used which give information at different alignment levels: one measures the accuracy of alignment mark placements and the other measures the amount of erroneous as-

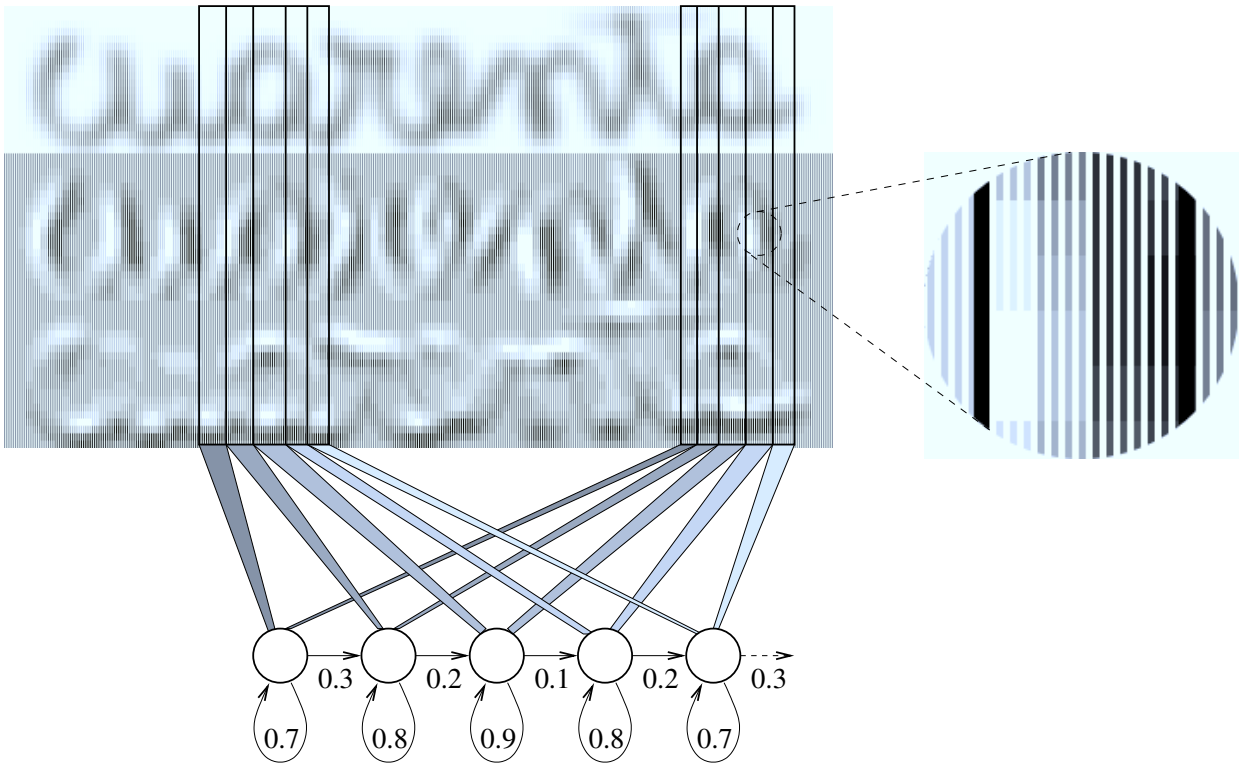


Figure 2: Example of 5-states HMM modeling (feature vectors sequences) of instances of the character "a" within the Spanish word "cuarenta" (forty). The states are shared among all instances of characters of the same class. The zones modelled by each state show graphically subsequences of feature vectors (see details in the magnifying-glass view) compounded by stacking the normalized grey level and its both derivatives features.

signments produced between word images and transcriptions (see section 4).

The remainder of this paper is organized as follows. First, the alignment framework is introduced and formalized in section 2. Then, an implemented prototype is described in section 3. The alignment evaluation metrics are presented in section 4. The experiments and results are commented in section 5. Finally, some conclusions are drawn in section 6.

2 HMM-based HTR and Viterbi alignment

HMM-based handwritten text recognition is briefly outlined in this section, followed by a more detailed presentation of the Viterbi alignment approach.

2.1 HMM HTR Basics

The traditional handwritten text recognition problem can be formulated as the problem of finding a most likely word sequence $\hat{\mathbf{w}} = \langle w_1, w_2, \ldots, w_n \rangle$, for a given handwritten sentence (or line) image represented by a feature vector sequence $\mathbf{x} = x_1^p = \langle x_1, x_2, \ldots, x_p \rangle$, that is:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} Pr(\mathbf{w}|\mathbf{x}) \\ &= \arg\max_{\mathbf{w}} Pr(\mathbf{x}|\mathbf{w}) \cdot Pr(\mathbf{w}) \end{aligned} \tag{1}$$

where $Pr(\mathbf{x}|\mathbf{w})$ is usually approximated by concatenated character Hidden Markov Models (HMMs) (Jelinek, 1998; Bazzi et al., 1999), whereas $Pr(\mathbf{w})$ is approximated typically by an n-gram word language model (Jelinek, 1998).

Thus, each character class is modeled by a continuous density left-to-right HMM, characterized by a set of states and a Gaussian mixture per state. The Gaussian mixture serves as a probabilistic law to model the emission of feature vectors by each HMM state. Figure 2 shows an example of how a HMM models a feature vector sequence corresponding to

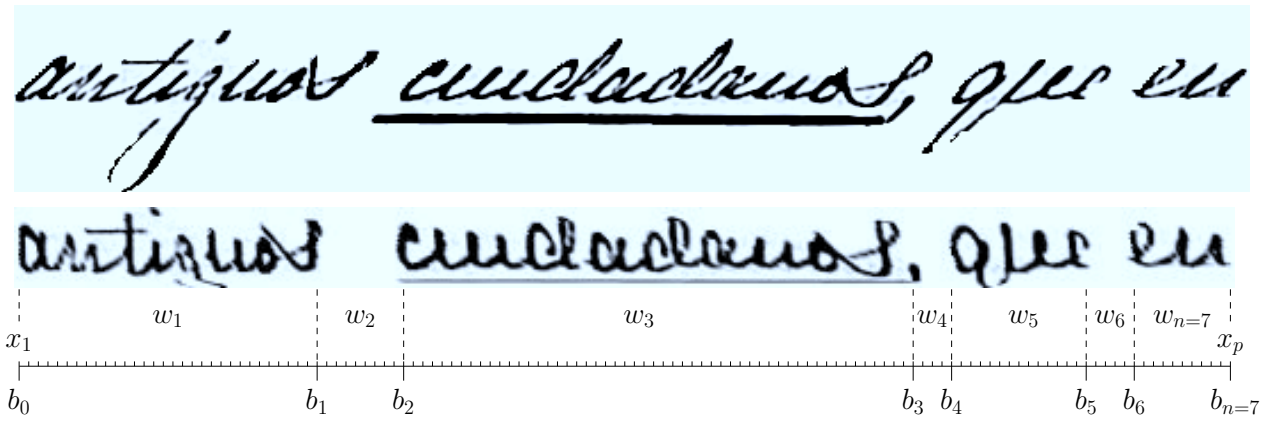


Figure 3: Example of segmented text line image along with its resulting deslanted and size-normalized image. Moreover, the alignment marks ($b_0 \dots b_8$) which delimit each of the words (including word-spaces) over the text image feature vectors sequence $\mathbf{x}$.

character "a". The process to obtain feature vector sequences from text images as well as the training of HMMs are explained in section 3.

HMMs as well as n-grams models can be represented by stochastic finite state networks (SFN), which are integrated into a single global SFN by replacing each word character of the n-gram model by the corresponding HMM. The search involved in the equation (1) to decode the input feature vectors sequence $\mathbf{x}$ into the more likely output word sequence $\hat{\mathbf{w}}$, is performed over this global SFN. This search problem is adequately solved by the Viterbi algorithm (Jelinek, 1998).

2.2 Viterbi Alignment

As a byproduct of the Viterbi solution to (1), the feature vectors subsequences of $\mathbf{x}$ aligned with each of the recognized words $w_1, w_2, \dots, w_n$ can be obtained. These implicit subsequences can be visualized into the equation (1) as follows:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{\mathbf{b}} Pr(\mathbf{x}, \mathbf{b}|\mathbf{w}) \cdot Pr(\mathbf{w}) \quad (2)$$

where $\mathbf{b}$ is an *alignment*; that is, an ordered sequence of $n+1$ marks $\langle b_0, b_1, \dots, b_n \rangle$, used to demarcate the subsequences belonging to each recognized word. The marks b_0 and b_n always point out to the first and last components of $\mathbf{x}$ (see figure 3).

Now, approximating the sum in (2) by the dominant term:

$$\hat{\mathbf{w}} \approx \arg\max_{\mathbf{w}} \max_{\mathbf{b}} Pr(\mathbf{x}, \mathbf{b}|\mathbf{w}) \cdot Pr(\mathbf{w}) \quad (3)$$

where $\hat{\mathbf{b}}$ is the optimal alignment. In our case, we are not really interested in proper text recognition because the transcription is known beforehand. Let $\tilde{\mathbf{w}}$ be the given transcription. Now, $Pr(\mathbf{w})$ in equation 3 is zero for all $\mathbf{w}$ except $\tilde{\mathbf{w}}$, for which $Pr(\tilde{\mathbf{w}}) = 1$. Therefore,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} Pr(\mathbf{x}, \mathbf{b}|\tilde{\mathbf{w}}) \quad (4)$$

which can be expanded to,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} Pr(\mathbf{x}, b_1|\tilde{\mathbf{w}}) Pr(\mathbf{x}, b_2|b_1, \tilde{\mathbf{w}}) \dots \dots Pr(\mathbf{x}, b_n|b_1 b_2 \dots b_{n-1}, \tilde{\mathbf{w}}) \quad (5)$$

Assuming independence of each b_i mark from $b_1 b_2 \dots b_{i-1}$ and assuming that each subsequence $x_{b_{i-1}}^{b_i}$ depends only of $\tilde{w}_i$, equation (5) can be rewritten as,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} Pr(x_{b_0}^{b_1}|\tilde{w}_1) \dots Pr(x_{b_{n-1}}^{b_n}|\tilde{w}_n) \quad (6)$$

This simpler Viterbi search problem is known as "forced recognition".

3 Overview of the Alignment Prototype

The implementation of the alignment prototype involved four different parts: document image preprocessing, line image feature extraction, HMMs training and alignment map generation.

Document image preprocessing encompasses the following steps: first, skew correction is carried out on each document page image; then background removal and noise reduction is performed by applying a bi-dimensional median filter (Kavallieratou and Stamatatos, 2006) on the whole page image. Next, a text line extraction process based on local minimums of the horizontal projection profile of page image, divides the page into separate line images (Marti and Bunke, 2001). In addition connected components has been used to solve the situations where local minimum values are greater than zero, making impossible to obtain a clear text line separation. Finally, slant correction and non-linear size normalization are applied (Toselli et al., 2004; Romero et al., 2006) on each extracted line image. An example of extracted text line image is shown in the top panel of figure 3, along with the resulting deslanted and size-normalized image. Note how non-linear normalization leads to reduced sizes of ascenders and descenders, as well as to a thiner underline of the word "ciudadanos".

As our alignment prototype is based on Hidden Markov Models (HMMs), each preprocessed line image is represented as a sequence of feature vectors. To do this, the feature extraction module applies a grid to divide line image into $N \times M$ squared cells. In this work, $N = 40$ is chosen empirically (using the corpus described further on) and M must satisfy the condition $M/N = \textit{original image aspect ratio}$. From each cell, three features are calculated: normalized gray level, horizontal gray level derivative and vertical gray level derivative. The way these three features are determined is described in (Toselli et al., 2004). Columns of cells or *frames* are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in its constituent cells.

Hence, at the end of this process, a sequence of M 120-dimensional feature vectors (40 normalized gray-level components, 40 horizontal and 40 vertical derivatives components) is obtained. An example of feature vectors sequence, representing an image of the Spanish word *"cuarenta"* (forty) is shown in figure 2.

As it was explained in section 2.1, characters are modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. This topology (number of HMM states and Gaussian densities per state) was determined by tuning empirically the system on the corpus described in section 5.1. Once a HMM "*topology*" has been adopted, the model parameters can be easily trained from images of continuously handwritten text (*without any kind of segmentation*) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called *forward-backward or Baum-Welch re-estimation* (Jelinek, 1998).

The last phase in the alignment process is the generation of the mapping proper by means of Viterbi "forced recognition", as discussed in section 2.2.

4 Alignment Evaluation Metrics

Two kinds of measures have been adopted to evaluate the quality of alignments. On the one hand, the average value and standard deviation (henceforward called MEAN-STD) of the absolute differences between the system-proposed word alignment marks and their corresponding (correct) references. This gives us an idea of the geometrical accuracy of the alignments obtained. On the other hand, the alignment error rate (AER), which measures the amount of erroneous assignments produced between word images and transcriptions.

Given a reference mark sequence $\mathbf{r} = \langle r_0, r_1, \ldots, r_n \rangle$ along with an associated tokens sequence $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$, and a segmentation marks sequence $\mathbf{b} = \langle b_0, b_1, \ldots, b_n \rangle$ (with $r_0 = b_0 \wedge r_n = b_n$), we define the MEAN-STD and AER metrics as follows:

MEAN-STD: The average value and standard deviation of absolute differences between reference and proposed alignment marks, are given by:

$$\mu = \frac{\sum_{i=1}^{n-1} d_i}{n-1} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n-1} (d_i - \mu)^2}{n-1}} \qquad (7)$$

where $d_i = |r_i - b_i|$.

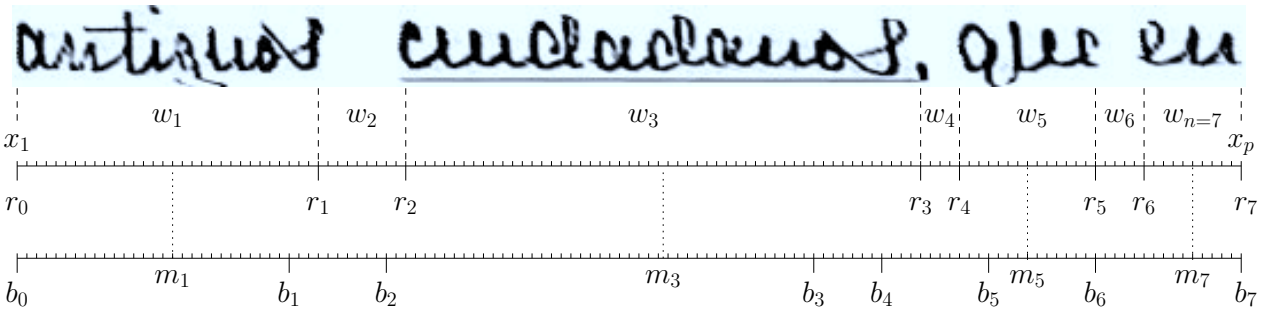


Figure 4: Example of AER computation. In this case $N = 4$ (only no word-space are considered: w_1, w_3, w_5, w_7) and w_5 is erroneously aligned with the subsequence $x_{b_5}^{b_6}$ ($m_5 \notin (b_4, b_5)$). The resulting AER is 25%.

AER: Defined as:

$$\text{AER}(\%) = \frac{100}{N} \sum_{j:w_j \neq b} e_j$$
$$e_j = \begin{cases} 0 & b_{j-1} < m_j < b_j \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

where b stands for the blank-space token, $N < n$ is the number of real words (i.e., tokens which are not b, and $m_j = (r_{j-1} + r_j)/2$.

A good alignment will have a μ value close to 0 and small σ. Thus, MEAN-STD gives us an idea of how accurate are the automatically computed alignment marks. On the other hand, AER assesses alignments at a higher level; that is, it measures mismatches between word-images and ASCII transcriptions (tokens), excluding word-space tokens. This is illustrated in figure 4, where the AER would be 25%.

5 Experiments

In order to test the effectiveness of the presented alignment approach, different experiments were carried out. The corpus used, as well as the experiments carried out and the obtained results, are reported in the following subsections.

5.1 Corpus description

The corpus was compiled from the legacy handwriting document identified as *Cristo-Salvador*, which was kindly provided by the *Biblioteca Valenciana Digital* (BIVALDI). It is composed of 53 text page images, scanned at 300dpi and written by only one writer. Some of these page images are shown in the figure 5.

As has been explained in section 3, the page images have been preprocessed and divided into lines, resulting in a data-set of 1,172 text line images. In this phase, around 4% of the automatically extracted line-separation marks were manually corrected. The transcriptions corresponding to each line image are also available, containing 10,911 running words with a vocabulary of 3,408 different words.

To test the quality of the computed alignments, 12 pages were randomly chosen from the whole corpus pages to be used as references. For these pages the true locations of alignment marks were set manually.

Table 1 summarized the basic statistics of this corpus and its reference pages.

Number of:	References	Total	Lexicon
pages	12	53	–
text lines	312	1,172	–
words	2,955	10,911	3,408
characters	16,893	62,159	78

Table 1: Basic statistics of the database

5.2 Experiments and Results

As mentioned above, experiments were carried out computing the alignments line-by-line. Two different HMM modeling schemes were employed. The first one models each of the 78 character classes using a different HMM per class. The second scheme uses 2 HMMs, one to model all the 77 no-blank character classes, and the other to model only the blank "character" class. The HMM topology was identical for all HMMs in both schemes: left-to-right with 6 states and 64 Gaussian mixture com-

Figure 5: Examples page images of the corpus "Cristo-Salvador", which show backgrounds of big variations and uneven illumination, spots due to the humidity, marks resulting from the ink that goes through the paper (called bleed-through), etc.

ponents per state.

As has been explained in section 4, two different measures have been adopted to evaluate the quality of the obtained alignments: the MEAN-STD and the AER. Table 2 shows the different alignment evaluation results obtained for the different schemes of HMM modeling.

	78-HMMs	2-HMMs
AER (%)	7.20	25.98
μ (mm)	1.15	2.95
σ (mm)	3.90	6.56

Table 2: Alignment evaluation results 78-HMMs and 2-HMMs.

From the results we can see that using the 78 HMMs scheme the best AER is obtained (7.20%). Moreover, the relative low values of μ and σ (in millimeters) show that the quality of the obtained alignments (marks) is quite acceptable, that is they are very close to their respective references. This is illustrated on the left histogram of figure 6.

The two typical alignment errors are known as over-segmentation and under-segmentation respectively. The over-segmentation error is when one word image is separated into two or more fragments. The under-segmentation error occurs when two or more images are grouped together and returned as one word. Figure 7 shows some of them.

6 Remarks and Conclusions

Given a manuscript and its transcription, we propose an alignment method to map every word image on the manuscript with its respective ASCII word on the transcript. This method takes advantage of the implicit alignment made by Viterbi decoding used in text recognition with HMMs.

The results reported in the last section should be considered preliminary.

Current work is under way to apply this alignment approach to the whole pages, which represents a more general case where the most corpora do not have transcriptions set at line level.

References

I. Bazzi, R. Schwartz, and J. Makhoul. 1999. An Omnifont Open-Vocabulary OCR System for English and

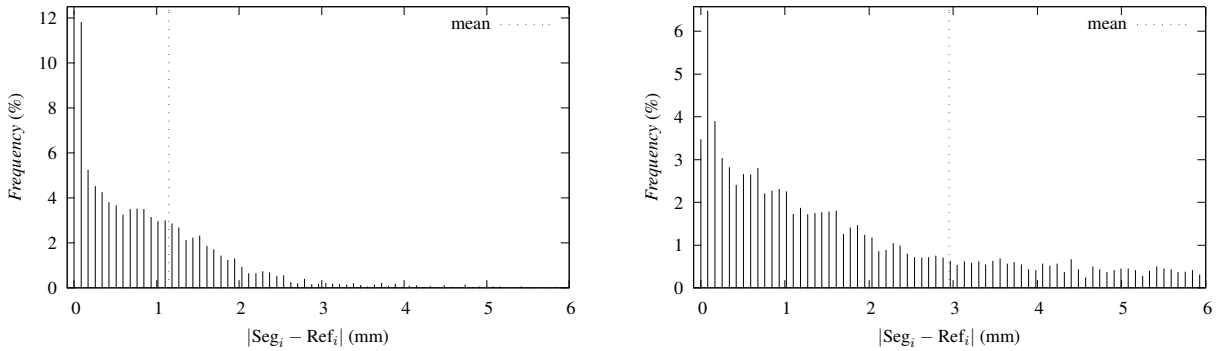


Figure 6: $|r_i - b_i|$ distribution histograms for 78-HMMs (left) and 2-HMMs (right) modelling schemes.

Figure 7: Word alignment for 6 lines of a particularly noisy part of the corpus. The four last words on the second line as well as the last line illustrate some of over-segmentation and under-segmentation error types.

Arabic. *IEEE Trans. on PAMI*, 21(6):495–504.

Chen Huang and Sargur N. Srihari. 2006. Mapping Transcripts to Handwritten Text. In Suvisoft Ltd., editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, pages 15–20, La Baule, France, October.

F. Jelinek. 1998. *Statistical Methods for Speech Recognition*. MIT Press.

Ergina Kavallieratou and Efstathios Stamatatos. 2006. Improving the quality of degraded document images. In *DIAL '06: Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06)*, pages 340–349, Washington, DC, USA. IEEE Computer Society.

E. M. Kornfield, R. Manmatha, and J. Allan. 2004. Text Alignment with Handwritten Documents. In *First International Workshop on Document Image Analysis for Libraries (DIAL)*, pages 195–209, Palo Alto, CA, USA, January.

U.-V. Marti and H. Bunke. 2001. Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. *Int. Journal of Pattern Recognition and Artificial In telligence*, 15(1):65–90.

V. Romero, M. Pastor, A. H. Toselli, and E. Vidal. 2006. Criteria for handwritten off-line text size normalization. In *Procc. of The Sixth IASTED international Conference on Visualization, Imaging, and Image Processing (VIIP 06)*, Palma de Mallorca, Spain, August.

A. H. Toselli, A. Juan, D. Keysers, J. Gonzlez, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta. 2004. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 18(4):519–539, June.

M. Zimmermann and H. Bunke. 2002. Automatic Segmentation of the IAM Off-Line Database for Handwritten English Text. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 4*, page 40035, Washington, DC, USA. IEEE Computer Society.

Retrieving lost information from textual databases: rediscovering expeditions from an animal specimen database

Marieke van Erp
Dept. of Language and Information Sciences
Tilburg University, P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands
M.G.J.vanErp@uvt.nl

Abstract

Importing large amounts of data into databases does not always go without the loss of important information. In this work, methods are presented that aim to rediscover this information by inferring it from the information that is available in the database. From and animal specimen database, the information to which expedition an animal that was found belongs is rediscovered. While the work is in an early stage, the obtained results are promising, and prove that it is possible to rediscover expedition information from the database.

1 Introduction

Databases made up of textual material tend to contain a wealth of information that remains unexplored with simple keyword-based search. Maintainers of the databases are often not aware of the possibilities offered by text mining methods to discover hidden information to enrich the basic data. In this work several machine learning methods are explored to investigate whether 'hidden information' can be extracted from an animal specimen database belonging to the Dutch National Museum for Natural History, Naturalis[1]. The database is a combination of information about objects in the museum collection from handwritten data sources in the museum, such as journal-like entries that are kept by biologists while collecting animal or plant specimens on expedition and tables that link the journal entries to the museum register. What is not preserved in the transition from the written sources to the database is the name of the expedition druing which an animal specimen was found.

By expedition, the following event is implied: a group of biologists went on expedition together in a country during a certain time period. Entries in the database that belong to this expedition can be collected by one or a subset of the participating biologists. For researchers at the natural history museum it would be helpful to have access to expedition information in their database, as for biodiversity research they sometimes need overviews of expeditions. It may also help further enrichment of the database and cleansing, because if the expedition information is available, missing information in certain fields, such as the country where a specimen was found, may be inferred from the information on other specimens found during the same expedition. Currently, if one wants to retrieve all objects from the database that belong to an expedition, one would have to create a database query that contains the exact data boundaries of the expeditions and the names of all collectors involved. Either one of these bits of information is not enough, as the same group of biologists may have participated in an expedition more than once, and the database may also contain expeditions that overlap in time. In this paper a series of experiments is described to find a way to infer expedition information from the information available in the database. To this end, three approaches are compared: supervised machine learning, unsupervised machine learning, and rule-based methods.

[1] http://www.naturalis.nl

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 17–24,
Prague, 28 June 2007.

The obtained results vary, but prove that it is possible to extract the expedition information from the data at hand.

2 Related Work

The field of data mining, which is concerned with the extraction of implicit, previously unknown and potentially useful information from data (Frawley et al., 1992), is a branch of research that has become quite important recently as every day the world is flooded with larger amounts of information that are impossible to analyse manually. Data mining can, for instance, help banks identify suspicious transactions among the millions of transactions that are executed daily (Fayyad and Uthurusamy, 1996), or automatically classify protein sequences in genome databases (Mewes et al., 1999), or aid a company in creating better customer profiles to present customers with personalised ads and notifications (Linden et al., 2003). Knowledge discovery approaches often rely on machine learning techniques as these are particularly well suited to process large amounts of data to find similarities or dissimilarities between instances (Mitchell, 1997).

Traditionally, governments and companies have been interested in gaining more insight into their data by applying data mining techniques. Only recently , digitisation of data in the cultural heritage domain has taken off, which means that there has not been much work done on knowledge discovery in this domain. Databases in this domain are often created and maintained manually and are thus often significantly smaller than automatically generated databases from, for example, customers' purchase information in a large company.

This means it is not clear whether data mining techniques, aimed at analysing enormous amounts of data, will work for the data at hand. This is investigated here. Manual data typically also contains more spelling variations/errors and other inconsistencies than automatically generated databases, due to different persons entering data into the database. Therefore, before one can start the actual process of knowledge discovery, it is very important to carefully select, clean and model the data one wants to use in order to avoid using data that is too sparse (Chapman, 2003). This applies in particular to databases that contain large amounts of textual information, which are quite prevalent in the cultural heritage domain. Examples of textual databases can be found freely on the internet, such as the databases of the Global Biodiversity Information Facility[2], the University of St. Andrews Photographic Collection[3], and the Internet Movie Database[4].

3 Data

The data that has been used in this experiment is an animal specimen database from the Dutch National Museum for Natural History. The database currently contains 16,870 entries that each represent an object stored in the museum's reptiles and amphibians collection. The entries provide a variety of information about the objects in 37 columns, such as the scientific name of the object, how the specimen is kept (in alcohol, stuffed, pinned) and under which registration number, where it was found, by whom and under which circumstances, the name of the person who determined the species of the animal and the name of the person who first described the species. Most fields are rather compact; they only contain a numeric value or a textual value consisting of one or several words. The database also contains fields of which the entries consist of longer stretches of text, such as the 'special remarks' field, describing anything about the object that did not fit in the other database fields and 'biotope', describing the biotic and abiotic components of the habitat from which the object was collected. Dutch is the most frequent language in the database, followed by English. Also some Portuguese and German entries occur. Taxonomic values, i.e., the scientific names of the animal specimens, are in a restricted type of Latin. A snippet of the database can be found in Figure 1.

3.1 Data Construction

In order to be able the measure the performance of the approaches used in the experiments, the database was annotated manually with expedition information. Adding this information was possible because there was access to the original field books from which the database is made up. Annotating 8166

[2] http://www.gbif.org/

[3] http://special.st-andrews.ac.uk/saspecial/

[4] http://www.imdb.com/

Collector	Coll. Date	Coll. #	Class	Genus	Species	Country	Expedition
Buttikofer, J.	30-07-1881	424	Reptilia	Lamprolepis	lineatus	132	buttikoferliberia1881
Buttikofer, J. & Sala	09-10-1881	504	Amphibia	Bufo	regularis	132	buttikoferliberia1881
M. Dachsel	02-05-1971	971-MSH186	Reptilia	Blanus	mettetali	156	mshbrazil71
Hoogmoed, M.S.	04-05-1971	1971-MSH187	Reptilia	Quendenfeldtia	trachyblepharus	156	mshbrazil71
Hoogmoed, M.S.	09-05-1971	1971-MSH202	Reptilia	Lacerta	hispanica	156	mshbrazil71
C. Schuil	14-03-1972	1972-MSH35	Amphibia	Ptychadaena	sp.	92	mshghana72
P. Lavelle	-03-1972	1972-MSH40	Reptilia	Crotaphopeltis	hotamboeia	92	mshghana72
Hoogmoed, M.S.	23-03-1972	1972-MSH55	Amphibia	Phrynobatrachus	plicatus	92	mshghana72

Figure 1: Snippet of the animal specimen database

entries with this information took one person about 2 days. There were 8704 entries to which no expedition is assigned, either because these specimens were not collected during an expedition or because it was not possible to determine the expedition. These entries were excluded from the experiments . Expeditions which contained 10 or fewer entries were also excluded because these would make the data set too sparse. A total of 7831 database entries were used in this work, divided into 60 expeditions. Although the 'smallest' expeditions were excluded from the experiments, the sizes of the expeditions still vary greatly: between 2170 and 11 items ($\sigma = 310.04$). This is mainly due to the fact that new items are still added to the database continuously, in a rather random order, hence some expeditions are more completely represented than others.

The database contains several fields that contain information that will probably not be that useful for this work. Information that was excluded was the specimen's sex, the number of specimens (in cases where one database entry refers to several specimens, for instance kept together in a jar), how the animal is preserved, and fields that contain information not on the specimen itself or how it was found but on the database (e.g., when the database entry was added and by whom). Values from the 'altitude' and 'coordinates' fields were also not included in the experiments as this is information is too often missing in the database to be of any use (altitude information is missing in 85% of the entries and coordinates in 96%).

Some information in the database is repetitive; there is for instance a field called 'country' containing the name of the country in which a specimen was found, but there is also a field called 'country-id' in which the same information is encoded as a numerical value. The latter is more often filled than the 'country' field, which also contains values in different languages, and thus it makes more sense to only include values from the 'country-id' field. A small conversion is applied to rule out that an algorithm will interpret the intervals between the different values as a measure of geographical proximity between the values, as the country values are chosen alphabetically and do not encode geographical location.

In some cases it seemed useful to have an algorithm employ interval relations between numbers. The fields 'registration number' and 'collection number' were used as such. These fields sometimes contain some alphabetical values: certain collectors, for instance, included their initials in their series of collection registration numbers. These were converted to a numeric code to obtain completely numeric values with preservation of the collector information. This also goes for the fields in the database that contain information on dates, i.e., the 'date of determination', the 'date the specimen came into the museum' and the 'collection date' fields. The collection date is the most important date here as this directly links to an expedition. The other dates might provide indirect information, for instance if the collection date is missing (which is the case in 14%). To aid clustering, the dates were normalised to a number, possibly the algorithm benefits from the fact that a small numerical interval means that the dates are close together.

Person names from the 'author', 'collector', 'determiner', and 'donator' fields were normalised to a 'first name - last name' format. From values from the taxonomic fields ('class', 'order', 'family', 'genus', 'species', and 'sub species'), and 'town/village' and 'province/state' fields, as well as from the person name fields, capitals, umlauts, accents and any other non-alphanumerical characters were removed.

It proved that certain database fields were not suitable for inclusion in the experiments. This goes for

the free text fields 'biotope', 'location' and 'special remarks'. Treating these values as they are will result in data that is too sparse, as their values are extremely varied. Preliminary experiments to see if it was possible to select only certain parts of these fields did not yield any satisfying results and was therefore abandoned.

This resulted in feature vectors containing 18 features, plus the manually assigned expedition class.

4 Methodology

The majority of the experiments that were carried out in an attempt to infer the expedition information from the database involved machine learning. Therefore in this section three algorithms for supervised learning are described, followed by a clustering algorithm for unsupervised learning. This section is concluded with a description of the evaluation metrics for clusters used by the different approaches.

