Experiments with geographic knowledge for information extraction

Dimitar Manov,
Atanas Kiryakov,
Borislav Popov
Ontotext Lab, Sirma AI Ltd
38A Christo Botev Blvd, Sofia 1000,
Bulgaria
{mitac,naso,borislav}@sirma.bg

Kalina Bontcheva,
Diana Maynard,
Hamish Cunningham
University of Sheffield
Regent Court, 211 Portobello St.,
Sheffield S1 4DP, UK
{kalina,diana,hamish}@dcs.shef.ac.uk

Abstract

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 text. 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 locations data is derived from number of publicly available resources and combined to assure best performance for domain-independent named-entity recognition in text. An evaluation and comparison to high performance IE application is given.

1 Introduction

Information Extraction (IE) research has focused mainly on the recognition of course-grained entities like Location, Organization, Person, etc. (Sundheim, 1998). The application of Information Extraction to new areas like the Semantic Web and knowledge management has posed new challenges, from which the most relevant here is the need for finer-grained recognition of entities, such as locations.

In this paper we present some experiments with building a reusable knowledge base of locations which is used as a component into an IE system, instead of a location gazetteer. This work is part of the Knowledge and Information Management (KIM) platform and still undergoing development and refinement.

With respect to coverage, the goal was to include the most important location categories for a wide range of applications and tasks, not specially related to geography or related areas. The locations data is derived from a number of publicly available resources and combined to assure best performance for named-entity recognition. An evaluation and comparison to high performance IE system using very small location gazetteers is given.

One important aspect of our work is that we choose to create a knowledge base of locations, structured according to an ontology and having relations between them, instead of having somewhat flat structures of gazetteer lists found in other IE systems. While a knowledge base can be plugged into an IE system instead of a flat gazetteer, it also has several unique advantages:

- the extra information, especially the transitive *subRegionOf* relation can be used for disambiguation and reasoning
- the location entities in the text can be recognised at the right level of granularity for the target application (i.e., as Location or as Country, City, etc).
- the ontology and knowledge base can be modified by the user and any changes are reflected immediately in the output of the IE system.

The paper is structured as follows. Section 2 puts our work in the context of previous research. Section 3 presents briefly the KIM platform, which contains the IE system and the location knowledge base. Then Section 4 describes the location knowledge base in more detail. The IE experiments are discussed in Section 5, followed by a discussion on problems and future work. The paper concludes by showing how such a knowledge base can be used to bootstrap a new IE system (Section 7).

2 Related work

In the context of this paper, the two most relevant areas of work are on large-scale gazetteers and location disambiguation. Here we present the Alexandria Digital Library Gazetteer because we used the ADL Feature Type Thesaurus as a basis of our location ontology. Related work on location disambiguation, like the one done in the Perseus Digital Library project, is relevant because in future work we will improve the location disambiguation mechanism in our system.

2.1 Alexandria Digital Library Gazetteer

The Alexandria Digital Library (ADL), an NSF-funded project at the University of California, Santa Barbara, has included gazetteer development from its beginning in 1994. Currently it contains approximately 4.4 million entries. The data is taken from various sources, including NIMA (National Imagery and Mapping Agency's of United States) Gazetteer, a set of countries and U.S. counties, set of U.S. topographic map quadrangle footprints, set of volcanoes, and set of earthquake epicenters. The Geographic Names Information System (GNIS) data from the U.S. Geological Survey has been partly added to the collection. The results as of today include thesaurus for feature types, Time Period data for the historical entries and spatial data with boundaries. The boundaries are defined as "satisficing" rectangles. The term "satisficing" is described in (Hill, 2000), and additional information about the project could also be found there as well as on the ADL gazetteer development page at http://alexandria.sdc.ucsb.edu/l̃hill/adlgaz/.

