



Spatially-Aware Information Retrieval on the Internet



SPIRIT is funded by EU IST Programme
Contract Number: IST-2001-35047

Extraction of semantic annotations from textual web pages

Deliverable number:	D15 6201
Deliverable type:	R
Contributing WP:	WP6
Contractual date of delivery:	01/01/04
Actual date of delivery:	04/01/04
Authors:	Paul Clough, Mark Sanderson and Hideo Joho University of Sheffield

Keywords: geographic mark-up, annotation, information extraction

Abstract: This report discusses our proposed ideas for extracting geographical locations from Web pages within the SPIRIT collection and grounding these using the SPIRIT ontology. The result of this work will be to annotate the SPIRIT collection with geographic references for use by other partners in the SPIRIT project. This report presents current work related to this task, and provides background information relevant to parsing and grounding geographic references.

Contents

1.	ABSTRACT	5
2.	INTRODUCTION	5
3.	IDENTIFYING NAMED ENTITIES	7
3.1.	Introduction	7
3.2.	Simple list lookup	9
3.3.	Incorporating context	10
3.4.	Methods/tools we plan to use for NER	10
3.5.	Exploiting Web page structure	12
4.	DEALING WITH GEO-REFERENCES	14
4.1.	Identifying geographic references	14
4.2.	Disambiguating the geo-reference	15
	Effectiveness of the approach.....	18
4.3.	Extent of a reference	18
	In-text references	18
	“Contact us” pages on Web sites.....	18
5.	GEOGRAPHIC RESOURCES	20
5.1.	Getty Thesaurus of Geographic Names (TGN)	20
5.2.	Code-Point®	20
5.3.	Gazetteer Plus	21
5.4.	GEOnet Names Server (GNS)	21
5.5.	Geographic Names Information System (GNIS)	21
6.	GEO-REFERENCE MARK-UP FORMAT	21
7.	BUILDING A GEO-REFERENCE TRAINING SET	22
8.	SUMMARY	23
9.	REFERENCES	24
APPENDIX I. REPORT ON HTL-NAACL 2003 WORKSHOP ON ANALYSIS OF GEOGRAPHIC REFERENCES		27
1.	INTRODUCTION	27
1.1.	Workshop overview	27
1.2.	Four conceptual stages	27
1.3.	The rest of the report	27
2.	PAPERS OVERVIEW	28

2.1.	Knowledge Based Ontology Design	28
2.2.	Design of gazetteer database.....	28
2.3.	Deriving geographical gazetteers	28
2.4.	Word sense disambiguation.....	29
2.5.	Spatially-aware search engines	29
2.6.	Visualising geographic references on map.....	30
2.7.	Back-transliteration of geographic entities.....	30
2.8.	Descriptive gazetteers and markup encoding	31
3.	DISCUSSION	31
3.1.	One sense in one discourse.....	31
3.2.	Ranking algorithm.....	31
3.3.	Ontology design.....	31
3.4.	Descriptive gazetteer	31
4.	CONCLUSION AND SUGGESTION	32
4.1.	Conclusion.....	32
4.2.	Workshop for GIS and IR.....	32
5.	REFERENCES	32
	APPENDIX I. <i>TOOLS AND RESOURCE</i>	33
	APPENDIX II. <i>ABSTRACT OF PRESENTED PAPERS AND DEMO</i>	34

Executive Summary

This report discusses our proposed ideas for extracting geographical locations from Web pages within the SPIRIT collection and grounding these using the SPIRIT ontology. The result of this work will be to annotate the SPIRIT collection with geographic references for use by other partners in the SPIRIT project. This report presents current work related to this task, and provides background information relevant to parsing and grounding geographic references.

D15 6201

Extraction of semantic annotations from textual Web pages

1. Abstract

The title of this deliverable, "Extraction of semantic annotations from textual Web pages", refers to the task in SPIRIT of identifying references in Web pages. As the ontology developed in SPIRIT mainly holds spatial information, the bulk of this task is to find in pages references to geographic locations. As the ontology also holds information on types of businesses within a location (e.g. hotels, train stations, tourist sites, etc), determination of the types of Web sites/pages is a further task of this work. Both tasks require similar stages of work:

- Identify within the text of a Web page a possible reference to an item or items in the ontology;
- If the reference could be to more than one item (if the reference is ambiguous; e.g. Sheffield is a city in South Yorkshire, a small village in southern England, or a city in Alabama, USA), then attempt to disambiguate that reference by examining the context in which the reference appears.
- Finally, it is necessary to determine the extent to which a geographic reference applies to surrounding text or to the Web pages linked to or linked from the Web page the reference appears in.

The rest of this document describes background work and the methods used to implement the stages, both for geo-references and for determination of business type. We also discuss the geographic resources, which will be used to aid the extraction process, and finally we describe the format of our proposed mark-up scheme.

2. Introduction

The task of this work package is to identify geographical references (*geo-references* or *toponyms*) in Web pages and assign to them geographic coordinates as provided by the SPIRIT ontology. These two tasks are commonly referred to as geoparsing and geocoding respectively (McCurley 2001) and (Larson 2000). Based on the framework proposed in the HLT-NAACL 2003 Workshop on Analysis of Geographical References¹ (see Appendix I), there are at least four conceptual stages in this SPIRIT task:

1. Geographic entity reference detection,
2. Contextual information gathering,
3. Disambiguation of entities, and
4. Grounding of identified entities.

¹ <http://kornai.com/NAACL/>

The first stage involves extracting geographic entities from texts and distinguishing them from other entities such as names of people or organisations, dates and times, or events. In language processing this is referred to as Named Entity Recognition (NER) and forms a core part of other text processing applications such as information extraction (IE) and information retrieval (IR). Existing NER software will be used to extract geo-references and business types, from texts. Most methods of identifying geo-references are an adaptation of existing NER algorithms (see proceedings of the HLT/NAACL 2003 workshop analysis of geographic references) and we plan to continue this trend in this SPiRiT task. The algorithms/categories will be adapted/extended to deal with more fine-grained categories, e.g. organisations will be classified into type, e.g. hotels, train stations, tourist sites, etc. It will be possible to perform this stage independently from the SPiRiT ontology to provide a more generic extraction tool, which will be able to also identify entities not contained within the ontology.

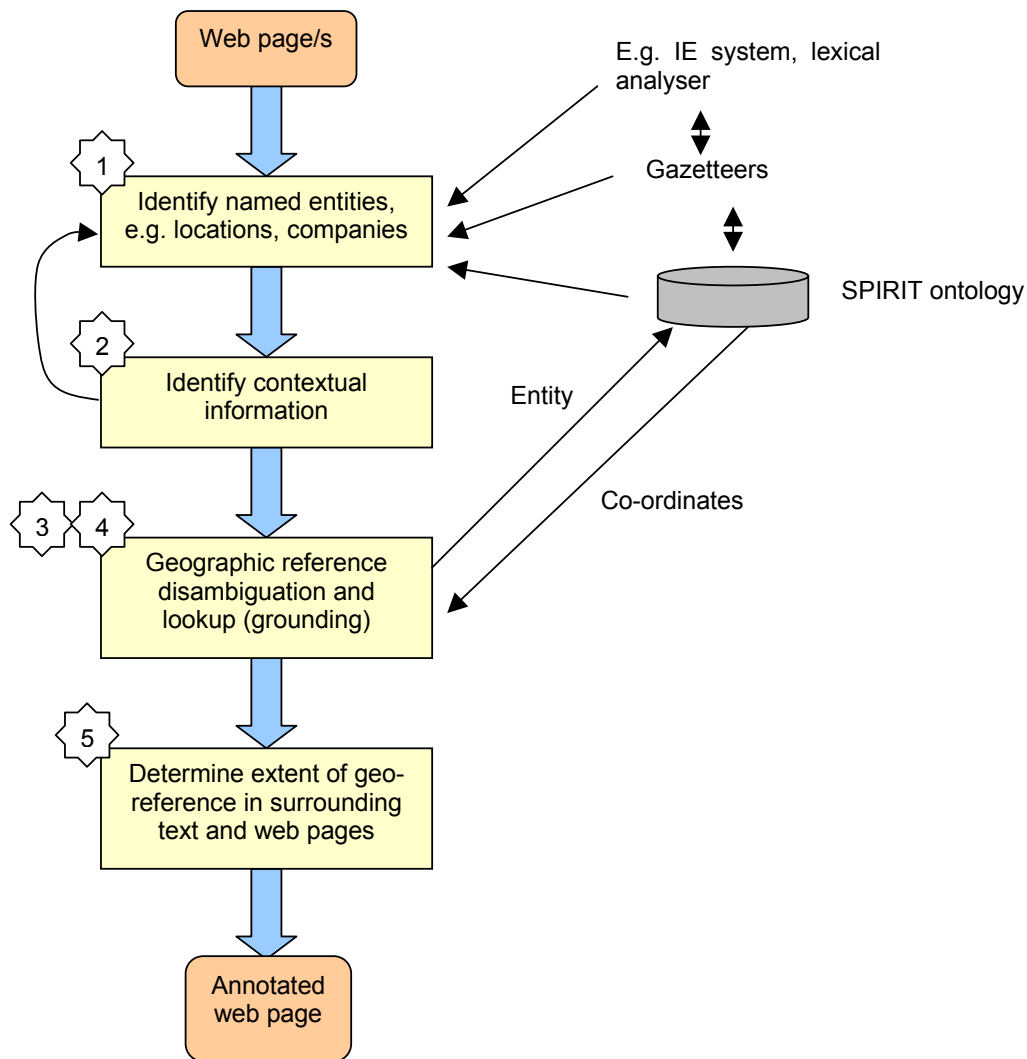


Fig. 1 The proposed method for extracting and annotating geographic references.

For many approaches of NER, implicit with disambiguating between different named entities is gathering context surrounding the entities (stage 2). Identifying additional geographic information surrounding a reference can be used in later stages to disambiguate and ground geographic references. For example, identifying 0114 as an STD code near the location

“Sheffield” would provide useful information in distinguishing between Sheffield in South Yorkshire, UK and Sheffield in Alabama, USA. Also contextual information can be derived from further attributes, which are specifically associated with Web pages, e.g. structured page mark-up, the link structure between pages and extraneous information, e.g. IP addresses. Additional information derived from the Web pages could also be used to extend the SPIRIT ontology and enhance existing resources.

The third stage deals with disambiguating geographic references. If the reference could be to more than one item (if the reference is ambiguous; e.g. Sheffield is a city in South Yorkshire, a small village in southern England, or a city in Alabama, USA), then we must attempt to disambiguate that reference by examining the context in which the reference appears. This is crucial prior to the fourth stage which is to assign geographical coordinates to the disambiguated entities. Geographical coordinates can be in the form of a point (e.g. longitude and latitude), or a polygon region. This information, provided by the SPIRIT ontology, is encoded into an annotated version of the Web page/s for use by other SPIRIT partners, e.g. re-ranking results and building a geographic index.

The process of extracting and annotating geographic references from Web pages is summarised in Fig. 1. The final stage shown in Fig. 1 (stage 5) determines the extent to which a geo-reference applies to the surrounding text or Web pages linked to or linked from that reference by examining the context in which the reference appears. This information is encoded into the geographic mark-up and discussed further in Section 4.