Algorithms

The first algorithm that was used is the ***k*-Nearest Neighbour** algorithm (*k*-NN) (Aha et al., 1991; Cover and Hart, 1967; DeVijver and Kittler, 1982). This algorithm is an example of a lazy learner: it does not model the training data it is given, but simply stores each instance of the training data in memory. During classification it compares the item it needs to classify to each item in its memory and assigns the majority class of the closest *k* (in these experiments *k*=1) instances to the new item. To determine which instances are closest, a variety of distance metrics can be applied. In this experiment the standard settings in the TiMBL implementation (Daelemans et al., 2004), developed at the ILK research group at Tilburg University, were used. The standard distance metric in the TiMLB implementation of *k*-NN is the Overlap metric, given in Equations1 and 2. Δ(X,Y) is the distance between instances X and Y, represented by *n* features, where δ is the distance between the features.

$$\Delta(X,Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{1}$$

where:

$$\delta(x_i, y_i) = \begin{cases} abs & \text{if numeric, else} \\ 0 & \text{if} x_i = y_i \\ 1 & \text{if} x_i \neq y_i \end{cases} \tag{2}$$

The second algorithm that was used is the **C4.5** decision tree algorithm (Quinlan, 1986). In the learning phase, it creates a decision tree in a recursive top-down process in which the database is partitioned according to the feature that separates the classes best; each node in the tree represents one partition. Deeper nodes represent more class-homogeneous partitions. During classification, C4.5 traverses the tree in a deterministic top-down pass until it meets a class-homogeneous end node, or a non-ending node when a feature-value test is not represented in the tree.

Naive Bayes is the third algorithm that was used in the experiments. It computes the probability of a certain expedition, given the observed training data according to the formula given in Equation 3. In this formula v_{NB} is the target expedition value, chosen from the maximally probably hypothesis ($\overset{argmax}{v_j \epsilon V} P(v_j)$, i.e., the expedition with the highest probability) given the product of the probabilities of the features ($\prod_i P(a_i|v_j)$).

$$v_{NB} = \overset{argmax}{v_j \epsilon V} P(v_j) \prod_i P(a_i|v_j) \tag{3}$$

For both the C4.5 algorithm and Naive Bayes the WEKA machine learning environment (Witten and Frank, 2005), that was developed at the University of Waikato, New Zealand, was used.

A quite different machine learning approach that was applied to try to identify expeditions in the reptiles and amphibians database is **clustering**. Clustering methods are unsupervised, i.e., they do not require annotated data, and in some cases not even the number of expeditions that are in the data. Items in the data set are grouped according to similarity. A maximum dissimilarity between the group members may be specified to steer the algorithm, but other than that it runs on its own. For an extensive overview of clustering methods see Jain et al., (1999). For this work, the options in choosing an implementation of a clustering algorithm were limited because many data mining tools are designed

only for numerical data, therefore the WEKA machine learning environment was also used for the clustering experiments. As clustering is computationally expensive, it was only possible to run experiments with WEKA's implementation of the Expectation Maximisation (**EM**) algorithm (Dempster et al., 1977). Preliminary experiments with other algorithms indicated execution times in the order of months. The EM algorithm iteratively tries to converge to a maximum likelihood by first computing an expectation of the likelihood of a certain clustering, then maximising this likelihood by computing the maximum likelihood estimates of the features. Termination of the algorithm occurs when the predefined number of iterations has been carried out, or when the overall likelihood (the measure of how 'good' a clustering is) does not increase significantly with each iteration.

Cluster Evaluation

Since the data is annotated with expedition information it was possible to use external quality measures (Steinbach et al., 2000). Three different evaluation measures were used: **accuracy**, **entropy** (Shannon, 1948), and the **F-measure** (van Rijsbergen, 1979).

The evaluation of results for the supervised learning algorithms was calculated in a straightforward way: because the classifier knows which expeditions there are and which entries belong to which expedition, it checks the expeditions it assigned to the database entries to the manually assigned expeditions and reports the overlap as accuracy.

It gets a little bit more complicated with entropy. Entropy is a measure of informativity, i.e., the minimum number of bits of information needed to encode the classification of each instance. If the expedition clusters are uniform, i.e., all items in the cluster are very similar, the entropy will be low. The main problem with using entropy for evaluation of clusters is that the best score (an entropy of 0) is reached when every cluster contains exactly on instance. Entropy is calculated as follows: first, the main class distribution, i.e., per cluster the probability that a member of that cluster belongs to a certain cluster, is computed. Using that distribution the entropy of each cluster is calculated via the formula in Equation 4. For a set of clusters the total entropy is then computed via the formula in Equation 5, in which *m* is the total number of clusters, s_y the size of cluster *y* and *n* the total number of instances.

$$E_y = -\sum_x P_{xy} log(P_{xy}) \tag{4}$$

$$E_{total} = \sum_{y-1}^{m} \frac{s_y \cdot E_y}{n} \tag{5}$$

The F-measure is the harmonic mean of precision and recall, and is commonly used in information retrieval. In information retrieval recall is the proportion of relevant documents retrieved out of the total set of relevant documents. When applied to clustering a 'relevant document' is an instance that is assigned correctly to a certain expedition, the set of all relevant documents is the set of all instances beloning to that expedition. Precision is the number of relevant documents retrieved from the total number of documents. So when applied to cluster evaluation this means the number of instances of an expedition that were retrieved from the total number of instances (Larsen and Aone, 1999). This boils down to Equations 6 and 7 in which *x* stands for expedition, *y* for cluster, n_{xy} for the number of instances belonging to expedition *x* that were assigned to cluster *y*, and n_x is the number of items in expedition *x*.

$$Recall(x, y) = \frac{n_{xy}}{n_x} \tag{6}$$

$$Precision(x, y) = \frac{n_{xy}}{n_y} \tag{7}$$

The F-measure for a cluster *y* with respect to expedition *x* is then computed via Equation 8. The F-measure of the entire set of clusters is computed through the function in Equation 9, which takes the weighted average of the maximum F-measure per expedition.

$$F(x, y) = \frac{2 \cdot Recall(x, y) \cdot Precision(x, y)}{Precision(x, y) + Recall(x, y)} \tag{8}$$

$$F = \sum_x \frac{n_x}{n} \max\{F(x, y)\} \tag{9}$$

5 Experiments and Results

First, two baselines were set to illustrate the situation if no machine learning or other techniques would be applied to the database. if one were to randomly assign one of the 60 expeditions t the entries this would go well in 1.7% of the cases. If all entries were labelled as belonging to the largest expedition this would yield an accuracy of 28%. In all machine learning experiments 10-fold cross validation was used for testing performance.

A series of supervised machine learning experiments was carried out first to investigate whether it is possible to extract the expeditions during which the animal specimens were found at all. Three learning algorithms were applied to the complete data set, which yielded accuracies between 88% and 98%. Feature selection experiments with the C4.5 decision tree algorithm indicated that features 'town/village', 'collection number', 'registration number', 'collector' and 'collection date' were considered most informative for this task, hence the experiments were repeated with a data set containing only those features. The results of both series of experiments are to be found in Table 1. For the C4.5 and Naive Bayes experiments the accuracy deteriorates significantly when using only the selected features ($\alpha = 0.05$, computing using McNemar's test (McNemar, 1962)), but it stays stable for the *k*-NN classifier. This indicates that not all data is needed to infer the expeditions, but that it matters greatly which approach is taken. However, as neither of the algorithm benefits from it, feature selection was not further explored.

Algorithm	All feat.	Sel. feat.
k-NN	95.9%	95.9%
C.4.5	98.3%	94.4%
NaiveBayes	88.1%	73.5%

Table 1: Accuracy of supervised machine learning experiments using all features and selected features

In these experiments all database entries were annotated with expedition information, which in a real setting is of course not the case. Through running a series of experiments with significantly smaller amounts of training data it was found that by using only as little as 5% of the training data (amounting to 392 instances) already an accuracy of 85% is reached. Annotating this amount of data with expedition information would take on person less than an hour. By only using 45% of the training data an accuracy of 97% is reached[5]. In Figure 2 the complete learning curve of the *k*-NN classifier is shown.

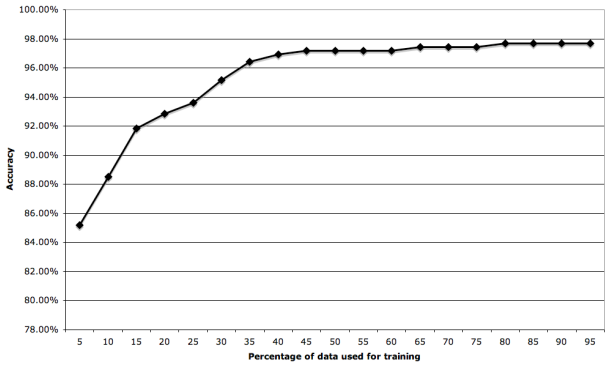


Figure 2: Accuracy of *k*-NN per percentage of training data

Ideally, one does not want to annotate data at all, therefore the use of a clustering algorithm was explored. For this, the EM algorithm from the WEKA machine learning environment was used. The result, as shown in Table 2, is not quite satisfying, but still well above the set baselines. As can be seen in Table 2, the clustering algorithm does not come up with anywhere near as many clusters as needed and unfortunately WEKA does not present the user with many options to remedy this. An intermediate experiment between completely supervised and unsupervised machine learning was attempted, i.e., prespecifying a number of clusters for the algorithm to define, but this was computationally too expensive to carry out.

Algorithm	# Clusters	Accuracy
EM	7	46.0%

Table 2: Result of clustering experiment

Since the clustering algorithm does not achieve an accuracy that is satisfying enough to use in a real setting and supervised learning requires annotated data, also a traditional, and quite different approach

[5]The slightly higher achieved accuracy in the learning curve experiments is due to the fact that the learning curve was not computed via cross-validation

was tried: namely finding expeditions via rules. Via a couple of simple rules the data set was split into possible expeditions using only information on collection dates, collector information and country information.

1. Sort dates in ascending order, start a new expedition when the distance between two sequential dates is greater than the average distance of the collection dates

2. First, sort collector information in ascending order, then sort collection dates in ascending order, start a new expeditions when the distance between two dates is greater than the average distance between dates or when a new collector is encountered

3. First, sort by country information, then by collector, and finally by collection date, start a new expedition when country or collectors change, or when the distance between two dates is greater than the average distance between dates

Surprisingly, only grouping collection dates already yields an F-measure of .83. This includes 1299 entries that contain no information on the collection date, leaving those out would increase precision on the entries whose collection date is not missing to an F-measure of .94. In Table 3 results of the rule-based experiments are shown. It is expected that when the database is further populated the date-rule will work less well as there will be more expeditions that overlap. The date+collector-rule should remedy this, although it does not work very well yet as spelling variations in the collector names are not taken into account at the moment.

Rules	# Clusters	F-measure	Entropy
1	78	.83	.16
2	199	.75	.15
3	216	.73	.11

Table 3: Results of the rule-based experiments

6 Conclusions and Future Work

In this work various approaches were presented to rediscover expedition information from an animal specimen database. As expected, the supervised learning algorithms performed best, but the disadvantage in using such an approach is the requirement to provide annotated data. However, a series of experiments to gain more insight into the quantities of data necessary for a supervised approach to perform well, indicate that only a small set of annotated data is required in this case to obtain very reasonable results. If no training data is available, a rule-based approach is a realistic alternative. Although it must be kept in mind that rules need to be created manually for every new data set. For this data set relatively simple rules already proved to be quite effective, but for other data sets deriving rules can be much more complicated and thus more expensive. This particular set of rules is also expected to behave differently when the database is extended with more entries from overlapping expeditions.

For the experiments presented in this work, only entries from the database of which the expedition they belonged to was known were used, which constitutes only half of the database entries. Researchers at Naturalis estimate that about 30% of the database entries do not belong to an expedition, while the other 20% not included here belong to unknown expeditions. The decision to exclude the expedition-less entries was made as these entries would imbalance the data and impair evaluation as it would not be possible to check predictions against a 'real value'. If all database entries would belong to a known expedition the performance of the approaches described in this paper that satisfactory results could be achieved over the complete data set. To prove this hypothesis one would need to test the approaches on other data sets which can be completely annotated. Performing such tests might provide more insight into how well the approaches would deal with a data set where all entries have an associated expedition. The natural history museum has several other similar (but smaller) data sets, which might be suitable for this task, and which will be tested as part of future work for evaluating the approaches described here. It may also be interesting to investigate what can be inferred from the other fields defined in other data sets.

A less satisfying aspect of the research described in this paper is that many of the intended experiments with unsupervised machine learning were too

computationally expensive to be executed. Potential workarounds to the limitation of certain implementations of clustering algorithms, in that they only work on numeric data, are sought in converting the textual data to numeric values and in the investigations into implementations of algorithms that can deal with textual data.

A particular peculiarity of textual data, from which the rule-based approach suffers, is the fact that the same name or meaning can be conveyed in several ways. Spelling variations and errors were for instance not normalised. Hence the approaches treated 'Hoogmoed' and 'M S Hoogmoed' as two different values whereas they may very well refer to the same entity.

From this work it can be concluded that the expedition information can definitely be reconstructed from the animal specimen database that was used here, but for it to be used in a real world application it still needs to be tested and fine-tuned on other data sets and extended to be able to deal with entries that are not associated with any expedition.

Acknowledgments

The research reported in this paper was funded by NWO, the Netherlands Organisation of Scientific Research as part of the CATCH programme. The author would like to thank the anonymous reviewers, and Antal van den Bosch and Caroline Sporleder for their helpful suggestions and comments.

References

David W. Aha, Dennis Kibler, and Mark K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Arthur D. Chapman. 2003. Notes on Environmental Data Quality-b. Data Cleaning Tools. Internal report, Centro de Referência em Informação Ambiental (CRIA).

T. M. Cover and P. E. Hart. 1967. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13:21–27.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. 2004. Timbl: Tilburg memory based learner, version 5.1, reference guide. Technical Report 04-02, ILK/Tilburg University.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodology)*, 39(1):1–38.

P. A. DeVijver and J. Kittler. 1982. *Pattern recognition. A statistical approach*. Prentice-Hall, London.

U. Fayyad and R. Uthurusamy. 1996. Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11):24–26.

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. 1992. Knowledge discovery in databases: An overview. *AI Magazine*, 13:57–70.

A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September.

Bjorner Larsen and Chinatsu Aone. 1999. Fast and effective text mining using linear-time document clustering. In *Proceedings of KDD-99*, San Diego, CA.

G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, Jan/Feb.

Q. McNemar. 1962. *Psychological Statistics*. Wiley, New York.

H. W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. 1999. Mips: a database for genomes and protein sequences. *Nucleic Acids Research*, 27(1):44–48.

Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. Technical report, Department of Computer Science, University of Minnesota.

Cornelis Joost van Rijsbergen. 1979. *Information Retrieval*. Buttersworth.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources

Tandeep Sidhu, Judith Klavans, and Jimmy Lin
College of Information Studies
University of Maryland
College Park, MD 20742
tsidhu@umiacs.umd.edu, {jklavans, jimmylin}@umd.edu

Abstract

We address the problem of mining text for relevant image metadata. Our work is situated in the art and architecture domain, where highly specialized technical vocabulary presents challenges for NLP techniques. To extract high quality metadata, the problem of word sense disambiguation must be addressed in order to avoid leading the searcher to the wrong image as a result of ambiguous — and thus faulty — metadata. In this paper, we present a disambiguation algorithm that attempts to select the correct sense of nouns in textual descriptions of art objects, with respect to a rich domain-specific thesaurus, the Art and Architecture Thesaurus (AAT). We performed a series of intrinsic evaluations using a data set of 600 subject terms extracted from an online National Gallery of Art (NGA) collection of images and text. Our results showed that the use of external knowledge sources shows an improvement over a baseline.

1. Introduction

We describe an algorithm that takes noun phrases and assigns a sense to the head noun or phrase, given a large domain-specific thesaurus, the Art and Architecture Thesaurus[1] (published by the Getty Research Institute). This research is part of the Computational Linguistics for Metadata Building (CLiMB) project (Klavans 2006, Klavans *in preparation*), which aims to improve image access by automatically extracting metadata from text associated with images. We present here a component of an overall architecture that automatically mines scholarly text for metadata terms. In order to filter and associate a term with a related concept, ambiguous terms must be clarified. The disambiguation of terms is a basic challenge in computational linguistics (Ide and Veronis 1990, Agirre and Edmonds 2006).

As more non-specialists in digital libraries search for images, the need for subject term access has increased. Subject terms enrich catalog records with valuable broad-reaching metadata and help improve image access (Layne 1994). Image seekers will receive more relevant results if image records contain terms that reflect conceptual, semantic, and ontological relationships. Furthermore, subject terms associated with hierarchical and faceted thesaural senses promise to further improve precision in image access. Such terms map to standardized thesaurus records that include the term's preferred, variant, and related names, including both broader and specific concepts, and other related concepts. This information can then be filtered, linked, and subsequently tested for usefulness in performing richer image access. As with other research on disambiguation, our hypothesis is that accurate assignment of senses to metadata index terms will results in higher precision for searchers. This hypothesis will be fully tested as we incorporate the disambiguation module in our end-to-end CLiMB Toolkit, and as we perform user studies.

Finding subject terms and mapping them to a thesaurus is a time-intensive task for catalogers

[1] http://www.getty.edu/research/conducting_research/vocabularies/aat/

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 25–32,
Prague, 28 June 2007.

(Rasmussen 1997, Ferguson and Intner 1998). Doing so typically involves reading image-related text or other sources to find subject terms. Even so, the lack of standard vocabulary in extensive subject indexing means that the enriched number of subject terms could be inadvertently offset by the vocabulary naming problem (Baca 2002).

This paper reports on our results using the subject terms in the AAT; the CLiMB project is also using the Thesaurus of Geographic Names (TGN) and the Union List of Artist Names (ULAN). Since the focus of this paper is on disambiguation of common nouns rather than proper nouns, the AAT is our primary resource.

2. Resources

2.1 Art and Architecture Thesaurus (AAT)

The AAT is a widely-used multi-faceted thesaurus of terms for the cataloging and indexing of art, architecture, artifactual, and archival materials. Since the AAT offers a controlled vocabulary for recording and retrieval of data in object, bibliographic, and visual databases, it is of interest to a wide community.

In the AAT, each concept is described through a record which has a unique ID, preferred name, record description, variant names, broader, narrower, and related terms. In total, AAT has 31,000 such records. For the purpose of this article, a record can be viewed as synonymous with sense. Within the AAT, there are 1,400 homonyms, *i.e.,* records with same preferred name. For example, the term *wings* has five senses in the AAT (see Figure 1 below).

Wings (5 senses):

- Sense#1: Used for accessories that project outward from the shoulder of a garment and are made of cloth or metal.
- Sense#2: Lateral parts or appendages of a work of art, such as those found on a triptych.
- Sense#3: The areas offstage and to the side of the acting area.
- Sense#4: The two forward extensions to the sides of the back on an easy chair.
- Sense#5: Subsidiary parts of buildings extending out from the main portion.

Figure 1: Selection of AAT records for term "wings"

Table 1 shows the breakdown of the AAT vocabulary by number of senses with a sample lexical item for each frequency.

# of Senses	# of Homonyms	Example
2	1097	bells
3	215	painting
4	50	alabaster
5	39	wings
6	9	boards
7	5	amber
8	2	emerald
9	1	plum
10	1	emerald green
11	1	magenta
12	1	ocher
13	1	carmine
14	2	slate

Table 1: Scope of the disambiguation problem in AAT

Note that there are potentially three tasks that could be addressed with our algorithm: (i) mapping a term to the correct sense in the AAT, (ii) selecting amongst closely related terms in the AAT, and (iii) mapping synonyms onto a single AAT entry. In this paper, our primary focus is on task (i); we handle task (ii) with a simple ranking approach; we do not address task (iii).

Table 1 shows that multiple senses per term makes mapping subject terms to AAT very challenging. Manual disambiguation would be slow, tedious, and unrealistic. Thus we explore automatic methods since, in order to identify the correct sense of a term in running text, each of these senses needs to be viewed in context.

2.2 The Test Collection

The data set of terms that we use for evaluation comes from the National Gallery of Art (NGA) online archive[2]. This collection covers paintings, sculpture, decorative arts, and works from the Middle Ages to the present. We randomly selected 20 images with corresponding text from this collection and extracted noun phrases to form the data set. The data set was divided into two categories: the training set and the test set. The training set consisted of 326 terms and was used

[2] http://www.nga.gov/home.htm

to develop the algorithm. The test set consisted of 275 terms and was used to evaluate.

Following standard procedure in word sense disambiguation tasks (Palmer et al. 2006), groundtruth for the data set was created manually by two labelers (referred to as Labeler 1 and Labeler 2 in Section 4 below). These labelers were part of the larger CLiMB project but they were not involved in the development of the disambiguation algorithm. The process of creating the groundtruth involved picking the correct AAT record for each of the terms in the data set. Terms not appearing in the AAT (as determined by the labelers) were given an AAT record value of zero. Each labeler worked independently on this task and had access to the online version of the AAT and the text where each term appeared. Interannotator agreement for the task was encouragingly high, at 85% providing a notional upper bound for automatic system performance (Gale et al. 1992).

Not all terms in this dataset required disambiguation; 128 terms (out of 326) under the training set and 96 terms (out of 275) under the test set required disambiguation, since they matched more than one AAT record. The dataset we selected was adequate to test our different approaches and to refine our techniques. We intend to run over more data as we collect and annotate more resources for evaluation.

2.3 SenseRelate AllWords[3] and WordNet[4]

SenseRelate AllWords (Banerjee and Pederson 2003, Patwardhan et al. 2003) is a Perl program that our algorithm employs to perform basic disambiguation of words. We have adapted SenseRelate for the purpose of disambiguating AAT senses.

Given a sentence, SenseRelate AllWords disambiguates all the words in that sentence. It uses word sense definitions from WordNet (in this case WordNet 2.1), a large lexical database of English nouns, verbs, adjectives, and adverbs. As an example, consider the text below:

With **more** than **fifty** individual scenes, the altarpiece was about fourteen feet wide.

The SenseRelate result is:

With **more#a#2** than **fifty#n#1** individual#n#1 scene#n#10 the altarpiece#n#1 be#v#1 about#r#1 fourteen#n#1 foot#n#2 wide#a#1

In the above example, more#a#2 means SenseRelate labeled *more* as an adjective and mapped it to second meaning of *more* (found in WordNet). fifty#n#1 means SenseRelate labeled *fifty* as a noun and mapped it to first meaning of *fifty* (found in WordNet). Note, that fifty#n#1 maps to a sense in WordNet, whereas in our algorithm it needs to map to an AAT sense. In Section 3, we show how we translate a WordNet sense to an AAT sense for use in our algorithm.

To perform disambiguation, SenseRelate requires that certain parameters be set: (1) the number of words around the target word (also known as the context window), and (2) the similarity measure. We used a value of 20 for the context window, which means that SenseRelate will use 10 words to the left and 10 words to the right of the target word to determine the correct sense. We used *lesk* as the similarity measure in our algorithm which is based on Lesk (1986). This decision was based on several experiments we did with various context window sizes and various similarity measures on a data set of 60 terms.

[3] http://sourceforge.net/projects/senserelate

[4] http://wordnet.princeton.edu/

3. Methodology

3.1 Disambiguation Algorithm

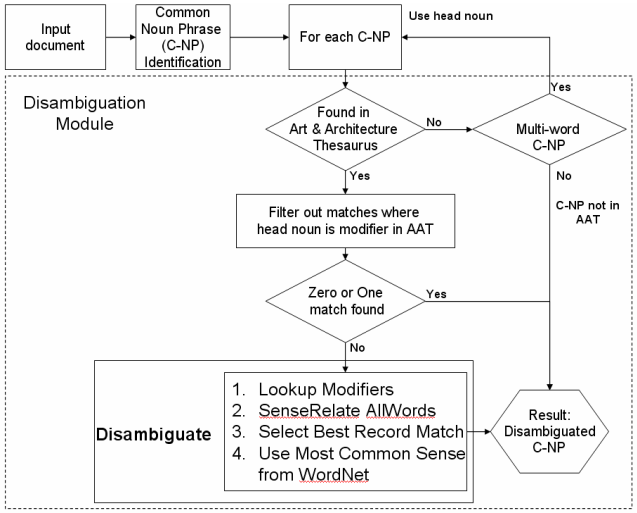


Figure 2: Disambiguation Algorithm

Figure 2 above shows that first we identify the noun phrases from the input document. Then we disambiguate each noun phrase independently by first looking it up in the AAT. If a record is found, we move on to the next step; otherwise we look up the head noun (as the noun phrase) in the AAT.

Second, we filter out any AAT records where the noun phrase (or the head noun) is used as an adjective (for a term like *painting* this would be *painting techniques*, *painting knives*, *painting equipment*, etc). Third, if zero records are found in the AAT, we label the term as "not found in AAT." If only one matching record is found, we label the term with the ID of this record. Fourth, if more than one record is found, we use the disambiguation techniques outlined in the next section to find the correct record.

3.2 Techniques for Disambiguation

For each of the terms, the following techniques were applied in the order they are given in this section. If a technique failed to disambiguate a term, we applied the next technique. If none of these techniques was able to disambiguate, we selected the first AAT record as the correct record. Findings for each technique are provided in the Results section below.

First, we used all modifiers that are in the noun phrase to find the correct AAT record. We searched for the modifiers in the record description, variant names, and the parent hierarchy names of all the matching AAT senses. If this technique narrowed down the option set to one record, then we found our correct record. For example, consider the term *ceiling coffers*. For this term we found two records: coffers (coffered ceiling components) and coffers (chests). The first record has the modifier *ceiling* in its record description, so we were able to determine that this was the correct record.

Second, we used SenseRelate AllWords and WordNet. This gave us the WordNet sense of our noun phrase (or its head noun). Using that sense definition from WordNet, we next examined which of the AAT senses best matches with the WordNet sense definition. For this, we used the word overlapping technique where we awarded a score of N to an AAT record where N words overlap with the sense that SenseRelate picked. The AAT record with the highest score was selected as the correct record. If none of the AAT records received any positive score (above a certain threshold), then it was decided that this technique could not find the one correct match.

As an example, consider finding the correct sense for the single word noun *bells* using SenseRelate:

1. Given the input sentence:

   "… city officials, and citizens were followed by women and children ringing **bells** for joy."

2. Search for AAT records. There are two records for the *bells* in AAT:

   a. **bells:** "Flared or bulbous terminals found on many open-ended aerophone tubes".

   b. **bells**: "Percussion vessels consisting of a hollow object, usually of metal but in some cultures of hard clay, wood, or glass, which when struck emits a sound by the vibration of most of its mass;…"

3. Submit the input sentence to SenseRelate, which provides a best guess for the corresponding WordNet senses for each word.

4. Get SenseRelate output, which indicates that the WordNet definition for *bells* is WordNet-Sense1, *i.e.,* "a hollow device made of metal that makes a ringing sound when struck"

SenseRelate output:

city#n#1 official#n#1 and citizen#n#1 be#v#1 follow#v#20 by#r#1 woman#n#1 and child#n#1 ringing#a#1 **bell#n#1** for joy#n#1

5. Find the correct AAT match using word overlap of the WordNet definition and the two AAT definitions for *bells*:

   WordNet: "a hollow device made of metal that makes a ringing sound when struck"
   compared with:
   AAT: "Flared or bulbous terminals found on many open-ended aerophone tubes"
   and *compared with*:
   AAT: "Percussion vessels consisting of a hollow object, usually of metal but in some cultures of hard clay, wood, or glass, which when struck emits a sound by the vibration of most of its mass;…"

6. The second AAT sense is the correct sense according to the word overlap (see Table 2 below):

Comparison	Score	Word Overlap
AAT – Definition 1 and WordNet Sense1	0	None
AAT – Definition 2 and WordNet Sense1	4	hollow, metal, sound, struck

Table 2: Word Overlap to Select AAT Definition

Notice that we only used the AAT record description for performing the word overlap. We experimented by including other information present in the AAT record (like variant names, parent AAT record names) also, but simply using the record description yielded the best results.

Third, we used AAT record names (preferred and variant) to find the one correct match. If one of the record names matched better than the other record names to the noun phrase name, that record was deemed to be the correct record. For example, the term *altar* more appropriately matches *altars* (religious building fixtures) than *altarpieces* (religious visual works). Another example is *children,* which better matches *children* (youth) than *offspring* (people by family relationship).

Fourth, if none of the above techniques succeeded in selecting one record, we used the most common sense definition for a term (taken from WordNet) in conjunction with the AAT results and word overlapping mentioned above to find the one correct record.

4. Results and Evaluation

4.1 Methodologies

We used three different evaluation methods to assess the performance of our algorithm. The first evaluation method computes whether our algorithm picked the correct AAT record (*i.e.,* the AAT sense picked is in agreement with the groundtruth). The second method computes whether the correct record is among the top three records picked by our algorithm. In Table 3 below, this is referred to as *Top3*. The third evaluation method computes whether the correct record is in top five records picked by our algorithm, *Top5*. The last two evaluations helped us determine the usability of our algorithm in situations where it does not pick the correct record but it still narrows down to top three or top five results.

We ranked the AAT records according to their preferred name for the baseline, given the absence of any other disambiguation algorithm. Thus, AAT records that exactly matched the term in question appear on top, followed by records that partially matched the term. For example, for term *feet*, the top three records were *feet* (terminal elements of objects), *French feet* (bracket feet), and *Spanish feet* (furniture components). For the noun *wings*, the top three records were *wings* (shoulder accessories), *wings* (visual works components), and *wings* (backstage spaces).

4.2 Overall Results

In this section, we present evaluation results for all the terms. In the next section, we present results for only those terms that required disambiguation.

Overall results for the training set (326 terms) are shown in Table 3. This table shows that overall accuracy of our algorithm is 76% and 68% for Labeler 1 and Labeler 2, respectively. The baseline accuracy is 69% for Labeler 1 and 62% for Labeler 2. The other two evaluations show much better results. The Top 3 and Top5 evaluations have accuracy of 84% and 88% for Labeler 1 and accuracy of 78% and 79% for Labeler 2. This argues for bringing in additional techniques to

enhance the SenseRelate approach in order to select from *Top3* or *Top5*.

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	76%	68%
Baseline Accuracy	69%	62%
Top3	84%	78%
Top5	88%	79%

Table 3: Results for Training Set (n=326 terms)

In contrast to Table 3 for the training set, Table 4 shows results for the test set. Labeler 1 shows an accuracy of 74% on the algorithm and 72% on the baseline; Labeler 2 has an accuracy of 73% on the algorithm and 69% on the baseline.

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	74%	73%
Baseline Accuracy	72%	69%
Top3	79%	79%
Top5	81%	80%

Table 4: Results for Test Set (n=275 terms)

4.3 Results for Ambiguous Terms

This section shows the results for the terms from the training set and the test set that required disambiguation. Table 5 below shows that our algorithm's accuracy for Labeler 1 is 55% compared to the baseline accuracy of 35%. For Labeler 2, the algorithm accuracy is 48% compared to baseline accuracy of 32%. This is significantly less than the overall accuracy of our algorithm. Top3 and Top5 evaluations have accuracy of 71% and 82% for Labeler 1 and 71% and 75% for Labeler 2.

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	55%	48%
Baseline Accuracy	35%	32%
Top3	71%	71%
Top5	82%	75%

Table 5: Ambiguous Terms for Training (n=128 terms)

Similar results can be seen for the test set (96 terms) in Table 6 below. Labeler 1 shows an accuracy of 50% on the algorithm and 42% on the baseline; Labeler 2 has an accuracy of 53% on the algorithm and 39% on the baseline.

Evaluation	Labeler 1	Labeler 2
Algorithm Accuracy	50%	53%
Baseline Accuracy	42%	39%
Top3	63%	68%
Top5	68%	71%

Table 6: Results for Ambiguous Terms under the Test Set (n=96 terms)

4.4 Analysis

Table 7 shows that SenseRelate is used for most of the AAT mappings, and provides a breakdown based upon the disambiguation technique used. Row One in Table 7 shows how few terms were disambiguated using the lookup modifier technique, just 1 in the training set and 3 in the test set.