2.2 Toponym-disambiguation in Perseus Digital Library project

A disambiguation system for historical place names for Perseus digital library is described in (Smith and Crane, 2001). The library is concentrated on representing historical data in the humanities from ancient Greece to nineteenth-century America. The authors present a procedure for disambiguation of such place names, based on internal and external evidence from the text. Internal evidence includes the use of honorifics, generic geographic labels, or linguistic environment. External evidence includes gazetteers, biographical information, and general linguistic knowledge. Evaluation of the performance of the system is given, using standard precision/recall methods for each of the five corpora: Greek, Roman, London, California, Upper Midwest. The system is best on Greek and worst on Upper Midwest corpus, and its overall performance for place names is higher than the most of other applications.

3 The KIM platform

The KIM Platform provides a novel Knowledge and Information Management (KIM[1]) infrastructure and services for automatic semantic annotation, indexing and retrieval of unstructured and semi-structured content. The ontologies and knowledge bases are kept in Semantic

[1]KIM, see http://www.ontotext.com/kim

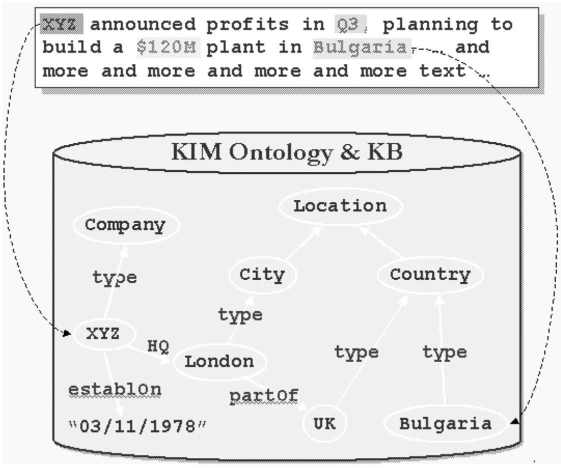

Figure 1: KIM Platform

repositories based on cutting edge Semantic Web technology and standards, including RDF(S) repositories[2], ontology middleware[3] (Kiryakov et al, 2002) and reasoning[4]. It provides a mature infrastructure for scalable and customizable information extraction as well as annotation and document management, based on GATE (Cunningham et al., 2002). GATE, a General Architecture for Text Engineering, is developed by the Sheffield NLP group and has been used in many language processing projects; in particular for Information Extraction in a variety of languages (Maynard and Cunningham, 2003).

An essential idea for KIM is the semantic (or entity) annotation, depicted on figure 1. It can be seen as a classical named-entity recognition and annotation process. However, in contrast to most of the existing IE system, KIM provides for each entity reference in the text (i) a pointer (URI) to the most specific class in the ontology and (ii) pointer to the specific instance in the knowledge base. The latest is (to the best of our knowledge) an unique KIM feature which allows further indexing and retrieval of documents with respect to entities.

For the end-user, the usage of a KIM-based application is straightforward and simple - one can highlight text in the browser and further explore the available knowledge for the entity, as shown in figure 3. A semantic query web user interface allows for queries such as "Organization-

[2]Sesame (http://sesame.aidministrator.nl/) is an open source RDF(S)-based repository and querying facility.
RDF, http://www.w3.org/RDF/. Resource Description Framework is an open standard for knowledge exchange over the Web, developed by W3C (www.w3.org).

[3]OMM, http://www.ontotext.com/omm. Ontology Middleware Module is an enterprise back-end for formal knowledge management.

[4]BOR, http://www.ontotext.com/bor/, is a DAML+OIL reasoner, compliant with the latest OWL specifications.

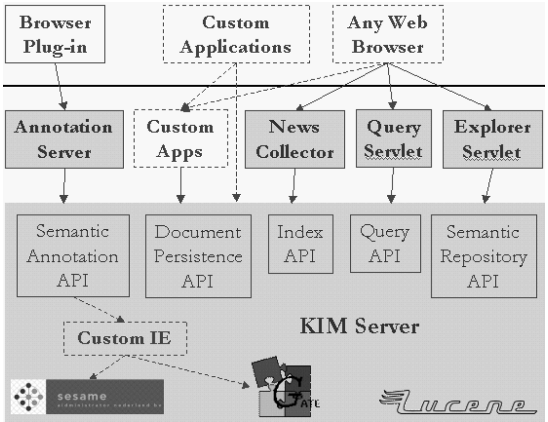

Figure 2: KIM architecture.

locatedIn-Country" to be executed.