3. Identifying named entities

Identifying potential geo-references and organisations will be performed in a similar manner (and probably at the same time) using generic NER tools adapted to deal with geographic entity types specific to the SPIRIT project. In this section we discuss generic approaches to NER, which will be used in the first stage of this SPIRIT task. We plan to investigate existing NER approaches and annotate more than geographical locations to provide a richer set of annotations for SPIRIT partners, e.g. business/organisation types (hotels, restaurants etc.).

3.1. Introduction

Many methods exist for Named Entity Recognition (or NER), the process of assigning every word or group of words to a set of pre-defined categories (including “not an entity”), and a number of them are considered here. NER was a subtask of the evaluation campaigns called the Message Understanding Conferences (or MUCs)² run by the American National Institute of Standards and Technology (NIST). These campaigns served to stimulate research in NER by providing a set of standard entity types to extract³ (Table 1), training data and an evaluation methodology.

On MUC-7 data, the subtask of recognising named-entities has achieved up to 96.6% of manual recognition (Zhou and Su 2002) using machine learning approaches. Further research has also shown that methods for NER can be language-independent (e.g. (Curran and Clark 2003)), which in the SPIRIT project would enable us to identify named entities in texts written in languages other than English. Most NER algorithms combine lists of known locations, organisations and people (called gazetteers) with rules, which capture elements of the surrounding context. Mikheev et al. (1999) have shown that NER performs reasonably well for most kinds of named-entities even *without* gazetteers (particularly people and organizations).

² Message Understanding Conferences (MUCs): http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

³ http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html

MUC entity type	Description
ORGANIZATION	named corporate, governmental, or other organizational entity
PERSON	named person or family
LOCATION	name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)
DATE	complete or partial date expression
TIME	complete or partial expression of time of day
MONEY	monetary expression
PERCENT	percentage

Table 1 Named entities defined in the MUC evaluation for NER

Most NER systems consist of at least three basic components: (1) a tokeniser, (2) gazetteer lists, and (3) a NE grammar. Tokenisation segments text into tokens, e.g. words, numbers, and punctuation. The gazetteer contains lists of named entities, e.g. towns, countries, cities, and lists of keywords such as titles and company designators (e.g. Plc and Ltd). The NE grammar consists of rules for NE recognition, which take into account surrounding context. Approaches of NER generally make use of two types of evidence: (1) internal and (2) external (McDonald, 1996). Names often have some kind of internal or phrasal structure, which suggests they are names. This information can be stored or guessed and includes, for example, capitalisation (CapWord), prefixes or suffixes (e.g. company designators) and lists of names. For example, the internal evidence for a location might be captured as:

CapWord + CapWord + {Inc, Plc, Ltd}
 e.g. British Telecommunications Plc

CapWord + {Street, Boulevard, Avenue, Crescent, Road}
 e.g. Portobello Street

Somewhere in the text there is likely to be external (or contextual) evidence, which makes clear what type of entity a word or phrase is. For example, consider the phrase “President Washington chopped the tree” (Stevenson, 2000). In this example “President” is clear external evidence that “Washington” is the name of a person and not a place. An approach based on a list of names alone uses only internal evidence and examples such as “Washington” illustrate limitations of this approach. Surrounding context is generally captured using rules expressed within a grammar and names are often used in predictive⁴ local contexts:

“to the” COMPASS “of” CapWord
 e.g. to the south of Loitokitok

“based in” CapWord
 e.g. based in Loitokitok

CapWord “is a” (ADJ)? GeoWord
 e.g. Loitokitok is a friendly city

⁴ These predictive contexts are identified manually or automatically using statistical methods.

Rules are often specified within terms of previously identified entities or annotations (e.g. ADJ, CapWord and GeoWord in the previous examples) according to pre-defined grammars. One such grammar is the Common Pattern Specification Language (CPSL) as defined in (Applet 1996) defined to match linear sequences of annotations. Matching is performed through regular expressions over annotations.

Identifying named entities is not as simple as one might first think. Whether the task is to identify only geo-references or a range of entities, similar problems prevail. For example, Tablan (2001) identifies the following problems in NER:

- Variation of NEs
 - Different forms of a name, e.g. John Smith, Mr. Smith, Smith, John.
 - Abbreviations, e.g. International Business Machines, IBM.
- Ambiguity of NE types
 - John Smith (company vs. person)
 - May (person vs. month)
 - Washington (person vs. location)
 - 1945 (date vs. time)
- Ambiguity with common words, e.g. "may"
- Issues of style, structure, domain, genre etc.
- Punctuation, spelling, spacing, and formatting.
- Ambiguously capitalized words (e.g. the first word in the sentence).
- Semantic ambiguity (e.g. Phillip Morris – organization or person?).
- Structural ambiguity (e.g. "Cable and Wireless" vs. "Microsoft" and "Dell").

3.2. Simple list lookup

The most straightforward means of identifying named entities is through matching input texts (i.e. the Web pages to be marked up) to previously generated lists of place names, addresses, postcodes, phone numbers etc. This method requires no grammars or even information about tokenisation and can also be language and domain independent. Input texts can be matched against the lists using pattern matching, e.g. UNIX tools such as `lex` or `flex`⁵. These match the input text to the longest matching item⁶ in the gazetteer quickly and accurately. However, McCurley (2001) suggests that approaches, which build finite state machines (such as `flex`), are not scalable beyond small phrase lists (e.g. 10,000 names). McCurley suggests the use of `fgrep` instead, which scales up to much larger lists.

Although simple, the pattern matching approach does suffer from at least three drawbacks: (1) the method is not able to identify new locations not found in the geo-reference lists, (2) it is not always correct to assume that a word is being used in a geographic context, e.g. Chicago can represent the US city, the name of a pop group, or the internal project name for Windows 95 (McCurley,2001), and (3) geo-references often appear in different forms, e.g. "UK" and "United Kingdom", which will not match unless within the list. However the advantage of such a technique is its speed and simplicity. Further problems include overlap between lists, their maintenance (e.g. new businesses may appear from day-to-day), and their finiteness.

⁵ <http://www.gnu.org/software/flex/>

⁶ For example if the input text contains the phrase "Old Kent Road" and the gazetteer contains the names "Old Kent Road", "Kent Road" and "Kent", pattern matching using this greedy approach will match against the three-word phrase in the input text over the single word units.

3.3. Incorporating context

Given the limitations of using lists of named entities alone, more sophisticated methods of NER make use of both lists and NE grammars to capture external evidence. Rules for NE recognition can be generated entirely by hand (knowledge-based) or automatically using machine learning (or statistical) techniques. The former method relies heavily on a knowledge expert; the latter aims to induce rules from a set of manually annotated texts using machine learning techniques such as Hidden Markov Models (HMMs), decision trees or Maximum Entropy models and model the probability of event sequences. For example, the HMM computes whether a word is part of a name or not by modelling it as a random event with a probability that can be estimated. One assumes an underlying finite state machine (not directly observable hence hidden), which changes state with each input word and reaches an end state when a sequence of words likely to be a name has been observed. The model is trained on previously identified names (i.e. a training corpus) and when the recogniser is run it computes the maximum likelihood path through the state model for the given word sequence, i.e. marking spans of words, which correspond to a name.

Two main concerns of the machine learning approach are: (1) the training data and (2) generalisation of rules. The first concern is how much data is sufficient for a machine learning approach to induce rules, which can be used reliably on new unseen texts. Careful selection of training set examples is often necessary to create a balanced and representative training set, which can often lead to a large amount of manual effort. The second concern deals with whether the induced rules will generalise across new unseen texts or whether they only operate successfully on examples found in the training set. This is often dependent on the features selected to represent rules. For example features based on the surface realisation of text will inevitably be problematic and prone to errors caused by data sparseness. However by generalising the features, e.g. by using morphologically derived root forms of words or part-of-speech class, problems of data sparseness can be overcome (see, e.g. Ciravenga 20001). Methods of wrapper induction are also ways of generalising rules to deal with variations on the surface realisation of texts (e.g. Borges et al. (2002) induce rules, which are based on the structure of HTML mark-up to learn addresses from Web pages).

Hand-crafting rules usually produce better NER performance than machine learning approaches, but typically only for restricted domains. The knowledge-based approach is therefore less portable than the machine learning method because to adapt the recogniser to new domains involves creating a new set of rules, which is often a time-consuming process requiring domain-specific knowledge. The machine learning approach has the benefit of being more portable and can be re-trained for different domains with minimal manual intervention. It may also be more robust than knowledge-based approaches because rules can be identified, which may have otherwise been missed if defined manually. One drawback with machine learning methods is that they require a large amount of good-quality training data. However, methods such as *bootstrapping* can be used reduce the amount of data required by combining a small amount of pre-classified data with a larger amount of unclassified data (Riloff and Jones, 1999).

3.4. Methods/tools we plan to use for NER

We plan to experiment with at least three methods to identify geo-references: (1) simple list lookup, (2) a rule-based approach to NER, and (3) a machine learning approach to NER to determine which one gives highest performance on a sample of Web data from the SPIRIT collection. The degree, to which gazetteers help identify named entities, is varied. For example Malouf (2002) found that gazetteers did not improve performance; whereas others (e.g. Carreras et al. (2002)) have gained significant improvements using gazetteers and trigger phrases. Mikheev et al. (1999) showed a NER could perform well even without gazetteers for most classes. However this was not the case for locations (51.7% F-measure without; 94.5% with). Mikheev et al. also showed that simple list lookup for locations performs reasonably well (precision of 90-94%; recall of 75-85%) with 5,000 locations collected from the CIA World Fact Book and evaluated on MUC-7 data. They point out that past experience from MUC has shown that the quality of the list has a more significant impact than its size. We plan to

experiment with high-quality gazetteer lists in combination with grammars to provide the best possible method of NER for the SPiRiT data.

The Natural Language Processing (NLP⁷) group at Sheffield have knowledge and experience in language processing. In particular the group have built tools for Information Extraction (IE), a task which aims to map natural language into predefined, structured representations, that when instantiated represent key information from the original source (Cowie and Leherter 1996, Gaizauskas et al. 1997). A core task of IE is identifying named entities and both rule-based and machine learning approaches to IE have been developed in the NLP group.