Row	Technique	Training Set(n=128)	Test Set (n=96)
One	Lookup Modifier	1	3
Two	SenseRelate	108	63
Three	Best Record Match	14	12
Four	Most Common Sense	5	18

Table 7: Breakdown of AAT mappings by Disambiguation Technique

Rows Two and Three show that most of the terms were disambiguated using the SenseRelate technique followed by the Best Record Match technique. The Most Common Sense technique (Row Four) accounted for the rest of the labelings.

Table 8 gives insight into the errors of our algorithm for the training set terms:

Technique	Reason for Error	Error Count
SenseRelate	SenseRelate picked wrong WordNet sense	16
	WordNet does not have the sense	8
	Definitions did not overlap	11
	Other reasons	10
Best Record Match		10
Lookup Modifier		0
Most Common Sense		3

Table 8: Breakdown of the errors in our algorithm under training set (58 total errors)

Table 8 shows the following:

(1) Out of the total of 58 errors, 16 errors were caused because SenseRelate picked the wrong WordNet sense.

(2) 8 errors were caused because WordNet did not contain the sense of the word in which it was

being used. For example, consider the term *workshop*. WordNet has two definitions of *workshop*:

i. "small workplace where handcrafts or manufacturing are done" and

ii. "a brief intensive course for a small group; emphasizes problem solving"

but AAT has an additional definition that was referred by term *workshop* in the NGA text:

> "In the context of visual and decorative arts, refers to groups of artists or craftsmen collaborating to produce works, usually under a master's name"

(3) 11 errors occurred because the AAT record definition and the WordNet sense definition did not overlap. Consider the term *figures* in the sentence, "As with The Holy Family, the style of the figures offers no clear distinguishing characteristic." Then examine the AAT and WordNet sense definitions below for *figures*:

> *AAT sense*: "Representations of humans or animals"
>
> *WordNet sense:* "a model of a bodily form (especially of a person)"

These definitions do not have any words in common, but they discuss the same concept.

(4) 10 errors occurred in the Best Record Match technique, 0 errors occurred under the Lookup Modifier Technique, and 3 errors occurred under the Most Common Sense technique.

5. Conclusion

We have shown that it is possible to create an automated program to perform word sense disambiguation in a field with specialized vocabulary. Such an application could have great potential in rapid development of metadata for digital collections. Still, much work must be done in order to integrate our disambiguation program into the CLiMB Toolkit, including the following:

(1) Our algorithm's disambiguation accuracy is between 48-55% (Table 5 and Table 6), and so there is room for improvement in the algorithm. Currently we depend on an external program (SenseRelate) to perform much of the disambiguation (Table 7). Furthermore, SenseRelate maps terms to WordNet and we then map the WordNet sense to an AAT sense. This extra step is overhead, and it causes errors in our algorithm. We can either explore the option of re-implementing concepts behind SenseRelate to directly map terms to the AAT, or we may need to find additional approaches to employ hybrid techniques (including machine learning) for disambiguation. At the same time, we may benefit from the fact that WordNet, as a general resource, is domain independent and thus offers wider coverage. We will need to explore the trade-off in precision between different configurations using these different resources.

(2) We need more and better groundtruth. Our current data set of noun phrases includes term like *favor*, *kind*, and *certain aspects*. These terms are unlikely to be used as meaningful subject terms by a cataloger and will never be mapped to AAT. Thus, we need to develop reliable heuristics to determine which noun phrases are potentially high value subject index terms. A simple frequency count does not achieve this purpose.

Currently we are evaluating based on groundtruth that our project members created. Instead, we would like to extend the study to a wider set of image catalogers as labelers, since they will be the primary users of the CLiMB tool. Image catalogers have experience in finding subject terms and mapping subject terms to the AAT. They can also help determine which terms are high quality subject terms.

In contrast to working with the highly experienced image cataloger, we also want to extend the study to include various groups with different user needs. For example, journalists have ongoing needs for images, and they tend to search by subject. Using participants like these for markup and evaluation promises to provide comparative results, ones which will enable us to effectively reach a broad audience.

We also would like to test our algorithm on more collections. This will help us ascertain what kind of improvements or additions would make CLiMB a more general tool.

6. Acknowledgements

We thank Rachel Wadsworth and Carolyn Sheffield. We also acknowledge Philip Resnik for valuable discussion.

7. References

Baca, Murtha, ed. 2002. Introduction to art image access: issues, tools, standards, strategies. Getty Research Institute.

Banerjee, S., and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on ArtificialIntelligence*, 805–810.

Ferguson, Bobby and Sheila Intner. 1998. Subject Analysis: Blitz Cataloging Workbook. Westport, CT:Libraries Unlimited Inc.

Gale, W. A., K. W. Church, and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 101-112, Montreal, Canada.

Ide, Nancy M. and Jean Veronis. 1990. Mapping Dictionaries: A Spreading Activation Approach. In *Proceedings of the 6th Annual Conference of the UW Centre for the New OED and Text Research*, 52-64 Waterloo, Ontario.

Lesk, Michael. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of ACM SIGDOC Conference*, 24-26, Toronto, Canada.

Klavans, Judith L. 2006. Computational Linguistics for Metadata Building (CLiMB). In *Procedings of the OntoImage Workshop,* G. Greffenstette, ed. *Language Resources and Evaluation Conference* (LREC), Genova, Italy.

Klavans, Judith L. (in preparation). Using Computational Linguistic Techniques and Thesauri for Enhancing Metadata Records in Image Search: The CLiMB Project.

Layne, Sara Shatford. 1994. Some issues in the indexing of images. *Journal of the American Society for Information Science*, 583-588.

Palmer, Martha, Hwee Tou Ng, & Hoa Trang Dang. 2006. Evaluation of WSD Systems. *Word Sense Disambiguation: Algorithms and Applications*. Eneko Agirre and Philip Edmonds, ed. 75-106. Dordrecht, The Netherlands:Springer.

Patwardhan, S., S. Banerjee, S. and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 241–257.

Rasmussen, Edie. M. 1997. Indexing images. *Annual Review of Information Science and Technology* (ARIST), 32, 169-196.

The Latin Dependency Treebank in a Cultural Heritage Digital Library

David Bamman
The Perseus Project
Tufts University
Medford, MA
david.bamman@tufts.edu

Gregory Crane
The Perseus Project
Tufts University
Medford, MA
gregory.crane@tufts.edu

Abstract

This paper describes the mutually beneficial relationship between a cultural heritage digital library and a historical treebank: an established digital library can provide the resources and structure necessary for efficiently building a treebank, while a treebank, as a language resource, is a valuable tool for audiences traditionally served by such libraries.

1 Introduction

The composition of historical treebanks is fundamentally different from that of modern ones. While modern treebanks are generally comprised of newspaper articles,[1] historical treebanks are built from texts that have been the focus of study for centuries, if not millennia. The Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000), for example, includes Chaucer's 14th-century *Parson's Tale*, while the York Poetry Corpus (Pintzuk and Leendert, 2001) includes the entire text of *Beowulf*. The scholarship that has attended these texts since their writing has produced a wealth of contextual materials, including commentaries, translations, and linguistic resources.

[1]To name just three, the Penn Treebank (Marcus et al., 1994) is comprised of texts from the *Wall Street Journal*; the German TIGER Treebank (Brants et al., 2002) is built from texts taken from the *Frankfurter Rundschau*; and the Prague Dependency Treebank (Hajič, 1998) includes articles from several daily newspapers (*Lidové noviny* and *Mladá fronta Dnes*), a business magazine (*Českomoravský Profit*) and a scientific journal (*Vesmír*).

For the past twenty years, the Perseus digital library (Crane, 1987; Crane et al., 2001) has collected materials of this sort to create an open reading environment for the study of Classical texts. This environment presents the Greek or Latin source text and contextualizes it with secondary publications (e.g., translations, commentaries, references in dictionaries), along with a morphological analysis of every word in the text and variant manuscript readings as well (when available).

We have recently begun work on syntactically annotating the texts in our collection to create a Latin Dependency Treebank. In the course of developing this treebank, the resources already invested in the digital library have been crucial: the digital library provides a modular structure on which to build additional services, contains a large corpus of Classical source texts, and provides a wealth of contextual information for annotators who are non-native speakers of the language.

In this the digital library has had a profound impact on the creation of our treebank, but the influence goes both ways. The digital library is a heavily trafficked website with a wide range of users, including professional scholars, students and hobbyists. By incorporating the treebank as a language resource into this digital library, we have the potential to introduce a fundamental NLP tool to an audience outside the traditional disciplines of computer science or computational linguistics that would normally use it. Students of the language can profit from the syntactic information encoded in a treebank, while traditional scholars can benefit from the textual searching it makes possible as well.

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 33–40,
Prague, 28 June 2007.

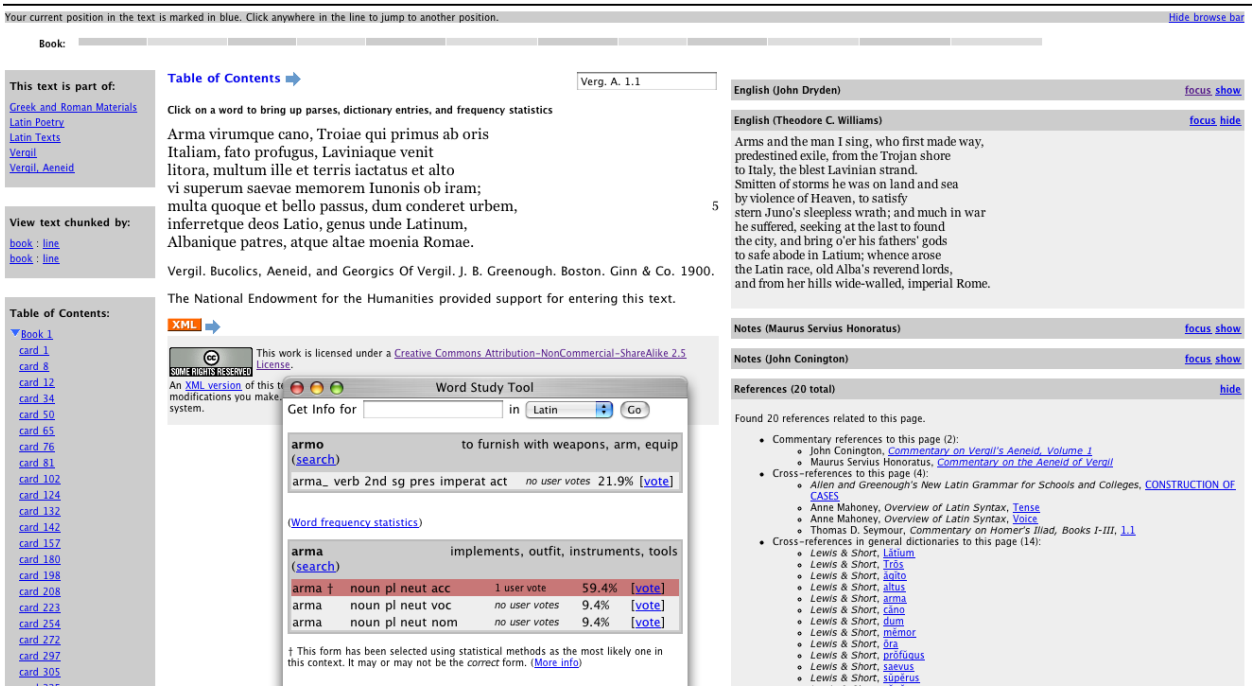


Figure 1: A screenshot of Vergil's *Aeneid* from the Perseus digital library.

2 The Perseus Digital Library

Figure 1 shows a screenshot from our digital library. In this view, the reader is looking at the first seven lines of Vergil's *Aeneid*. The source text is provided in the middle, with contextualizing information filling the right column. This information includes:

- Translations. Here two English translations are provided, one by the 17th-century English poet John Dryden and a more modern one by Theodore Williams.
- Commentaries. Two commentaries are also provided, one in Latin by the Roman grammarian Servius, and one in English by the 19th-century scholar John Conington.
- Citations in reference works. Classical reference works such as grammars and lexica often cite particular passages in literary works as examples of use. Here, all of the citations to any word or phrase in these seven lines are presented at the right.

Additionally, every word in the source text is linked to its morphological analysis, which lists every lemma and morphological feature associated with that particular word form. Here the reader has clicked on *arma* in the source text. This tool reveals that the word can be derived from two lemmas (the verb *armo* and the noun *arma*), and gives a full morphological analysis for each. A recommender system automatically selects the most probable analysis for a word given its surrounding context, and users can also vote for the form they think is correct.[2]

3 Latin Dependency Treebank

Now in version 1.3, the Latin Dependency Treebank is comprised of excerpts from four texts: Cicero's *Oratio in Catilinam*, Caesar's *Commentarii de Bello Gallico*, Vergil's *Aeneid* and Jerome's *Vulgate*.

Since Latin has a highly flexible word order, we have based our annotation style on the dependency grammar used by the Prague Dependency Treebank (PDT) (Hajič, 1998) for Czech (another non-projective language) while tailoring it for Latin via

[2]These user contributions have the potential to significantly improve the morphological tagging of these texts: any single user vote assigns the correct morphological analysis to a word 89% of the time, while the recommender system does so with an accuracy of 76% (Crane et al., 2006).

Date	Author	Words
63 BCE	Cicero	1,189
51 BCE	Caesar	1,486
19 BCE	Vergil	2,647
405 CE	Jerome	8,382
	Total:	13,683

Table 1: Treebank composition by author.

the grammar of Pinkster (1990).[3]

In addition to the index of its syntactic head and the type of relation to it, each word in the treebank is also annotated with the lemma from which it is inflected and its morphological code. We plan to release the treebank incrementally with each new major textual addition (so that version 1.4, for instance, will include the treebank of 1.3 plus Sallust's *Bellum Catilinae*, the text currently in production).

4 The Influence of a Digital Library

A cultural heritage digital library has provided a fertile ground for our historical treebank in two fundamental ways: by providing a structure on which to build new services and by providing reading support to expedite the process of annotation.

4.1 Structure

By anchoring the treebank in a cultural heritage digital library, we are able to take advantage of a structured reading environment with canonical standards for the presentation of text and a large body of digitized resources, which include XML source texts, morphological analyzers, machine-readable dictionaries, and an online user interface.

Texts. Our digital library contains 3.4 million words of Latin source texts (along with 4.9 million words of Greek). The texts are all public-domain materials that have been scanned, OCR'd and formatted into TEI-compliant XML. The value of this prior labor is twofold: most immediately, the existence of clean, digital editions of these texts has saved us a considerable amount of time and resources, as we would otherwise have to create them before annotating them syntactically; but their encoding as repurposeable XML documents in a larger library also allows us to refer to them under standardized citations. The passage of Vergil displayed in Figure 1 is not simply a string of unstructured text; it is a subdocument (*Book=1:card=1*) that is itself part of a larger document object (*Perseus:text:1999.02.0055*), with sisters (*Book=1:card=8*) and children of its own (e.g., *line=4*). This XML structure allows us to situate any given treebank sentence within its larger context.

Morphological Analysis. As a highly inflected language, Latin has an intricate morphological system, in which a full morphological analysis is the product of nine features: part of speech, person, number, tense, mood, voice, gender, case and degree. Our digital library has included a morphological analyzer from its beginning. This resource maps an inflected form of a word (such as *arma* above) to all of the possible analyses for all of the dictionary entries associated with it. In addition to providing a common morphological standard, this mapping greatly helps to constrain the problem of morphological tagging (selecting the correct form from all possible forms), since a statistical tagger only needs to consider the morphological analyses licensed by the inflection rather than all possible combinations.

User interface. The user interface of our library is designed to be modular, since different texts have different contextual resources associated with them (while some have translations, others may have commentaries). This modularity allows us to easily introduce new features, since the underlying architecture of the page doesn't change – a new feature can simply be added.

Figure 2 presents a screenshot of the digital library with an annotation tool built into the interface. In the widget on the right, the source text in view (the first chunk of Tacitus' *Annales*) has been automatically segmented into sentences; an annotator can click on any sentence to assign it a syntactic annotation. Here the user has clicked on the first sentence (*Vrbem Romam a principio reges habuere*); this action brings up an annotation screen in which a partial automatic parse is provided, along with the most likely morphological analysis for each word. The annotator can then correct this automatic output

[3] We are also collaborating with other Latin treebanks (notably the Index Thomisticus on the works of Thomas Aquinas) to create a common set of annotation guidelines to be used as a standard for Latin of any period (Bamman et al., 2007).

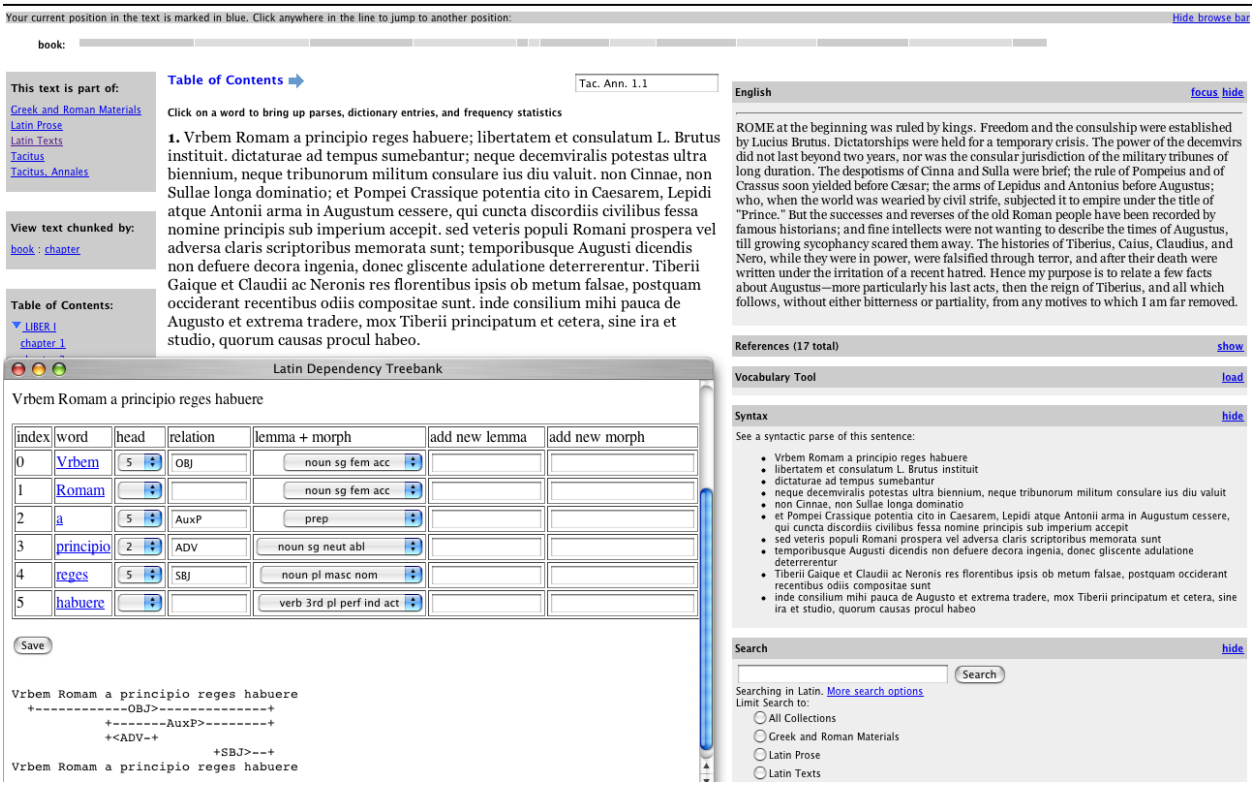


Figure 2: A screenshot of Tacitus' *Annales* from the Perseus digital library.

and move on to the next segmented sentence, with all of the contextual resources still in view.

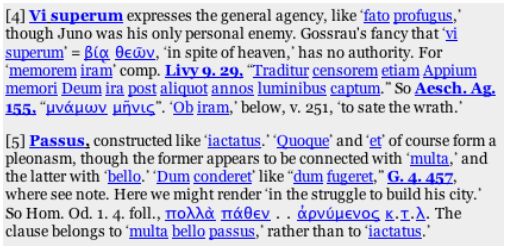
4.2 Reading support

Modern treebanks also differ from historical ones in the fluency of their annotators. The efficient annotation of historical languages is hindered by the fact that no native speakers exist, and this is especially true of Latin, a difficult language with a high degree of non-projectivity. While the Penn Treebank can report a productivity rate of between 750 and 1000 words per hour for their annotators after four months of training (Taylor et al., 2003) and the Penn Chinese treebank can report a rate of 240-480 words per hour (Chiou et al., 2001), our annotation speeds are significantly slower, ranging from 90 words per hour to 281. Our best approach for Latin is to develop strategies that can speed up the annotation process, and here the resources found in a digital library are crucial. There are three varieties of contextual resources in our digital library that aid in the understanding of a text: translations, commentaries, and dictionaries. These resources shed light on a text, from the level of sentences to that of individual words.

Translations. Translations provide reading support on a large scale: while loose translations may not be able to inform readers about the meaning and syntactic role of any single word, they do provide a broad description of the action taking place, and this can often help to establish the semantic structure of the sentence – who did what to whom, and how. In a language with a free word order (and with poetry especially), this kind of high-level structure can be important for establishing a quick initial understanding of the sentence before narrowing down to individual syntactic roles.

Commentaries. Classical commentaries provide information about the specific use of individual words, often noting morphological information (such as case) for ambiguous words or giving explanatory information for unusual structures. This information often comes at crucial decision points

in the annotation process, and represents judgments by authorities in the field with expertise in that particular text.



[4] **Vi superum** expresses the general agency, like 'fato profugus,' though Juno was his only personal enemy. Gossrau's fancy that 'vi superum' = βίᾳ θεῶν, 'in spite of heaven,' has no authority. For 'memorem iram' comp. **Livy 9. 29.** "Traditur censorem etiam Appium memori Deum ira post aliquot annos luminibus captum." So **Aesch. Ag. 155.** "μνάμων μῆνις". 'Ob iram,' below, v. 251, 'to sate the wrath.'

[5] **Passus,** constructed like 'iactatus.' 'Quoque' and 'et' of course form a pleonasm, though the former appears to be connected with 'multa,' and the latter with 'bello.' 'Dum conderet' like "dum fugeret," **G. 4. 457**, where see note. Here we might render 'in the struggle to build his city.' So Hom. Od. 1. 4. foll., πολλὰ πάθεν . . ἀρνύμενος κ.τ.λ. The clause belongs to 'multa bello passus,' rather than to 'iactatus.'

Figure 3: An excerpt from Conington's commentary on Vergil's *Aeneid* (Conington, 1876), here referring to Book 1, lines 4 and 5.

Machine-Readable Dictionaries. In addition to providing lists of stems for morphological analyzers, machine-readable dictionaries also provide valuable reading support for the process of lemma selection. Every available morphological analysis for a word is paired with the word stem (a lemma) from which it is derived, but analyses are often ambiguous between different lemmas. The extremely common form *est*, for example, is a third person singular present indicative active verb, but can be inflected from two different lemmas: the verb *sum* (to be) and the verb *edo* (to eat). In this case, we can use the text already tagged to suggest a more probable form (*sum* appears much more frequently and is therefore the likelier candidate), but in less dominant cases, we can use the dictionary: since the word stems involved in morphological analysis have been derived from the dictionary lemmas, we can map each analysis to a dictionary definition, so that, for instance, if an annotator is unfamiliar with the distinction between the lemmas *occido1* (to strike down) and *occido2* (to fall), their respective definitions can clarify it.

Machine-readable dictionaries, however, are also a valuable annotation resource in that they often provide exemplary syntactic information as part of their definitions. Consider, for example, the following line from Book 6, line 2 of Vergil's *Aeneid*: *et tandem Euboicis Cumarum adlabitur oris* ("and at last it glides to the Euboean shores of Cumae"). The noun *oris* (shores) here is technically ambiguous, and can be derived from a single lemma (*ora*) as a noun in either the dative or ablative case. The dictionary definition of *allabor* (to glide), however, disambiguates this for us, since it notes that the verb is often constructed with either the dative or the accusative case.

al-lābor (**adl-**), lapsus, 3, v. dep.,

I. *to glide to* or *toward* something, *to come to, to fly, fall, flow, slide*, and the like; constr. with dat. or acc. (**poet.**—oftenest in Verg.— "or in more elevated prose): viro adlapsa sagitta est," **Verg. A. 12, 319**: "fama adlabitur aurīs," *id. ib. 9, 474*: Curetum adlabimur oris, *we land upon*, etc., *id. ib. 3, 131*; cf. *id. ib. 3, 569*: "mare crescenti adlabitur aestu," *rolls up with increasing wave, id. ib. 10, 292*: "adlapsus genibus," *falling down at his knees, Sen. Hippol. 666.*—In prose: umor adlapsus extrinsecus, * *Cic. Div. 2, 27, 58*: "angues duo ex occulto adlapsi," **Liv. 25, 16**.

Figure 4: Definition of *allabor* (the dictionary entry for *adlabitur*) from Lewis and Short (1879).

Every word in our digital library is linked to a list of its possible morphological analyses, and each of those analyses is linked to its respective dictionary entry. The place of a treebank in a digital library allows for this tight level of integration.

5 The Impact of a Historical Treebank

The traffic in our library currently exceeds 10 million page views by 400,000 distinct users per month (as approximated by unique IP addresses). These users are not computational linguists or computer scientists who would typically make use of a treebank; they are a mix of Classical scholars, students, and amateurs. These different audiences have equally different uses for a large corpus of syntactically annotated sentences: for one group it can provide additional reading support, and for the other a scholarly resource to be queried.

5.1 Treebank as Reading Support

Our digital library is predominantly a reading environment: source texts in Greek and Latin are presented with attendant materials to help facilitate their understanding. The broadest of these materials are translations, which present sentence-level equivalents of the original; commentaries provide a more detailed analysis of individual words and phrases. A

treebank has the potential to be a valuable contextual resource by providing syntactic information for every word in a sentence, not simply those chosen by a commentator for discussion.

5.2 Treebank as a Scholarly Resource

For Classical scholars, a treebank can also be used as a scholarly resource. Not all Classicists are programmers, however, and many of those who would like to use such a resource would profit little from an XML source file. We have already released version 1.3 of the Latin Dependency Treebank in its XML source, but we also plan to incorporate it into the digital library as an object to be queried. This will yield a powerful range of search options, including lemmatized and morpho-syntactic searching, and will be especially valuable for research involving lexicography and semantic classification.

Lemmatized searching. The ability to conduct a lemma-based textual search has long been a desideratum in Classics,[4] where any given Latin word form has 3.1 possible analyses on average.[5] Locating all inflections of *edo* (to eat) in the texts of Caesar, for example, would involve two things:

1. Searching for all possible inflections of the root word. This amounts to 202 different word forms attested in our texts (including compounds with enclitics).

2. Eliminating all results that are homonyms derived from a different lemma. Since several inflections of *edo* are homonyms with inflections of the far more common *sum* (to be), many of the found results will be false positives and have to be discarded.

This is a laborious process and, as such, is rarely undertaken by Classical scholars: the lack of such a resource has constrained the set of questions we can ask about a text. Since a treebank encodes each word's lemma in addition to its morphological and syntactic analysis, this information is now free for the taking.

Morpho-syntactic searching. A treebank's major contribution to scholarship is that it encodes the syntax of a sentence, along with a morphological analysis of each word. These two together can be combined into elaborate searches. Treebanks allow scholars to find all instances of any particular construction. For example:

- When the conjunction *cum* is the head of a subordinate clause whose verb is indicative, it is often recognized as a temporal clause, qualifying the time of the main clause's action;

- When that verb is subjunctive, however, the clause retains a different meaning, as either circumstantial, causal, or adversative.

These different clause types can be found by querying the treebank: in the first case, by searching for indicative verbs that syntactically depend on *cum*; in the second, for subjunctive verbs that depend on it. In version 1.3 of the Latin Dependency Treebank, *cum* is the head of a subordinate clause 38 times: in 7 of these clauses an indicative verb depends on it, while in 31 of them a subjunctive one does. This type of searching allows us to gather statistical data while also locating all instances for further qualitative analysis.[6]

Lexicography. Searching for a combination of lemma and morpho-syntactic information can yield powerful results, which we can illustrate with a question from Latin lexicography: how does the meaning of a word change across authors and over time? If we take a single verb – *libero* (to free, liberate) – we can chart its use in various authors by asking a more specific question: what do different Latin authors want to be liberated from? We can imagine that an orator of the republic has little need to speak of liberation from eternal death, while an apostolic father is just as unlikely to speak of being freed from another's monetary debt.

[4] Both the Perseus Project and the Thesaurus Linguae Graecae (http://www.tlg.uci.edu) allow users to search for all inflected forms of a lemma in their texts, but neither filters results that are homonyms derived from different lemmas.

[5] Based on the average number of lemma + morphology combinations for all unique word tokens in our 3.4 million word corpus. The word form *amor*, for example, has 3 analyses: as a first-person singular present indicative passive verb derived from the lemma *amo* (to love) and as either a nominative or vocative masculine singular noun derived from *amor* (love).

[6] For the importance of a treebank in expediting morpho-syntactic research in Latin rhetoric and historical linguistics, see Bamman and Crane (2006).

We can answer this more general question by transforming it into a syntactic one: what are the most common complements of the lemma *libero* that are expressed in oblique cases (e.g., ablative, genitive, etc.) or as prepositional phrases? In a small test of 100 instances of the lemma in Cicero and Jerome, we find an interesting answer, presented in Table 2.

Cicero		Jerome	
periculo	14	manu	22
metu	8	morte	3
cura	6	ore	3
aere	3	latronibus	2
scelere	3	inimico	2
suspicione	3	bello	2

Table 2: Count of objects *liberated from* in Cicero and Jerome that occur with frequency greater than 1 in a corpus of 100 sentences from each author containing any inflected form of the verb *libero*.

The most common entities that Cicero speaks of being liberated from clearly reflect the cares of an orator of the republic: *periculo* (danger), *metu* (fear), *cura* (care), and *aere* (debt). Jerome, however, uses *libero* to speak of liberation from a very different set of things: his actors speak of deliverance from *manu* (e.g., the hand of the Egyptians), from *ore* (e.g., the mouth of the lion) and from *morte* (death). A treebank encoded with lemma and morpho-syntactic information lets us quantify these typical arguments and thereby identify the use of the word at any given time.

Named entity labeling. Our treebank's place in a digital library also means that complex searches can draw on the resources that already lie therein. Two of our major reference works include Smith's *Dictionary of Greek and Roman Geography* (1854), which contains 11,564 place names, and Smith's *Dictionary of Greek and Roman Biography and Mythology* (1873), which contains 20,336 personal names. By mapping the lemmas in our treebank to the entries in these dictionaries, we can determine each lemma's broad semantic class. After supplementing the Classical Dictionary with names from the Vulgate, we find that the most common people in the treebank are *Iesus*, *Aeneas*, *Caesar*, *Catilina*, *Satanas*, *Sibylla*, *Phoebus*, *Misenus* and *Iohannes*; the most common place names are *Gallia*, *Babylon*, *Troia*, *Hierusalem*, *Avernus* and *Sardis*.

One use of such classification is to search for verbs that are typically found with sentient agents. We can find this by simply searching the treebank for all active verbs with subjects known to be people (i.e., subjects whose lemmas can be mapped to an entry in Smith's *Dictionary*). An excerpt of the list that results is given in Table 3.

mitto	to send
iubeo	to order
duco	to lead
impono	to place
amo	to love
incipio	to begin
condo	to hide

Table 3: Common verbs with people as subjects in the Latin Dependency Treebank 1.3.

Aside from its intrinsic value of providing a catalogue of such verbs, a list like this is also useful for classifying common nouns: if a verb is frequently found with a person as its subject, all of its subjects in general will likely be sentient as well. Table 4 presents a complete list of subjects of the active voice of the verb *mitto* (to send) as attested in our treebank.

angelus	angel
Caesar	Caesar
deus	God
diabolus	devil
Remi	Gallic tribe
serpens	serpent
ficus	fig tree

Table 4: Subjects of active *mitto* in the Latin Dependency Treebank 1.3.