Information retrieval functionality is available, based on Lucene[5], which is adapted to measure relevance to entities instead of tokens and stems. The full architecture is shown in figure 2. It is important to note that KIM as a software platform is domain and task independent.

3.1 The ontology

KIM Ontology (KIMO) covers the most general 250 classes of entities and 40 relations. The main classes are *Entity*, *EntitySource* and *LexicalResource*. The most important class in the ontology is *Entity*, further specialized into *Object*, *Abstract* and *Happening*. *LexicalResource* class and its subclasses are used for different IE-related information. The instances of the *Alias* class represent different names of instances of *Entity*. *hasAlias* relation is used to link *Entity* to its aliases (one-to-many relation). The *hasMainAlias* links to the main alias (the official name). Each instance of *Entity* is linked to an instance of *EntitySource* via *generatedBy* relation. There are two types of *EntitySource* - *Trusted* and *Recognized*. The "trusted" entities are those pre-defined. The recognized are the ones which were recognized from text as part of the IE tasks.
The upper part of the ontology can be seen on the same figure 3 in the left frame.
For ontology representation we choose RDF(S), mainly because it allows easy extension to OWL[6] (Lite).

Location sub-ontology

Because the Geographic features (Locations) form a large part of the entities of general importance, we developed a *Location* sub-ontology as part of the KIM ontology. The goal was to include the most important and frequently used types of Locations (which are specializations of *Entity*), including relations between them (such as *hasCapital*, *subRegionOf* (more specific than *part-of*)), relations between Locations and other Entities (Organization *locatedIn* Location) and various attributes.

The Location entity denotes an area in 3D space[7], which includes geographic entities with physical boundaries, such as geographical areas and landmasses, bodies of water, geological formations and also politically defined areas (e.g. "U.S. Administered areas").

The classification hierarchy (consisting of 97 classes) is based on the ADL Feature Type Thesaurus version 070203. The differences target simplicity; a number of distinctions and unnecessary levels of abstraction were removed where irrelevant to general (non-geographic) context, as we wanted the ontology to be easy to understand for an average user. Examples of sub-classes omitted: Territorial waters, Tribal areas, Administrative Areas (its sub-types are put directly under Location).

The Location ontology provides the following additional information:

- the exact type of a feature, for example to be able to recognize a geographic feature as *CountryCapital* instead of just *Location*.

- relations between geographic feature and other entities (e.g. "Diego Garcia" is a *MilitaryBase*, located somewhere in the Indian Ocean and it is *subRegionOf* USA).

- the different names of a location ("Peking" and "Beijing" are two aliases for one location).

- the transitive *subRegionOf* relation allows one to search for Entities located in a continent (e.g. "Morgan Stanley" - locatedIn - "New York" - subRegionOf - "NY" - subRegionOf - "USA" - subRegionOf - "North America")

- "trusted" vs "recognized" sources in *generatedBy* property of a Location is an extra hint in disambiguation tasks. The class hierarchy is shown in figure 5.

[5] Lucene, http://jakarta.apache.org/lucene/, high performance full text search engine

[6] Ontology Web Language (OWL), http://www.w3.org/TR/owl-semantics/

[7] Actually, the instances of Location are Entities with spatial identity criteria (Guarino and Welty, 2000). For instance a building can be considered as Property, Location or Cultural Artifact, but the focus in the ontology is placed on the Location aspect.