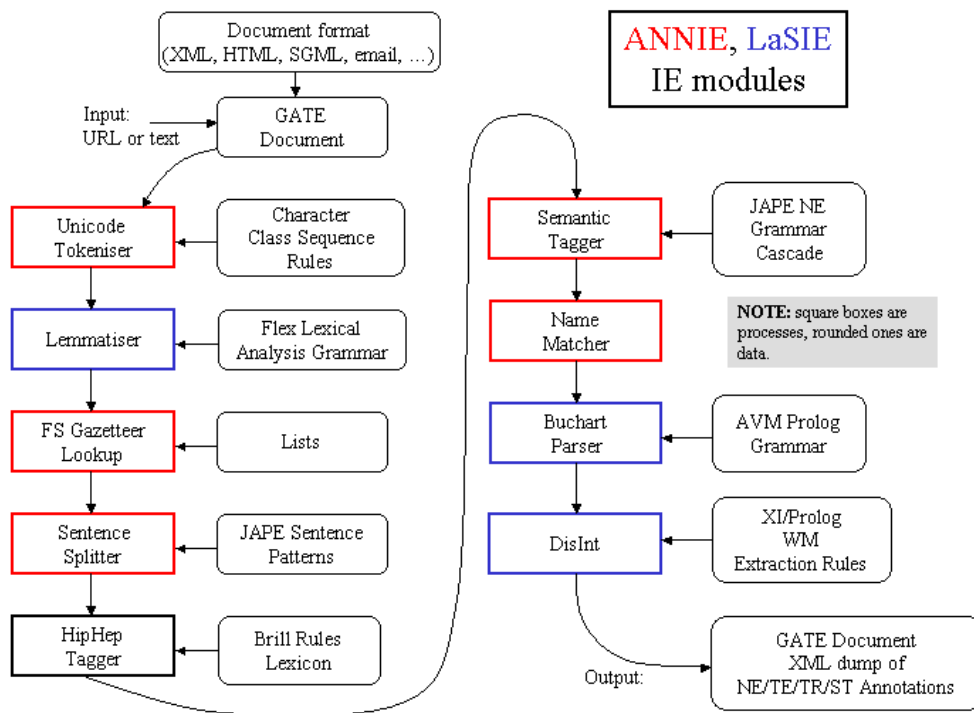


Fig 3. Components of the ANNIE information extraction system (Cunningham 2003)

The GATE⁸ system (Cunningham et al., 2003) has been used for several NER tasks in different domains (e.g. MUSE⁹ and MUMIS¹⁰) and operates a rule-based approach. Manov et al. (2003) showed how a geographic ontology could be integrated within the GATE system to provide a platform for automatic semantic annotation, indexing and retrieval of unstructured (i.e. plain or free-running text) and semi-structured text (e.g. a Web page encoded in HTML). The GATE system provides a framework (in Java) within which to develop custom language engineering applications. GATE provides a Collection of REusable Objects for Language Engineering (CREOLE), a set of resources integrated into GATE and packaged as a Java Archive (or JAR) file and XML configuration data. The system consists of a reusable family of language and processing resources such as ANNIE (A Nearly New Information Extraction system), a default IE system, which includes a tokeniser, gazetteer, sentence splitter, part-of-speech tagger, semantic tagger and a co-reference module (see Fig. 3). Using this IE system we will be able to create our own SPiRiT-specific named-entity recogniser.

⁷ NLP group at Sheffield University: <http://nlp.shef.ac.uk>

⁸ General Architecture for Text Engineering (GATE) is a rule-based IE system: <http://gate.ac.uk/>

⁹ MUlti Source Entity finder: <http://www.dcs.shef.ac.uk/nlp/muse/>

¹⁰ MUlti-Media Indexing and Searching Environment: <http://www.dcs.shef.ac.uk/nlp/mumis/>

GATE provides a grammar called JAPE¹¹, which is compiled into a finite state transducer for pattern matching (similar to `flex` and `lex`). Rules can be defined within terms of entities (or annotations) identified within GATE, which may or may not depend on previous stages in the IE process. Custom entities can be defined within JAPE enabling us to create SPIRIT-specific annotations, e.g. rather than a generic location we could annotate cities, addresses etc. Surrounding context can be defined in terms of entities, word Part-Of-Speech, case information, gazetteer lists and specific keywords, e.g. for geo-references words/phrases such as “north of”, “east of” etc. could be identified and used.

As an alternative/complement to the rule-based approach we also plan to use machine learning methods to automatically learn NE grammars from previously classified texts. This is particularly relevant in the SPIRIT project where we aim to reliably annotate Web pages of varying quality, domain and style. Creating a reliable NE grammar on these kinds of text may lend itself to an adaptive approach rather than one based on a static set of rules. Example machine learning methods used in NER include Hidden Markov Models (or HMMs) (Miller 1998, Yu 1998 and Zhou and Su 2002), decision trees (Sekine 1998, Bennett 1997), and Maximum Entropy (Borthwick 1998, Curran and Clark 2003). The most successful methods so far appear to be HMMs (Zhou and Su 2002) and Maximum Entropy (Curran and Clark 2003). Curran and Clark use a Maximum Entropy tagger to perform language independent NER because the tagger uses features which appear across language. Features include POS tag, a history of preceding and following NEs, orthographic information and gazetteers (only for first and last names). On data for the CoNLL-2003 NE task, Curran and Clark achieve F1 scores of 87.7% (English), 71% (German) and 83.2% (Dutch) for identifying locations. Their approach is currently being used by developers in the Geo-X-Walk¹² project for their geo-parser.

In IE, machine learning methods are used in adaptive IE to create systems, which are more portable and adaptable to a variety of text types. Adaptive IE is also a focus of the NLP group at Sheffield where a number of supervised and unsupervised tools have been developed specifically for information extraction from Web pages and annotation of Semantic Web (SW) sources. In particular, the Amilcare¹³ tool is an adaptive IE system developed to support document annotation in the SW. Machine learning algorithms are used to induce rules to extract information by generalising over examples in the training set. A further tool called Melita¹⁴ provides an interface for human annotators to speed up the annotation process and improve reliability by focusing the user towards the validation of previously unseen rules, which are induced without user intervention by Amilcare. A final tool called Armadillo¹⁵ combines different information sources to exploit redundancy on the Web (i.e. the same information in different superficial formats) and thereby derive annotations with minimal manual intervention. It is possible to use multiple occurrences to bootstrap recognisers that when generalised will retrieve other further information (see, e.g. Dingli et al. 2003).

3.5. Exploiting Web page structure

In most previous work on NER, particularly geographic references, texts used for training and testing are grammatically well structured, of a similar style, from a particular domain and independent from other texts. However, in the SPIRIT project the texts to be annotated are derived from the Web. It is likely that existing methods of NER, which rely on well-formed texts, e.g. rule-based approaches, will be less successful than methods, which are able to adapt to different domains/scenarios. HTML tags, document formatting and ungrammatical language may be found in texts from the SPIRIT collection. However despite these problems, Web pages do exhibit structure in the form of document mark-up, which can be exploited during IE, and in particular in identifying some forms of geographic reference.

¹¹ Java Patterns Annotations Engine (JAPE) – see, e.g. (Maynard, 2003)

¹² The Geo-X-Walk project: <http://hds.essex.ac.uk/geo-X-walk/>

¹³ The Amilcare adaptive IT system: <http://nlp.shef.ac.uk/amilcare/>

¹⁴ The Melita adaptive annotation system: <http://nlp.shef.ac.uk/melita/>

¹⁵ The Armadillo system: <http://www.dcs.shef.ac.uk/~alexiei/WebSite/University/Armadillo/index.html>

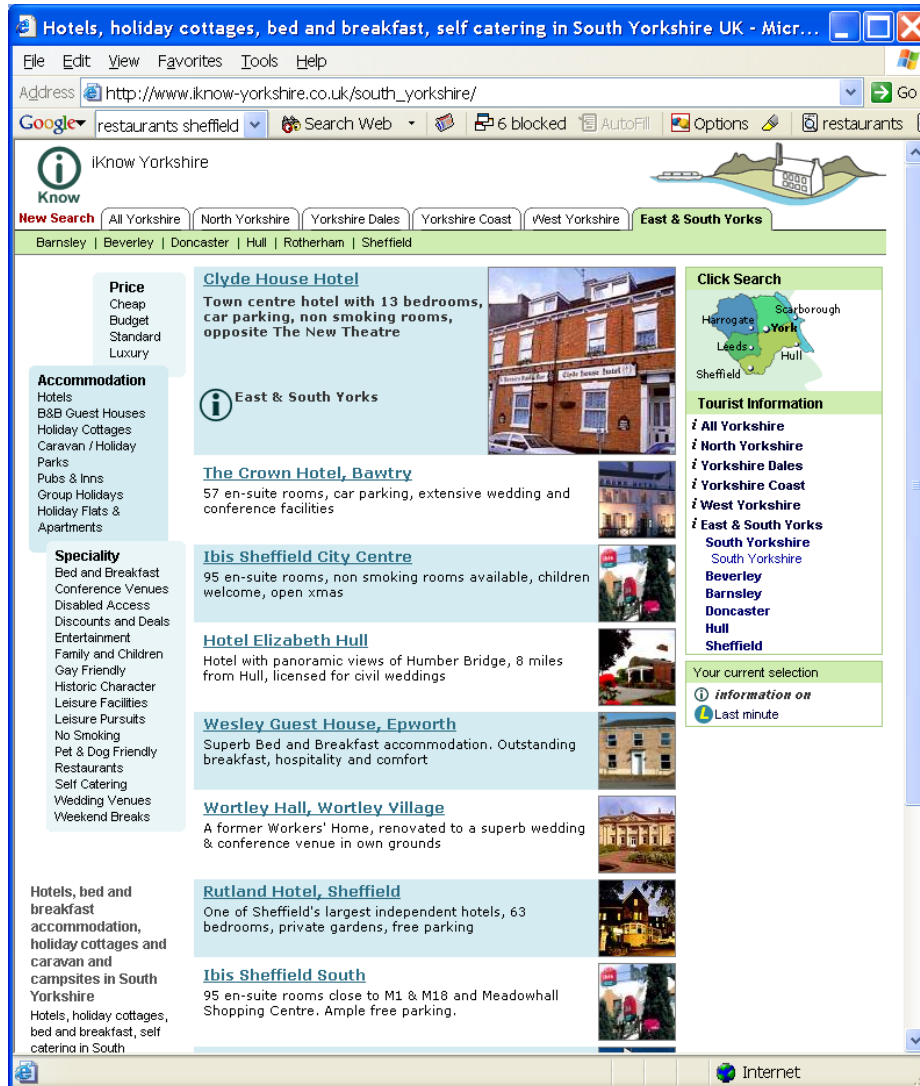


Fig 4. An example list of geographic locations

One method, which has been developed to deal specifically with Internet-derived texts, is wrapper induction (Kusmerik et al.1997). Wrappers are IE systems, which rely heavily on document formatting to induce patterns for IE, e.g. in HTML this includes tables, lists etc. Algorithms such as the (LP)² induction algorithm developed by Ciravegna (2000) and used in Amilcare, Melita and Armadillo exploit document mark-up (as well as use NLP techniques) to induce symbolic rules to generalise across text format, style, domain etc. Borges et al. (2003) use wrapper induction to extract addresses from Web pages. Example pages with addresses annotated by users are used to create general purpose wrappers by deriving patterns based on the structure of the Web page and HTML markup surrounding the address (see, e.g. Fig. 4). A set of regular expressions are induced which can be used to extract further addresses from unseen examples. This form of rule generalization (i.e. extracting rules, which rely not on the address itself but its structure and format) were successfully used to extract data from pages in the same Web source, 65 pages from selected sites about hotels, restaurants, pubs, museums and other cultural attractions. Further papers discussing the extraction of geographic references specifically from Web pages include (Watters and Amoudi, 2002), (McCurley, 2001), (Buyukkokten et al., 1999), (Borges et al., 2003) and (Morimoto et al., 2002).