Only two of these subjects are proper names (*Caesar* and *Remi*) that can be found in Smith's *Dictionary*, but almost all of these nouns clearly belong to the same semantic class – *angelus*, *deus*, *diabolus* and *serpens* (at least in this text) are entities with cognition.

Inducing semantic relationships of this sort is the typical domain of clustering techniques such as la-

tent semantic analysis (Deerwester et al., 1990), but those methods generally work best on large corpora. By embedding this syntactic resource in a digital library and linking it to external resources such as reference works, we can find similar semantic relationships with a much smaller corpus.

6 Conclusion

Treebanks already fill a niche in the NLP community by providing valuable datasets for automatic processes such as parsing and grammar induction. Their utility, however, does not end there. The linguistic information that treebanks encode is of value to a wide range of potential users, including professional scholars, students and amateurs, and we must encourage the use of these resources by making them available to such a diverse community. The digital library described in this paper has proved to be crucial for the development and deployment of our treebank: since the natural intuitions of native speakers are hard to come by for historical languagues, it is all the more important to leverage the cultural heritage resources we already have.

7 Acknowledgments

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484) and the National Science Foundation (BCS-0616521) provided support for this work.

References

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Technical report, Tufts Digital Library, Medford.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol.

Fu-Dong Chiou, David Chiang, and Martha Palmer. 2001. Facilitating treebank annotation using a statistical parser. In *Proceedings of the First International Conference on Human Language Technology Research HLT '01*, pages 1–4.

John Conington, editor. 1876. *P. Vergili Maronis Opera. The Works of Virgil, with Commentary*. Whittaker and Co, London.

Gregory Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. 2001. Drudgery and deep thought: Designing digital libraries for the humanities. *Communications of the ACM*, 44(5):34–40.

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David M. Mimno, Adrian Packel, David Sculley, and Gabriel Weaver. 2006. Beyond digital incunabula: Modeling the next generation of digital libraries. In *ECDL 2006*, pages 353–366.

Gregory Crane. 1987. From the old to the new: Integrating hypertext into traditional scholarship. In *Hypertext '87: Proceedings of the 1st ACM conference on Hypertext*, pages 51–56. ACM Press.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.

A. Kroch and A. Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/ppcme2-release-2/.

Charles T. Lewis and Charles Short, editors. 1879. *A Latin Dictionary*. Clarendon Press, Oxford.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Harm Pinkster. 1990. *Latin Syntax and Semantics*. Routledge, London.

Susan Pintzuk and Plug Leendert. 2001. York-Helsinki Parsed Corpus of Old English Poetry.

William Smith. 1854. *A Dictionary of Greek and Roman Geography*. Walton and Maberly, London.

William Smith. 1873. *A Dictionary of Greek and Roman Biography and Mythology*. Spottiswoode, London.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22. Kluwer Academic Publishers.

Cultural Heritage Digital Resources: from Extraction to Querying

Michel Généreux
Natural Language Technology Group
University of Brighton
Umited Kingdom
M.Genereux@brighton.ac.uk

Abstract

This article presents a method to extract and query Cultural Heritage (CH) textual digital resources. The extraction and querying phases are linked by a common ontological representation (CIDOC-CRM). A transport format (RDF) allows the ontology to be queried in a suitable query language (SPARQL), on top of which an interface makes it possible to formulate queries in Natural Language (NL). The extraction phase exploits the propositional nature of the ontology. The query interface is based on the Generate and Select principle, where potentially suitable queries are *generated* to match the user input, only for the most semantically similar candidate to be *selected*. In the process we evaluate data extracted from the description of a medieval city (Wolfenbüttel), transform and develop two methods of computing similarity between sentences based on WordNet. Experiments are described that compare the pros and cons of the similarity measures and evaluate them.

1 Introduction

The CIDOC-CRM (DOERR, 2005) ontology is an ISO standard created to describe in a formal language the explicit and implicit concepts and relations underlying the documentation produced in CH. The ontology aims at accommodating a wide variety of data from the CH domain, but its sheer complexity may make it difficult for non-experts to learn it quickly, let alone use it efficiently. For others, it may even be simpler to find a way to translate automatically their data from the storage mechanism already in place into CIDOC-CRM. For practitioners unfamiliar with strict formalisms, it may be more natural to describe collections in natural language (e.g. English), and there is already an unprecedented wealth of information available on-line in natural language for almost anything, including CH. Wouldn't it be practical to be able to describe a collection of artifacts in plain English, with little or no knowledge of the CIDOC-CRM formalism, and let language technology take over and produce a CIDOC-CRM database? The principle behind that idea is based on the observation that the building blocks of the CIDOC-CRM ontology, the *triples*, have a predicative nature, which is structurally consistent with the way many natural languages are built (DOERR, 2005):

> The domain class is analogous to the grammatical subject of the phrase for which the property is analogous to the verb. Property names in the CRM are designed to be semantically meaningful and grammatically correct when read from domain to range. In addition, the inverse property name, normally given in parentheses, is also designed to be semantically meaningful and grammatically correct when read from range to domain.

A triple is defined as:

DOMAIN PROPERTY RANGE

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 41–48,
Prague, 28 June 2007.

The domain is the class (or entity) for which a property is formally defined. Subclasses of the domain class inherit that property. The range is the class that comprises all potential values of a property. Through inheritance, subclasses of the range class can also be values for that property. Example 1 is somewhat artificial, but it illustrates how triples can be extracted from natural language, where entities E48 and E53 are *Place Name* and *Place* respectively, while P1 is the property *identify*.

(1) *Rome* *identifies* *the capital of Italy.*
DOM E41 PROP P1 RANGE E1
E48 P1 E53
'Rome identifies the capital of Italy.'

The task of the natural language processing tool is to map relevant parts of texts to entities and properties in such a way that triples can be constructed (see also (SHETH, 2003; SCHUTZ, 2005; DAGAN, 2006)). In a nutshell, the Noun Clauses (NC) *Rome* and *the capital of Italy* are mapped to *Entity 48* and *Entity 53*, themselves subclasses of the domain E41 and range E1 respectively, while the Verb Clause (VC) *identifies* is mapped to *Property P1*.

On the other hand, a natural language interface (ANDROUTSOPOULOS, 1995) to query structurally complex and semantically intertwined data such as those that can be found in the archaeological domain can lighten a great deal the tasks of browsing and searching. This state of affairs is even more true for people not familiar with formal languages, as is often the case in archaeology in particular and cultural heritage in general. With the Semantic Web[1] in full development and ontologies such as CIDOC-CRM teaming together to render semantic navigation a realistic prospect, natural language interfaces can offer a welcomed simplified view of the underlying data.

One of the most important and Semantic Web oriented conceptual model available today is the CIDOC-CRM, which is becoming the new standard model to document CH data: the creation of tools ready to manage CIDOC-CRM compliant archives will be one of the most important goals of the coming years (HERMON, 2000). The full implementation of the CIDOC-CRM model is simplified today by a family of languages developed by the World Wide Web Consortium[2] and XML-based (LI, 2006). One of its most important representative is RDF[3], on top of which sits a query language such as SPARQL[4].

2 Extraction

2.1 Methodology

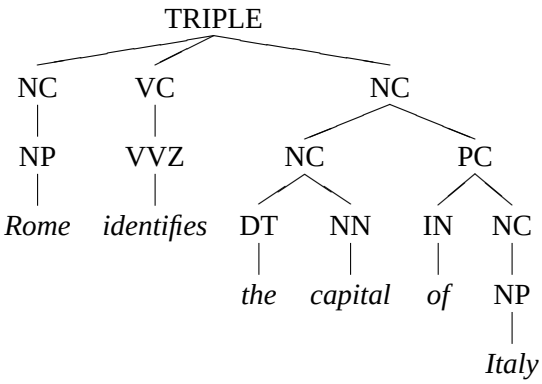


Figure 1: Linguistic parse tree for example 1.

Figure 1 suggests that all pairs of NC separated by a VC (and possibly other elements) are potentially valid CIDOC-CRM triples. Part-of-speeches (POS) and phrasal clauses can be obtained with a POS tagger and chunker[5]. To validate the triples, we must first make sure that the predicate is relevant by extracting the main verb of the verbal clause (VC) and see if its meaning is similar (synonym) to at least one of the CIDOC-CRM properties. For example, it is possible to use the verb *describe* instead of *identify*. Once a set of possible properties is identified, we must verify if the noun clauses (NC) surrounding the property are related to the DOMAIN and the RANGE of that property. To establish the relation, the first step is to identify the semantics of each NC clause. For English, a good indicator of the NC semantics is the rightmost NN in the clause, excluding any attached PC. The rightmost NN is usually the most significant: for example, in the NC *the museum artifact*, the main focus point is *artifact*, not *museum*. In figure 1 the rightmost NN of *the capital*

[1] http://www.w3.org/2001/sw/

[2] W3C: http://www.w3.org/

[3] http://www.w3.org/RDF/

[4] http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406/

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

of Italy is *capital* (excluding the attached PC); this tells us that we are dealing with an object of type *capital*. The second step is to see if the type is a subclass of the DOMAIN or RANGE. Because *entity* (E1) is a hypernym of *capital*, then we conclude that the clause *the capital of Italy* is a subclass of E1:CRM Entity. What if the NC has no NN? One possibility[6] is that the clause is made up of at least one proper noun (*Rome*). To establish the type of a proper noun, we use the Web as corpus and compute a measure of *semantic association* (CHURCH, 1989) with all CIDOC-CRM classes and choose the most similar as being the type of the NC clause. This would yield the following triple:

E41	P1	*E1*
Rome		*the capital of Italy*

where E1 and E41 are the entities *Appellation* and *CRM Entity* respectively.

2.2 Extracting a triple from free text

The following experiment shows the result of extracting a triple from a textual description of the medieval city of Wolfenbüttel based on the method described previously. The document was 3922 words long with 173 sentences. The system extracted 197 intermediate triples and 79 final triples. Table 1 shows a few processing steps for the following fragment of text:

> The street's particular charm lies in its broad-faced half-timbered buildings.

In step ①, an intermediate triple is extracted from texts, then we use synonyms and hypernyms in step ② to find mappings with domains (D), properties (P) and ranges (R) of the ontology. The final triples appears in step ③. For example, *consist* is a synonym for *lie*, and *object* is an hypernym of *building*. In each case, we extracted from WordNet[7] (PEDERSEN, 2004) the synonyms and hypernyms of the three most common uses for each word (verb, noun).

[6]The other possibility, *pronouns*, is omitted for simplicity.

[7]http://wordnet.princeton.edu/

①	D	[The street's particular charm]
	P	lies in
	R	[its broad-faced half-timbered buildings]
②	D	[attribute, charm, entity, language, object]
	P	[consist]
	R	[activity, building, creation, entity, event, object]
③	D	[e13:Attribute Assignment]
	P	p9:consists of
	R	[e7:Activity]

Table 1: A triple extracted from free text.

3 Querying

3.1 NL Interface to SPARQL Querying

Our approach to the problem of mapping a query in natural language to a query expressed in a particular query language (here SPARQL) is to *generate* (BURKE, 1997) the most likely candidates and *select* the item which shows maximum semantic similarity with the input string. We now explain both steps in turn.

3.1.1 Generation

We start from two parallel grammars describing both the target query language and one or more natural languages. Here is an excerpt from one query language (SPARQL),

$$SelectQuery \rightarrow Select \left\{ \begin{array}{c} Var+ \\ Star \end{array} \right\} DC?\ WC\ SM?$$
$$DC \rightarrow From\ Table$$
$$WC \rightarrow Where?\ \{\ Filter\ \}$$
$$SM \rightarrow OrderBy\ Modifier$$
$$Star \rightarrow \text{'*'}$$
$$Select \rightarrow \text{'select'}$$
$$From \rightarrow \text{'from'}$$
$$Table \rightarrow \text{'}<\ OneTable\ >\text{'}$$

and part of its equivalent in natural language (here English):

$$Select \rightarrow \left\{ \begin{array}{c} \text{'select'} \\ \text{'show'} \end{array} \right\} \text{'the'?}$$
$$From \rightarrow \text{'from'}$$

$$Star \rightarrow \left\{ \begin{array}{l} \textit{'all records'} \\ \textit{'everything'} \end{array} \right\}$$
$$OneTable \rightarrow \textit{'clients'}$$

Therefore, for a SPARQL query such as *select * from <clients> {}*, we are able to generate the equivalent in natural language: *select all records from clients*. The generation space of SPARQL and natural languages can be very large (in fact it can be infinitely large in both cases), so generation must be constrained in some way (it is in fact constrained by the size of the input string). More specifically, the grammar generates candidate strings of length to be contained between a fraction *f1* shorter and a fraction *f2* longer than the size (in meaningful words) of the input strings. Meaningful words are limited to be adjectives (tag J), nouns (tag N), verbs (tag V) and adverbs (tag RB), partly because they can be compared against each other using WordNet. The values of *f2* is usually less than the value of *f1*, but the exact values are to be determined empirically. The idea behind this is based on the general observation that queries expressed in natural languages are more likely to be redundant than underspecified. Let's look at example 2, a particular example of a user query.

(2) Could/MD you/PP show/VVP me/PP all/PDT the/DT records/NNS you/PP have/VHP please/VV ./SENT

There are three (*show*, *records* and *have*) meaningful words in 2. Assuming that we have 0.4 and 0.1 for the values of *f1* and *f2* respectively, the generator would then be constrained to produce candidate strings having a length in the range [3-0.4*3, 3+0.1*3] or [1.8, 3.3], i.e. between two and three words after rounding. The generative process must also be informed on possible values employed by the user for the sake of filtering. For example, in queries such as *Show me everything that has a salary above 500* and *Select people named Smith*, the value of the fields *salary* and *name* are respectively specified as being above *500* and *Smith*. These values are used by the generator. They are assume to be found as symbols (SYM), foreign words (FW), nouns (N), cardinal numbers (CD) or adjectives (J) in the input string. The whole generative process can be summarised as follows:

1. Compute the input query strings length I in meaningful word tokens and detect potential field values (SYM, FW, N, CD or J)
2. Generates candidate strings of a given language L with length in the range [I-f1*I, I+f2*I]
3. For each candidate string, generate the equivalent SPARQL query

The candidate strings in language L from step 2 are passed on to the *selection* process.

3.1.2 Selection

The *selection* process is based on a measure of similarity between the input string and candidates issued from generation. The two similarity measures we are presenting are based on an available semantic resource for English, *Wordnet*. Both measures assume that two sentences are semantically similar if they share words with similar meanings. This assumption is certainly not true in general but, in the case of database querying, we can assume that the use of metaphors, irony or contextualised expressions is relatively rare. There are different approaches to compute similarity, but we are constrained by the fact that the system must potentially analyse and compare a large number of sentences with varying lengths. The so-called *Levenshtein distance* or *edit distance* is a simple option based on dynamic programming (RISTAD, 1998). It can be transformed to become a *semantic distance*, that is, the semantic distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion (cost 1), deletion (cost 1), or substitution of a single word (as opposed to letters in the original edit-distance). The exact cost of substitution is given by how dissimilar a pair of words is according to WordNet (from 0 to 2). Two strings are therefore similar if they have words semantically related, with a preference for the *same word order*. This last requirement is not always acceptable for natural language, as can be illustrated by examples 3 and 4, which are clear semantic equivalent, although a measure based on the *Levenhstein distance* would be unduly penalising because of a different word order.

(3) Show me the name and salary of all clients.

(4) Look into clients and show me their name and salary.

However, the *edit distance* is computationally attractive and it is not clear whether word-order is such an important factor when querying database in natural language.

One way to have more control over word-order is to built a similarity matrix. A similarity matrix provides a pairwise measure of similarity between each word of two sentences. Let's say we want to compare the similarity of a user's sentence 5 with a candidate query 6 generated by system.

(5) Show me salaries for names Smith.

(6) Select the salary where name is Smith.

The corresponding similarity matrix is shown as table 2. Each word is assigned a part-of-speech and transformed to its base-form to simplify comparison using WordNet. The similarity values in the table are

Similarity	show	salary	name	Smith
select	25	0	0	0
salary	0	100	8	6
name	0	8	100	17
be	33	0	0	0
Smith	0	6	17	100

Table 2: Similarity matrix between two sentences

in the [0,100] range. They are computed using simple edge counting in WordNet, a technique similar to computing how two people are genetically related through their common ancestors (BUDANITSKY, 2001). Only nouns, verbs, adjectives and adverbs can be semantically related by WordNet, therefore strings are initially stripped of all other grammatical categories. For example, table 2 shows that the word *select* has a degree of similarity of 25 with show. This approach does not take on board word-order *at all*, and we introduce a slight correction for the value of each entry in the table: similarity is decreased when words appear in different positions in a string. This is a sensible compromise to consider word-order without undue penalties. This approach can be expressed as follows: similarity values are decreased by a maximum of MaxDecrease only when a pair of words are significantly distant (by factor SigDistant) in their respective position within each string. This is expressed by the following formula:

$$IF \ \frac{\mid l - c \mid}{L} > SigDistant \ THEN$$

$$Sim \leftarrow Sim * \left(1 - \frac{\mid l - c \mid}{L} * MaxDecrease\right)$$

where *l* and *c* are the line and column numbers respectively and *L* is the size of the longest string. If we set the values of SigDistant and MaxDecrease to 0.2, then table 2 is transformed to 3. In table 3,

Similarity	show	salary	name	Smith
select	25	0	0	0
salary	0	100	7	5
name	0	7	100	17
be	28	0	0	0
Smith	0	5	16	100

Table 3: Transf. sim. matrix between two sentences

we can see that the similarity between *show* and *be* has been reduced from 33 to 28. Once we have the transformed similarity matrix, we can compute the similarity between the two sentences as such. This is achieved by the following four steps:

1. Generate all possible squared (k*k) sub-matrices from the transformed similarity matrix. There are $C_n^k = \frac{n!}{k!(n-k)!}$ such matrices where k is the size of the shortest sentence and n the longest

2. Generate all possible word-pairings for each sub-matrices. This amounts to selecting elements being on a different row and column. There are k! such pairings for each $C_n^k = \frac{n!}{k!(n-k)!}$ squared sub-matrices

3. Compute the similarity of each k! word-pairs for all C_n^k sub-matrices by adding their similarity value

4. The similarity of the transformed matrix is taken to be the same as the highest among the k! word-pairs * C_n^k sub-matrices, divided (normalised) by the size of the longest string n

For our running example in table 3, step 1 yields five 4*4 sub-matrices. For each sub-matrix, there are 24

word-pairings (step 2). It is easy to see which word pairing from table 2 gives the highest similarity: (be-show,28), (salary-salary,100), (name-name,100) and (Smith-Smith,100), for a total of 328, normalised to the length of the longest string (5): 328/5 = 66. For comparison, the semantic similarity distance between the same two sentences using the edit-distance is 250, and this must be normalised to the added length of the shortest and the longest sentence, 250/(5+4) = 28. Since Levenhstein gives us a distance, we have 1-distance for similarity. The normalising factor is (longest+shortest = 5+4), since two strings completely different would necessitate k replacements and n-k insertions. The maximum cost is therefore k*2 + (n-k) = k+n.

We can get a flavour of the computational complexities involved in both measures in terms of the number of semantic similarity computations between two words (the most costly computation). The ration between these numbers for *Matrix* (n!/(n-k)!) and *Edit* (k*n) is (n-1)!/k(n-k)!. This ratio is equal or greater than 1 in all cases except when n=k=2 and n=k=3, which confirms the expected greater complexity of the *Matrix* method. For example, when two strings of 8 words (n=k=8) are compared, complexity is 64 for *Edit* and 40320 for *Matrix*.

3.2 Comparative Evaluation

In this experiment[8] we aim at evaluating and comparing the two (word-based) measures of semantic similarity between sentences previously described and based on WordNet. We will refer to these measures as *Edit* and *Matrix*. We need a reference corpus where phrases are paired as *paraphrases*, so we used the Microsoft Research Paraphrase Corpus (QUIRK, 2004), which is described by the authors as:

> . . . a text file containing 5800 pairs of sentences which have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship.

[8]Values of parameters for the methods: cost of substitution = 2, word-pairings are centred, contiguous and do not exceed 7, MaxDecrease=0.2, SigDistant=0.2, method for similarity = count of edges

One of two levels of quality is assigned to each paraphrase (0 or 1). For example, phrases 7 are better paraphrases (annotated “1”) than 8 (annotated “0”).

(7) Amrozi accused his brother, whom he called ”the witness”, of deliberately distorting his evidence./ Referring to him as only ”the witness”, Amrozi accused his brother of deliberately distorting his evidence.

(8) Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for $2.5 billion./ Yucaipa bought Dominick’s in 1995 for $693 million and sold it to Safeway for $1.8 billion in 1998.

We selected random subsets of 100 pairs of good paraphrases (i.e. annotated with “1”), 100 pairs of less good paraphrases (annotated with “0”) and 100 pairs of phrases not paraphrases of each other. We computed semantic similarity for each subset using both methods. Results are presented in table 4. For each method the minimum and maximum values of similarity are reported. Variance is relatively low and both methods appear to correlate. As expected, paraphrases have higher similarity values, with type “1” values slightly ahead. Moreover, average values for paraphrases are significantly higher than for non-paraphrases, which is a sign that both methods can discriminate between semantically related sentences. When querying databases, we cannot always expect a clear front runner, but a continuum of more or less likely valuable candidates, more in line with the case of paraphrases “0”.

Compar.	Min	Avg	Max	Var	Cor
No(E)	5	12	24	0.2	0.7
No(M)	3	14	30	0.4	0.7
“0”(E)	21	57	86	1.2	0.8
“0”(M)	20	54	84	3.3	0.8
“1”(E)	35	69	94	1.9	0.6
“1”(M)	34	61	84	2.4	0.6

Table 4: Compar. eval. of the (E)dit and (M)atrix methods for types “0”, “1” and (No) paraphrases.

2-best pairs In this last experiment, 40 sets of 9 phrases are submitted to each method for evaluation. Each set includes only one pair of paraphrases: sets 1 to 20 include type “0” paraphrases, while sets 21 to 40 include type “1” paraphrases. There was no indication in the corpus that two phrases were not paraphrase of each other, so we assumed that

phrases not paired as being paraphrases were not. Therefore, our random selection of non-paraphrases can be more or less dissimilar. Table 5 show the results, where underlined similarity scores are those of the actual paraphrases, and columns BEST and SECOND give the actual measures of similarity for the best match (the pair the system thinks are paraphrases) and its closest follower respectively. We can see that all 40 paraphrases were selected as the best by both methods (M and E). Numbers in bold indicate cases where methods have selected different second best. The differences between type "0" and "1" are consistent with those observed in table 4. These are very encouraging results that suggest both methods could be used in a real system.

SET	Type 0				Type 1				SET
	Best		Second		Best		Second		
	M	E	M	E	M	E	M	E	
1	43	54	19	20	59	48	26	17	21
2	39	38	19	14	62	94	24	17	22
3	40	59	32	21	74	90	**16**	**18**	23
4	46	65	**24**	**20**	57	86	**39**	**21**	24
5	51	57	**33**	**31**	47	47	25	19	25
6	39	43	**19**	**15**	53	54	15	11	26
7	54	70	41	39	46	60	**16**	**15**	27
8	50	59	13	9	51	79	**12**	**10**	28
9	72	78	17	20	52	62	**21**	**14**	29
10	60	67	33	23	56	60	42	29	30
11	56	78	**17**	**15**	56	52	27	26	31
12	36	50	**15**	**14**	84	79	**21**	**17**	32
13	72	80	**18**	**16**	48	60	29	27	33
14	66	68	**29**	**25**	80	79	16	13	34
15	39	65	15	12	84	87	**34**	**29**	35
16	52	58	**10**	**9**	52	77	22	14	36
17	75	71	**23**	**21**	84	82	21	18	37
18	48	53	22	19	84	87	**21**	**17**	38
19	60	60	27	19	69	71	**15**	**13**	39
20	84	63	**18**	**14**	55	80	**18**	**18**	40

Table 5: Similarity scores for each of the 2 most similar pairs of phrases as computed by M and E

3.3 Conclusions and Future Work

It is difficult to have a comprehensive evaluation of the extraction phase through standard metrics (precision, recall), since there is no benchmark for this type of analysis. A good benchmark would be a CIDOC-CRM human-annotated text. Yet we can give some evidence of the performance of the system. In our experiment, we have collected 79 final triples from a 173 sentences long document describing buildings and places of interest in a medieval city. The data was relatively clean, although punctuation was heavily used throughout the document, confusing the chunker. Despite modest results, there is no doubt that a system like this gives a head start to anyone wishing to build a collection using the CIDOC-CRM ontology. A first pass in the documentation gives a good idea of what the textual documentation is about. However, a fuller interpretation will often involve combining many triples together to form paths. Because of time restriction, we have decided to process the three most common meanings of each word that we looked up in WordNet (avoiding the need to select the correct meaning among many); this may have the side effect of lowering accuracy. Speed was not an issue without access to the Web, not an absolute necessity if we have a good thesaurus for proper nouns. Finally, we have tuned the CRM to analyse impressions of a city, which is not a domain for which the CRM is optimally intended. We conjecture that texts about museum catalogues would have yielded better results.

The approach to database querying presented in this paper demonstrates that more and more semantic resources can be used to render natural language interfaces more efficient. The semantic web provides the backbone and the technology to support complex querying of naturally complex data. Lexical resources such as WordNet makes it possible to compute semantic similarity between sentences, allowing researchers to develop original ways to semantic parsing of natural languages. Our experiments show that it is possible to map English queries to a subset of SPARQL with high level of precision and recall. The main drawback of the *Edit* method is its overemphasis on *word-order*, making it less suitable for some languages (e.g. German). The *Matrix* method is computationally greedy, and future research must investigate efficient ways of cutting down the large search space. Perhaps step 2 should limit the number of word-pairings by taking only adjacent combinations.

Another improvement might include less uncon-

ventional methods for generating the sentences such as FUF/Surge or the realiser of the LKB system, as well as the use of a corpus more relevant to CH. At this point we concede that the generation space may be problematic as input gets longer, but we conjecture that user's input should in most cases be of manageable length. Finally, more standards evaluation metrics could serve to situate the two similarity measures that are being presented with regards to more standard approaches used for the same purpose (KAUCHAK, 2006).

Finally, we have avoided the issue raised by polysemic words by considering only the most common senses found in WordNet, so the approach would be well complemented by contribution from the field of Word-Sense Disambiguation (WSD).

Acknowledgement

This work has been conducted as part of the EPOCH network of excellence (IST-2002-507382) within the IST (Information Society Technologies) section of the Sixth Framework Programme of the European Commission. Thank you to the reviewers for useful comments.

References

ANDROUTSOPOULOS I., RITCHIE G., THANISCH P. (1995). *Natural language interfaces to databases - an introduction*. Journal of Language Engineering, 1(1), 29.

BUDANITSKY A., HIRST G. (2001). *Semantic distance in wordnet : an experimental, application-oriented evaluation of five measures*. In NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh.

BURKE R.D., HAMMOND K.J., KULYUKIN V., LYTINEN S.L., TOMURO N., SCHOENBERG. S. *Question answering from frequently asked question files - experiences with the faq finder system*. AI Magazine, 18(2), 57.

CHURCH K.W., HANKS P. (1989) *Word association norms, mutual information, and lexicography*. In Proc. of the 27th. Annual Meeting of the ACL Vanvouver, B.V., 1989), pp. 76-83.

CRESCIOLI M., D'ANDREA A., NICCOLUCCI F. (2002). *XML Encoding of Archaeological Unstructured Data. In G. Burenhault (ed.), Archaeological Informatics : Pushing the envelope*. In Proc. of the 29th CAA Conference, Gotland April 2001, BAR International Series 1016, Oxford 2002, 267-275.

DAGAN I., GLICKMAN O., MAGNINI B. (2006). *The PASCAL Recognising Textual Entailment Challenge*. Lecture Notes in Computer Science, Volume 3944, Jan 2006, Pages 177 - 190.

DOERR M. (2005) *The CIDOC CRM, an Ontological Approach to Schema Heterogeneity*. Semantic Interoperability and Integration. Dagstuhl Seminar Proceedings, pp. 1862-4405. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

HERMON S., NICCOLUCCI F. (2000). *The Impact of Web-shared Knowledge on Archaeological Scientific Research*. Proc. of Intl CRIS 2000 Conf., Helsinki, Finland, 2000.

KAUCHAK D., BARZILAY R. (2006). *Paraphrasing for Automatic Evaluation*. In Proc. of NAACL/HLT, 2006.

LI Y., YANG H., JAGADISH H. (2006). *Constructing a generic natural language interface for an xml database*. International Conference on Extending Database Technology

PEDERSEN T., PATWARDHAN S.,MICHELIZZI J.(2004). *Wordnet::Similarity - Measuring the Relatedness of Concepts*. In Nineteenth National Conference on Artificial Intelligence (AAAI-04), San Jose, CA. (Intelligent Systems Demonstration).

QUIRK C., BROCKETT C., DOLAN W.B. (2004). *Monolingual Machine Translation for Paraphrase Generation*. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona Spain.

RISTAD E.S., YIANILOS P. N. (1998). *Learning string-edit distance*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(5), 522.

SCHUTZ A., BUITELAR P. (2005). *RelExt: A Tool for Relation Extraction in Ontology Extension*. In: Proc. of the 4th International Semantic Web Conference, Galway, Ireland, Nov. 2005.

SHETH A. (2003) *Capturing and applying existing knowledge to semantic applications*. Invited Talk "Sharing the Knowledge" - International CIDOC CRM Symposium. March 2003. Washington DC.

All web references visited on 02-05-2007.

Dynamic Path Prediction and Recommendation in a Museum Environment

Karl Grieser[†‡], Timothy Baldwin[†] and Steven Bird[†]

† CSSE
University of Melbourne
VIC 3010, Australia

‡ DIS
University of Melbourne
VIC 3010, Australia

{kgrieser,tim,sb}@csse.unimelb.edu.au

Abstract

This research is concerned with making recommendations to museum visitors based on their history within the physical environment, and textual information associated with each item in their history. We investigate a method of providing such recommendations to users through a combination of language modelling techniques, geospatial modelling of the physical space, and observation of sequences of locations visited by other users in the past. This study compares and analyses different methods of path prediction including an adapted naive Bayes method, document similarity, visitor feedback and measures of lexical similarity.

1 Introduction

Visitors to an information rich environment such as a museum, are invariably there for a reason, be it entertainment or education. The visitor has paid their admission fee, and we can assume they intend to get the most out of their visit. As with other information rich environments and systems, first-time visitors to the museum are at a disadvantage as they are not familiar with every aspect of the collection. Conversely, the museum is severely restricted in the amount of information it can convey to the visitor in the physical space.

The use of a dynamic, intuitive interface can overcome some of these issues (Filippini, 2003; Benford et al., 2001). Such an interface would conventionally take the form of a tour guide, audio tour, or a curator stationed at points throughout the museum. This research is built around the assumption that the museum visitor has access to a digital device such as a PDA and that it is possible for automatic systems to interact with the user via this device. In this way we aim to be able to deliver relevant content to the museum visitor based on observation of their movements within the physical museum space, as well as make recommendations of what exhibits they might like to visit next and why. At present, we are focusing exclusively on the task of recommendation.

Recommendations can be used to convey predictions about what theme or topic a given visitor is interested in. They can also help to communicate unexpected connections between exhibits (Hitzeman et al., 1997), or explicitly introduce variety into the visit. For the purposes of this research, we focus on this first task of providing recommendations consistent with the visitor's observed behaviour to that point. We investigate different factors which we hypothesise impact on the determination of what exhibits a given visitor will visit, namely: the physical proximity of exhibits, the conceptual similarity of exhibits, and the relative sequence in which other visitors have visited exhibits.