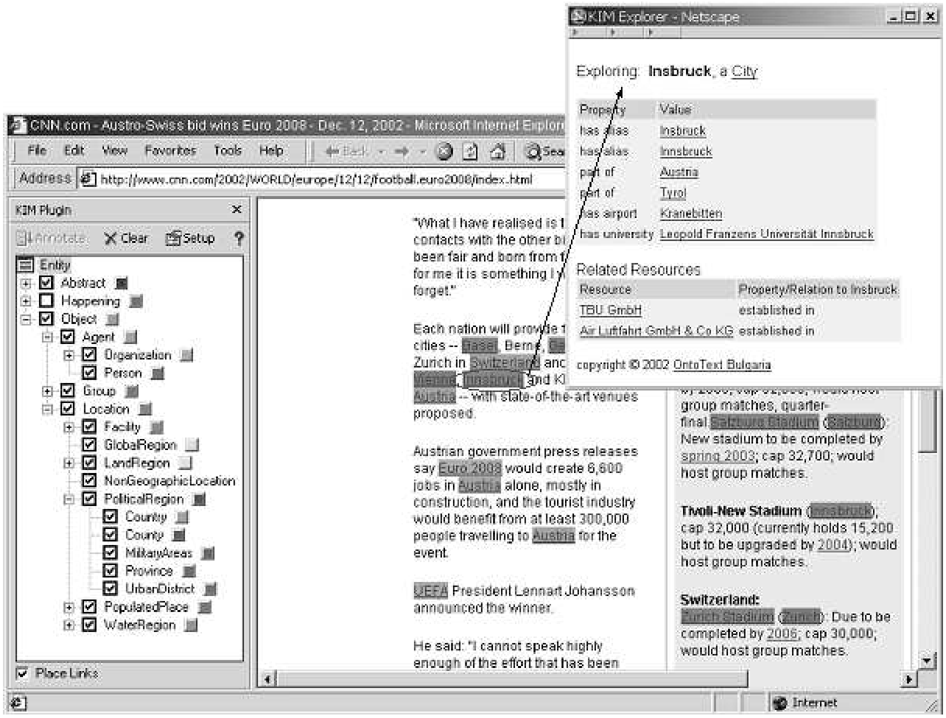

Figure 3: KIM usage - highlight and explore. The upper part of KIM ontology (KIMO) is shown in the left frame.

3.2 The knowledge base

Geographic information usually introduces a high level of ambiguity between named entities, for the following three reasons:

- there could be several Locations with the same name (this includes sharing common alias);
- a name of a Location could match a common English word (e.g. "Has", "The");
- other named entities (Company, Person, even Date or Numeric data) could share a common alias with a Location (examples: "*Paris* Corporation", "O'Brian" county, "10" district, "Departamento de Nueve de Julio" with alias "9 de Julio").

In order to allow easy bootstrapping of applications based on KIM and to eliminate the need for them to write a Geo-gazetteer, the KIM knowledge base provides exhaustive coverage of entities of general importance. By limiting the Locations to only "important" ones, we also keep the system as generic, domain- and task-independent as possible. The term "importance" of a location is hard to define, and part of the problem is that it is dependent on the domain where the IE tasks are focused. Yet it is common sense that such locations include continents, countries, big cities, some rivers, mountains, etc. In addition to the above predefined locations, KIM:

- learns from the texts it analyses;
- has a comprehensive set of rules and patterns helping it to recognize unknown entities;
- has a Hidden Markov Model learner, capable of correcting symbolic patterns.

As a test domain, KIM uses political and economic news articles from leading newswires[8].

4 Populating the location knowledge base

As a main source of geographic knowledge we used NIMA's GEOnet Names Server (GNS) data. GNS database is the official repository of foreign place-name decisions approved by the U.S. Board on Geographic Names (US BGN) and contains approximately 3.9 million features with 5.37 million names. Approximately 20,000 of the database's features are updated monthly. The data is available for download in standard formatted text files, which contain: unique feature index (UFI), several names per Location (the official name, short name, sometimes different transcriptions of the name), geographic coordinates (one point; no bounding rectangle). Geographic coverage of the data is worldwide, excluding United States and Antarctica. For U.S. geographic