4. Dealing with geo-references

The stages for marking up a Web site or Web page with a reference to a location within the SPIRIT ontology are described here.

4.1. Identifying geographic references

The first stage is to identify a potential geographic reference within a text. A location may be identifiable in a number of ways; five common ones are:

- the name of the place (e.g. Sheffield);
- an address (e.g. Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK);
- an address fragment (e.g. Ross lived in Dalmeny Street, Edinburgh).
- a postcode on its own (e.g. S1 4DP);
- a phone number (e.g. most Sheffield located phone numbers start with the prefix 0114);

As we saw in Table 1, most previous NER work has been to extract fairly course-grained entities. In the SPIRIT project we plan to create more fine-grained categories, which will identify the specific type of location, e.g. `POSTCODE`, `ADDRESS`, `PLACENAME` (e.g. `CITY`, `COUNTY`), and `PHONENUMBER`. More detailed location classes have been successfully used in previous geographic research, e.g. (Uryupina, 2003) and (Axelrod,2003) and we will decide on the granularity of the classes as we begin to identify locations and after discussion with the requirements of the SPIRIT partners. We plan to experiment with a variety of methods for geoparsing including simple list lookup and rule-based NER to perform identification of locations at whatever granularity we decide.

Using the SPIRIT ontology, we have built a gazetteer lookup tool using `flex` to perform fast pattern matching. A list of candidate geographical references was compiled from 5016 Unique IDs or UIDs (out of 10275) up to (and including) the fourth level¹⁶ of the hierarchy in the ontology (see Table 2). After removing stopwords¹⁷, we extracted 4665 names from which to build a NE recognizer. On an interim collection of 9010 texts (85Mb to index) the recognizer matched 726432 occurrences (76.68 per document) in 6-7 seconds. This is not a completed test collection and therefore we are unable to determine the coverage of locations in documents or the precision of recognition (e.g. names of people matched as locations).

UK354	/United Kingdom/England/London
UK226	/United Kingdom/England/London/the City of London
UK353	/United Kingdom/Wales/Cardiff
UK8937	/United Kingdom/Wales/Cardiff/Trowbridge
UK273	/United Kingdom/England/Sheffield
UK10256	/United Kingdom/England/Sheffield/Broomhill
UK193	/United Kingdom/Scotland/Glasgow
UK10398	/United Kingdom/Scotland/Glasgow/Hillhead

Table 2 Example locations within the forth level of the SPIRIT ontology

¹⁶ Limited locations up to the forth level in the SPIRIT hierarchy because of `flex` restrictions.

¹⁷ Stopwords consist of the most frequent names occurring in the SPIRIT ontology, e.g. over, university, send, link, and college.

From Table 2 we observe the limitations of matching exactly those names found in the ontology, e.g. United Kingdom rather than UK, and “the City of London” rather than “City of London”. However, given comprehensive gazetteers and without limitations on the number of patterns matched, this method is likely to perform well (see, e.g. Axelrod (2003) and (Mikheev et al. 1999)).

Using an IE system would provide us with a framework in which to develop an extraction system for geo-references, and indeed Manov et al. (2003) show how geographic information can be integrated within the GATE system to provide a platform for automatic semantic annotation, indexing and retrieval of unstructured (i.e. plain or free-running text) and semi-structured text (e.g. a Web page encoded in HTML). Using the extensive geographic resources available to us in SPIRIT we expect to obtain good results for geo-parsing using GATE. It is likely that most of the work will be in dealing with poorly structured texts and location disambiguation. After initialisation, entities can be recognised quickly in GATE and the system also has a feature enabling it to be run over a collection of texts in a way which utilises memory consumption thereby making it feasible to use GATE on the 1Tbyte SPIRIT collection. We can also make use of other parts of the GATE system such as the co-reference module, which may provide useful in identifying the extent of a geographic reference.

We plan to also experiment with machine learning approaches based on a training examples from the SPIRIT collection, and adaptive IE methods to extract NEs from structured HTML markup (e.g. using the Amilcare and Armadillo tools).

4.2. Disambiguating the geo-reference

Once identified as a reference, if more than one item in the ontology is potentially referred to, it has to be disambiguated. There are two main ambiguities in geo-references: (1) referent ambiguity and (2) reference ambiguity. The former occurs when the same name is used for more than one location, e.g. in the Getty Thesaurus of Geographic Names (TGN) the name “Chapeltown” refers to a location in South Yorkshire (UK), Lancashire (UK), Kent County (USA) and Panola County (USA). This also includes a geographical reference being used for other entities (e.g. the name of a person or company), which is called referent class ambiguity. Smith and Mann (2003) point out that this is particularly a problem in Britain where surnames are taken from place names, and in the USA where places are named after prominent or obscure people.

Reference ambiguity occurs when the same location can have more than one name, e.g. due to the historical deviation of a location name over time (Smith and Mann, 2003), transliteration (Kwok and Deng, 2003) and implicit formats and character encoding (Axelrod, 2003). Smith and Mann (2003) analysed the TGN for the proportion of places that have multiple names (reference ambiguity) and the number of names, which refer to more than one place (referent ambiguity). While the former is relatively consistent over the continents (e.g. North & Cent. America 11.5%, Asia 32.7%, and Europe 18.2%); the latter is in great variance (e.g. North & Cent. America 57.1%, Asia 20.3%, Europe 16.6%). In the SPIRIT-related task, reference ambiguity is more likely due to variants of a place name rather than historical variation because of the contemporary nature of the SPIRIT ontology. For example, the ontology contains the name “South Yorkshire”, but often this is expressed as “South Yorks”, “S. Yorkshire” etc. From the 10275 standard names in the SPIRIT ontology, 10% of these maps to more than one UID (i.e. need disambiguating). Around 7% of names map to 2 UIDs, 1.5% to 3 UIDs and 0.6% to 4 UIDs.

A great deal of research in disambiguation, particularly word sense disambiguation or WSD (determining what dictionary sense a word in a text refers to), has been conducted in the past (Veronis and Ide, 2000). A survey of WSD techniques used in information retrieval systems was conducted by Sanderson (2000). More recently, Dill et al. (2003) presented research describing identification and disambiguation of references to items in an ontology. With all the disambiguation techniques described in the papers, in order to determine which item an

ambiguous reference refers to, elements of the item's context need to be examined and compared to a pre-stored list of contextual elements that provide clues to the disambiguation system about which item the reference refers to. Dealing with the ambiguity of geo-references has also received much attention; indeed the majority of papers at the HLT/NAACL 2003 workshop on analysis of geographic references discuss this problem. Typical methods, which have been used to disambiguate the referent, include the following (summarised from the HLT/NAACL workshop by Yamaguchi¹⁸):

- **One referent per document** – assume that reference to a place name throughout a document refers to the same location ((Smith and Mann, 2003) and (Leidner et al., 2003)). This is similar to previous WSD work in which senses are assumed to have “one sense per discourse” (Gale et al., 1992).
- **Contextual pattern matching** – look for geographical keywords (e.g. “X” + “city”) and contextual patterns (e.g. “X” miles north of “Y”), which can be used to disambiguate between instances of a geo-reference. For example, see (Li et al., 2003) and (Rauch et al., 2003).
- **Proximity of place names** – assume that a place name geographically closest to other unambiguous place names in a document is more plausible than others. GIS is used to compute physical distance between candidates (see, e.g. (Li et al., 2003) and (Rauch et al., 2003)).
- **Default sense** – use a default location for an ambiguous place name, e.g. the most commonly occurring place (Smith and Mann, 2003), by population of the place name (Rauch et al., 2003) or by semi-automatic extraction from the Web (Li et al., 2003).

Methods to disambiguate referent class ambiguity include (plus methods of NER from section 3):

- **Statistical methods** – use tagged corpora to accumulate statistics of phrases surrounding an ambiguous name to support the judgement of a class (Rauch et al., 2003), e.g. the occurrence of words describing hotels is more likely to indicate an ambiguous term refers to a location rather than a person or company.
- **Contextual patterns** - specific words or phrases surrounding an ambiguous term as external evidence (e.g. “north of X” or “views of Y”) as used by Manov et al. (2003).

Methods to disambiguate reference ambiguity include:

- **Structured and feature rich gazetteers** – the use of hierarchical naming schemes with transitive sub-region capability (Waldinger et al., 2003) (Manov et al., 2003), and name alias (Axelord, 2003) (Deng, 2003).
- **Bootstrapping for gazetteer entry collection** – start with seeds to obtain phrase patterns for geographic types (e.g. city, region, county, and river) and use the statistics for collection and classification of other geo-references from the Internet (e.g. (Uryupina, 2003)).

For SPIRIT, texts marked up with references to known ontology items will either be obtained through manual mark-up or through an automatic bootstrapping process using gazetteers. Here the containment relations often found within gazetteers can be exploited. For example,

¹⁸ http://www.stanford.edu/~shuji/reports/Geo_Reference_Disambiguation.pdf

the Getty TGN states that UK city Sheffield is to be found in the county of South Yorkshire, whereas the city of Sheffield in the United States is the in the state of Alabama. It can be assumed that texts containing the words “Sheffield” and “South Yorkshire” in close proximity to each other hold a reference to the UK city Sheffield and not its US namesake. It is assumed that a sufficient number of texts will always be found to provide an initial list of contextual elements for the disambiguator¹⁹. A list of elements for each item in the ontology is created through a training phase where the contexts of references in texts, already marked up with references to ontology items, are examined. Words, HTML tags, and other Web page properties, commonly found in the context of a reference to a particular item are assumed to be contextual clues to identify that item and are added to that item’s contextual element list in the ontology.



Fig. 5 Context surrounding “Sheffield” in a UK Web page

Fig. 5 shows an example Web page describing Sheffield City Council (UK). This page is would appear to be well constructed and at least four sources of evidence for surrounding context can be used to deduce the location:

- The URL itself (i.e. gov.uk).
- Links to related Web pages (e.g. Sheffield in bloom 2004).

¹⁹ If it turns out this assumption is false and the collection of texts to be marked up has insufficient documents, then either searching on a large web search engine (e.g. Google) for texts, or as a last resort, manually marking up texts will be used. It should also be noted that a further assumption is being made here: that a word found to match with one or more items in the ontology can only refer to the item or items and nothing else. However, “Sheffield” is the name of a company in Auckland, New Zealand (www.sheffield.co.nz). The extent to which such cases will cause problems remains to be seen, it is assumed for now that such cases are exceptional.

- A cluster of locations in close geographical proximity (e.g. Barnsley, Rotherham and Doncaster) and further geographical references such as South Yorkshire in close proximity with Sheffield (i.e. within the same paragraph).
- A “contact us” link, which contains address details, phone numbers and email addresses.

We plan to experiment with a variety of approaches for the SPiRiT disambiguation task. An initial task will be the creation of training set (section 8) in which will provide manually annotated and disambiguated geo-references. This will enable us to determine disambiguation problems in the SPiRiT collection and focus our approach. An important step will be determining a default geographic location for a given place name (i.e. default word sense) and will also act as a baseline in our evaluations.