Recommendation systems in physical environments are notoriously hard to evaluate, as the recommendation system is only one of many stimuli which go to determine the actual behaviour of the visitor. In order to evaluate the relative impact of different factors in determining actual visitor behaviour, we separate the stimuli present into a range of predictive methods. In this paper we target the task of user prediction, that is prediction of what exhibit a visitor will visit next based on

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 49–56,
Prague, 28 June 2007.

their previous history. Language based models are intended to simulate a potentially unobservable source of information: the visitor's thought process. In order to identify the reason for the visitor's interest in the multiple part exhibits we parallel this problem with the task of word sense disambiguation (WSD). Determining the visitor's reason for visiting an exhibit allows a predictive system to more accurately model the visitor's future path.

This study aims to arrive at accurate methods of predicting how a user will act in an information-rich museum. The space focused on in this research is the Australia Gallery Collection of the Melbourne Museum, at Carlton Gardens in Melbourne, Australia. The predictions take the form of which exhibits a visitor will visit given a history of previously visited exhibits. This study analyses and compares the effectiveness of supervised and unsupervised learning methods in the museum domain, drawing on a range of linguistic and geospatial features. A core contribution of this study is its focus on the relative import of heterogeneous information sources a user makes use of in selecting the next exhibit to visit.

2 Problem Description

In order to recommend exhibits to visitors while they are going through a museum, the recommendations need to be accurate/pertinent to the goals that the visitor has in mind. Without accurate recomme ndations, recommendations given to a visitor are essentially useless, and might as well not have been recommended at all.

Building a recommender system based on contextual information (Resnick and Varian, 1997) is the ultimate goal of this research. However the environment in this circumstance is physical, and the actions of visitors are expected to vary within such a space, as opposed to the usual online or digital domain of recommender systems. Studies such as HIPS (Benelli et al., 1999) and the Equator project[1] have analysed the importance and difficulty of integrating the virtual environment into the physical, as well as identifying how non-physical navigation systems can relate to similar physical systems. For the purpose of this study, it is sufficient to acknowledge the effect of the physical environment by scaling all recommendations against their distances from one another.

The common information that museum exhibits contain is key in determining how each individual relates to each other exhibit in the collection. At the most basic level, the exhibits are simply isolated elements that share no relationship with one another, their only similarity being that they occur together in visitor paths. This interpretation disregards any meaning or content that each exhibit contains. But museum exhibits are created with the goal of providing information, and to disregard the content of an exhibit is to disregard its purpose.

An exhibit in a museum may be many kinds of things, and hence most exhibits will differ in presentation and content. The target audience of a museum is one indicator of the type of content that can be expected within each exhibit. An art gallery is comprised of mainly paintings and sculptures: single component exhibits with brief descriptions. A children's museum will contain a high proportion of interactive exhibits, and much audio and visual content. In these two cases the reason for visiting the exhibit differs greatly.

Given the diversity of information contained within each exhibit and the greater diversity of a museum collection, it can be difficult to see why visitors only examine certain exhibits during their tours. It is very difficult to perceive what a visitor's intention is without constant feedback, making the problem of providing relevant recommendations a question of predicting what a visitor is interested in based on characteristics of exhibits the visitor has already seen. The use of both physical attributes and exhibit information content are used in conjunction in an effort to account for multiple possible reasons for visiting as exhibit. Connections between physical attributes of an exhibit are easier to identify than connections based on information content. This is due to the large quantity of information associated with each exhibit, and the difficulty in determining what the visitor liked (or disliked) about the exhibit.

In order to make prediction based on a visitor's history, the importance of the exhibits in the visitors path must be known. This is difficult to obtain directly without the aid of real-time feedback from

[1] `http://www.equator.ac.uk`

the user themselves. In an effort to emulate the difficulty of observing mental processes adopted by each visitor, language based predictive models are employed.

3 Resources

The domain in which all experimentation takes place is the Australia Gallery of the Melbourne Museum. This exhibition provides a history of the city of Melbourne Melbourne, from its settlement up to the present day, and includes such exhibits as the taxidermised coat of Phar Lap (Australia's most famous race horse) and CSIRAC (Australia's first, and the world's fourth, computer). The Gallery contains enough variation so that not all exhibits can be classified into a single category, but is sufficiently specialised to offer much interaction and commonality between the exhibits.

The exhibits within the Australia Gallery take a wide variety of forms, from single items with a description plaque, to multiple component displays with interactivity and audio-visual enhancement; note, for our purposes in experimentation, we do not differentiate between exhibit types or modalities. The movement of visitors within an exhibition can be restricted if the positioning of the exhibits require visitors to take a set path (Peponis et al., 2004), which can alter how a visitor chooses between exhibits to view. In the case of the Australia Gallery, however, the collection is spread out over a sizeable area, and has an open plan design such that visitor movement is not restricted or funnelled through certain areas and there is no predetermined sequence or selection of exhibits that a given visitor can be expected to spend time at.

We used several techniques to represent the different aspects of each exhibit. We categorised each exhibit by way of its **physical attributes** (e.g. size) and taxonomic information about the **exhibit content** (e.g. clothing or animal). We also described each exhibit by way of its **physical location** within the Australia Gallery, relative to a floorplan of the Gallery.

The Melbourne Museum also has a sizable web-site[2] which contains much detailed information about the exhibits within the Australia Gallery. This data is extremely useful in that it provides a rich vocabulary of information based on the content of each exhibit. Each exhibit identified within the Australia Gallery has a corresponding web-page describing it. The information content of an exhibit is made up of the text in its corresponding web-page combined with its attributes. By having a large source of natural language information associated with the exhibit, linguistic based predictive methods can more accurately identify the associations made by visitors.

The dataset that forms that basis of this research is a database of 60 visitor paths through the Australia Gallery, which was collected by Melbourne Museum staff over a period of four months towards the end of 2001. The Australia Gallery contains a total of fifty-three exhibits. This data is used to evaluate both physical and conceptual predictive methods. If predictive methods are able to accurately describe how a visitor travels in a museum, then the predictive method creates an accurate model of visitor behaviour.

Exhibit components can be combined to form a description for each exhibit. For this purpose, the Natural Language Toolkit [3] (Bird, 2005) was used to analyse and compare the lexical content associated with each exhibit, so that relationships between exhibits can be identified.

4 Methodology

Analysis of user history as a method of prediction (or recommendation) has been examined in Chalmers et al. (1998). Also discussed is the role that user history plays in anticipating user goals. This approach can be adapted to a physical environment by simply substituting in locations visited in place of web pages visited. Data gathered from the paths of previous visitors also forms a valid means of predicting other visitors' paths (Zukerman and Albrecht, 2001). This approach operates under the assumption that all visitors behave in a similar fashion when visiting a museum. However visitors' goals in visiting a museum can differ widely. For example, the goals of a student researching a project will differ to those of a family with young children on a weekend outing.

[2] http://www.museum.vic.gov.au/

[3] http://nltk.sourceforge.net/

A conceptual model of the exhibition space is created by visitors with a specific task in mind. Interpretation of this conceptual model is key to creating accurate recommendations. The building of such a conceptual model takes place from the moment a visitor enters an exhibition, until the time they leave, and skews the visitor towards groups of conceptual locations and categories.

The representation of these intrinsically dynamic models is directly related to the task the visitor has in mind. Students will form a conceptual model based around their course requirements, children around the most visually attractive exhibits, and so forth. This necessitates the need for multiple exhibit similarity measures, however in the absence of express knowledge of the 'type' of each visitor in the sample data, a broad-coverage recommendation system that functions best in all circumstances is the desired goal. It is hoped that in future, reevaluation of the data to classify visitors into broad categories (e.g. information seeking, entertainment seeking) will allow for the development of specialised models tailored to visitor types.

The models of exhibit representation we examine in this research are exhibit proximity, text-based exhibit information content, and exhibit popularity (based on the previous visitor data provided by the Melbourne Museum), as well as combinations of the three. Exhibit information content is a two part representation: primarily each exhibit has a large body of text describing the exhibit drawn from the Melbourne Museum website. It is fortunate that this information is curated, and managed from a central source, so that inconsistencies between exhibit information are extremely rare. The authors were unable to find any contradictory information in the web-pages used for experimentation, as may be the case with larger non-curated document bodies. The second component of the information content is a small set of key terms describing the attributes of the exhibit. Textual content as a means of determining exhibit similarity has been analysed previously (Green et al., 1999), both in terms of keyword attributes and bodies of explanatory text.

In order to form a prediction about which exhibit a visitor will next visit, the probability of the transition of the visitor from their current location to every other exhibit in the collection must be known. Prediction of the next exhibit by proximity simply means choosing the closest not-yet-visited exhibit to the visitor's current location. In terms of information content, each exhibit is related to all other exhibits to a certain degree. To express this we use the attribute keywords as a query to find the exhibit most similar. We use the attribute keywords associated with each document to search the document space of the exhibits to find the exhibit that is most similar to the exhibit the visitor is currently located at. To do this we use a simple tf·idf scheme, using the attribute keywords as the queries, and the exhibit associated web pages as the document space. The score of each query over each document is normalised into a transitional probability array such that $\sum_j P(q|d_j) = 1$ for a query (q) over the j exhibit documents (d_j).

In order to determine the popularity of an exhibit, the visitor paths provided by the Melbourne Museum were used to form another matrix of transitional probabilities based on the likelihood that a visitor will travel to an exhibit from the exhibit they are currently at. I.e. for each exhibit e an array of transitional probabilities is formed such that $\sum_j P(e|c_j) = 1$ where $c_j \in C' = C/\{e\}$, i.e. all exhibits other than e. In both cases Laplacian smoothing was used to remove zero probabilities.

The methods of exhibit popularity and physical proximity are superficial in scope and do not extend into the conceptual space adopted by the visitors. They do however give insight into how a physical space affects a visitors' mental representation of the conceptual areas associated with specific exhibit collections, and are more easily observable. Visitor reaction to exhibit information content is harder to observe and more problematic to predict. Any accurate recommender systems produced in this fashion will need to take into account the limitations these two methods place on the thought processes of visitors.

Connections that visitors make between exhibits are more fluid, and are harder to represent in terms of similarity measures. Specifically it is difficult to see why visitors make connections between exhibits as there can be multiple similarities between two exhibits. To this end we have equated this problem with the task of Word Sense Disambiguation (WSD). The path that a visitor takes can be seen as a sentence of exhibits, and each exhibit in the

sentence has an associated meaning. WSD is used to determine the meaning of the next exhibit based on the meanings of previous exhibits in the path. For each word in the keyword set of each exhibit, the WordNet (Fellbaum, 1998) similarity is calculated against each word in another exhibit. The similarity is the sum of the WordNet similarities between all attribute keywords in the two exhibits (K_1, K_2), normalised over the length of both keyword sets:

$$\frac{\sum_{k_1 \in K_1} \sum_{k_2 \in K_2} WNsim(k_1, k_2)}{|K_1||K_2|}$$

For the purposes of this experiment we have chosen to use three WordNet similarity/relatedness measures to simulate the conceptual connections that visitors make between exhibits. The Lin (Lin, 1998) and Leacock-Chodorow (Leacock et al., 1998) similarity measures and the Banerjee-Pedersen (Patwardhan and Pedersen, 2003) relatedness measures were used. The similarities were normalised and transformed into probability matrices such that $\sum_j P_{WNsim}(e|c_j) = 1$ for each next exhibit c_i. The use of WordNet measures is intended to simulate the mental connections that visitors make between exhibit content, given that each visit can interpret content in a number of different ways.

The history of the visitor at any given time is essential in keeping the visitor's conceptual model of the exhibit space current. The recency of a given exhibit within a visitor's history is inversely proportional to how long ago the exhibit was encountered.

To take into account the visitor history, the collaborative data, proximity, document vectors, and conceptual WordNet similarity, we adapt the naive Bayes approach. The conditional probabilities of each method are combined along with the temporal recency of an exhibit to produce a predictive exhibit recommender. The resultant recommendation to a visitor can be described as follows:

$$\hat{c} = \arg\max_{c_i} P(c_i) \sum_{j=1}^{t} P(A_j|c_i) \times 2^{-(t-j+1)} + \frac{2^{-t}}{t}$$

where t is the length of the visitor's history, $A_j \in C$ is an exhibit at time j in the visitor history (and C is the full set of exhibits), and $c_i \in C' = C/\{A_j\}$ is each unvisited exhibit. The most probable next exhibit ($\hat{c}$) is selected from all possible next exhibits (c_i). Any selections made must be compared against the visitor's history. In this, we assume that a previously visited exhibit has already been seen, and hence should not be recommended again.

The effectiveness of these methods was tested in multiple combinations, both with history modeling and without (only the exhibit the visitor is currently at is considered). Testing was carried out using the sixty visitor paths supplied by the Melbourne Museum. For each method two tests were carried out:

- Predict the next exhibit in the visitor's path.
- Only make a prediction if the probability of the prediction is above a given threshold.

Each path was analysed independently of the others, and the resulting recommendations evaluated as a whole. The measures of precision and recall in the evaluation of recommender systems has been applied effectively in previous studies (Raskutti et al., 1997; Basu et al., 1998). In the second test precision is the measure we are primarily concerned with: it is not the aim of this recommender system to predict all elements of a visitor's path in the correct order. The correctness of the exhibits predicted is more important than the quantity of the predictions the visitor visits, hence only exhibits predicted with a (relatively) high probability are included in the final list of predicted exhibits for that visitor.

The thresholds are designed to increase the correctness of the predictions, by only making a prediction if there is a high probability of the visitor travelling to the exhibit. As all predictive methods choose the most probable transition from all possible transitions, the transition with the highest probability is always selected. The threshold values simply cut off all probabilities below a certain value.

5 Results and Evaluation

The first tests carried out were done only using the simple probability matrices described in Section 4, and hence only use the information associated with the visitor's current location and not the entirety of their history. The baseline method being used in all testing is the naive method of moving to the closest not-yet-visited exhibit.

Method	BOE	Accuracy
Proximity (baseline)	0.270	0.192
Popularity	0.406	0.313
Tf·Idf	0.130	0.018
Lin	0.129	0.039
Leacock-Chodorow	0.116	0.024
Banerjee-Pedersen	0.181	0.072
Popularity - Tf·Idf	0.196	0.093
Popularity - Lin	0.225	0.114
Popularity - Leacock-Chodorow	0.242	0.130
Popularity - Banerjee-Pedersen	0.163	0.064
Proximity - Tf·Idf	0.205	0.084
Proximity - Lin	0.180	0.114
Proximity - Leacock-Chodorow	0.220	0.151
Proximity - Banerjee-Pedersen	0.205	0.105
Proximity - Popularity	0.232	0.129

Table 1: Single exhibit history using individual and combined transitional probabilities

In order to prevent specialisation of the methods over the training data (the aforementioned 60 visitor paths), 60 fold cross-validation was used. With the path being used as the test case removed from the training data at each iteration.

The results of prediction using only the current exhibit as information can be seen in Table 1. Combinations of predictive methods are also included to add physical environment factors to conceptual similarity methods. For example, if two exhibits may be highly related conceptually but on opposite sides of the exhibit space, a visitor may forgo the distant exhibit in favour of a closer exhibit that is slightly less relevant.

Due to the lengths of the recommendation sets made for each visitor (a recommendation is made for each exhibit visited), precision and recall are identical. The measure of Bag Of Exhibits (BOE) describes the percentage of exhibits that were visited by the visitor, but not necessarily in the same order as they were recommended. The BOE measure is the same as measuring precision and recall for the purposes of this evaluation. With the introduction of thresholds to improve precision, precision and recall are measured as separate entities.

As seen in Table 1 the performance of the conceptual or information similarity methods (the tf·idf method, Lin, Leacock-Chodorow and Banerjee-Pedersen) is worse than that of the methods based on static features of the exhibits, and all perform worse than the baseline. In order to produce a higher percentage of correct recommendations, thresholds were introduced. Using thresholds, a recommendation is only made if the probability of a visitor visiting an exhibit next is above a given percentage. The thresholds used in Table 2 are arbitrary, and were arrived at after experimentation.

It is worth noting that in both tests, with and without thresholds, the method of exhibit popularity based on visitor paths is the most successful. One expects this trend to continue with the introduction of the history based model described in Section 4. Each transitional probability matrix was used in conjunction with the history model, the results of this experimentation can be seen in Table 3.

Only single transitional probability matrices are used in conjunction with the history model. The physical distance to an exhibit is only relevant to the current prediction, the distance travelled in the past from exhibit to exhibit is irrelevant, and so physical conceptual combinations are not necessary. A model such as this describes the evolution of a thought process, or is able to identify the common conceptual thread linking the exhibits in a visitor's path. This is only true *if* the visitor has a conceptual model in mind when touring the museum. Without the aid of a common information thread, conceptual predictive methods based on exhibit information content will always perform poorly.

6 Discussion

The visitor paths supplied by the Melbourne Museum represent sequential lists of exhibits, and each visitor is a black box travelling from exhibit to exhibit. It is this token vs. type problem that does not allow us to select an appropriate predictive method with which to make recommendations. Instead a broad coverage method is necessary. Use of history models to analyse entire visitor paths are less successful than analysis of solely the current location of the visitor. This can be attributed to the fact that a majority of the visitors tracked may not have had preconceived tasks in mind when they entered the museum space, and just moved from one visually impressive exhibit to the next. The visitors do not consider their entire history as being relevant, and only take into account their current

Method	Threshold	Precision	Recall	F-score
Proximity	0.03	0.271	0.270	0.270
Popularity	0.06	0.521	0.090	0.153
Tf·Idf	0.06	0.133	0.122	0.128
Lin	0.01	0.129	0.129	0.129
Leacock-Chodorow	0.01	0.117	0.117	0.117
Banerjee-Pedersen	0.01	0.182	0.180	0.181
Popularity - Tf·Idf	0.001	0.176	0.154	0.164
Popularity - Lin	0.0005	0.383	0.316	0.348
Popularity - Leacock-Chodorow	0.0005	0.430	0.349	0.385
Popularity - Banerjee-Pedersen	0.001	0.236	0.151	0.184
Proximity - Tf·Idf	0.001	0.189	0.174	0.181
Proximity - Lin	0.0005	0.239	0.237	0.238
Proximity - Leacock-Chodorow	0.0005	0.252	0.250	0.251
Proximity - Banerjee-Pedersen	0.0005	0.182	0.180	0.181
Proximity - Popularity	0.001	0.262	0.144	0.186

Table 2: Single exhibit history predictive methods using thresholds

Method	BOE	Accuracy
Proximity	0.066	0.0
Popularity	0.016	0.0
Tf·Idf	0.033	0.0
Lin	0.064	0.0
Leacock-Chodorow	0.036	0.0
Banerjee-Pedersen	0.036	0.0

Table 3: Entire visitor history predictive methods.

context. This also explains the relative success of the predictive method built from analysis of the visitor paths, presenting a marked improvement over the baseline of nearest exhibit. In the best case (as seen in Table 2) the exhibit popularity predictive method was able to give relevant recommendations 52% of the time.

The interaction between predictive methods here is highly simplified. The assumption made is that all aspects of the visitor's conceptual model are independent, or only interact on a superficial level (see the lower halves of Tables 1–2). More complex methods of prediction need to be explored fully take into account the interaction between predictive methods.

Representations based on physical proximity take into account little of how a visitor conceptualises a museum space. They do however describe the fact that closer exhibits are more visible to visitors, and are hence more likely to be visited. Proximity can be used as an augmentation to a conceptual model designed to be used within a physical space.

Any exhibit is best described by the information it contains. Visitors with a specific task in mind when entering an exhibition already have a pre-initialised conceptual model, relating to a theme. The visitors seek out content related to their conceptual model, and separate the bulk of the collection content from the information they require. The representation of the content within each exhibit as a vocabulary of terms allows us to find similarity between exhibits. The data available at the time of this testing does not make the distinction between user types, and so only broad coverage methods result in a improvements.

With the introduction of user types to the data supplied by the museum, specific predictive methods can be applied to each individual user. This additional information can be significantly beneficial as the specialisation of predictive types to visitors is expected to produce much more accurate predictions and recommendations. Currently the only method available to discern the user type is to analyse the length of time the visitor spends at each each exhibit. This data is yet to be adapted and annotated from the raw data supplied by the Melbourne Museum.

7 Conclusion

The above methods are intended to represent baseline components of possible conceptual models that represent how a visitor is able to selectively assess the dynamic context of museum visits. The model that a visitor generates for themselves is unique, and is difficult to represent in terms of physical attributes of exhibits.

Being able to predict future actions of a user within a given environment allows a recommender system to influence a user's choices. Key to the prediction of future actions, is the idea that a user has a conceptual model of how they see content within the environment in relation to a task. With respect to a museum environment, the majority of users have no preconceived conceptual model upon entering an exhibition and must build one as they explore the environment. Users with a preconceived task will more often than not stick to exhibits surrounding a particular theme. Use of a language-based conceptual model based on the information contained within an exhibit can be combined with conceptual models based on geospatial attributes of the exhibit to create a representation of how a user will react to an exhibit. The use of heterogeneous information contained within the exhibit space is only relevant when the visitor has an information-centric task in mind.

7.1 Future Work

The methods dealing with a language-based conceptual model given here are very basic, and the overall accuracy and precision of the recommender system components require improvement. Additional annotation of the paths of visitors to the museum will enable proper evaluation of conceptual information based predictive methods. On-site testing of predictive methods at the Melbourne Museum is the ultimate goal of this project, and testing the effects of visitor feedback on recommendations will also be analysed. In order to gain more insight into visitor behaviour, the current small-scale set of visitors needs to be expanded to include multiple visitor types, as well as tasks.

Acknowledgments

This research was supported by Australian Research Council DP grant no. DP0770931. The authors wish to thank the staff of the Melbourne Museum for their help in this study. Special thanks goes to Carolyn Meehan and Alexa Reynolds for their gathering of data, and helpful suggestions throughout this study. Thanks also goes to Ingrid Zukerman and Liz Sonenberg for their input on this research.

References

Chumki Basu, Haym Hirsh, and William Cohen. 1998. Recommendations as classification: Using social and content-based information in recommendation. In *Proceedings of the National Conference of Artificial Intelligence*, pages 714–720, Madison, United States.

Giuliano Benelli, Alberto Bianchi, Patrizia Marti, David Sennati, and Elena Not. 1999. HIPS: Hyper-Interaction within Physical Space. In *ICMCS '99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume 2, page 1075. IEEE Computer Society.

Steve Benford, John Bowers, Paul Chandler, Luigina Ciolfi, Martin Flintham, Mike Fraser, Chris Greenhalgh, Tony Hall, Sten-Olof Hellstrom, Shahram Izadi, Tom Rodden, Holger Schnadelbach, and Ian Taylor. 2001. Unearthing virtual history: using diverse interfaces to reveal hidden worlds. In *Proc Ubicomp*, pages 1–6. ACM.

Steven Bird. 2005. NLTK-Lite: Efficient scripting for natural language processing. In *Proceedings of the 4th International Conference on Natural Language Processing (ICON)*, pages 11–18, Kanpur, India.

Matthew Chalmers, Kerry Rodden, and Dominique Brodbeck. 1998. The Order of Things: Activity-Centred Information Access. *Computer Networks and ISDN Systems*, 30:1–7.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Silvia Filippini. 2003. Personalisation through IT in museums: Does it really work? Presentation at ICHIM 2003.

Stephen J. Green, Maria Milosavljevic, Robert Dale, and Cecile Paris. 1999. When virtual documents meet the real world. In *Proc. of WWW8 Workshop: Virtual Documents, Hypertext Functionality and the Web*.

Janet Hitzeman, Chris Mellish, and Jon Oberlander. 1997. Dynamic generation of museum web pages: The intelligent labelling explorer. *Archives and Museum Informatics*, 11(2):117–115.

Claudia Leacock, Martin Chodorow, and George A Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–65.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *(CoLING)-(ACL)*, pages 768–774, Montreal, Canada.

Siddharth Patwardhan and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, Mexico.

John Peponis, Ruth Conroy Dalton, Jean Wineman, and Nick Dalton. 2004. Measuring the effect of layout on visitors' spatial behaviors in open plan exhibition settings. *Environment and Planning B: Planning and Design*, 31:453–473.

Bhavani Raskutti, Anthony Beitz, and Belinda Ward. 1997. A feature-based approach to recommending selections based on past preferences. *User Modelling and User Adaption*, 7(3):179–218.

Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM*, 40(3):56–58.

Ingrid Zukerman and David W Albrecht. 2001. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1–2):5–18.

Anchoring Dutch Cultural Heritage Thesauri to WordNet: two case studies

Véronique Malaisé and **Antoine Isaac**
Vrije Universiteit
Amsterdam
The Netherlands
{vmalaise, aisaac}@few.vu.nl

Luit Gazendam
Telematica Instituut
Enschede
The Netherlands
Luit.Gazendam@telin.nl

Hennie Brugman
Max Planck Institute
for Psycholinguistics, Nijmegen
The Netherlands
Hennie.Brugman@mpi.nl

Abstract

In this paper, we argue on the interest of anchoring Dutch Cultural Heritage controlled vocabularies to WordNet, and demonstrate a reusable methodology for achieving this anchoring. We test it on two controlled vocabularies, namely the GTAA thesaurus, used at the Netherlands Institute for Sound and Vision (the Dutch radio and television archives), and the GTT thesaurus, used to index books of the Dutch National Library. We evaluate the two anchorings having in mind a concrete use case, namely generic alignment scenarios where concepts from one thesaurus must be aligned to concepts from the other.

1 Introduction

Cultural Heritage Institutions are the keepers of large collections of data. To optimize the core tasks of indexing and searching through these collections, controlled vocabularies like thesauri are often used. These vocabularies are structured concept networks[1] and help indexers to select proper subjects for description, and users to formulate queries or to browse collections using the concepts that appear in the metadata.

The Netherlands Institute for Sound and Vision[2], for example, uses the GTAA thesaurus for indexing public radio and TV programs – GTAA is a Dutch abbreviation for "Common Thesaurus [for] Audiovisual Archives". Its hierarchy of subjects contains about 3800 Preferred Terms and 2000 Non Preferred terms. A second example is the GTT thesaurus, which contains 35000 concepts, gathering 50000 preferred and non-preferred Dutch terms. This thesaurus is used to index and retrieve books from the Dutch National Library[3] – GTT is a Dutch abbreviation for "GOO keyword thesaurus", GOO referring to the Joint Subject Indexing system used by many Dutch libraries.

Besides this classic scenario, thesauri can also allow for (semi-)automatic optimization of search processes, like query expansion exploiting their hierarchical structure. But the available structure might not be rich and regular enough for such purposes. In fact, it has been shown that a mapping to a richer and sounder terminology, like the English WordNet (Fellbaum, 1998), would enable more sophisticated query expansion or other inferencing possibilities (Voorhees, 1994; Hollink, 2006). This will become especially true now that WordNet exists in the form of an RDF ontology (van Assem et al., 2006).

Mapping Cultural Heritage controlled vocabular-

[1] The typical semantic relationships found between elements from thesauri are **Broader Term** linking a specialized concept to a more general one, **Narrower Term**, its inverse relationship, and **Related Term**, which denotes a general associative link. Thesauri also contain lexical information, where the *preferred terms* used for description are given *synonyms* or *non-preferred terms* (**Use** and **Used for** links), as well as general **scope notes** giving indexers instructions regarding the use of a term.

[2] http://www.beeldengeluid.nl

[3] http://www.kb.nl

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 57–64,
Prague, 28 June 2007.

ies in Dutch to WordNet can also be beneficial for sharing information across institutions, which is difficult when the metadata attached to the different documents come from different thesauri. This issue can be solved by building equivalence links between the elements from these different vocabularies, as in (van Gendt et al., 2006). This *vocabulary alignment* problem is comparable to the *ontology matching* one, and techniques similar to the ones developed by the Semantic Web research community can be applied here. As found e.g. in (Euzenat, 2004), the existing methods are quite diverse, and proposed strategies often mix several individual techniques:

- lexical techniques, trying to compare the labels found in vocabularies;
- structural techniques, assessing similarities between concepts from the structure of vocabularies (e.g. hierarchical links);
- instance-based techniques, looking at the objects that are actually populating the ontologies to infer from their similarities correspondences between the concepts they instantiate.
- techniques making use of some background knowledge source, by trying to derive from the information found there relations between the elements from the original vocabularies.

Here, we are interested in the last kind of techniques. In these approaches, concepts from the vocabularies to be aligned are first attached – "anchored" – to the concepts from a third vocabulary (Aleksovski, 2006). Then, these anchors in the background vocabulary are compared together. When a relation is found between them[4], a similar relation can be inferred between the elements from the vocabularies to be aligned. This is especially interesting when the lexical overlap between the vocabularies is low or when the vocabularies are quite poorly structured: it is expected then that the background knowledge will alleviate these shortcomings. The choice of this knowledge is therefore crucial, and WordNet, which has a rich structure and a broad coverage, has been exploited in many existing alignment methods (Giunchiglia et al., 2005; Castano et al., 2005).

For these reasons – even if this paper will only focus on the alignment scenario – we wanted to experiment the anchoring of two aforementioned Dutch thesauri to WordNet. Unlike literature about linking English thesauri to WordNet, we propose in this paper an anchoring method for vocabularies in other languages, and experiment it on these two thesauri, testing its usefulness in terms of possibilities for vocabulary alignment. The remainder of the paper is organized as follows: in section 2, we present the general anchoring methodology. The anchoring experiment is described in section 3: first the GTAA case (section 3.1) and then the GTT one (section 3.2), as a reusability test. We evaluate the two anchoring processes in section 3.3 and conclude on general reflexions about this method. Then, we show examples of such anchorings in the context of a possible alignment between GTAA and GTT in section 4. We conclude on perspectives to this research in section 5.

2 Anchoring methodology

The anchoring experiment presented in this paper is based on a comparison of lexical descriptions of the thesaurus terms with the ones of WordNet synsets, the *glosses*: WordNet is a lexical database of English, which entries are grouped "into sets of cognitive synonyms (synsets), each expressing a distinct concept"[5]. In contrast to many anchoring methods, like the one in (Khan and Hovy, 1997), we do not compare the terms from our thesauri to the labels of synsets, but measure the lexical overlap of their descriptions. The same approach has already been followed, for example, by (Knight and Luk, 1994).

As the thesauri we focus on in this paper are in Dutch, we first need to map their terms to English descriptions, and possibly translations, to make a comparison with the English glosses. Given the fact that these thesauri cover a broad range of topics, we hypothesize that using a general language bilingual dic-

[4]The reader can turn to (Budanitsky and Hirst, 2006) for an overview of the different methods that have been proposed in this field.

[5]http://wordnet.princeton.edu/

tionary will lead to a good coverage of their content. Additionally, it might give on top of the definitions – *i.e.* the natural language descriptions of a term's meaning – useful information such as term translations and Part Of Speech (POS) tags – their grammatical category: noun, verb, etc. For each thesaurus term which has been associated to an English definition, the rest of the anchoring procedure consists in checking the overlap between the lexical content of the definitions and the one of the different WordNet glosses, considered as bags of words. The hypothesis is that the closest gloss should give us a pointer to a synset semantically equivalent to the intended meaning of a thesaurus term.

3 Anchoring feasibility experiments and evaluations

3.1 Anchoring GTAA concepts

First step: Finding English definitions for GTAA terms The first step in mapping Dutch terms from the GTAA to WordNet was to select an online dictionary that would cover a significant part of the thesaurus entries and that would allow automatic queries for these terms. We have tested the bilingual dictionary LookWAYup[6], which returned a 2222 results – definitions and translations – on our query set.

This query set consisted in the list of GTAA Preferred terms (3800), Non preferred terms (2000) and their singular forms[7] (3200). These singular forms were computed in the context of a MultimediaN project[8], on the basis of linguistic derivational rules and a manual correction.

Given the fact that most of the thesaurus terms are in plural form, but not all of them[9], and knowing that the dictionary entries are only standard lemma forms (most of the time in singular), we first assumed that queries on the dictionary with a plural form would not generate a result, and simply added the singular forms to the singular ones in the query set. It turned out that the dictionary gave result for some plural forms, creating noise: some plural forms corresponded to lemmas of verbs, and a spelling correction facility provided definitions for some plural forms.