[8]See News Collector, http://news.ontotext.com

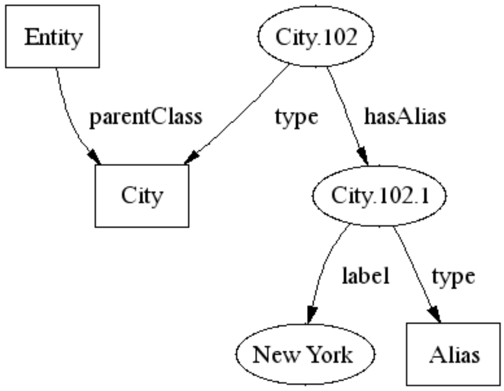

Figure 4: RDF representation of a *Location*.

data we used partially USGS/GNIS data[9], which follows similar format as GNS data. For country names we followed FIPS[10], which was natural choice since GNS data is structured that way. A list of big cities was obtained from UN Statistics site, which covers city data (http://unstats.un.org/unsd/citydata/).

We then created a mapping between our location classes and GNS feature designators. Some of the features were completely ignored (e.g. "abandoned populated places", "drainage ditch"), other were combined into one (e.g. "ADM2", "ADMD" into *County*).

There is some inconsistency in the way the data is entered for different countries, mostly because of improper usage of designators (using different designators for similar geographic features and vice versa). This made creation of the mapping a bit harder, as we needed to include more designators mapped to one class. The per-country files were almost consistently entered (with some exceptions, for example in UK, "England", "Scotland", "Northern Ireland" and "Wales" are entered as AREA, which hints the same importance as the other 40 areas in UK). We expect that a per-country mapping instead of a global one will lead to better performance results, yet we haven't experimented with this as it will require manual tuning for about 250 countries.

The different names of the geographic features are mapped to aliases of the Location entities, with a main alias pointing to the official name. The RDF representation of a Location is shown in figure 4. Because these names sometimes match common English words and Person names a list of stop words is created and the aliases are filtered.

The import procedure uses the mapping described above but can also be restricted by list of countries and classes to be imported. Currently imported classes are: *Continent*, *GlobalRegion*, *Country*, *Province*, *County*, *CountryCapital*, *LocalCapital*, *City*, *Ocean*, *Sea*, *Gulf*, *OilField*, *Monument*, *Bridge*, *Plateau*, *Mountain*, *MountainRange*, *Plain*. These classes were selected as "important", based on common sense and statistical information derived from GNS data.

The GNS data has three main problems when it comes to extracting only geographical entities of global importance and the relations between them:

- There is no way to tell the importance of a location (e.g. is Chirpan a big city or a small town);
- The only part-of relations available are between a location and its country, but not province or county;
- Some locations are not country-specific (e. g. oceans, seas, mountains) but are listed as separate locations with different identifiers in different per-country lists.

We addressed the first problem by limiting the types of locations to a small subset of important ones (as explained above). The importance of cities was determined by using a list of all big cities (with population over 100,000). We attempted to solve the second problem by using an algorithm to calculate the distance between a location and all provinces/counties in this country, and then to create a part-of relation with the nearest one. However, our experiments showed that the accuracy of the results was not satisfactory. This is mostly due to the fact that in GNS data only the location footprint is given, but not the extent. Comparing the geographic coordinates of the locations with a common alias and type and then combining the matching ones into a single entity in the knowledge base solved the third problem.

Currently the KB contains about 50,000 Locations grouped into 6 Continents, 27 GlobalRegions (such as "Caribbean" or "Eastern Europe"), 282 Countries, all country capitals and 4,700 Cities (including all the cities with population over 100,000). Each location has several aliases (usually including English, French and sometimes the local transcription of the location), geographic coordinates, the designator (DSG) and Unique Feature Index (UFI), according to GNS. The figures for entities of global importance in KIM KB are shown in table 1.