Effectiveness of the approach

Word sense disambiguators that use such a fully automatic approach are known to determine word senses correctly much more often than incorrectly. However, a substantial number of errors are still made (10%-20%) including the disambiguation of geographic references. Smith and Mann (2003) train Naive Bayes classifiers on a variety of text types and achieve 87% accuracy on two months of news stories from the AP newswire and 69% for a collection of historical documents. Watters and Amoudi (2002) achieve 80% success in assigning locations to URLs from a random selection of 100 links generated by the Yahoo Random Link Generator²⁰. Despite proposed methods of disambiguation in past research, there appears to be very little evaluation of disambiguation of geographic references, particularly on a “standard” test collection.

It is anticipated that errors in this generally untested domain will be lower: many of the errors in sense disambiguation are due to words whose senses are closely related. For example, the word French can refer to the French people or the French language, both senses can be expected to appear in similar contexts and therefore be relatively hard to disambiguate. Experiments in disambiguation evaluation exercises, such as SENSEVAL, have confirmed this (see Kilgariff, 2000). Geographic terms tend to be un-related: Sheffield, South Yorkshire and Sheffield, Alabama are likely to appear in different contexts and so be easier to distinguish than a word like “French”.

4.3. Extent of a reference

Once a reference in a text has been located and, if required, disambiguated, it will be necessary to determine, the extent of that reference: how much of the rest of the text on the Web page or Web site that the reference is found on is itself referring to aspects of the identified location. Based on discussions with Mirago²¹ and an internal analysis of Web pages, it was determined that two main types of geographic references appear to exist: simple in-text references to locations; and “Contact us” Web pages.

In-text references

The default treatment of ontology items referred to in texts will be to assume the extent of the reference to cover the entire Web page it is found on. An alternative approach will be to assume the extent covers a window of text around the reference. The size of the passage maybe determined by the presence of references to any other ontology items found within the same text and by preventing overlapping passages. Which of these possibilities is better for SPiRiT’s geographic searching will be determined in testing of the prototype.

“Contact us” pages on Web sites

A common feature of the Web sites of many organisations is the “Contact us” Web page, which holds contact details for the organisation including a postal address. With such pages, it would be wrong to limit the extent of the address’s reference to a passage of text or the page it

²⁰ Yahoo Random Link Generator: <http://random.yahoo.com/fast/ryl>

²¹ Mirago: <http://www.mirago.co.uk/>

appears on. Such references need a wider extent, exactly how much wider will be determined in testing of the SPIRIT prototype. The entire Web site the “Contact us” page is found on or pages linking to the contact page are anticipated to be the appropriate extents of this form of reference. For example, Fig. 6 shows the “Contact us” Web page of Sheffield City Council as linked to from the page in Fig. 5.

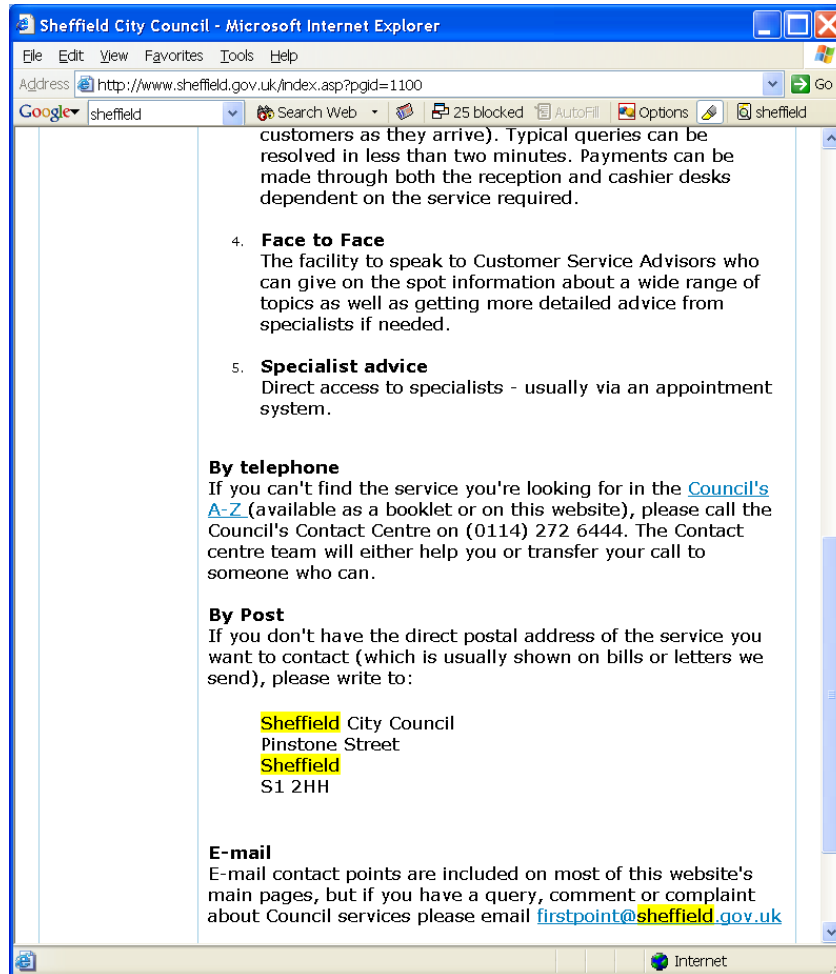


Fig. 6 The “Contact us” link for Sheffield City Council

In order to work with such pages, it will be necessary to build a “Contact us” Web page classifier. As with disambiguation, building such a system requires training data. It is anticipated that Mirago will provide such data in the form of a large list of company Web sites (some 200,000) from which the geographic location of the company has already been determined mainly through manual classification. The armadillo tool as described in Dingli et al. (2003), and in Ciravegna (2001) will be used to learn how to recognise such Web pages from the training data.


5. Geographic resources

Various geo-reference lists will be used in conjunction with the SPIRIT ontology and we discuss a variety of resources, which cover locations in the UK, as well as those worldwide. These resources will be used to identify named entities, as well as provide additional information to aid the disambiguation of geo-references.

5.1. Getty Thesaurus of Geographic Names (TGN)

TGN²² is a compiled resource, which contains more than 1 million names and other information about places across all continents (including political and historical places). The emphasis in TGN is on art and architecture. Each place is represented by a unique ID (UID) and linked to the record for each place is its location within a geographical hierarchy (i.e. parent and child geographic locations), the longitude and latitude of the location, and additional information such as historical name variants (see, e.g. Fig. 7). We have bought a license to use this resource in SPIRIT.

ID: 7010692 **Record Type:** [administrative](#)

 **Sheffield (inhabited place)**

Coordinates:
 Lat: 53 23 00 N *degrees minutes* Lat: 53.3833 *decimal degrees*
 Long: 001 30 00 W *degrees minutes* Long: -1.5000 *decimal degrees*

Note: Located near River Don; center for famous cutlery since 14th cen.; metallurgical innovations by Henry Bessemer in 19th cen. made town famous for steel; food-processing & clothing also important; University of Sheffield known for programs in metallurgy.

Hierarchical








-  [World](#) (facet (hierarchical))
-  [Europe](#) (continent)
-  [United Kingdom](#) (nation)
-  [England](#) (country)
-  [South Yorkshire](#) (county)
-  [Sheffield](#) (unitary authority)
-  [Sheffield](#) (inhabited place)

Fig. 7 TGN record for Sheffield, South Yorkshire, UK

5.2. Code-Point®

Code-Point is produced by Ordnance Survey and provides a National Grid reference for each unit postcode in Great Britain. We have access to this data through Digimap, a service provided to selected academic institutions by Edinburgh University Data Library (EDINA²³). Each co-ordinate expressed within Code-Point is stated within a resolution of one metre. The

²² Getty TGN: http://www.getty.edu/research/conducting_research/vocabularies/tgn/

²³ EDINA Digimap: <http://digimap.edina.ac.uk/main/download.jsp>

postcode data is sourced, from among other resources, Address Point, which contains 26 million addresses recorded in the Royal Mail Postcode Address List (PAF). The entire postcode information can be downloaded and used for research use only at the University of Sheffield.

5.3. Gazetteer Plus

Also available from the EDINA service, this product (also from Ordnance Survey) provides 260,000 names in Great Britain from the current and previous years' 1:50,000 Scale gazetteer. The same place names can be found on Ordnance Survey's Landranger® map series and we are able to download the entire gazetteer for research use within the University of Sheffield. An example entry is:

264:SK3282:Abbeydale:SK28:53:20.3:1:30.7:382500:432500:W:SP:Sheff:Sheffield:O:01-MAR-1993:l:110:111:0

5.4. GEOnet Names Server (GNS)

The GEOnet Names Server²⁴ (GNS) provides access to the National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names' (US BGN) database of foreign geographic feature names. The database is the official repository of foreign place name decisions approved by the US BGN. Approximately 20,000 of the database's features are updated monthly. The GNS contains 4.5 million names (approximate) with 3.97 million features covering locations worldwide (excluding the United States and Antarctica – the GNIS covers these regions). The coordinate system for data served by GNS is WGS84 and the data can be downloaded (either the entire database, or names for each country) and stored locally.

5.5. Geographic Names Information System (GNIS)

The Geographic Names Information System²⁵ (GNIS) contains information about almost 2 million physical and cultural geographic features in the United States and its territories. The Federally recognised name of each feature is described and references are made to a feature's location by State, county, and geographic coordinates. The GNIS is America's official repository of domestic geographic names information. This data can be downloaded and used locally.

6. Geo-reference mark-up format

Once references to items in the SPIRIT ontology have been determined, disambiguated, and their extent defined, information about these items needs to be recorded for other parts of the SPIRIT system to use. An XML format has been created for this purpose, a sample of which is shown below. A key point about the mark-up is that it will not be inserted into the texts of the pages being analysed, instead it will be stored as a separate *stand-off mark-up* file; one file per page marked up.

```
<DOC>
<DOCID>SPRT-xxxx-xxxx-xxxx</DOCID>
<MKDUPDATE>Thu Dec 18 14:15:11 GMT 2003</MKDUPDATE>
<GEO>
  <TERM type="string">London</TERM>
  <OFFSET type="int">66</OFFSET>
```

²⁴ GNS: <http://earth-info.nga.mil/gns/html/>

²⁵ GNIS: <http://geonames.usgs.gov/>

```
<LENGTH type="int">12</LENGTH>
<TGNID type="int">200030303</TGNID>
<TGNCONF type="float">50.0</TGNCONF>
<ONTID type="string">UK125</ONTID>
<ONTCONF type="float">50.0</ONTCONF>
<EXTENT type="string">Sentence</EXTENT>
  </GEO>
<GEO>
...
</GEO>
</DOC>
```

The Web page being referred to is identified by the <DOCID> tag. This is followed by a list of <GEO> tags that list all of the identified references to the SPIRIT ontology. The location of the identified reference is specified by the <OFFSET> and <LENGTH> tags. The ID of the item in the TGN and SPIRIT ontologies is identified by the <TGNID> and <ONTID> tags respectively. If the disambiguator is unable to determine which item in the SPIRIT ontology is referred to more than one ID tag may appear. The <TGNCONF> and <ONTCONF> tags will reflect the confidence of the disambiguator in its choice of one item over another. The extent of the reference is defined in the <EXTENT> tag. If the extent of a reference is judged to range beyond an individual Web page, the <GEO> tag for that reference will be duplicated in the other Web pages it is judged to extend to.