Removing doubles We cleaned manually the first set of errors, and automatically the last one, based on POS tag information. In the future, we will avoid introducing duplicate lemmas in our the query set.

After cleaning, 1748 terms had one or more translation in English together with their associated POS tag(s) and definition(s)[10]. This low number, compared with the original set of 5800 distinct thesaurus terms can be explained by the fact that our vocabulary contains numerous multi-words terms and also compound entries, both of which are rarely dictionary entries. We discuss possible solutions to this shortcoming in section 3.3.

POS tag-based cleaning We did then a rough manual evaluation of these candidate definitions. The evaluation was conducted by three people and took about one day each. It turned out that some of the definitions were irrelevant for our task: the Dutch *Bij* was associated with the English *Bee* and *Honey bee*, but also with the preposition *by*. We used again the information given by the POS tag to remove these irrelevant definitions: we kept only definitions of Nouns and (relevant) Verbs. After this last cleaning, some terms still had more then one definition.

Cleaning based on thesaurus relationships We used the hierarchical relationship in the thesaurus to check the intended meaning of these terms: for example, *Universiteit* (*University*) had a Broader Term relationship with *Wetenschappelijk onderwijs* (*Scientific education*), so its meaning is restricted to the "Educational aspect", and it should not be used to describe TV programs about University buildings for instance. We used this information to restrict the

[6]Built by RES Inc., Canada, online at the URL: `http://lookwayup.com/free/`.

[7]Following the recommendations of the ISO standard, most of GTAA terms are in plural form.

[8]MultimediaN Project 5 – Semantic Multimedia Access, `http://monetdb.cwi.nl/projects/trecvid/MN5/index.php/Main_Page`, transformation done by Gijs Geleijnse, from the Philips Research group.

[9]For example, the term corresponding to *Baptism* is in singular form.

[10]1299 terms have more than one definition.

Step	Result
Gathering query set	3800 + 2000 + 3200 terms
Querying dictionary	2222 defined terms
Removing doubles	1748 different defined terms
POS tag-based cleaning	1655 def. terms, 7530 definitions
Thesaurus-based cleaning	
Anchoring to WordNet	1060 anchored concepts

Table 1: GTAA term anchoring experiment

number of valid candidate definitions associated with every GTAA term. But in some cases the distinction was hard to make between the different definitions, or no clue was provided by the thesaurus to dismabiguate the senses of the term: sometimes it did not have any relationship to other concepts nor explanatory text (Scope Note).

Conclusion of the first step As a final result, as summarized in table 1, 1655 GTAA terms had one or more English equivalent and their related candidate definitions (7530). We decided to postpone a more in-depth validation to the evaluation of anchoring results with WordNet: we kept all candidate definitions and translations that were not obviously incorrect, and checked the WordNet anchoring result to see if some further refinement had to be done. The idea was that the anchoring process would only work for parts of the definitions, so we wanted to keep as many data as possible.

Second step: Anchoring to WordNet synsets We stemmed the candidate definitions of GTAA terms and the glosses from WordNet with the Porter stemmer to augment mapping possibilities. Stemming is the operation of reducing words to a root, for example by removing the "s' character at the end of an (English) word in plural form. This process can reduce different unrelated words to a same root, and hence should be handled with care, but it requires less resources then a full fledged lemmatizing and helps comparing a larger number of words then on the basis of the graphical forms only. As announced, in order to map synset to GTAA terms, we compared their lexical descriptions: we compared the different sets of stems in a simple bag-of-words approach. We actually found out that the definitions of the online dictionary were exact matches with WordNet glosses, thus all defined terms could be straightforwardly anchored to one or more synsets. In the end, 1060 concepts from GTAA are successfully anchored to a synset, which represents 28% of the total number of concepts.

Evaluation of the results We evaluated the number of semantically relevant anchorings for a random representative part of the the 1655 GTAA terms that had one or more WordNet anchor: we evaluated 1789 mappings out of 7530. On these 1789 mappings, 85 were not equivalence links: 5 out of these 85 links were relating Related Terms (like *zeerov* anchored to *corsair*, the first being in GTAA a profession and the second a ship in Wordnet), 17 pointed to Broader Terms, and the others were mapping a term with a correct translation that was correct *per se* but did not correspond to the intended meaning of the term in GTAA. For example, two anchorings were proposed for *Vrouwen*: *married_woman* and *female_person*, the latter one being the only valid for our thesaurus. The first cases (RT and BT relationships between the original term and its anchoring) still provide useful information for aligning vocabularies, but we took only equivalence relationships into account in this experiment.

An additional evaluation that was also performed on a sample set was to check that non-preferred terms that were given a definition were pointing to the same synset as their related preferred terms. It turned to be correct for the evaluated pairs.

On a qualitative perspective, we found different types of mappings:

- some GTAA terms had more then one translation, all of them pointing to the same synset: this was the confirmation that the mapping from the term to the synset was correct;
- some GTAA terms had more then one translation, pointing to different but close synsets: nothing in the thesaurus content could help us distinguish between the different synsets, thus we kept the different possibilities;

- some different GTAA terms pointed to a same synset and, although they were not linked in the thesaurus, they had a semantic relationship. This information can be used to enrich the structure of the GTAA.

We can conclude that the anchoring was quite successful: only 4.7% of the anchorings were incorrect in the test sample. And this was due to cases where multiple senses were linked to a same term, which would not cause a big problem in a semi-automated anchoring process. Moreover, this process can bring an additional value to the thesaurus structure itself, on top of the possible applications mentioned in the introduction.

3.2 Anchoring GTT concepts

Setting We carried out for GTT the same experiment as for GTAA, but did not compute singular forms, although GTT terms are generally in plural form. Also, because GTT had 70% of its concepts already translated to English by human experts, we decided that we would measure the global performance of our method based on this translation gold standard, additionally to manually assess the relevance of the produced anchorings from GTT to WordNet.

Results Out of the 35194 GTT general subjects, only 2458 were given some English definition and translation by the dictionary service we used. For the set of 25775 concepts for which there was already a translation, the figure drops down to 2279, slightly less than 9%.

As said, we tested the validity of these definitions and translations by comparing them to the expert translations. Our assumption was that an English definition for a concept would prove to be correct if its associated term matched one of the expert translations of the concept[11]. We found that 1479 of the 2279 concepts being given both expert and automatic translations had the expert translation confirming one of the automatically found ones, *i.e.* a precision rate of 65% in terms of defined concepts.

When measuring accuracy of the found English definitions for the 2279 defined concepts, we saw that out of a total 3813 English definitions associated to a concept, 2626 – 69% – had an associated term confirmed by the expert translation.

We also tried to assess the quality of the translations associated to the concepts of this set by our method: out of 5747 terms proposed as translations, 1479 matched the expert translation. This precision rate is low (25.7%) but it actually highlights one of the problem of the expert translations found in the thesaurus: the manual translation had a very low lexical coverage, having provided with very few synonyms for the "preferred" translations. The set of 25775 translated GTT concepts only brings 26954 English terms in total...

The evaluation by comparison to the expert translation brings useful information, but it has some drawbacks, especially the limited coverage of the translation work and a correctness assumption bringing a (small) error rate. To complete it, we carried out a manual investigation, inspired by what had been done for the GTAA thesaurus.

For this, we selected the 179 concepts that were translated by our method but had not previously been assigned English labels by experts. For this subset, 441 glosses had been assigned. Of these, 172 were correct, concerning 138 concepts. We therefore obtain a 77% precision rate in terms of anchored concepts. However, if we aim at assessing the quality of the method and its potential to be used in a semi-automatic anchoring process, we have to consider the obtained glosses themselves. And here precision falls to 39%, which is a far less satisfactory figure.

Feasibility of the proposed method in GTT case Some of the previously mentioned anchorings to wrong glosses could have been successfully found by applying the heuristics mentioned in section 3.1. The use of POS tags and the checking of the singular form of terms allowed to manually spot 41 obviously wrong results. The other irrelevant glosses were mainly found using the thesaurus information:

[11] A manual checking of this assumption on the first 150 concepts matching the criterion demonstrated an error rate of 4%: 4% of the concepts had no correct definition in their associated glosses while there was a match between the expert translation and one of the terms linked to the definitions.

Comparison with expert Gold Standard	
Concepts with expert translation	25775
Concepts with a definition	2279
Concepts with def. confirmed by GS	1479
Total definitions given	3813
Definitions confirmed by GS	2626
Total translations given	5747
Translations confirmed by GS	1479
Manual evaluation	
Concepts	179
Concepts with correct definition	138
Total definitions given	441
Correct definitions	172
Global results	
Total GTT concepts	35194
Concepts with a definition	2458
Concepts with correct definition	1617
Total definitions given	4254
Correct definitions	2798

Table 2: GTT term anchoring evaluation

the Broader Term information helped to discriminate 68 cases, compared with 6 for Related Term, 6 for synonyms and 15 for scope notes.

It is however still uncertain whether these different kinds of information can be used in a more automatised setting. If we could count on translation of broader and related terms to be done by the process we have applied, taking into account scope notes would require more effort. And the poor structure of thesauri such as GTT – some 20000 concepts have no parents at all – makes such validations by semantic links difficult. It is also important to notice that in 14 cases, it was necessary to check the books which have been indexed by a concept to find out its precise meaning.

This could yet be compensated by an interesting result we have observed: the anchoring method gave us material for inferring new semantic links, as in the GTAA case. Amongst the translated GTT concepts, 689 concepts are sharing at least one synset and are not connected by a thesaurus link. We found interesting matches, such as *gratie* (pardon) and *absolutie* (absolution) or between *honger* (hunger) and *dorst* (thirst). This potential for enriching thesauri could actually be used to spark some positive feedback loop for the anchoring process itself: a richer vocabulary enables for example to use with greater profit the selection strategies based on thesaurus structure.

An important problem for the implementation of such strategies remains to deal with disambiguation (when several English definitions are found, which one shall be selected?) in a context of fine-grained vocabularies. Both GTT and WordNet have a high level of precision, but they are focused on different matters. Especially, for a same GTT term the dictionary pointed at several meanings that were very close, but considered as different synsets in WordNet. A typical example is the distinction made between the gloss attached to moderation and temperance, "the trait of avoiding excesses", and the one attached to moderateness and moderation, "quality of being moderate and avoiding extremes". Looking at the books indexed by the concepts which these glosses were attached to, it was not clear whether the indexers systematically considered such a distinction.

Finally, we made rough estimattions of recall – the number of concepts that were correctly anchored compared to the number of concepts anchored in the ideal case. If we compare the 1479 correctly defined concepts to the 25775 concepts being given an expert translation, we find a very disappointing recall rate of 5.7%. This very low performance is in fact largely due to three recurrent situations in which the online dictionary could not give any translation:

- terms containing some special Dutch characters – especially the so-called Dutch *ij*, where *i* and *j* make a single character – and which occurs for more than 2000 concepts;
- specialized scientific terms, like *kwantumhalleffect*;
- complex notions, rendered in Dutch by compound words (e.g. *gebruikersinterfaces* for *user interfaces*), multi words (*Algemene kosten* for *general costs*) or a mixture of the two (*Grafische gebruikersinterfaces* for *graphic user interfaces*).

Whereas the encoding problem appears fairly simple, the last ones are more serious – they were indeed also encountered in the GTAA case – and shall be discussed further.

3.3 Conclusion on the anchoring methodology

As just mentioned, a drawback of our anchoring method is the fact that there are very few multi-word entries in dictionaries but they compose a large part of thesauri, and particularly thesauri in Dutch. Previous work about assigning a semantic relationship between a multi-word term and its components (see (Ibekwe, 2005)) could be used in order to give elements of solution to this problem. Using this preprocessing, we could apply our method to the single-word part that corresponds to the generic meaning of the original multi-word term, and try to anchor the single-word corresponding to the semantic root of the thesaurus' multi-word term (*Kosten* for *Algemene kosten* – *Cost* for *General cost* – for instance).

From a more conceptual point of view, however, further effort would be needed to adapt our anchoring method – and the subsequent alignment of one vocabulary with the other – to the cases where a concept from one vocabulary should be anchored to more than one element from WordNet. More complex heuristics come closer to traditional anchoring problems cases – without translation – and could be solved using existing solutions, as proposed by (Giunchiglia et al., 2005; Castano et al., 2005).

The last problem encountered in the anchoring process was the fact that specialized notions, that also appear in general purpose thesauri, have usually no definition in a general language dictionary. Specialized dictionaries should be used as a complementary resource.

These different shortcomings reduced the coverage of the anchoring, but our method has still positive points: the number of obviously wrong anchors was rather low for the found pairs and additional links could be provided for both of the source thesauri. This method also provides a starting point for anchoring complex and large vocabularies to WordNet, which is also a large lexical resource, and both are hard to grasp completely by a human expert.

4 GTAA and GTT alignment using WordNet anchoring: a qualitative evaluation

Once the anchoring is performed, the synsets corresponding to the terms from the different thesauri can be compared, in order to infer from them equivalences between the original concepts, as is done in classical alignment techniques using background knowledge. In this section, we present some examples illustrating the kind of alignment results one can expect from a proper anchoring of our Dutch controlled vocabularies.

First, we can confirm alignments of equal Dutch labels: *gtaa:arbeiders* is aligned to *gtt:arbeiders* since they are both anchored to the synset "someone who works with their hand, someone engaged in manual labor". In some cases, though, a first stemming or lemmatizing process would have been needed to achieve alignment, as in the example of *gtaa:bekeringen* and *gtt:bekering* (*Conversion*, respectively in plural and singular form), or *gtaa:biljart* and *gtt:biljartspel*[12] (*Billiard* and *Billiard game*).

Nevertheless, the more interesting cases are the ones involving concepts with large semantic overlap but a small lexical one, as in the case of *gtaa:plant* (*Plant*) and *gtt:begroeiing* (*Excessive growth of vegetation*) via the WordNet *flora* synset. *Begroeiing* is actually semantically related in the GTT to the concept *Planting*. Here, the translation process compensates for the lack of lexical coverage in the respective vocabularies, which precisely corresponds to one of the traditional features background knowledge-based techniques boast. We can also derive general conceptual similarity relationship based on the overlap between glosses, such as the one between *gtaa:drank* and *gtt:alcohol*, which are not direct matches but for which our method has found some common glosses like "an alcoholic beverage that is distilled rather than fermented".

[12]Notice that substring-based matching could also give these results, but this method is usually very noisy for alignment processes and therefore must be used cautiously.

5 Conclusion and perspectives

Our experiments showed that the partial anchoring of large Dutch controlled vocabularies to WordNet can be done via a bilingual dictionary, even though there is an obvious loss in information: not every thesaurus concept can easily be found in a general language bilingual dictionary, and a preprocessing of multi-word and compound thesaurus entries has to be done. Yet, a significant part of the GTAA thesaurus could be anchored, and with some improvement to the method this could be true for GTT too. Besides multi-word and compound words processing, useful extensions should also take into account specialized dictionaries and have a closer look at methodologies for anchoring a thesaurus term to multiple WordNet synsets with close meanings. We plan to test such strategies in future experiments, and hope to obtain a better coverage of the thesauri.

In this paper, we have sketched a way to use of these anchorings in a vocabulary alignment scenario, and underlined the potential gains on test examples. Even if the number of results given by the current implementation of our method is quite low, the reader should notice that the process can already, as is, suggest new relationships between concepts of the source thesauri. Moreover, proposed strategies in the alignment field often advocate using combined methods: combined contributions can be used to proceed with some cross validation if they overlap, or to provide with larger number of candidate for further (semi-)automatic selection. In such a setting, every contribution of candidate links is welcome. In this respect, what is useful here is the ability of a WordNet-based method to provide with results that could not be obtained with other techniques because of the lack of explicit semantic information and hierarchical structure in the original vocabularies.

Finally, as mentioned in the introduction, there are other motivating use cases that we plan to experiment with. Especially interesting is the way a mapping with WordNet can enhance the existing access to document collections of the Dutch Cultural Heritage Institutes by providing with query refinement services and browsing possibilities.

Acknowledgements

This research was carried out in the context of the CATCH projects CHOICE and STITCH, funded by NWO, the Dutch organization for scientific research.

References

Aleksovski Z. 2006. *Matching Unstructured Vocabularies using a Background Ontology*. 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2006).

van Assem M., Gangemi A. and Schreiber G. 2006. *RDF/OWL Representation ofWordNet*. W3C Working Draft, 19 June 2006. http://www.w3.org/TR/wordnet-rdf/

Budanitsky A. and Hirst G. 2006. *EvaluatingWordNet-based Measures of Lexical Semantic Relatedness*, volume 32(1). Computational Linguistics, 13–47.

Castano S., Ferrara A. and Montanelli S. 2005. *Matching Ontologies in Open Networked Systems: Techniques and Applications*, volume 5. Journal on Data Semantics (JoDS).

Euzenat J., coordinator. 2004. *State of the art on ontology alignment*. KnowledgeWeb Deliverable 2.2.3.

Fellbaum C. 1998. *WordNet An Electronic Lexical Database*. MIT Press.

van Gendt M., Isaac A., van der Meij L. and Schlobach S. 2006. *Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study*. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), 426–437.

Giunchiglia F., Shvaiko P., and Yatskevich M. 2005. *Semantic Schema Matching*. 13th International Conference on Cooperative Information Systems (CoopIS 2005).

Hollink L. 2006. *Semantic annotation for retrieval of visual resources*. PHD Thesis, Vrije Universiteit Amsterdam.

Ibekwe-SanJuan F. 2005. *Clustering semantic relations for constructing and maintaining knowledge organization tools.* volume 62 (2). Journal of Documentation, Emerald Publishing Group, 229–250.

Khan L. R. and Hovy E. 1997. *Improving the Precision of Lexicon-to-Ontology Alignment Algorithm*. AMTA/SIG-IL First Workshop on Interlinguas, San Diego, CA, October 28.

Knight K. and Luk S. 1994. *Building a Large-Scale Knowledge Base for Machine Translation*. In Proceedings of the AAAI-94 Conference.

Voorhees E. 1994. *Query expansion using lexical-semantic relations*. 17th International ACM/SIGIR Conference on Research and Development in Information Retrieval, 61–69.

Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation

Idan Szpektor, Ido Dagan
Dept. of Computer Science
Bar Ilan University
szpekti@cs.biu.ac.il

Alon Lavie
Language Technologies Inst.
Carnegie Mellon University
alavie+@cs.cmu.edu

Danny Shacham, Shuly Wintner
Dept. of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

Abstract

We describe a system which enhances the experience of museum visits by providing users with language-technology-based information retrieval capabilities. The system consists of a cross-lingual search engine, augmented by state of the art semantic expansion technology, specifically designed for the domain of the museum (history and archaeology of Israel). We discuss the technology incorporated in the system, its adaptation to the specific domain and its contribution to cultural heritage appreciation.

1 Introduction

Museum visits are enriching experiences: they provide stimulation to the senses, and through them to the mind. But the experience does not have to end when the visit ends: further exploration of the artifacts and their influence on the visitor is possible *after* the visit, either on location or elsewhere. One common means of exploration is Information Retrieval (IR) via a Search Engine. For example, a museum could implement a search engine over a collection of documents relating to the topics exhibited in the museum.

However, such document collections are usually much smaller than general collections, in particular the World Wide Web. Consequently, phenomena inherent to natural languages may severely hamper the performance of human language technology when applied to small collections. One such phenomenon is the semantic *variability* of natural languages, the ability to express a specific meaning in many different ways. For example, the expression *"Archaeologists found a new tomb"* can be expressed also by *"Archaeologists discovered a tomb"* or *"A sarcophagus was dug up by Egyptian Researchers"*. On top of monolingual variability, the same information can also be expressed in different languages. Ignoring natural language variability may result in lower recall of relevant documents for a given query, especially in small document collections.

This paper describes a system that attempts to cope with semantic variability through the use of state of the art human language technology. The system provides both semantic expansion and cross lingual IR (and presentation of information) in the domain of archaeology and history of Israel. It was specifically developed for the Hecht Museum in Haifa, Israel, which contains a small but unique collection of artifacts in this domain. The system provides different users with different capabilities, bridging over language divides; it addresses semantic variation in novel ways; and it thereby complements the visit to the museum with long-lasting instillation of information.

The main component of the system is a domain-specific search engine that enables users to specify queries and retrieve information pertaining to the domain of the museum. The engine is enriched by linguistic capabilities which embody an array of means for addressing semantic variation. Queries are expanded using two main techniques: semantic expansion based on textual entailment; and cross-lingual expansion based on translation of Hebrew queries to English and vice versa. Retrieved documents are presented as links with associated snippets; the system also translates snippets from Hebrew to English.

The main contribution of this work is, of course, the system itself, which was recently demonstrated

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 65–72,
Prague, 28 June 2007.

successfully at the museum and which we believe could be useful to a variety of museum visitor types, from children to experts. For example, the system provides Hebrew speakers access to English documents pertaining to the domain of the museum, and vice versa, thereby expanding the availability of multilingual material to museum visitors. More generally, it is an instance of adaptation of state of the art human language technology to the domain of cultural heritage appreciation, demonstrating how general resources and tools are adapted to a specific domain, thereby improving their accuracy and usability. Finally, it provides a test-bed for evaluating the contribution of language technology in general, as well as specific components and resources, to a large-scale natural language processing system.

2 Background and Motivation

Internet search is hampered by the complexity of natural languages. The two main characteristics of this complexity are *ambiguity* and *variability*: the former refers to the fact that a given text can be interpreted in more than one way; the latter indicates that the same meaning can be linguistically expressed in several ways. The two phenomena make simple search techniques too weak for unsophisticated users, as existing search engines perform only direct keyword matching, with very limited linguistic processing of the texts they retrieve.

Specifically, IR systems that do not address the variability in languages may suffer from lower recall, especially in restricted domains and small document locations. We next describe two prominent types of variability that we think should be addressed in IR systems.

2.1 Textual Entailment and Entailment Rules

In many NLP applications, such as Question Answering (QA), Information Extraction (IE) and Information Retrieval (IR), it is crucial to recognize that a specific target meaning can be inferred from different text variants. For example, a QA system needs to induce that *"Mendelssohn wrote incidental music"* can be inferred from *"Mendelssohn composed incidental music"* in order to answer the question *"Who wrote incidental music?"*. This type of reasoning has been identified as a core semantic inference task by the generic *textual entailment* framework (Dagan et al., 2006; Bar-Haim et al., 2006).

The typical way to address variability in IR is to use lexical query expansion (Lytinen et al., 2000; Zukerman and Raskutti, 2002). However, there are variability patterns that cannot be described using just constant phrase to phrase entailment. Another important type of knowledge representation is *entailment rules* and paraphrases. An entailment rule is a directional relation between two *templates*, text patterns with variables, e.g., 'X compose Y $\rightarrow$ X write Y'. The left hand side is assumed to entail the right hand side in certain contexts, under the same variable instantiation. Paraphrases can be viewed as bidirectional entailment rules. Such rules capture basic inferences in the language, and are used as building blocks for more complex entailment inference. For example, given the above entailment rule, a QA system can identify the answer *"Mendelssohn"* in the above example. This need sparked intensive research on automatic acquisition of paraphrase and entailment rules.

Although knowledge-bases of entailment-rules and paraphrases learned by acquisition algorithms were used in other NLP applications, such as QA (Lin and Pantel, 2001; Ravichandran and Hovy, 2002) and IE (Sudo et al., 2003; Romano et al., 2006), to the best of our knowledge the output of such algorithms was never applied to IR before.

2.2 Cross Lingual Information Retrieval

The difficulties caused by variability are amplified when the user is not a native speaker of the language in which the retrieved texts are written. For example, while most Israelis can read English documents, fewer are comfortable with the specification of English queries. In a museum setting, some visitors may be able to read Hebrew documents but still be relatively poor at searching for them. Other visitors may be unable to read Hebrew texts, but still benefit from non-textual information that are contained in Hebrew documents (e.g., pictures, maps, audio and video files, external links, etc.)

This problem is addressed by the paradigm of Cross-Lingual Information Retrieval (CLIR). This paradigm has become a very active research area in recent years, addressing the needs of multilingual and non-English speaking communities, such as the

European Union, East-Asian nations and Spanish speaking communities in the US (Hull and Grefenstette, 1996; Ballesteros and Croft, 1997; Carbonell et al., 1997). The common approach for CLIR is to translate a query in a source language to another target language and then issue the translated query to retrieve target language documents. As explained above, CLIR research has to address various generic problems caused by the variability and ambiguity of natural languages, as well as specific problems related to the particular languages being addressed.

3 Coping with Semantic Variability in IR

We describe a search engine that is capable of performing: (a) semantic English information retrieval; and (b) cross-lingual (Hebrew-English and English-Hebrew) information retrieval, allowing users to pose queries in either of the two languages and retrieve documents in both. This is achieved by two sub-processes of the search engine: first, the engine performs shallow semantic linguistic inference and supports the retrieval of documents which contain phrases that imply the meaning of the translated query, even when no exact match of the translated keywords is found. This is enabled by automatic acquisition of semantic variability patterns that are frequent in the language, which extend traditional lexical query expansion techniques. Second, the engine translates the original or expanded query to the target language, based on several linguistic processes and a machine readable bilingual dictionary. The result is a semantic expansion of a given query to a variety of alternative wordings in which an answer to this query may be expressed in the target language of the retrieved documents.

These enhancements are facilitated via a specification of the domain. As our system is specifically designed to work in the domain of the history and archaeology, we could focus our attention on resources and tools that are dedicated to this domain. Thus, for example, lexicons and dictionaries, whose preparation is always costly and time consuming, were developed with the specific domain in mind; and textual entailment and paraphrase patterns were extracted for the specific domain. While the resulting system is focused on visiting the Hecht Museum, the methodology which we used and discuss here can be adapted to other areas of cultural heritage, as well as to other narrow domains, in the same way.

3.1 Setting Up a Basic Retrieval Application

We created a basic retrieval system in two steps: first, we collected relevant documents; then, we implemented a search engine over the collected documents.

In order to construct a local corpus, an archaeology expert searched the Web for relevant sites and pages. We then downloaded all the documents linked from those pages using a crawler. The expert looked for documents in both English and Hebrew. In total, we collected a non-comparable bilingual corpus for Archaeology containing several thousand documents in English and Hebrew.

We implemented our enhanced retrieval modules on top of the basic Jakarta Lucene indexing and search engine[1]. All documents were indexed using Lucene, but instead of inflected words, we indexed the lemma of each word (see detailed description of our Hebrew lemmatization in Section 3.3). In order to match the indexed terms, query terms (either Hebrew or English) were also lemmatized before the index was searched, in a manner similar to lemmatizing the documents.

3.2 Query Expansion Using Entailment Rules

As described in Section 2.1, entailment rules had not been used as a knowledge resource for expanding IR queries, prior to our work. In this paper we use this resource instead of the typical lexical expansion in order to test its benefit. Most entailment rules capture relations between different predicates. We thus focus on documents retrieved for queries that contain a predicate over one or two entities, which we term here *Relational IR*. We would like to retrieve only documents that describe an occurrence of that predicate, but possibly in words different than the ones used in the query. In this section we describe in detail how we learn entailment rules and how we apply them in query expansion.

Automatically Learning Entailment Rules from the Web Many algorithms for automatically learning paraphrases and entailment rules have been explored in recent years (Lin and Pantel, 2001;

[1] http://jakarta.apache.org/lucene/docs/index.html

Ravichandran and Hovy, 2002; Shinyama et al., 2002; Barzilay and Lee, 2003; Sudo et al., 2003; Szpektor et al., 2004; Satoshi, 2005). In this paper we use TEASE (Szpektor et al., 2004), a state-of-the-art unsupervised acquisition algorithm for lexical-syntactic entailment rules.

TEASE acquires entailment relations for a given input template from the Web. It first retrieves from the Web sentences that match the input template. From these sentences it extracts the variable instantiations, termed *anchor-sets*, which are identified as being characteristic for the input template based on statistical criteria.

Next, TEASE retrieves from the Web sentences that contain the extracted anchor-sets. The retrieved sentences are parsed and the anchors found in each sentence are replaced with their corresponding variables. Finally, from this retrieved corpus of parsed sentences, templates that are assumed to entail or be entailed by the input template are learned. The learned templates are ranked by the number of occurrences they were learned from.

Entailment Rules for Domain Specific Query Expansion Our goal is to use the knowledge-base of entailment rules learned by TEASE in order to perform query expansion. The two subtasks that arise are: (a) acquiring an appropriate knowledge-base of rules; and (b) expanding a query given such a knowledge-base.

TEASE learns entailment rules for a given input template. As our document collection is domain specific, a list of such relevant input templates can be prepared. In our case, we used an archaeology expert to generate a list of verbs and verb phrases that relate to archaeology, such as: *'excavate'*, *'invade'*, *'build'*, *'reconstruct'*, *'grow'* and *'be located in'*. We then executed TEASE on each of the templates representing these verbs in order to learn from the Web rules in which the input templates participate. An example for such rules is presented in Table 1. We learned approximately 3900 rules for 80 input templates.

Since TEASE learns lexical-syntactic rules, we need a syntactic representation of the query. We parse each query using the Minipar dependency parser (Lin, 1998). We next try to match the left hand side template of every rule in the learned knowledge-base. Since TEASE does not identify the direction of the relation learned between two templates, we try both directional rules that are induced from a learned relation. Whenever a match is found, a new query is generated, in which the constant terms of the matched left hand side template are replaced with the constant terms of the right hand side template. For example, given the query *"excavations of Jerusalem by archaeologists"* and a learned rule *'excavation of* Y *by* $X \rightarrow X$ *dig in* Y*'*, a new query is generated, containing the terms *'archaeologists dig in Jerusalem'*. Finally, we retrieve all the documents that contain all the terms of at least one of the expanded queries (including the original query). The basic search engine provides a score for each document. We re-score each document as the sum of scores it obtained from the different queries that it matched. Figure 1 shows an example of our query expansion, where the first retrieved documents do not contain the words used to describe the predicate in the query, but other ways to describe it.

All the templates learned by TEASE contain two variables, and thus the rules that are learned can only be applied to queries that contain predicates over two terms. In order to broaden the coverage of the learned rules, we automatically generate also all the partial templates of a learned template. These are templates that contain just one of variables in the original template. We then generate rules between these partial templates that correspond to the original rules. With partial templates/rules, expansion for the query in Figure 1 becomes possible.

3.3 Cross-lingual IR

Until very recently, linguistic resources for Hebrew were few and far between (Wintner, 2004). The last few years, however, have seen a proliferation of resources and tools for this language. In this work we utilize a relatively large-scale lexicon of over 22,000 entries (Itai et al., 2006); a finite-state based morphological analyzer of Hebrew that is directly linked to the lexicon (Yona and Wintner, 2007); a medium-size bilingual dictionary of some 24,000 word pairs; and a rudimentary Hebrew to English machine translation system (Lavie et al., 2004). All these resources had to be adapted to the domain of the Hecht museum.