5 Experiments with direct use for IE

The locations KB is used for Information Extraction (IE) as part of the KIM system, combining symbolic and stochastic approaches, based on the ANNIE IE components from GATE. As a baseline, using a gazetteer module, the aliases of the entities (including all locations) are

[9] US Geological Survey (UGCS); Geographic Names Information System (GNIS)

[10] Federal Information Processing Standards, http://www.itl.nist.gov/fipspubs/

Entities	77,561
Aliases	110,308
Locations	49,348
Cities	4,720
Companies	7,906
Public companies	5,150
Key people	5,500
Organizations	8,365

Table 1: Instances per subclass of Entity.

being looked up in the text. Further, unknown or not precisely matching entities are recognized with pattern-based grammars:

- using location pre/post keys to identify locations, e.g. ”The River Thames”
- using location pre/post keys + Location, e.g. ”north Egypt”, ”south Wales”
- context-based recognition, such as: ”in” + Token-with-first-uppercase Number of disambiguation problems (mostly in the case of Location names occurring in the composite name of other Entities) are also detected and resolved:
- ambiguity between Person and Organization, e.g. ”U.S. Navy” (this would normally be recognized as a Person name from the pattern ”two initials + Family name”, but in this case the initials match a location alias)
- occurrence of locations in person names, e.g. ”Jack London” (disambiguated because in the KB there is *LexicalResource* ”Jack” is a first name of Person)
- occurrence of locations in Organization names, e.g. ”Scotland Yard” (disambiguated because in the KB there is such Organization)

Finally, some of the recognized Entities (including Locations), which are not marked as noun by the part of speech tagger are discarded.

Some of the newly recognized Locations appear frequently in the analyzed texts. Those, which could be found in the GNS data are potential candidates to be entered in the knowledge base, because there is an extra evidence for their importance. This is a way to extend the knowledge base and make it contain all the ”important” Locations in the sense of frequently used in the one or more application domain(s).

The performance of the KIM system was measured on a news corpus using GATE’s evaluation tools. The system was also compared to an high-precision named entity recognition system, which uses small flat gazetteer lists.

Entity	Number
Location	792
Organisation	773
Person	764
Date	603
Percent	54
Money	94

Table 2: Distribution of entities in the corpus

5.1 Evaluation Corpus

The corpus was collected from 3 online English newspapers: the Independent, the Guardian and the Financial Times. In total it contains 101 documents with 56,221 words. The corpus was manually annotated with entities. Table 2 shows the number of entities of each type in the corpus.

5.2 Corpus Benchmark Tool

The Corpus Benchmark Tool(CBT) is one of the components in GATE which enables automatic evaluation of an application in terms of Precision, Recall and F-measure, against a set of ground truths. Furthermore, it also enables two versions of a system to be compared against each other (e.g. for regression testing) or two different systems to be compared. Each system is evaluated by comparing the annotations produced with a set of key annotations (produced manually) and producing a score – two systems can therefore be compared with each other and indications are given as to where they differ from each other.

5.3 MUSE

MUSE is an information extraction system developed within GATE which aims to perform named entity recognition on different types of text (Maynard et al, 2002). MUSE recognises the standard MUC entity types of Person, Location, Organisation, Date, Time, Percent, and some additional types such as Addresses and Identifiers. The system is based on ANNIE, the default IE system within GATE, but has been extended to deal with a variety of text sources and genres, and incorporates a mechanism for automatically selecting the most appropriate set of resources depending on the text type.

MUSE uses flat-list gazetteers which primarily contain contextual clues that help with the identification of named entities, e.g., company designators (such as Ltd, GmbH), job titles, person titles (such as Mr, Mrs), common first names, typical organisation types (e.g., Ministry, University). In addition, MUSE has lists enumerating concrete types of locations which have about 27 500 entries, including 25,000 UK ones. Further breakdown is given in Table 3:

global regions (including continents)	71
aliases of countries	450
provinces	1215
mountains	5
water regions (oceans, lakes, etc)	15
cities world wide	1900
UK regions (such as East Sussex, Essex)	140
cities in UK	23792
UK rivers	3

Table 3: MUSE Location gazetteer entries

As can be seen from the location entries in the MUSE gazetteers, the system is specifically tailored to recognise UK locations with high recall and precision, whereas the KIM locations KB is not skewed towards any particular country.