By separating the new (geo) mark-up from the pages being annotated, it is easy to manage multiple versions of mark up, which while the annotation tools for SPIRIT are being developed will be a valuable feature to have. Documents that are marked up are stored along side the collection.

7. Building a geo-reference training set

Using a sample of Web pages derived from the 1Tbyte Web collection held at Sheffield, we will create a training and evaluation set. This will be used to analyse geographic references found within Web pages, and to train and evaluate methods for named entity recognition and disambiguation of place names. We intend to identify all geo-references within the test collection (including organisations), which can be used to test the coverage and accuracy of the NER. Further to this we will disambiguate ambiguous geo-references to be able to test the disambiguation method. One issue to consider is that the SPIRIT ontology only contains references to European locations; therefore we will need to decide whether to disambiguate only these locations, or include only European Web pages in the training set. However it is more likely that we will find ambiguity between locations *across* countries (e.g. UK and USA). Within the SPIRIT ontology we find that only 10% of the names are ambiguous and unless a more detailed ontology that spans countries is produced we will not run into a great deal of ambiguity.

To start building a training set, we have generated an interim collection 9010 documents (approximately 85MB of text) using the following manner:

1. Choose a set of queries
2. Do a search for the query set
3. Record the top N documents from each query
4. Generate a document pool from the previous stage and remove any duplicates
5. Fetch the full-texts for the unique documents
6. Index the fetched texts

For the interim1 collection, the query set was compiled from the names of the top 200 largest (by population) cities and towns in UK²⁶. For each of the 200 names (or query), the top 50 documents were recorded, although several names did not return a full of 50 documents. The current interim collection contains a variety of texts and we may wish to focus on a particular domain or Web site for further evaluation.

8. Summary

In summary, this SPiRiT task will involve the following work:

1. Create a suitable test collection based on the SPiRiT collection.
 - a. Decide on how to sample the data.
 - b. Decide on the level of granularity for geo-reference annotations.
 - c. Manually identify locations within texts.
 - d. Manually disambiguate locations (with respect to the SPiRiT ontology).
2. Investigate various methods of NER and evaluate performance.
 - a. Compute the precision and recall of NER methods
 - b. Determine the impact of gazetteers on NER
3. Investigate methods of geo-reference disambiguation
 - a. Evaluate various methods of disambiguation
 - b. Exploit Web pages and structure for disambiguation
4. Automatically annotate texts in the SPiRiT collection.
 - a. Identify named entities and annotate
 - b. Disambiguate and geocode locations and businesses
 - c. Determine the extent of geo-references

The main research areas of this SPiRiT task are:

1. To create a test collection for evaluation of geoparsing and geocoding evaluation based on the SPiRiT collection. We plan to create a publicly-available resource for the research community, providing a standardised resource for geographic name recognition and disambiguation.
2. To compare various methods of geoparsing and geocoding approaches and determine their success on Web data. To date most evaluation has been on either newswire and historical texts, or small sets of Web pages.
3. To investigate how the structure and interconnectivity of Web pages can be exploited for both geoparsing and geocoding.
4. To determine which “features” are best for identification of named entities (i.e. to use within the NE grammars), and as context for disambiguation.

²⁶ http://www.citymayors.com/gratis/uk_topcities.html

9. References

- Appelt, D.E. (1996) The Common Pattern Specification Language. Technical report, SRI International, Artificial Intelligence Center, 1996.
- Axelrod, A. E. (2003) On building a high performance gazetteer database. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 63-68.
- Bennett S.W., Aone C. and Lovell C. (1997) Learning to Tag Multilingual Texts through Observation. In *Proceedings of the Second Conference on Empirical Methods in NLP*, 109-116.
- Borges, B., Laender, A., Bauzer Medeiros, C., Silva, A. and Davis Jr., C. (2003) The Web as a Data Source for Spatial Databases. In *Proceedings of V Brazilian Geoinformatics Symposium GEOINFO 2003*.
- Borthwick, A., Sterling, J., Agichtein, E. and Grishman, R. (1998) NYU: Description of the MENE Named Entity System as Used in MUC-7, In *Proceedings of the MUC-7*.
- Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L. and Shivakumar, N. (1999) Exploiting geographical location information of Web pages. In *Proceedings of Workshop on Web Databases (WebDB'99)* held in conjunction with ACM SIGMOD'99, June 1999.
- Carreras, X., Màrquez, L. and Padró, L. (2002) Named entity recognition using AdaBoost. In *Proceedings of the 2002 CoNLL Workshop*, Taipei, Taiwan, 167-170.
- Ciravegna, F. (2001) Adaptive Information Extraction from Text by Rule Induction and Generalisation. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*.
- Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.
- Curran, J.R. and Clark, S. (2003) Language Independent NER using a Maximum Entropy Tagger. In *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, 164-167.
- Cowie, J. and Lehnert, W. (1996) Information Extraction. *Communications of the ACM*, 39(1), 80-91.
- Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., and Zien, J.Y. (2003) SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the WWW Conference*.
- Dingli, A., Ciravegna, F., Wilks, Y. (2003) Automatic Semantic Annotation using Unsupervised Information Extraction and Integration, In *Proceedings of the K-CAP 2003 Workshop on Knowledge Markup and Semantic Annotation*
- Gaizauskas, R. Humphreys, K., Azzam, S. and Wilks, Y. (1997) Conception vs. Lexicons: An Architecture of Multilingual Information Extraction. In Paziienza, M. (Ed) *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. LNCS, Springer-Verlag, Vol 1299. 28-43.
- Gale, W., Church, K. and Yarowsky, D. (1992) One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Harriman, NY, February 1992, 233-237.

- Ide, N. and Veronis, J. (1998) Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, Computational Linguistics.
- Kilgarriff, A., Palmer (2000) Special Issue of the Journal *Computers and the Humanities*, 34(1-2), Kilgarriff and Palmer (eds).
- Kwok, K. L. and Deng, Q. (2003) "GeoName: a system for back-transliterating pinyin place names". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 26-30.
- Leidner, J. L., et al. (2003) Grounding spatial named entities for information extraction and question answering. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 31-38.
- McCurley, S.K. (2001) Geospatial mapping and navigation of the web. In *Proceedings of the Tenth International WWW Conference*, Hong Kong, 1-5 May, 221-229.
- McDonald, D. (1996) Internal and external evidence in the identification and semantic categorisation of proper names. In B. Boguraev and J. Pustejovsky (Eds) *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA, chapter 2, 21-39.
- Mikheev A., Moens M. and Grover C. (1999) *Named Entity recognition without gazetteers*. In *Proceedings of the Annual Meeting of the European Association for Computational Linguistics EACL'99*, Bergen, Norway, 1-8.
- Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R. and the 1998 Annotation Group (1998). Algorithms that Learn to Extract Information; BBN: Description of the SIFT System as Used for MUC-7. In *Proceedings of the MUC-7*.
- Larson, R.R. (1996) Geographic Information Retrieval and Spatial Browsing. In *GIS and Libraries: Patrons, Maps and Spatial Information*, Linda Smith and Myke Gluck, Eds., University of Illinois.
- Li, H., et al. (2003) InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 39-44.
- Malouf, R. (2002) Markov models for language-independent named entity recognition. In *Proceedings of the 2002 CoNLL Workshop*, Taipei, Taiwan, 187-190.
- Manov, D., Kiryakov, A. and Popov, B. (2003) Experiments with geographic knowledge for information extraction. In: Kornai, A. and Sundheim, B. (Eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 1-9.
- Morimoto, Y., Aono, M., Houle, M. E. and McCurley, K. S. (2003) Extracting Spatial Knowledge from the Web, In *Proceedings of 2003 Symposium on Applications and the Internet (SAINT 2003)*, 27-31 January 2003, Orlando, FL, USA, 326-333.
- Sanderson, M. (2000) Retrieving with good sense. *Information Retrieval*, 2(1), 49-69.
- Stevenson, M. and Gaizauskas, R. (2000) Improving Named Entity Recognition using Annotated Corpora. In *Proceedings of the LREC Workshop: "Information Extraction meets Corpus Linguistics"*, Athens, Greece.
- Rauch, E., et al. (2003) A confidence-based framework for disambiguating geographic terms. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 50-54.

- Riloff, E. and Jones, R. (1999) Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 474-479.
- Sekine, S. (1998) Description of the Japanese NE System Used for MET-2, In *Proceedings of the MUC-7*.
- Smith, D. A. and Mann, G. S. (2003) Bootstrapping toponym classifiers. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 45-49.
- Tablan, V. (2001) Software Architecture for Language Engineering. Presentation at Eurolan-2001 conference: <http://www.racai.ro/EUROLAN-2001/page/resources/profs/Tablan/SALE/eurolan2001/>
- Yu, S., Bai, S. and Wu, P. (1998) Description of the Kent Ridge Digital Labs System Used for MUC-7. In *Proceedings of the MUC-7*.
- Uryupina, O. (2003) Semi-supervised learning of geographical gazetteers from the internet. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 18-25.
- Waldinger, R., et al. (2003) Pointing to places in a deductive geospatial theory. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 10-17.
- Watters, C. R. and Amoudi, G. (2003) Geosearcher: Location-based Ranking of Search Engine Results, *Journal of American Society for Information Science*, 54(2), 140-151.
- Zhou, G. and Su, J. (2002) Named Entity Recognition Using a HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, Philadelphia, July 2002, 473-480.

Appendices

Appendix I. Report on HTL-NAACL 2003 Workshop on Analysis of Geographic References

1. Introduction

1.1. Workshop overview

I attended a workshop on analysis of geographical references to learn the state-of-the-arts advance in this area, and exploit the potential inputs to SPIRIT project. The workshop was held as a part of HLT-NAACL 2003 Conference, Edmonton, Canada, on 31st May 2003. It was co-chaired by Andras Kornai (Metacarta Inc.) and Beth Sundheim (US Navy SPAWAR Systems Center). Eleven papers, one demo system, and two invited talks were presented.

The first session of the workshop focused on the semantics of geographic references such as feature types, ontologies, and disambiguation, and the second session focused on system and gazetteer development.

More information about the workshop can be accessed from the homepage:

<http://kornai.com/NAACL/>

1.2. Four conceptual stages

The workshop defined four conceptual stages in the analysis of geographical references, which would help us to get an overview of this workshop and papers presented. They are:

1. geographic entity reference detection,
2. contextual information gathering,
3. disambiguation of entities, and
4. grounding of identified entities.

The first stage involves the extraction of geographic entities from texts. The second stage aims to help identify the feature type and approximate location of geographic entities. The third stage is the actual disambiguation of the entities in terms both of type and location. The fourth stage plays a role of assigning the geographical coordinates to the disambiguated entities.

The first and third stage can be very similar in a real system. However, at least in the context of this workshop, the first stage includes the process of separating a geographic entity from other types of named entities (e.g. people or organizations) that are found in texts, while the third stage focused on those which are identified as a geographic reference.