Cross-lingual language technology is utilized in

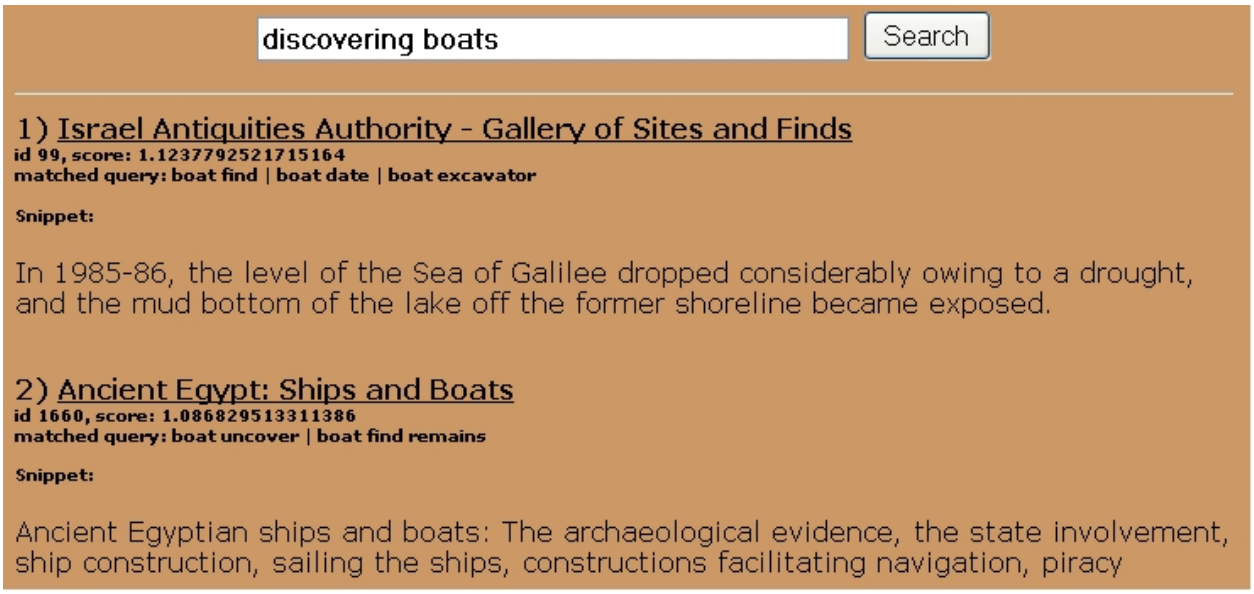


Figure 1: Semantic expansion example. Note that the expanded queries that were generated in the first two retrieved texts (listed under '**matched query**') do not contain the original query.

three different components of the system: Hebrew documents are morphologically processed to provide better indexing; query terms in English are translated to Hebrew and vice versa; and Hebrew snippets are translated to English. We discuss each of these components in this section.

Linguistically-aware indexing The correct level of indexing for morphologically-rich language has been a matter of some debate in the information retrieval literature. When Arabic is concerned, Darwish and Oard (2002) conclude that "Character n-grams or lightly stemmed words were found to typically yield near-optimal retrieval effectiveness". Since Hebrew is even more morphologically (and orthographically) ambiguous than Arabic, and especially in light of the various prefix particles which can be attached to Hebrew words, we opted for full morphological analysis of Hebrew documents before they are indexed, followed by indexing on the lexeme.

We use the HAMSAH morphological analyzer (Yona and Wintner, 2007), which was recently rewritten in Java and is therefore more portable and efficient (Wintner, 2007). We processed the entire domain specific corpus described above and used the resulting lexemes to index documents. This preprocessing brought to the foreground several omissions of the analyzer, mostly due to domain-specific terms missing in the lexicon. We selected the one thousand most frequent words with no morphological analysis and added their lexemes to the lexicon. While we do not have quantitative evaluation metrics, the coverage of the system improved in a very evident way.

Query translation When users submit a query in one language they are provided with the option to request a translation of the query to the other language, thereby retrieving documents in the other language. The motivation behind this capability is that users who may be able to read documents in a language may find the specification of queries in that language too challenging; also, retrieving documents in a foreign language may be useful due to the non-textual information in the retrieved documents, especially in a museum environment.

In order to support cross-lingual query specification we capitalized on a medium-size bilingual dictionary that was already used for Hebrew to English machine translation. Since the coverage of the dictionary was rather limited, and many domain-specific items were missing, we chose the one thousand most frequent lexemes which had no transla-

Input Template	Learned Template
X excavate Y	X discover Y, X find Y, X uncover Y, X examine Y, X unearth Y, X explore Y
X construct Y	X build Y, X develop Y, X create Y, X establish Y
X contribute to Y	X cause Y, X linked to Y, X involve in Y
date X to Y	X built in Y, X began in Y, X go back to Y
X cover Y	X bury Y, X provide coverage for Y
X invade Y	X occupy Y, X attack Y, X raid Y, X move into Y
X restore Y	X protect Y, X preserve Y, X save Y, X conserve Y

Table 1: Examples for correct templates that were learned by TEASE for input templates.

tions and translated them manually, augmenting the lexicon with missing Hebrew lexemes where necessary and expanding the bilingual dictionary to cover this domain.

In order to translate query terms we use the Hebrew English dictionary also as an English-Hebrew dictionary. While this is known to be sub-optimal, our current results support such an adaptation in lieu of dedicated directional bilingual dictionaries.

Translating a query from one language to another may introduce ambiguity where none exists. For example, the query term *spinh* 'vessel' is unambiguous in Hebrew, but once translated into English will result in retrieving documents on both senses of the English word. Usually, this problem is overcome since users tend to specify multi-term queries, and the terms disambiguate each other. However, a more systematic solution can be offered since we have access to semantic expansion capabilities (in a single language). That is, expanding the query in the source language will result in more query terms which, when translated, are more likely to disambiguate the context. We leave such an extension for future work.

Snippet translation When Hebrew documents are retrieved, we augment the (Hebrew) snippet which the system produces by an English translation. We use an extended, improved version of a rudimentary Hebrew to English MT system developed by Lavie et al. (2004). Extensions include an improved morphological analysis of the input, an extended bilingual dictionary and a revised set of transfer rules, as well as a more modern transfer engine and a much larger language model for generating the target (English) sentences.

The MT system is transfer based: it performs linguistic pre-processing of the source language (in our case, morphological analysis) and post-processing of the target (generation of English word forms), and uses a small set of transfer rules to translate local structures from the source to the target and create translation hypotheses, which are stored in a lattice. A statistical language model is used to decode the lattice and select the best hypotheses.

The benefit of this architecture is that domain specific adaptation of the system is relatively easy, and does not require a domain specific parallel corpus (which we do not have). The system has access to our domain-specific lexicon and bilingual dictionary, and we even refined some transfer rules due to peculiarities of the domain. One advantage of the transfer-based approach is that it enables us to treat out-of-lexicon items in a unique way. We consider such items proper names, and transfer rules process them as such. As an example, Figure 2 depicts the translation of a Hebrew snippet meaning *A jar from the early bronze period with seashells from the Nile*. The word *nilws* 'Nile' is missing from the lexicon, but this does not prevent the system from producing a legible translation, using the transliterated form where an English equivalent is unavailable.

4 Conclusions

We described a system for cross-lingual and semantically-enhanced retrieval of information in the cultural heritage domain, obtained by adapting existing state-of-the-art tools and resources to the domain. The system enhances the experience of museum visits, using language technology as a vehicle for long-lasting instillation of information. Due to the novelty of this application and the dearth of available multilingual annotated resources in this domain, we are unable to provide a robust, quan-

Figure 2: Translation example

Query	Without Expansion		With Expansion	
	Relevant in Top 10	Total Retrieved	Relevant in Top 10	Total Retrieved
discovering boats	2	2	5	86
growing vineyards	0	0	6	8
Persian invasions	5	5	8	22
excavations of the Byzantine period	10	37	10	100
restoring mosaics	0	0	3	69

Table 2: Analysis of the number of relevant documents out of the top 10 and the total number of retrieved documents (up to 100) for a sample of queries.

titative evaluation of the approach. A preliminary analysis of a sample of queries is presented in Table 2. It illustrates the potential of expansion for document collections of narrow domain. In what follows we provide some qualitative impressions.

We observed that the system was able to learn many expansion rules that cannot be induced from manually constructed lexical resources, such as thesauri or WordNet (Fellbaum, 1998). This is especially true for rules that are specific for a narrow domain, e.g. 'X restore $Y \rightarrow X$ preserve Y'. Furthermore, the system learned lexical syntactic rules that cannot be expressed by a mere lexical substitution, but include also a syntactic transformation. For example, 'date X to $Y \leftrightarrow X$ go back to Y'.

In addition, since rules are acquired by searching the Web, they are not necessarily restricted to learning from the target domain, but can be learned from similar terminology in other domains. For example, the rule 'X *discover* $Y \leftrightarrow X$ *find* Y' was learned from contexts such as $\{X=$'*astronomers*';$Y=$'*new planets*'$\}$ and $\{X=$'*zoologists*';$Y=$'*new species*'$\}$.

The quality of the rules that were automatically acquired is mediocre. We found that although many rules were useful for expansion, they had to be manually filtered in order to retain only rules that achieved high precision.

Finally, we note that applying semantic query expansion (using entailment rules), followed by English to Hebrew query translation, results in query expansion for Hebrew using techniques that were so far applicable only to resource-rich languages, such as English.

Acknowledgements

This research was supported by the Israel Internet Association; by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06 and grant No. 1095/05); by the Caesarea Rothschild Institute for Interdisciplinary Application of Computer Science at the University of Haifa; by the ITC-irst/University of Haifa collaboration; and by the US National Science Foundation (grants IIS-0121631, IIS-0534217, and the Office of International Science and Engineering).

We wish to thank the Hebrew Knowledge Center at the Technion for providing resources for Hebrew. We are grateful to Oliviero Stock, Martin Golumbic, Alon Itai, Dalia Bojan, Erik Peterson, Nurit Melnik, Yaniv Eytani and Noam Ordan for their help and support.

References

Lisa Ballesteros and W. Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Second PASCAL Challenge Workshop for Recognizing Textual Entailment*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*.

Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. 1997. Translingual information retrieval: A comparative evaluation. In *IJCAI (1)*, pages 708–715.

Ido Dagan, Oren Glickman, and Bernardo. Magnini. 2006. The pascal recognising textual entailment challenge. In *Lecture Notes in Computer Science, Volume 3944*, volume 3944, pages 177–190.

Kareem Darwish and Douglas W. Oard. 2002. Term selection for searching printed Arabic. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 261–268, New York, NY, USA. ACM Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

D. A. Hull and G. Grefenstette. 1996. Querying across languages. a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th ACM SIGIR Conference*, pages 49–57.

Alon Itai, Shuly Wintner, and Shlomo Yona. 2006. A computational lexicon of contemporary Hebrew. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC-2006)*.

Alon Lavie, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. In *Natural Language Engineering*, volume 7(4), pages 343–360.

Dekang Lin. 1998. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on Evaluation of Parsing Systems at LREC*.

S. Lytinen, N. Tomuro, and T. Repede. 2000. The use of wordnet sense tagging in faqfinder. In *Proceedings of the AAAI00 Workshop on AI and Web Search*.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL*.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL*.

Sekine Satoshi. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of IWP*.

Yusuke Shinyama, Satoshi Sekine, Sudo Kiyoshi, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*.

Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. 2003. An improved extraction pattern representation model for automatic ie pattern acquisition. In *Proceedings of ACL*.

Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP*.

Shuly Wintner. 2004. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 21(2):113–138.

Shuly Wintner. 2007. Finite-state technology as a programming environment. In Alexander Gelbukh, editor, *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007)*, volume 4394 of *Lecture Notes in Computer Science*, pages 97–106, Berlin and Heidelberg, February. Springer.

Shlomo Yona and Shuly Wintner. 2007. A finite-state morphological grammar of Hebrew. *Natural Language Engineering*. To appear.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *Proceedings of ACL*.

Deriving a Domain Specific Test Collection from a Query Log

Avi Arampatzis[1] **Jaap Kamps**[1,2] **Marijn Koolen**[1] **Nir Nussbaum**[2]
[1] Archives and Information Science, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

Abstract

Cultural heritage, and other special domains, pose a particular problem for information retrieval: evaluation requires a dedicated test collection that takes the particular documents and information requests into account, but building such a test collection requires substantial human effort. This paper investigates methods of generating a document retrieval test collection from a search engine's transaction log, based on submitted queries and user-click data. We test our methods on a museum's search log file, and compare the quality of the generated test collections against a collection with manually generated and judged known-item topics. Our main findings are the following. First, the test collection derived from a transaction log corresponds well to the actual search experience of real users. Second, the ranking of systems based on the derived judgments corresponds well to the ranking based on the manual topics. Third, deriving pseudo-relevance judgments from a transaction log file is an attractive option in domains where dedicated test collections are not readily available.

1 Introduction

Cultural heritage, and other special domains, pose a particular problem for information retrieval. Progress in information retrieval depends heavily on the availability of suitable test collections consisting of a set of documents; a set of search topics; and (human) relevance judgments. Standard benchmarks, such as those developed at TREC (2007), have been developed using newspaper and newswire data. Whilst these test collections are immensely useful to evaluate generic properties of retrieval systems, such as fundamental ranking principles, they do not capture the specific context of particular domains (Ingwersen and Järvelin, 2005). To take cultural heritage as an example, the documents are cultural heritage descriptions which are different in character from newspaper articles, and also the search requests and relevance judgments about art are more subjective than factual queries about news (Koolen et al., 2007). As a result, special domains like cultural heritage require a dedicated test collection that takes the particular documents and information requests into account, but building such a test collection requires substantial human effort.

We opt for a different approach. Search engines commonly store the actions of users in transaction logs, which allow an unobtrusive way of studying user behaviour. Logs contain valuable information such as what searchers are looking for, what results they find interesting enough to click on, etc. In this paper, we investigate methods of extracting queries and user-clicks (on the search result items) from transaction logs in order to create a quality test collection for Document Retrieval.

A quality test collection for Document Retrieval is traditionally considered as a set of queries on a document collection with *complete* and *reliable* relevance judgements. Complete in the sense that all documents are judged for relevance against all queries, and reliable in the sense that judgements are sta-

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 73–80,
Prague, 28 June 2007.

ble across a majority of human assessors. Nevertheless, considering the fact that a test collection is used *"as a mechanism for comparing system performance"* (Voorhees, 2002), the requirements for completeness and reliability may be relaxed somewhat.

The Text REtrial Conference (TREC) has traditionally used incomplete judgements for comparing system effectiveness via the "pooling" method (Jones and van Rijsbergen, 1975), and it is also well-known that human assessor agreement is relatively low (Voorhees and Harman, 2005). Consequently, test collections which *preserve* the effectiveness ranking of several systems can be considered of equivalent quality in the context of comparing system effectiveness. In order to evaluate the quality of test collections extracted in various ways from a transaction log, it would be sufficient to compare their ability to rank several retrieval systems against a reference system ranking produced by an already known good test collection not produced from the log.

One can think of several ways of extracting queries and clicks from a transaction log and turning them into a set of queries with relevance judgments. A simple (and naive) way would be to treat every query typed by a user as a topic, and every result that the user clicked on as a positive relevance judgment. However, such an approach may not lead to a good test set. Previous research on user click behaviour has shown that clicks on search engine results do not directly correspond to explicit, absolute relevance judgments, but can be considered as *relative* relevance judgments (Joachims et al., 2005), i.e., if a user skips result *a* and clicks on result *b*, than the user preference reflects $rank(b) > rank(a)$. Moreover, the occurrence frequencies of queries and the numbers of retrieved items vary significantly across queries which may lead to wide variation in effectiveness.

The challenge we take up has several dimensions which can be summarized in the following questions:

- How can we derive topics and pseudo-relevance judgments from a transaction log file, and how does this impact the quality of the generated test collection?
- How does system effectiveness on the automatically generated test collection compare to the effectiveness on a set of manually constructed known-item topics?

If automatic methods of building test collections are indeed feasible, this opens up a whole new dimension of possibilities for Information Retrieval evaluation: there is an enormous lengths of transaction logs generated daily at numerous web-sites and at on-line search engines.

The rest of this paper is organized as follows. Next, in Section 2 we discuss transaction logs in general, and the specific transaction log from a museum that we'll use in the case study of this paper. Section 3 details how we have extracted topics and pseudo-relevance judgments from a museum's log file, and their evaluation. Then, in Section 4, we evaluate the merits of the derived test collection in comparison to human generated and judged topics. We end with Section 5 in which we summarize our findings.

2 Transaction Logs

2.1 Previous Work

There has been substantial interest in using click-through data from transaction logs as a form of implicit feedback (Dumais et al., 2003). A range of implicit feedback techniques have been used for query expansion and user profiling in information retrieval tasks (Oard and Kim, 2001; Kelly and Teevan, 2003). Joachims et al. (2005, p.160) conclude that "the implicit feedback generated from clicks shows reasonable agreement with the explicit judgments of the pages".

Transaction logs have been analysed to study user search behaviour in Web search engines (Chau et al., 2005) and digital libraries (Jones et al., 2000), amongst others (Jansen, 2006). In Chau et al. (2005), user behaviour is studied using the transaction log of a website's search engine and is compared to that of general purpose search engines. They find that the number of query terms used for website search engines is comparable to queries submitted to general purpose search engines, but the search topics and terms are different.

In this paper, we go one step further and try to exploit the user behaviour implicit in the data to con-

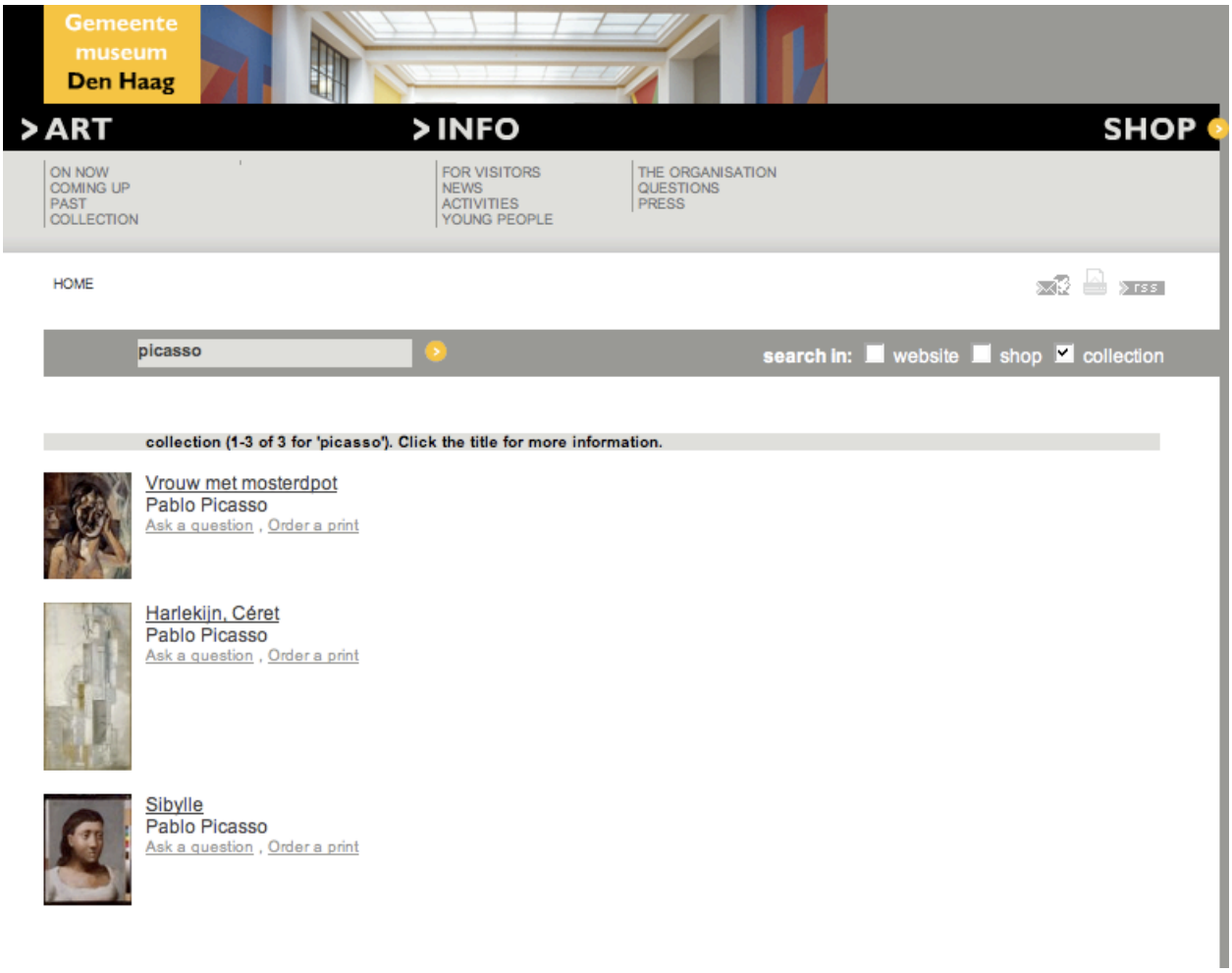


Figure 1: The search engine of the Gemeentemuseum's website.

struct a test set with real user needs, queries and judgments.

2.2 A Website's Search Engine

The website of the *Haags Gemeentemuseum*[1] in the Hague, the Netherlands, offers a search engine for three different parts of the *Gemeentemuseum*, the website content, the on-line shop, and the highlights of the museum's object collection (see Figure 2.1). The searchable on-line collection consists of 1,127 objects, the highlights of the museum, from a total database of 116,493 museum objects. The metadata of these objects are stored in a legacy system, and queries are matched against the title and creator fields (Koolen et al., 2007). The descriptions contain many more fields, however. The objects database treats the query as a Boolean AND query, and returns a warning if there is no object description containing all terms in one field. Although the database allows a drop-back to the individual terms, the website search engine retains a strict Boolean AND query and returns an empty result list.

The transaction log contains the transactions from the server side. The website uses a Java script to interact with the search engine. The query itself is not stored in the transaction log. If a user clicks on a result that leads to another web page in the domain, or to an item in the shop, this click is registered in the transaction, but the actual query is not. If a user clicks on a result from the object collection however, the database query is stored in the transaction log, from which we can extract the actual user query, and the object that user wants to see.

[1] http://www.gemeentemuseum.nl

This has an effect on the queries found in the log file. Queries containing both title and creator names often lead to an empty result list, as there is no single field containing both creator and title terms. The database looks for all the terms in one field at a time, and will not match with any object. With an empty result list, users cannot click on an object and hence, the query is not logged. Another effect is that all the results that users can click on have all the query terms in either the title or creator field. Although end users sometimes express their information needs in terms different from the terms chosen by indexers, i.e. the curators in the museum (Markkula and Sormunen, 2000), this discrepancy cannot be observed in the log-file data.

This may lead to the concern that the topics that can be extracted from the transaction log are "easy" topics, since the relevant descriptions necessarily contain all the query terms. It is unclear whether this affects the extracted topic set significantly, since we will look only at the relative ranking of systems over a set of queries. We will compare the ability to rank systems of our automatically generated topic sets with the system ranking ability of a manual topic set. If the extracted topic sets preserve the system ranking of the manual topic set, the bias in the topic sets towards "easy" topics has no negative influence on the quality of the topic sets.

3 Experiments and setup

We have obtained the log files covering a period of one and a half years, between September 14, 2005 and February 26, 2007.

From the transaction log, we extracted the queries and the object identifiers from the database query, and turned them into Qrels, i.e., the object is relevant for the query.

We use the following terminology:

- **User**: the client side of the transaction, identified by ip-address.
- **Transaction**: any exchange between client (user) and server (system), corresponding to a line in the transaction log.
- **Session**: A sequence of transactions by the same user, where the maximum interval between transaction n and $n + 1$ is 1 hour. More than 1 hour of inactivity signals a session boundary.
- **Query**: the string typed by the user as it appears in the transaction log.
- **Result**: the identifier of the museum object, used to retrieve the object data from the object database.

Topic set	# Topics	Query length average	Query length median	Avg. # rel. docs
Raw	7,531	1.18	1	2.38
Union	1,183	1.38	1	3.86
Intersection	974	1.42	1	1.41
Manual	150	2.38	2	1.00

Table 1: Statistics on the extracted topic sets.

3.1 Extraction methods

We used 3 extraction methods to construct a test set:

1. **Raw queries**: each query appearing in the log is used, i.e. the bag of queries. Here, a topic consist of a query and the corresponding clicked results from one session. If the same user types the same query in another session, this is treated as a new topic.
2. **Unique union**: All unique queries are used, i.e. the set of queries. All the results clicked by all users typing the same query are considered relevant documents.
3. **Unique intersection**: All unique queries are used, i.e. the set of queries. The intersection of the results clicked by all users typing the same query are considered relevant documents. Thus, a result is relevant only if all users who typed the query, clicked on that result.

Table 1 shows statistics on the resulting topic sets. In calculating these numbers, stop words were removed from the queries. As most queries are in Dutch, we used the standard Snowball stopword list for Dutch (Snowball, 2007). The queries are very short on average. For the Raw, Union and Intersection topic sets, the queries with 1 term form 84%, 70% and 68% of the query sets respectively. There

are 1,183 unique queries, and on average, 3.86 results are clicked by at least one user. Understandably, the Intersection set has less topics than the Union set, as there are queries with no single result clicked on by all users. Also, the average number of relevant documents per topic is lower for the intersection set.

We created 150 Known-Item topics by hand and used this test set, referred to as KI-topics, on the same collection and include the results as a comparison with the new test sets. Table 1 shows the statistics of these human generated topics in the last row. These search request have more verbose topic statements with a median length of 2, compared to a median length of 1 for the query log topics. Also the number of relevant documents differs considerably, with a unique relevant page for the human known-item topics, and several "clicked" pages per query for the transaction log.

3.2 Retrieval system

To see if our test sets lead to a stable system ranking, we need a number of retrieval systems to compare their ranking on the different test collections. To get a number of different systems, we simply use a standard retrieval model with different parameter settings to create different runs.

We use a standard language model (Hiemstra, 2001). Our system is an extension to Lucene (ILPS, 2005) and uses Jelinek-Mercer smoothing, controlled by the parameter λ, and a length prior, controlled by the parameter β, i.e., for a collection D, document d and query q:

$$P(d|q) = P(d) \cdot \prod_{t \in q} \left((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d) \right), \quad (1)$$

where

$$P(t|d) = \frac{tf_{t,d}}{|d|} \quad (2)$$

$$P(t|D) = \frac{\text{doc\_freq}(t, D)}{\sum_{t' \in D} \text{doc\_freq}(t', D)} \quad (3)$$

$$P(d) = \frac{|d|}{\sum_{d' \in D} |d|} \quad (4)$$

We assign a prior probability to an document d relative to its length in the following manner:

$$P(d) = \frac{|d|^{\beta}}{\sum_{d} |d|^{\beta}}, \quad (5)$$

System	λ	β
A	0.10	0
B	0.50	0
C	0.90	0
D	0.10	1
E	0.50	1
F	0.90	1
G	0.10	2
H	0.50	2
I	0.90	2

Table 2: Parameter settings for the different systems.

where $|d|$ is the length of a document d. The β parameter introduces a length bias which is proportional to the document length with $\beta = 1$ (the default setting). For more details on language models and smoothing, see (Hiemstra, 2001). For details on the effect of the length parameter, see (Kamps et al., 2004).

3.3 Experimental Set-up

In our experiments we will emulate a set of different retrieval systems by using arbitrary parameter settings for smoothing (λ) and length prior (β). This will result in a range of different rankings of documents, and we can compare their retrieval effectiveness on our various topic sets. In this way, we can compare the system ranking of the automatically generated topic sets with the system ranking of a manually crafted topic set.

We made 9 different runs with each topic set, using 3 different values (0.10, 0.50 and 0.90) for the smoothing parameter λ, corresponding to heavy, average and little smoothing respectively, and 3 different values (0, 1 and 2) for the length prior β corresponding to no length normalization and length normalization proportional to the document length.

To measure the correlation of the system rankings resulting from the different topic sets, we look at Kendall's tau coefficient.

4 Results

Table 3 shows the detailed results for all runs over all topics sets. As noted above, we will focus on the relative system rankings over topic sets. We limit our analysis to the performance in terms of mean-

Topics	# Topics	MRR	Success@10
Raw topics $\beta = 0, \lambda = 0.10$	7,527	0.5974	0.8023
Raw topics $\beta = 0, \lambda = 0.50$	7,527	0.5970	0.8030
Raw topics $\beta = 0, \lambda = 0.90$	7,527	0.5970	0.8031
Raw topics $\beta = 1, \lambda = 0.10$	7,527	0.5673	0.7506
Raw topics $\beta = 1, \lambda = 0.50$	7,527	0.5765	0.7574
Raw topics $\beta = 1, \lambda = 0.90$	7,527	0.5767	0.7574
Raw topics $\beta = 2, \lambda = 0.10$	7,527	0.5531	0.7427
Raw topics $\beta = 2, \lambda = 0.50$	7,527	0.5618	0.7468
Raw topics $\beta = 2, \lambda = 0.90$	7,527	0.5644	0.7474
Union $\beta = 0, \lambda = 0.10$	1,183	0.6908	0.8191
Union $\beta = 0, \lambda = 0.50$	1,183	0.6925	0.8233
Union $\beta = 0, \lambda = 0.90$	1,183	0.6927	0.8233
Union $\beta = 1, \lambda = 0.10$	1,183	0.6622	0.7887
Union $\beta = 1, \lambda = 0.50$	1,183	0.6772	0.8005
Union $\beta = 1, \lambda = 0.90$	1,183	0.6782	0.8005
Union $\beta = 2, \lambda = 0.10$	1,183	0.6216	0.7566
Union $\beta = 2, \lambda = 0.50$	1,183	0.6477	0.7828
Union $\beta = 2, \lambda = 0.90$	1,183	0.6515	0.7870
Intersection $\beta = 0, \lambda = 0.10$	974	0.6481	0.8008
Intersection $\beta = 0, \lambda = 0.50$	974	0.6505	0.8049
Intersection $\beta = 0, \lambda = 0.90$	974	0.6506	0.8049
Intersection $\beta = 1, \lambda = 0.10$	974	0.6187	0.7690
Intersection $\beta = 1, \lambda = 0.50$	974	0.6329	0.7793
Intersection $\beta = 1, \lambda = 0.90$	974	0.6341	0.7793
Intersection $\beta = 2, \lambda = 0.10$	974	0.5783	0.7310
Intersection $\beta = 2, \lambda = 0.50$	974	0.6053	0.7618
Intersection $\beta = 2, \lambda = 0.90$	974	0.6093	0.7659
KI-topics $\beta = 0.0\lambda = 0.10$	150	0.5446	0.7067
KI-topics $\beta = 0.0\lambda = 0.50$	150	0.5590	0.7267
KI-topics $\beta = 0.0\lambda = 0.90$	150	0.5608	0.7200
KI-topics $\beta = 1.0\lambda = 0.10$	150	0.5253	0.7067
KI-topics $\beta = 1.0\lambda = 0.50$	150	0.5465	0.7200
KI-topics $\beta = 1.0\lambda = 0.90$	150	0.5516	0.7200
KI-topics $\beta = 2.0\lambda = 0.10$	150	0.4602	0.6667
KI-topics $\beta = 2.0\lambda = 0.50$	150	0.5196	0.7133
KI-topics $\beta = 2.0\lambda = 0.90$	150	0.5292	0.7133

Table 3: Mean Reciprocal Rank and Success@10 for all topic sets on the web site objects.

Topic set	System ranking
Raw	$A \succ B \succeq C \succ F \succ E \succ D \succ I \succ H \succ G$
Union	$C \succ B \succ A \succ F \succ E \succ D \succ I \succ H \succ G$
Intersection	$C \succ B \succ A \succ F \succ E \succ D \succ I \succ H \succ G$
KI-topics	$C \succ B \succ F \succ E \succ A \succ I \succ D \succ H \succ G$

Table 4: Systems rankings of the 4 topic sets.

	KI-topics	Raw	Union	Intersect.
KI-topics	1.00			
Raw	0.67	1.00		
Union	0.83	0.83	1.00	
Intersection	0.83	0.83	1.00	1.00

Table 5: Rank correlation coefficients between the topic sets.

reciprocal rank (i.e., 1 over the rank at which the first relevant document is found). The rankings over the four different topic sets are given in Table 4 (based on the labeling introduced in Table 2).

The results show that ranking based on the Raw Topic set deviates slightly from ranking based on the Union and Intersection topic sets. The Union and Intersection topic sets result in exactly the same ranking. There is a clear grouping of systems with the same length prior. The systems without a length prior (A,B and C) outrank the systems with a length prior $\beta = 1$ (D, E and F), which in turn outrank the systems with length prior $\beta = 2$ (systems G, H and I). Within these groups, the system ranks correspond to the smoothing parameter settings. A higher λ value corresponds to a higher rank. The only deviation is observed in the ranking based on the Raw Topic set. Here, the lowest value for λ leads to the best performance for the systems with no length prior.