We ran the MUSE system over our test corpus to see how KIM matched up to it.

5.4 Results

MUSE vs KIM performance comparison is given in table 4. When interpreting these results one also must bear in mind that the high-performance IE system is only tagging geographical entities as locations, whereas the GNS-based system is actually disambiguating them with respect to their specific type (e.g., City, Province, Country). Investigation of the reasons behind the lower recall shows that:

- the KB is too coarse-grained, i.e., there are no "smaller" locations, such as small towns/counties in UK, we do not import military bases in KB from GNS data ("Diego Garcia"), etc.
- The application was not specifically tuned for the corpus/news texts, e.g. we do not use the fact, that the texts often clarify the locations when they are first mentioned (e.g., Aberdeen, UK).
- there are not any historical Locations, such as "Soviet Union".

It is expected that the first two problems will be fixed with enhancement of the KB with regard to domain targeting of a KIM-based application. To check this assumption we did another experiment. Because the corpus contains a lot of UK-related information (the articles are from three English newspapers) and MUSE is specifically tailored to UK locations, we needed extra UK-specific information in the KB. As we mentioned earlier the import procedure is flexible to the extend that allowed to add all the locations from UK GNS data. The performance of this enhanced KB is shown in table 5. The recall is higher than in MUSE (increased to 95% vs 93%).

The precision is 10% behind MUSE (85% vs 95%). An obvious reason is that we have more entities in KB, and we do not control the aliases (except for stop words list), while all the locations in MUSE gazetteer lists are manually entered and therefore produce very little ambiguity.

6 Discussion

We produced a KB of locations with world wide coverage using GNS data. The size of about 50,000 Location is more than most other IE systems have. It is not big (compared to 4M locations in ADL Gazetteer), but provides good coverage of Locations (91%). Because the KB was not tuned for the test corpus specifics we could expect similar coverage for other corpora.

Our flexible import procedure allows for domain-targeted versions of the KB (by means of importing more Location types) to be produced, which is expected to have good-enough coverage on locations.

The impact of the location KB on the IE performance is still under evaluation and improvement. We are working on improvements in two directions: i) decreasing the amount of GNS-data entered in KB - for both locations and their aliases; ii) changing the way in which the IE system uses the KB to improve precision. On the latter, we are currently experimenting with applying the regular named entity recognition grammars first and then using the location KB to lookup only the unclassified entities, instead of using it as a gazetteer prior to named entity recognition as we do now.

7 Bootstrapping IE for new languages from the KB

We were able to make use of the KB as part of the TIDES Surprise Language Exercise, a collaborative effort between a number of sites to develop resources and tools for various language engineering tasks on an unknown language. A dry run of this program took place in March 2003, whereby participants were given a week from the time the language was announced, to collect tools and resources for processing that language. The language chosen was Cebuano, spoken by 24% of the population in the Phillipines. The University of Sheffield developed a Named Entity recognition system for Cebuano, to which we contributed a list of locations from the Philippines. This was particularly useful as this kind of information was not readibly available from the Internet, and time was of the essence. The NE system (developed within a week) achieved scores for the recognition of locations at 73%

System	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
MUSE	744	9	54	37	0.947	0.928	0.937
KIM	726	24	61	113	0.855	0.910	0.881

Table 4: MUSE vs KIM performance comparison

System	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
MUSE	744	9	54	37	0.947	0.928	0.937
KIM-UK	759	28	27	167	0.810	0.950	0.874

Table 5: MUSE vs KB with all UK locations

Precision, 78% Recall and 76% F-measure. We predict that this kind of information will be very useful for the full Surprise Language Program in June, where participants will have more time (a month) to create resources on another surprise language – not only for Information Extraction but also for tasks such as Cross-Language Information Retrieval and Machine Translation.