1.3. The rest of the report

This report first provides an overview of each paper presented in the workshop, followed by the discussions highlights the points that are particularly related to SPIRIT project. The report

concludes with a recommendation. Appendix I contains a comprehensive list of tools and resources found in the proceedings and those that have been actually used in research projects or products. Appendix II contains the abstracts of all papers and demo presented in the workshop for reference.

2. Papers overview

2.1. Knowledge Based Ontology Design

Manov et al. [1] introduced an ontology, which is a part of their knowledge and information management (KIM) platform. The platform provides the infrastructure and services for automatic semantic annotation, indexing and retrieval of unstructured and semi-structured content. Their ontology is represented by Resource Description Framework (RDF), and contains about 60000 locations grouped into 6 continents, 27 global regions, 282 countries, all country capitals and 4700 cities. Each location has several aliases (in different languages).

The relationships used in their ontology is a modified version of Alexandria Digital Library feature type thesaurus. In addition, features such as *CountryCapital*, and more transitive *subRegionOf* (i.e. more specific than *part-of*) has been added. They also regard the cities populated over 100000 as a set of 'important' cities. Their geographical knowledge is from multiple sources such as NIMA's GEOnet Names Server (GNS) data, Federal Information Processing Standards (FIPS), UN Statistics site, etc. Their disambiguation is mainly based on the combination of lookup of their data sources.

Their evaluation on disambiguation using UK news articles revealed that their ontology is very sensitive to the data source of their geographical knowledge. An ad-hoc attempt of identification of unknown names had little help. Nevertheless the details of their ontology design (and the screenshot in the appendix of the paper) might be of our interest.

2.2. Design of gazetteer database

Axelord's paper [10] describes a design of gazetteer database, which takes multiple data sources. GazDB, a core data unit for MetaCarta system, consists of three components such as conversion scripts, database, and export scripts. A set of conversion scripts enables GazDB to store the original data that are organised in different formats, and the export scripts are used to provide the data structure of a gazetteer, which can be customized for a different request.

The paper provides a number of practical advices on designing a gazetteer database from the basic relation structure, three different representations of an entry such as spatial, functional, and administrative, to an issue of character encoding, deletion/update of records, authoritativeness of entries. A consideration for a temporally-sensitive gazetteer is also given.

2.3. Deriving geographical gazetteers

Uryupina [3] presented an approach to the automatic generation of a gazetteer that covers feature types such as city, island, river, mountain, country, and region using 1260 seed names randomly sampled from World Atlases. Her approach is simple and interesting, and overall performance seems acceptable.

Her approach submits a seed name as a query to AltaVista and fetches the top 100 pages. Extract a string that contains the seed name and two words before and after the name from the pages. For each name, the top 20 patterns are ranked based on the frequency and a score from Ripper classifier previously trained with 520 names. Submit the patterns as a query to AltaVista again and evaluate the performance on the top 200 page.

Overall, her system works well for island (96.4% of accuracy), river (89.6%), mountain (88.8%), and country (99.6%), but marginal for region (82.6%) and less impressive for city (62.0%).

2.4. Word sense disambiguation

Almost all papers presented in the workshop involved with some degree an issue of disambiguation of geographical references. Among them, there are two papers that address this issue in particular.

The first one is by Li et al. [6]. Their approach is a hybrid of a gazetteer, pattern matching, and graph search algorithm. Although much of the paper discuss and report the graph search algorithm, they also commented that “the default senses play a significant role in location normalization”. Their previous paper ([15]) describes an approach to determine the default sense using the co-occurrence data from a search engine.

Their experiment shows that the default sense achieves 90% of accuracy while the combination with other components outperforms it by 6%. A similar result has been found in IR. Sanderson and van Rijsbergen [14] show that the frequency of occurrence of a word tends to be skewed to its primary sense. This highlights the importance of determining the default sense of geographical references.

The second paper is by Smith and Mann [7]. They examined Naïve Bayes classifiers to find a U.S. state or country to which a place belongs, using three different corpora (i.e. one News archive and two historical documents). The training was carried out both on a tagged and non-tagged version of the corpora of the historical documents, and only non-tagged one was used for the News archive.

The evaluation shows that the classifier works at an acceptable accuracy of 87.38% for the non-tagged News with 87.10% for seen names and 69.72% for new names (i.e. did not appear in the training data). However it was found that the historical documents were much harder to disambiguate even if the tagged data was used for training. In particular the performance of new names was as low as 9.38% for Civil War corpora (tagged). It is likely that the training on the tagged data could cause the performance. Also the News archive was found to be easier to disambiguate.

They provide an interesting survey based on Getty, which counts the number of places with multiple names and names applied to more than one place. While the former is relatively consistent over the continents (e.g. North & Cent. America 11.5%, Asia 32.7%, Europe 18.2%), the latter is in great variance (e.g. North & Cent. America 57.1%, Asia 20.3%, Europe 16.6%).

2.5. Spatially-aware search engines

There are two papers that present a searching mechanism specifically designed for spatially-aware queries.

The first one is by Rauch, et al. [8] who present a confidence-based framework used in MetaCarta system (<http://www.metacarta.com/>). Geographical confidence is based on several criteria. They are 1) a default sense of an entity derived from a training corpus, 2) local context such as an entry followed by ‘city of’ giving a positive score and ‘Mr’ giving a negative score, 3) textual proximity as described below, and 4) population of the entity provided by a gazetteer.

An interesting comment found in their paper is about a correlation between the geographical distance and proximity in texts: “We have found that there is a high degree of spatial correlation in geographic references that are in textual proximity. This applies not only to points that are nearby, such as Madison and Milwaukee, but also to the situation when points are enclosed by regions, e.g. Madison and Wisconsin. (p.52)”

The final ranking of documents based both on textual relevance and geographical confidence is calculated in a simple way:

$$Score = (1 - W)R_g + W \cdot R_w$$

where W is an IDF-based term weighting for a query, R_g is a relevance based on the geographical confidence, and R_w is a relevance based on BM25-like scoring. Instead of document length, though, they normalise the score based on the number of geographical references found in a document. See Robertson, et al. [16] for the detail of IDF and BM25.

Another paper is by Bilhaut, et al. [9], which is based on passage retrieval using a small tagged corpus consisting of 200 pages in Education. This paper is relatively hard to read on, but the sample queries shown in the paper might be of our interest.

- Educational difficulties in west of France since the 50's
- Variations of the number of pupils in secondary school in Paris area
- Variations of the number of pupils in rural areas
- Transfers of the teaching staff to southern districts

One of the interesting characteristics of their work is to address the issues of *zone* and *quantification* (in addition to types) expressed in a query. For example, *some* seaboard towns, the *quarter of* districts of *north of* France, or *big* cities in France. "Some", "quarter of", or "big" are the examples of quantification, and "north of" is an example of zone.

They propose some heuristics to rank a set of retrieved passages based on the analysis of these three aspects, but no experiment is given. Also it is not clear if many users are willing to provide the zone and quantity in their queries.

2.6. Visualising geographic references on map

Leidner et al. [5] presented a work to visualise the geographic references associated with news articles on a map. They follow a development of an event in a news article. For example, an article was describing a story about a baby who was flown to Scotland to have a surgery. The story contains Scotland, Tooting, London, Glasgow, London, Glasgow, Northolt, etc. as the event was developed. They then look up a gazetteer (UN/LOCODE) to determine a set of references, which will yield the minimum area by their coordinates.

By plotting the area determined by the geographic references they could visualise the area involved in the story on the map using Generic Map Tools (GMT). The GMT and MapIt is free software that allows us to plot dots or polygon on the map by providing latitudes/longitudes. These tools might be useful for our development of SPIRIT system although we may have a more sophisticated tool and map at a later stage.

2.7. Back-transliteration of geographic entities

Kwok and Deng [4] discussed an issue of transliteration between Chinese and English. This back-transliteration means to retrieve the original Chinese characters from geographical references written in English. For example, *Beijin* could be written in 460 possible ways in Chinese. Although this could be a specific issue between English and Asian languages, a part of their approach to determine the correct Chinese character might be of our interest.

They submit the English description (e.g. *Beijin*) as a query to Google and retrieve only the documents written in a Chinese encoding such as *GB*. If a candidate name from a set of pre-filtered list is found, give a weight for the name. Alternatively they also mentioned to submit the

English and a candidate Chinese, and simply weight them based on the number of documents retrieved for the pair.

2.8. Descriptive gazetteers and markup encoding

Southall [11] introduces a project that focused on marking-up several descriptive gazetteers published in the mid-19th century in UK. A descriptive gazetteer is a kind of encyclopedia for places. He explains in detail his method to extract and relate geographic entities found in these semi-structured texts.

He also pointed out that there are several standards for marking-up geographical references in free texts, but little is used. He showed how to mark-up texts using an extended version of Text Encoding Initiative (TEI), which was adopted in the project.

3. Discussion

3.1. One sense in one discourse

The empirical notion of “one sense in one discourse” (Gale et al. [13]) was mentioned several times as a reasonable rule to accept (e.g. [4] and [5]). Following this rule, the same sense of a name can be used throughout a document once it is disambiguated, hence, we can focus on a case where we have more confidence to determine for a name.

Also the importance of determining the default sense of a geographical reference was highlighted by Li et al. [6] who showed 90% of accuracy solely on it. Although more research will be needed to evaluate their approach, a simple co-occurrence analysis using a gazetteer and a large document collection could be seen as a good starting point.

3.2. Ranking algorithm

Clearly the document ranking algorithm that considers geographical information was not the central topic in the workshop.

Rauch, et al. [8] was the only paper, which discussed an integration of the geographical information into relevance ranking. However the effective use of textual and geographical relevance needs to be exploited in further studies. As they indicate there might be a high correlation between textual proximity in a document and geographical distance. This would help making use of the geographical footprint of documents in our relevance ranking. One also should examine the correlation of rankings based on textual proximity with the one based on geographical attributes.

3.3. Ontology design

Manov et al. [1] and Axelord [10] discussed several practical aspects of designing and building a geographical ontology from multiple data resources. It was interesting that while Axelord [10] tries to be as general and flexible as possible to adopt many sources written in a different structure, Manov et al. [1] attempt to gather only *important* entities, features, or relations. Obviously how to determine the importance of data is an open question and they, for example, used the population information for an entity.

Although it is likely that the design of ontologies is strongly determined by its primary purpose, the notion of *subRegionOf* (as opposed to *part-of* relation) discussed in Manov et al. [1] seems to be useful in the context of SPiRiT project.

3.4. Descriptive gazetteer

Lastly, although this could be less relevant to SPiRiT, I felt there was some scope for developing an approach to automatic generation of a descriptive gazetteer based on a set of

retrieved documents. As shown in Southall [11] and a short discussion with one of the authors of Densham and Reid [12] it was evident that the descriptive gazetteers have been generated manually for a limited region.

However a short description of unfamiliar geographical references in a document might help users in their information access process (e.g. "What is the Winter Garden in Sheffield?" – "Winter Garden is one of the largest temperate glasshouses to be built in the UK during the last hundred years"). Such a question is not necessary be their initial and main interest of search but could be emerged during the process.