If we compare the three automatically generated topic sets to the manual known-item topic set, we see some more differences. For the manual topics, systems E and F, which have a unit length prior, outrank system A, which has no length prior. A possible explanation for this is that the higher λ of systems E and F help the longer queries of the manual topic set. In the other topic sets, most of the queries have only one term, so smoothing has very little influence. This same effect might explain why system I outranks system D.

If we look at the correlation coefficient (Table 5), we see a positive correlation between all topic sets. As the Union and Intersection topic sets lead to the same system ranking, they have a correlation of 1. The system ranking of the Raw topic set shows the lowest correlation with the other topic sets, but the correlation with the manual topic set is still high, indicating that all the extraction methods lead to topic sets that have an ability to rank system similar to that of a manually constructed topic set. Of course, the number of known-item topics is much smaller than the other topic sets, but these initial results point out that the automatic generation of test collections from transaction logs makes sense.

5 Discussion and Conclusions

Cultural heritage, and other special domains, pose a particular problem for information retrieval: evaluation requires a dedicated test collection that takes the particular documents and information requests into account, but building such a test collection requires substantial human effort. We have investigated methods of generating a document retrieval test collection from a search engine's transaction log, based on submitted queries and user-click data. We tested our methods on a museum's search log file, and compared the quality of the generated test collections against a collection with manually generated and judged known-item topics.

Our main findings are the following. First, the test collection derived from a transaction log corresponds well to the actual search experience of real users. An important criterion of bench-marks is that they correspond well to the real-world phenomenon that they are supposed to measure. By basing the test collection directly on a large sample of real end-user interaction, with real information needs, we can ensure that the test collection reflects the information seeking behaviors of users well. This is of particular importance for domain-specific test collections, where results may be impacted by the particular type of information available, and the particular sorts of search requests that are likely to be issued.

Second, the ranking of systems based on the derived judgments corresponds well to the ranking based on the manual topics. We extracted three different sets of topics and corresponding pseudo-relevance judgments from the transaction log. All three sets result in very similar system rankings, indicating that the results are robust against particular choices in the extraction phase. The system rankings are corresponding well to a ranking based on human generated known-item topics. Given the promising initial results, we are currently working on a more

rigorous comparative evaluation, with more human topics, and more diverse systems to be ranked, aiming to understand better the exact conditions under which the extracted test collections behave similar to human generated test collections—and when they behave differently.

Third, deriving pseudo-relevance judgments from a transaction log file is an attractive option in domains where dedicated test collections are not readily available. The results in the paper should not be interpreted as a claim to replace human relevance judgments with extracted topics and pseudo-relevance judgments. There are however many domains and tasks where no suitable test collection is available, and creating a new human test collection might be either impractical or even impossible. Recall that creating human judged test collections requires considerable effort: it is usually a community effort where a number of participating teams provide a diverse set of runs needed for pooling, or even engage in peer-assessments. Hence, deriving a test collection from a transaction log—if available—can be an attractive alternative.

Acknowledgments

This research is part of the MUSEUM (MUltiple-collection SEarching Using Metadata; `http://www.nwo.nl/catch/museum/`) project of the CATCH (Continuous Access To Cultural Heritage) research program in the Netherlands.

The authors were supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMatch contract IST-033104).

References

Michael Chau, Xiao Fang, and Olivia R. Liu Sheng. 2005. Analysis of the query logs of a web site search engine. *J. Am. Soc. Inf. Sci. Technol.*, 56(13):1363–1376.

Susan Dumais, Thorsten Joachims, Krishna Bharat, and Andreas Weigend. 2003. SIGIR 2003 workshop report: implicit measures of user interests and preferences. *SIGIR Forum*, 37:50–54.

Djoerd Hiemstra. 2001. *Using Language Models for Information Retrieval*. Thesis, University of Twente.

ILPS. 2005. The *ilps* extension of the *lucene* search engine. `http://ilps.science.uva.nl/Resources/`.

Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. The Kluwer International Series on Information Retrieval. Springer Verlag, Heidelberg.

Bernard J. Jansen. 2006. Search log analysis: What is it; what's been done; how to do it. *Library and Information Science Research*, 28(3):407–432.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM Press, New York, NY, USA.

Karen Sparck Jones and C. van Rijsbergen. 1975. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development report 5266, Computer Laboratory, University of Cambridge.

Steve Jones, Sally Jo Cunningham, Rodger J. McNab, and Stefan J. Boddie. 2000. A transaction log analysis of a digital library. *Int. j. on Digital Libraries*, 3(2):152–169. URL `citeseer.ist.psu.edu/jones00transaction.html`.

Jaap Kamps, Maarten de Rijke, and Börkur Sigurbjörnsson. 2004. Length normalization in xml retrieval. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–87. ACM Press, New York, NY, USA.

Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37:18–28.

Marijn Koolen, Avi Arampatzis, Jaap Kamps, Nir Nussbaum, and Vincent de Keijzer. 2007. Unified access to heterogeneous data in cultural heritage. To appear.

Marjo Markkula and Eero Sormunen. 2000. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1:259–285.

Douglas W. Oard and Jinmook Kim. 2001. Modeling information content using observable behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*, pages 38–45.

Snowball. 2007. Stemming algorithms for use in information retrieval. `http://www.snowball.tartarus.org/`.

TREC. 2007. Text REtrieval Conference. `http://trec.nist.gov/`.

Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer.

Ellen M. Voorhees and Donna K. Harman, editors. 2005. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press.

Multilingual Search for Cultural Heritage Archives via Combining Multiple Translation Resources

Gareth J. F. Jones, Ying Zhang, Eamonn Newman, Fabio Fantino
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
{gjones,yzhang,enewman,ffantino}
@computing.dcu.ie

Franca Debole
ISTI-CNR
Pisa
Italy
franca.debole
@isti.cnr.it

Abstract

The linguistic features of material in Cultural Heritage (CH) archives may be in various languages requiring a facility for effective multilingual search. The specialised language often associated with CH content introduces problems for automatic translation to support search applications. The MultiMatch project is focused on enabling users to interact with CH content across different media types and languages. We present results from a MultiMatch study exploring various translation techniques for the CH domain. Our experiments examine translation techniques for the English language CLEF 2006 Cross-Language Speech Retrieval (CL-SR) task using Spanish, French and German queries. Results compare effectiveness of our query translation against a monolingual baseline and show improvement when combining a domain-specific translation lexicon with a standard machine translation system.

1 Introduction

Online Cultural Heritage (CH) content is being produced in many countries by organisations such as national libraries, museums, galleries and audiovisual archives. Additionally, there are increasing amounts of CH relevant content available more generally on the World Wide Web. While some of this material concerns national or regional content only of local interest, much material relates to items involving multiple nations and languages, for example concerning events in Europe or Asia. In order to gain a full understanding of such events, including details contained in different collections and exploring different cultural perspectives requires effective multilingual search technologies. Facilitating search of this type requires translation tools to cross the language barrier between users and the available information sources.

CH content encompasses various different media, including of course text documents, images, videos, and audio recordings. Search of text documents between languages forms the focus of cross-language information retrieval (CLIR) research, while search for images is the concern of content-based image retrieval. However, whatever the media of the items they are accompanied by metadata. Such metadata may include simple factual details such as date of creation, but also descriptive details relating to the contents of the item. Multilingual searching using metadata content requires that either the metadata be translated into a language with which the user is able to search or that the search query be translated into the language of the metadata. This alternative of document or query translation is a well rehearsed argument in CLIR, which has generally concerned itself with full text document searching. However, the features of metadata require a more careful analysis. Metadata is typically dense in search terms, while lacking the linguistic structure and information redundancy of full text documents. The absence of linguistic structure makes precise translation of content problematic, while the lack of redundancy means that accurate translation of individual words

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 81–88,
Prague, 28 June 2007.

and phrases is vital to minimise mismatch between query and document terms. Furthermore, CH content is typically in specialised domains requiring domain specific resources for accurate translation. Developing reliable and robust approaches to translation for metadata search is thus an important component of search for many CH archives.

The EU FP6 MultiMatch[1] project is concerned with information access for multimedia and multilingual content for a range of European languages. In the investigation reported in this paper we introduce the first stage multilingual search functionality of the MultiMatch system, and describe its use in an investigation for multilingual metadata search. Since at present we do not have a search test collection specifically developed for MultiMatch we use data from the CLEF 2006 Cross-Language Speech Retrieval (CL-SR) task for our experiments (Oard et al., 2006).

The remainder of this paper is organised as follows: Section 2 gives an overview of the MultiMatch search architecture, Section 3 outlines the experimental search task, Section 4 describes the translation resources used for this study, Section 5 and 6 concern our experimental setup and results, and finally Section 7 summarises our conclusions and gives details of our ongoing work.

2 MultiMatch Search System

The MultiMatch search system is centered on the MILOS Multimedia Repository system (Amato et al., 2004) which incorporates free-text search using Lucene (Hatcher and Gospodnetic, 2004) and image search using an open source image retrieval system GIFT (Müller et al., 2001). In order to support multilingual searching a number of translation tools are being developed based on standard online machine translation tools and dictionaries augmented with domain-specific resources gathered from the WWW and elsewhere. In this section we briefly introduce the relevant details of MILOS and Lucene. Since this paper focuses on text search within MultiMatch, we do not describe the multimedia features of the MultiMatch system.

[1] www.multimatch.org

2.1 MILOS: Multimedia Repository

MILOS (Multimedia dIgital Library for On-line Search) is a repository system conceived to support the distributed storage and retrieval of multimedia objects. This Multimedia Content Management System (MCMS) is able to manage not only structured data, as in databases, but also textual data (using information retrieval technologies), semi-structured data (typically in XML), mixed-mode data, and multimedia data. In MultiMatch, we use MILOS as a metadata repository to enable querying on the structure of the data stored.

MILOS has a three-tier architecture composed of three main components:

1. the XML Search Engine (XMLSE) component which manages the metadata;
2. the MultiMedia Server (MMS) component which manages the documents; and
3. the MultiMedia Digital Library service (MMDLS) component MMDLS which provides application developers with a uniform and integrated way of accessing MMS and XMLSE.

Each of these components is implemented using solutions providing flexibility, scalability, and efficiency.

2.1.1 XMLSE

XMLSE is an enhanced native XML database/repository system with special features for digital library applications. This is especially justified by the well known and accepted advantages of representing metadata as XML documents. Metadata represented with XML can have arbitrary complex structures, which allows it to handle with complex metadata schemas, and can easily be exported and imported. Our XML database can store and retrieve any valid XML document. No metadata schema or XML schema definition is needed before inserting an XML document, except optional index definitions for performance boosting. Once an arbitrary XML document has been inserted in the database it can be immediately retrieved using XQuery. This allows digital library applications to use arbitrary (XML encoded) metadata schemas

and to deal with heterogeneous metadata, without any constraint on schema design and/or overhead due to metadata translation. Thus, the native XML database/repository system is simpler than a general purpose XML database system, but offers significant improvements in specific areas: it supports standard XML query languages such as XPath and XQuery, and offers advanced search and indexing functionality on arbitrary XML documents. It supports high performance search and retrieval on heavily structured XML documents, relying on specific index structures.

Moreover XMLSE provides the possibility of using particular indexes. For example, using the configuration file of XMLSE the system administrator can associate the `<abstract>` elements of a document with a full-text index and to the MPEG-7 `<VisualDescriptor>` elements can be associated with a similarity search index. XMLSE uses Apache Lucene[2] to provide partial (or approximate) text string matching, effectively providing information retrieval functionality within MILOS. This allows XMLSE to use the ranked searching and wildcard queries of Lucene to solve queries like "find all the articles whose title contains the word XML" and so on. This application allows users to interrogate the dataset combining full text, and exact or partial match search. For example the user can look for documents whose `<metadata>` element contains the word "Switzerland". MILOS generates and submits to XMLSE the following XQuery query:

```
for $a in /document where
    $a//metadata ~ 'Switzerland'
return
    <result>
        {$a//title}, {$a//author}
    </result>
```

The query will return a list of results which consist of the title and author of all documents whose metadata contains the term "Switzerland".

2.2 Lucene

Full text search in MILOS is provided by using Lucene as a plugin. Ranked retrieval uses the standard $tf \times idf$ vector-space method provided in Lucene (Hatcher and Gospodnetic, 2004). Lucene also provides additional functionality to improve retrieval effectiveness by providing various query expansion services using techniques such as relevance feedback, although these are not used in the current investigation. Documents and search requests are preprocessed to remove stop words and stemming is applied using the standard resources supplied with Lucene.

[2] `http://lucene.apache.org`

3 Evaluation Task

The MultiMatch system will enable search from a number of CH repository sources including formally published documents, images and video, as well as material gathered from relevant WWW sources. However, in order to explore metadata search issues and evaluate our approaches to addressing related translation problems, a test collection including sample user search topics and relevance judgements is required. Since MultiMatch does not yet have such a collection available, for our current experiments we made use of the data provided for the CLEF 2006 CL-SR track (Oard et al., 2006).

The document collection comprises 8104 English documents that are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totaling 589 hours of speech. Several automatic speech recognition transcripts are available for these interviews. However, for this study we focus on the metadata fields provided for each document: two sets of 20 automatically assigned keywords (`<AUTOKEYWORD2004A1>` and `<AUTOKEYWORD2004A2>`) determined using two different kNN classifiers, denoted by AKW1 and AKW2 respectively; a set of a varying number of manually-assigned keywords (`<MANUALKEYWORD>`), denoted by MKW; and a manual three-sentence summary written by an expert in the field (`<SUMMARY>`), denoted by SUMMARY.

The CLEF collection includes a set of 33 search topics in standard TREC format created in English, and translated into Czech, German, French, and Spanish by native speakers. Since we wish to investigate topics with minimal redundancy, for our experiments we used only the topic Title fields as our search request. Relevance judgments were generated using a search guided procedure and standard pooling methods were also provided with the collec-

tion. Full details of the this collection can be found in (Oard et al., 2006; White et al., 2005).

To explore metadata field search, we used various methods, described in the next section, to automatically translate the French, German, and Spanish topics into English[3].

4 Translation Techniques

The MultiMatch translation resources are based on the WorldLingo machine translation system augmented with domain-specific dictionary resources gathered automatically from the WWW. This section briefly reviews WorldLingo[4], and then describes construction of our augmentation translation lexicons and their application for query translation in multilingual metadata search.

4.1 Machine translation system

There are a number of commercial machine translation systems currently available. After evaluation of several candidate systems, WorldLingo was selected for the MultiMatch project because it generally gives good translation well between the English, Spanish, Italian, and Dutch, languages relevant to the Multimatch project[5]. In addition, it provides a useful API that can be used to translate queries on the fly via HTTP transfer protocol. The usefulness of such a system is that it can be integrated into any application and present translations in real-time. It allows users to select the source/target languages and specify the text format (e.g. plain text file or html file) of their input files. The WorldLingo translation system also provides various domain-specific dictionaries that can be integrated with translation system. A particularly useful feature of WorldLingo with respect to for MultiMatch, and potentially applications within CH in general, is that to improve the quality of translations, additional locally developed customized dictionaries can be uploaded. This enables the WorldLingo dictionaries to be extended to contain special terms for a specific domain.

[3]Due to a lack of translation resources, we did not use the Czech translations in these experiments

[4]`http://www.worldlingo.com/`

[5]Additionally, it translates well between French and English, as used in this paper

4.2 Translation lexicon construction

To extend the standard dictionaries provided with WorldLingo we used the current online *wikipedia*. Wikipedia[6] is the largest multilingual free-content encyclopedia on the Internet. As of March 21 2007, there are approximately 6.8 million articles written in 250 languages available on the web, according to *Wiki Stats*[7]. Wikipedia is structured as an interconnected network of articles. Each wikipedia page can hyperlink to several other wikipedia pages. Wikipedia page titles in one language are also linked to a multilingual database of corresponding terms. Unlike the web, most hyperlinks in wikipedia have a more consistent and semantically meaningful interpretation and purpose. The comprehensive literature review presented by Adafre and Rijke (2005) describes the link structure of wikipedia. As a multilingual hypertext medium, wikipedia presents a valuable new source of translation information. Recently, researchers have proposed techniques to exploit this opportunity. Adafre and Rijke (2006) developed a technique to identify similar text across multiple languages in wikipedia using page content-based features. Boumaet et al. (2006) utilized wikipedia for term recognition and translation in order to enhance multilingual question answering systems. Declerck et al. (2006) showed how the wikipedia resource can be used to support the supervised translation of ontology labels.

In order to improve the effectiveness of multilingual metadata search, we mine wikipedia pages as a translation source and construct translation lexicons that can be used to reduce the errors introduced by unknown terms (single words and multiword phrases) during query translation. The major difference in our proposal is that the translations are extracted on the basis of hyperlinks, meta keywords, and emphasized concepts — e.g. anchor text, boldface text, italics text, and text within special punctuation marks — appearing in the first paragraph of wikipedia articles.

Meta keywords Wikipedia pages typically contain meta keywords assigned by page editors. This meta keywords can be used to assist in the iden-

[6]`http://www.wikipedia.org/`

[7]`http://s23.org/wikistats/wikipedias_html.php?sort=good_desc`

tification of the associated terms on the same topic.

Emphasized concepts In common with standard summarization studies, we observed that the first paragraph of a wikipedia document is usually a concise introduction to the article. Thus, concepts emphasized in the introductory section are likely to be semantically related to the title of the page.

In our study we seek to use these features from multilingual wikipedia pages to compile a domain-specific word and phrase translation lexicon. Our method in using this data is to augment the queries with topically related terms in the document language through a process of *post-translation query expansion*. This procedure was performed as follows:

1. An English vocabulary for the domain of the test collection was constructed by performing a limited crawl of the English wikipedia[8], Category:World War II. This category contains links to pages and subcategories concerning events, persons, places, and organizations pertaining to war crimes or crimes against humanity especially during WWII. It should be noted that this process was neither an exhaustive crawl nor a focused crawl. The purpose of our current study is to explore the effect of translation expansion on metadata retrieval effectiveness. In total, we collected 7431 English web pages.

2. For each English wikipedia page, we extracted its hyperlinks to German, Spanish, and French. The basename of each hyperlink is considered as a term (single word or multi-word phrase that should be translated as a unit). This provided a total of 4446 German terms, 3338 Spanish terms, and 4062 French terms. As an alternative way of collecting terms in German, Spanish, and French, we are able to crawl the wikipedia in a specific language. However, a page with no link pointing to its English counterpart will not provide enough translation information.

[8] en.wikipedia.org

RUN ID	Augmented lexicon using all terms appearing in the following fields		
	Title terms	Meta keywords	Emphasized concepts
RUN_{mt+t}	$\surd$	$\times$	$\times$
RUN_{mt+m}	$\times$	$\surd$	$\times$
RUN_{mt+c}	$\times$	$\times$	$\surd$
RUN_{mt+m+c}	$\times$	$\surd$	$\surd$

Table 1: Run descriptions.

3. For each of the German, Spanish, and French terms obtained, we used the title term, the meta keywords, and the emphasized concepts obtained from the same English wikipedia page as its potential translations.

For example, consider an English page titled as "World War II"[9]. The title term, the meta keywords, the emphasized concepts in English, and the hyperlinks (to German, Spanish, and French) associated are shown in Figure 1. We first extract the basenames "Zweiter Weltkrieg" (in German), "Segunda Guerra Mundial" (in Spanish), and "Seconde Guerre mondiale" (in French) using the hyperlink feature. To translate these terms into English, we replace them using the English title term, all the English meta keywords and/or all the English emphasized concepts occurring in the same English wikipedia page. This is a straightforward approach to automatic post-translation query expansion by using meta keywords and/or emphasized concepts as expanded terms. The effects of the features described above are investigated in this work, both separately and in combination, as shown in Table 1,

5 Experimental Setup

In this section we outline the design of our experiments. We established a monolingual reference (RUN_{mono}) against which we can measure multilingual retrieval effectiveness. To provide a baseline for our multilingual results, we used the standard WorldLingo to translate the queries (RUN_{mt}). We then tested the MT integrated with different lexicons compiled using wikipedia. Results of these experiments, shown in Table 1, enable us gauge the effect of each of our additional translation resources generated using wikipedia.

[9] http://en.wikipedia.org/wiki/World_War_II

Title: World War II

Hyperlink to German: http://de.wikipedia.org/wiki/Zweiter_Weltkrieg

Hyperlink to Spanish: http://es.wikipedia.org/wiki/Segunda_Guerra_Mundial

Hyperlink to French: http://fr.wikipedia.org/wiki/Seconde_Guerre_mondiale

Meta keywords:

World War II, WWII history by nation, WWII history by nation, 101st Airborne Division, 11th SS Volunteer Panzergrenadier Division Nordland, 15th Army Group, 1937, 1939, 1940

Emphasized concepts:

World War II (abbreviated **WWII**), or the **Second World War**, was a worldwide conflict which lasted from 1939 to 1945. World War II was the amalgamation of two conflicts, one starting in Asia as the Second Sino-Japanese War, and the other beginning in Europe with the Invasion of Poland. The war was caused by the expansionist and hegemonic ambitions of Germany, Italy, and Japan and economic tensions between all major powers.

Figure 1: Title, hyperlinks, meta keywords, and emphasized concepts (underlined terms) extracted from the English wikipedia page http://en.wikipedia.org/wiki/World_War_II.

The focus of this paper is not on optimising absolute retrieval performance, but rather to explore the usefulness of our translation resources. Thus we do not apply retrieval enhancement techniques such as relevance feedback which would make it more difficult to observe the impact of differences in behaviour of the translation resources. The experiments use the SUMMARY field, as an example of concise natural language descriptions of CH objects; and the AKW1 and AKW2 fields as examples of automatically assigned keyword labels without linguistic structure, with the MKW field providing similar manually assigned for keyword labels. Retrieval effectiveness is evaluated using standard TREC mean average precision (MAP) and the precision at rank 10 (P@10).

6 Results and Discussion

The results of our query translation experiments are shown in Table 2, 3, 4, and 5. For search using SUMMARY and MKW fields, the lexicon compiled using title terms provided an improvement of $7 \sim 9\%$, $7 \sim 19\%$, and $20 \sim 30\%$, in German–English, Spanish–English, and French–English retrieval task, respectively. These improvements are statistically significant at the 95% confidence level, and emphasize the importance of a good domain-specific translation lexicon.

The addition of meta keywords or emphasized concepts also improves results in most cases relative to the RUN$_{mt}$ results. However, we can see that retrieval performance degrades when the query is expanded to contain terms from both meta keywords and emphasized concepts. This occurs despite the fact that the additional terms are often closely related to the original query terms. While the addition of all these terms generally produces an increase in the number of retrieved documents, there is little or no increase in the number of relevant documents retrieved, and the combination of the two sets of terms in the queries leads on average to a slight reduce in the rank of relevant documents.

The results show that RUN$_{mt+t}$ runs provide the best results when averaged across a query set. However, when analysed at the level of individual queries different combined translation resources are more effective for different queries, examples of this effect are shown in Table 6. This suggests that it may be possible to develop a more sophisticated translation expansion methods to select the best terms from different lexicons. At the very least, it should be possible to use "context-sensitive filtering" and "combination of evidence" (Smets, 1990) approaches to improve the overall translation quality. We plan to explore this method in further investigations.

7 Conclusion and Future Work

This paper reports experiments with techniques developed for domain-specific lexicon construction to facilitate multilingual metadata search for a CH re-

RUN ID	German–English		Spanish–English		French–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN_{mt}	0.0750	0.1233	0.0756	0.1250	0.0652	0.1152
RUN_{mt+t}	**0.0815**	**0.1516**	**0.0899**	**0.1545**	**0.0783**	**0.1333**
RUN_{mt+m}	0.0775	0.1266	0.0797	0.1364	0.0690	0.1030
RUN_{mt+c}	0.0669	0.1000	0.0793	0.1303	0.0770	0.1152
RUN_{mt+m+c}	0.0668	0.0968	0.0737	0.1212	0.0646	0.0970
RUN_{mono}	MAP = 0.1049		P@10 = 0.1818			

Table 2: Results for SUMMARY field search. (RUN_{mt+t} run provides the best results in all cases.)

RUN ID	German–English		French–English		Spanish–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN_{mt}	0.1158	0.1750	0.1000	0.1677	0.0903	0.1677
RUN_{mt+t}	**0.1235**	**0.2100**	**0.1071**	**0.2031**	**0.1171**	**0.2194**
RUN_{mt+m}	0.1171	0.1393	0.1023	0.2000	0.0983	0.1903
RUN_{mt+c}	0.1084	0.1500	0.0958	0.1636	0.1089	0.1667
RUN_{mt+m+c}	0.1069	0.1600	0.0947	0.1727	0.0940	0.1742
RUN_{mono}	MAP = 0.1596		P@10 = 0.2812			

Table 3: Results for MKW field search. (RUN_{mt+t} run provides the best results in all cases.)

RUN ID	German–English		French–English		Spanish–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN_{mt}	0.0264	0.0731	0.0247	0.0548	0.0316	0.0767
RUN_{mt+t}	**0.0273**	**0.0828**	**0.0274**	**0.0656**	**0.0406**	**0.0867**
RUN_{mt+m}	0.0268	0.0633	0.0258	0.0606	0.0357	0.0613
RUN_{mt+c}	0.0266	0.0667	0.0266	0.0636	0.0383	0.0839
RUN_{mt+m+c}	0.0259	0.0633	0.0260	0.0606	0.0328	0.0677
RUN_{mono}	MAP = 0.0388		P@10 = 0.1000			

Table 4: Results for AKW1 field search. (RUN_{mt+t} run provides the best results in all cases.)

RUN ID	German–English		French–English		Spanish–English	
	MAP	P@10	MAP	P@10	MAP	P@10
RUN_{mt}	0.0279	0.0375	0.0347	0.0625	0.0205	0.0483
RUN_{mt+t}	0.0279	**0.0481**	0.0351	**0.0680**	0.0238	0.0433
RUN_{mt+m}	**0.0302**	0.0448	**0.0361**	0.0556	0.0223	0.0484
RUN_{mt+c}	0.0275	0.0414	0.0332	0.0593	0.0268	0.0548
RUN_{mt+m+c}	0.0299	0.0448	0.0351	0.0536	**0.0273**	**0.0581**
RUN_{mono}	MAP = 0.0420		P@10 = 0.0821			

Table 5: Results for AKW2 field search. (The best results are in bold.)

trieval tasks. The results show that our techniques can provide a statistically significant improvement in the retrieval effectiveness. Using a tailored translation lexicon enables us to achieve $(77\%, 78\%)$, $(86\%, 67\%)$ and $(75\%, 63\%)$ of the monolingual effectiveness in German–English, Spanish–English, and French–English multilingual metadata SUMMARY, MKW field search tasks. In addition, the multilingual wikipedia proved to be a rich resource of translations for domain-specific terms.

Intuitively, document translation is superior to query translation. Documents provide more context for resolving ambiguities (Oard, 1998) and the translation of source documents into all the languages supported by the retrieval system effectively reduces CLIR to a monolingual IR task. Furthermore, it has the added advantage that document content is accessible to users in their native languages. In our future work, we will compare the effectiveness of these two approaches to metadata search in a multilingual environment.

	Query ID	MT WorldLingo	Augmented lexicon using all terms appearing in the following fields			
			Title terms	Meta keyword	Emphasized concepts	Meta keyword + Emphasized concepts
German–English	*1133*	0.6000	0.6000	0.6195	0.6092	**0.6400**
	1325	0.0000	0.0003	**0.0020**	**0.0020**	0.0018
	1623	0.2210	0.2210	**0.3203**	0.0450	0.0763
	3007	0.0000	0.0003	0.0025	0.0047	**0.0054**
	3012	0.0087	0.0087	0.0073	0.0073	**0.0097**
	3025	0.0052	0.0052	**0.0060**	0.0052	**0.0060**
Spanish–English	*1623*	0.0063	0.0063	**0.1014**	0.0084	0.0334
	3007	0.0000	0.0004	0.0028	0.0048	**0.0057**
French–English	*1133*	0.6000	0.6000	0.6195	0.6092	**0.6400**
	1345	0.0600	0.0667	**0.0809**	0.0495	0.0420
	1623	0.0750	0.0798	**0.1810**	0.0228	0.0528
	3005	0.0200	0.0232	0.0226	**0.2709**	0.1063
	3007	0.0003	0.0003	0.0024	0.0025	**0.0037**
	3025	0.0173	0.0173	**0.0178**	0.0173	**0.0178**

Table 6: Examples of MAP values obtained using different translation combinations for SUMMARY field search. (The best results are in bold.)

Acknowledgement

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST- 033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, Chicago, Illinois. ACM Press.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69, Trento, Italy.

Giuseppe Amato, Claudio Gennaro, Fausto Rabitti, and Pasquale Savino. 2004. Milos: A multimedia content management system for digital library applications. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 14–25. Springer-Verlag.

Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jorg Tiedemann. 2006. The university of groningen at QA@CLEF 2006 using syntactic knowledge for QA. In *Working Notes for the Cross Language Evaluation Forum 2006 Workshop*, Alicante, Spain.

Thierry Declerck, Asunciòn Gòmez Pèrez, Ovidiu Vela, Zeno Gantner, and David Manzano-Macho. 2006. Multilingual lexical semantic resources for ontology translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Erik Hatcher and Otis Gospodnetic. 2004. *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA.

Henning Müller, Wolfgang Müller, and David McG. Squire. 2001. Automated benchmarking in content-based image retrieval. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo*, Tokyo, Japan. IEEE Computer Society.

Douglas W. Oard, Jianqiang Wang, Gareth J. F. Jones, Ryen W. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. 2006. Overview of the CLEF-2006 cross-language speech retrieval track. In *Working Notes for the Cross Language Evaluation Forum 2006 Workshop*, Alicante, Spain.

Douglas W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 472–483, London, UK. Springer-Verlag.

Philippe Smets. 1990. The combination of evidence in the transferable belief model. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(5):447–458.

Ryen W. White, Douglas W. Oard, Gareth J. F. Jones, Dagobert Soergel, and Xiaoli Huang. 2005. Overview of the CLEF-2005 cross-language speech retrievaltrack. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 744–759. Springer.

Lessons from the MALACH Project:
Applying new technologies to improve intellectual access to large oral history collections

Douglas W. Oard, University of Maryland, USA

Abstract:

In this talk I will describe the goals of the MALACH project (Multilingual Access to Large Spoken Archives) and our research results. I'll begin by describing the unique characteristics of the oral history collection that we used, in which Holocaust survivors, witnesses and rescuers were interviewed in several languages. Each interview has been digitized and extensively catalogued by subject matter experts, thus producing a remarkably rich collection for the application of machine learning techniques. Automatic speech recognition techniques originally developed for the domain of conversational telephone speech were adapted to process these materials with word error rates that are adequate to provide useful features to support interactive search and automated clustering, boundary detection, and topic classification tasks. As I describe our results, I will focus particularly on the evaluation methods that that we have used to assess the potential utility of this technology. I'll conclude with some remarks about possible future directions for research on applying new technologies to improve intellectual access to oral history and other spoken word collections. This is joint work with Charles University (Prague), IBM Research (T.J. Watson), the Johns Hopkins University (Baltimore), the University of Southern California (Los Angeles), an the University of West Bohemia (Pilsen),

About the Speaker:

Douglas Oard is Associate Dean for Research at the College of Information Studies of the University of Maryland, College Park, where he holds joint appointments as Associate Professor in the College of Information Studies and in the Institute for Advanced Computer Studies. He earned his Ph.D. in Electrical Engineering from the University of Maryland, and his research interests center around the use of emerging technologies to support information seeking by end users. Dr. Oard's recent work has focused on interactive techniques for cross-language information retrieval, searching conversational media, and leveraging observable behavior to improve user modeling. Additional information is available at http://www.glue.umd.edu/~oard/.

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), page 89,
Prague, 28 June 2007.

Author Index


ACL 2007
PRAGUE

MultiMatch