8 Conclusion and future work

This paper presented work on the creation of a locations knowledge base and its use for information extraction. In order to allow easy bootstrapping of IE to different languages and applications, we are building a knowledge base (KB) with entities of general importance, including geographic locations. The aim is to include the most important and frequently used types of Locations. An evaluation and comparison to high performance IE application was given.

The system is still under development and future improvements are envisaged, mainly related to implementing better disambiguation techniques (e.g., like those described in (Smith and Crane, 2001)) and experimenting with new ways of using the KB from the IE application.

Acknowledgements

Work on GATE has been supported by the Engineering and Physical Sciences Research Council (EPSRC) under grants GR/K25267 and GR/M31699, and by several smaller grants. The last author is currently supported by the EPSRC-funded AKT project (http://www.aktors.org) grant GR/N15764/01.

References

Atanas Kiryakov, Kiril Simov, Damyan Ognyanov. 2002. *Ontology Middleware and Reasoning* In the "Towards the Semantic Web: Ontology-Driven Knowledge Management", editors John Davies, Dieter Fensel, Frank van Harmelen. John Wiley & Sons, Europe, 2002.

Beth Sundheim, editor. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. ARPA, Morgan Kaufmann, 1998.

David A. Smith and Gregory Crane 2001. *Disambiguating Geographic Names in a Historical Digital Library*. In Proceedings of ECDL, pages 127-136, Darmstadt, 4-9 September 2001.

Diana Maynard, Valentin Tablan, Hamish Cunningham, Cristian Ursu, Horacio Saggion, Kalina Bontcheva, Yorick Wilks 2002. *Architectural Elements of Language Engineering Robustness*. In Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data, 8 (1) pp 257-274

Diana Maynard and Hamish Cunningham. 2003. *Multilingual Adaptations of a Reusable Information Extraction Tool*. In Proceedings of EACL 2003, Budapest, Hungary, 2003.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva and Valentin Tablan. 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, 2002.

Linda L. Hill. 2000. *Core elements of digital gazetteers: placenames, categories, and footprints*. In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000 (pp. 280-290). Berlin: Springer.

Nicola Guarino and Christopher Welty. 2000. *Towards a methodology for ontology-based model engineering*. In Proceedings of ECOOP-2000 Workshop on Model Engineering. Cannes, France.

Apendix A. Ontology screenshots

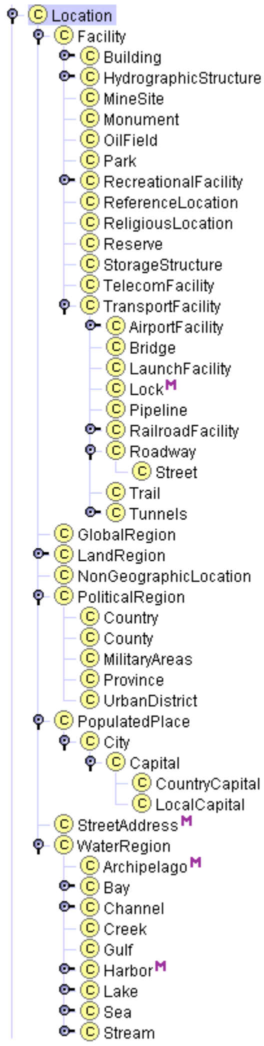

Figure 5: Location sub-ontology.

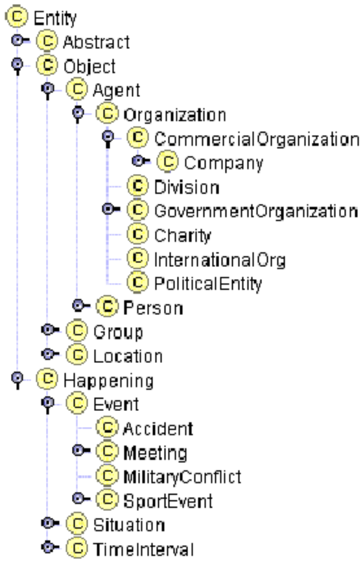

Figure 6: Upper level of KIM ontology.