4. Conclusion and Suggestion

4.1. Conclusion

This report started with an overview of HLT-NAACL 2003 Workshop on Analysis of Geographical References, and summarized individual papers presented. Some significant points were highlighted in the discussion to link the findings and SPiRiT project.

4.2. Workshop for GIS and IR

As the preface of the proceedings stated, the GIS community was missing from the workshop. Also I felt that the workshop was covering only limited aspects of the issues that have been addressed by SPiRiT Project. Thus, we are currently working on a potential workshop based on SPiRiT Project in SIGIR 2004, which is going to be held at Sheffield. This workshop will be aimed to facilitate the exchange of ideas and technologies among the research fields such as IR, GIS, NLP, and other related areas.

5. References

Note that the abstract of the papers presented in the workshop is available in Appendix II.

[1] Manov, D., et al. (2003) "Experiments with geographic knowledge for information extraction". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 1-9, Alberta, Canada: ACL.

[2] Waldinger, R., et al. (2003) "Pointing to places in a deductive geospatial theory". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 10-17, Alberta, Canada: ACL.

[3] Uryupina, O. (2003) "Semi-supervised learning of geographical gazetteers from the internet". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 18-25, Alberta, Canada: ACL.

[4] Kwok, K. L. and Deng, Q. (2003) "GeoName: a system for back-transliterating pinyin place names". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 26-30, Alberta, Canada: ACL.

[5] Leidner, J. L., et al. (2003) "Grounding spatial named entities for information extraction and question answering". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 31-38, Alberta, Canada: ACL.

[6] Li, H., et al. (2003) "InfoXtract location normalization: a hybrid approach to geographic references in information extraction". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 39-44, Alberta, Canada: ACL.

[7] Smith, D. A. and Mann, G. S. (2003) "Bootstrapping toponym classifiers". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 45-49, Alberta, Canada: ACL.

- [8] Rauch, E., et al. (2003) "A confidence-based framework for disambiguating geographic terms". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 50-54, Alberta, Canada: ACL.
- [9] Bilhaut, F., et al. (2003) "Geographic reference analysis for geographic document querying". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 55-62, Alberta, Canada: ACL.
- [10] Axelrod, A. E. (2003) "On building a high performance gazetteer database". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 63-68, Alberta, Canada: ACL.
- [11] Southall, H. (2003) "Defining and identifying the roles of geographic references within text: Examples from the Great Britain Historical GIS project". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 69-78, Alberta, Canada: ACL.
- [12] Densham, I. and Reid, J. (2003) "A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service". In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 79-80, Alberta, Canada: ACL.
- [13] Gale, W. A., Church, K. W., and Yarowsky, D. (1992) "One sense Per Discourse". In: *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, 233-237.
- [14] Sanderson, M., van Rijsbergen, C. J. (1999) "The impact on retrieval effectiveness of skewed frequency distributions." *ACM Transactions on Information Systems* 17(4): 440-465.
- [15] Li, H., Srihari, R. K., Niu, C., Li, W. (2002) "Location Normalization for Information Extraction". In: *Proceedings of COLING 2002*, Taipei, Taiwan.
- [16] Robertson, S. E. and Sparck Jones, K. (1997). "Simple, proven approaches to text retrieval". Technical Report. London, Department of Information Science, City University. Available from <http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/>. [Accessed 25/06/2003].

Appendix I. Tools and Resource

General Resource List

- GIS WWW Resource List <http://www.geo.ed.ac.uk/home/giswww.html>

Automatic Mapping Tool

- MapIt <http://stellwagen.er.usgs.gov/mapit/>
- Generic Mapping Tools (GMT) <http://gmt.soest.hawaii.edu/>

Working Geo Search Engine

- MetaCarta <http://www.metacarta.com/>
- GeoXwalk <http://www.geoXwalk.ac.uk/>

Gazetteers

- Alexandria Digital Library Gazetteers <http://www.alexandria.ucsb.edu/>
- UN/LOCODE <http://www.unece.org/cefact/locode/service/main.htm>
- Online Britain Historical Gazetteers <http://www.qbhgis.org/>

Descriptive Gazetteers

- Descriptive Gazetteers for Scotland <http://www.geo.ed.ac.uk/scotgaz/>

Disambiguation System

- InfoXtract http://www.cymfony.com/tech_infox.html

- Perseus Digital Library <http://www.perseus.tufts.edu/>

Knowledge and Information Management (KIM) infrastructure

- Sesame (RDF(s) repositories) <http://sesame.aidnavigator.nl>
- OMM (Ontology Middleware Module) <http://www.ontotext.com/omm>
- BOR (DAML+OIL reasoner) <http://www.ontotext.com/bor>

Standards

- Ontology Web Language (OWL) <http://www.w3.org/owl-semantics/>
- DARPA Agent Markup Language (DAML) for spatial data <http://www.daml.org/listarchive/daml-spatial/0005.html>
- Text Encoding Initiative (TEI) <http://tei-c.org/>

Appendix II. Abstract of presented papers and demo**Manov, D., et al. "Experiments with geographic knowledge for information extraction"**

Here we present work on using spatial knowledge in conjunction with information extraction (IE). Considerable volume of location data was imported in a knowledge base (KB) with entities of general importance used for semantic annotation, indexing, and retrieval of texts. The semantic Web knowledge representation standards are used, namely RDF(S). An extensive upper-level ontology with more than two hundred classes is designed. With respect to the locations, the goal was to include the most important categories considering public and tasks not specially related to geography or related areas. The location data is derived from number of publicly available resources and combined to assure best performance for domain-independent named-entity recognition in texts. An evaluation and comparison to high performance IE application is given.

Waldinger, R., et al. "Pointing to places in a deductive geospatial theory"

Issues in the description of places are discussed in the context of a logical geospatial theory. This theory lies at the core the system GeoLogica, which deduces answers to geographical questions based on knowledge provided by multiple agents.

Uryupina, O. "Semi-supervised learning of geographical gazetteers from the internet"

In this paper we present an approach to the acquisition of geographical gazetteers. Instead of creating these resources manually, we propose to extract gazetteers from the World Wide Web, using Data Mining techniques.

The bootstrapping approach, investigated in our study; allows us to create new gazetteers using only a small seed database (1260 words). In addition to gazetteers, the system produces classifiers. They can be used online to determine a class (CITY, ISLAND, RIVER, MOUNTAIN, REGION, COUNTRY) of any geographical name. Our classifiers perform with the average accuracy of 86.5%.

Kwok, K. L. and Deng, Q. "GeoName: a system for back-transliterating pinyin place names"

To be unambiguous about a Chinese geographic name represented in English text as Pinyin, one needs to recover the name in Chinese characters. We present our approach to this back-transliteration problem based on processes such as bilingual geographic name lookup, name suggestion using place name character and pair frequencies, and confirmation via a collection of monolingual names or the WWW. Evaluation shows that about 48% to 72% of the correct names can be recovered as the top candidate, and 82% to 8% within top ten, depending on the process employed.

Leidner, J. L., et al. "Grounding spatial named entities for information extraction and question answering"

The task of named entity annotation of unseen text has recently been successfully automated with near-human performance. But the full task involves more than annotation, i.e. identifying the scope of each (continuous) text span and its class (such as place name). It also involves grounding the named entity (i.e. establishing its denotation with respect to the world or a model). The latter aspect has so far been neglected.

In this paper, we show how geo-spatial named entities can be grounded using geographic coordinates, and how the results can be visualized using off-the-shell software. We use this to compare a "textual surrogate" of a newspaper story, with a "visual surrogate" based geographic coordinates.

Li, H., et al. "InfoXtract location normalization: a hybrid approach to geographic references in information extraction"

Ambiguity is very high for location names. For example, there are 23 cities named 'Buffalo' in the U.S. Based on our previous work, this paper presents a refined hybrid approach to geographic references using our information extraction engine InfoXtract. The infoXtract location normalization module consists of local pattern matching and discourse co-occurrence analysis as well as default senses. Multiple knowledge sources are used in a number of ways: (i) pattern matching driven by local context, (ii) maximum spanning tree search for discourse analysis, and (iii) applying default sense heuristics and extracting default senses from the web. The results are benchmarked with 96% accuracy on our test collections that consists of both news articles and tourist guides. The performance contribution for each component of the module is also benchmarked and discussed.

Smith, D. A. and Mann, G. S. "Bootstrapping toponym classifiers"

We present minimally supervised methods for training and testing geographic name disambiguation (GND) systems. We train data-driven place name classifiers using toponyms already disambiguated in the training text -- by such existing cues as "Nashville, Tenn." or "Springfield, MA" -- and test the system on texts where these cues have been stripped out and on hand-tagged historical texts. We experiment on three English-language corpora of varying provenance and complexity: newsfeed from the 1990s, personal narratives from the 19th century American west, and memories and records of the U.S. Civil War. Disambiguation accuracy ranges from 87% for news to 69% for some historical collections.

Rauch, E., et al. "A confidence-based framework for disambiguating geographic terms"

We describe a purely confidence-based geographic term disambiguation system that crucially relies on the notion of "positive" and "negative" context and methods for combining confidence-based disambiguation with measures of relevance to a user's query.

Bilhaut, F., et al. "Geographic reference analysis for geographic document querying"

The work presented in this paper concerns Information Retrieval from geographical documents, i.e. documents with a major geographic component. The final aim, in response to an informational query of the user, is to return a ranked list of relevant passages in selected documents, allowing text browsing within them. We consider in this paper the spatial component of the texts and the queries. The idea is to perform an off-line linguistic analysis of the document, extracting spatial expressions (i.e. expressions denoting geographical localisations). The point is that such expressions are (in general) much more complex than simple place names. We present a linguistic analyser, which recognises them, performing a semantic analysis and computing symbolic representations of their "content". These representations, stored in the text thanks to XML annotation, will act as indexes of passages with which queries are compared. The matching of queries with text expressions is a complex process, needing several kinds of numeric and symbolic computations. A prospective outline of it is described.

Axelrod, A. E. "On building a high performance gazetteer database"

We define a data model for storing geographic information from multiple sources that enables the efficient production of customizable gazetteers. The GazDB separates names from features while storing the relationships between them. Geographic names are stored in a variety of resolutions to allow for i18n and for multiplicity of naming. Geographic features are categorized along several axes to facilitate selection and filtering.

Southall, H. "Defining and identifying the roles of geographic references within text: Examples from the Great Britain Historical GIS project"

Reliably recognizing, disambiguating, normalizing, storing, and displaying geographical names poses many challenges. However, associating each name with a geographical point location cannot be the final stage. We also need to understand each name's role within the document, and its association with adjacent text. The paper develops these points through a discussion of two different types of historical texts, both rich in geographic names: descriptive gazetteer entries and travellers' narratives. It concludes by discussing the limitations of existing mark-up systems in this area.

Densham, I. and Reid, J. "Demo: A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service"

We describe a basic Geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. The development of a geo-parser comes from the need to explicitly georeference large resource collections such as the Statistical Accounts of Scotland, which currently only contains implicit georeferences in the form of place names thus making such collections inherently geographically searchable.