The Week at a Glance - Cross-language Cross-document Information Extraction and Translation

Jim Cowie, Yevgeny Ludovik, Hugo Molina-Salgado, and Sergei Nirenburg
Computing Research Laboratory
New Mexico State University
Las Cruces, New Mexico
{jcowie, eugene, hsalgado, sergei}@crl.nmsu.edu

Abstract

Work on the production of texts in English describing instances of a particular event type from multiple news sources will be described. A system has been developed which extracts events , such as meetings, from texts in English, Russian, Spanish, and Japanese. The extraction is currently carried out using only ontological information. The results of a set of such extractions were combined to produce a table of event instances, date stamped, with links back to the original documents. The original documents can then be summarized and translated by the system on demand.

By using techniques from information retrieval, information extraction, summarization, and machine translation, in a multi-lingual environment, new documents can be produced which provide "at a glance" access to news on events from multiple sources.

The paper concludes with a discussion of the key resources which need to be developed to enhance the accuracy and coverage of the techniques used in our experiment.

1 Introduction

Multi-lingual information extraction (MUC6), summarization (AAAI 98), and cross-language information retrieval (Harman, 1995) have over the past three or four years become emerging technologies which are being driven in part by the development of the WWW, and also by the general availability of machine readable texts in many languages. It appeared to the authors that an interesting application could be built which demonstrated an integrated use of these technologies in combination with a variety of machine translation techniques.

The system we describe here has as its central component a multi-lingual information extraction engine, which uses an ontology as its main controlling element.

Extraction based summarization produces summaries, either in tabular form, or by generating sentences using structured information derived from texts. Summaries of this type are focused on whatever events, the underlying extraction system handles. The summaries are informative in nature. That is they provide specific facts which may allow a user to gain sufficient information without reference to the original documents. The potential applications are: producing personal profiles, assuming a series of documents on an individual are available over time; tracking complex events, assuming a script is available which describes the event in terms of simpler events; and monitoring for single event types in a data stream. This is the application we focus on in this paper.

The method is particularly promising for texts in multiple languages as the structured information produced by information extraction is relatively easy to translate. The principle drawback is that an information extraction system of this kind needs such expensive resources as ontology (one for all languages) and otological lexicons (one per language). The development of the system itself is not expensive, if we can get these resources.

The extraction method described here is based on the Mikrokosmos ontology (Mahesh, 1995), and uses the concepts in the ontology both to define an extraction template and to control the extraction process. At present the only information used from the Mikrokosmos lexicons, which supply language specific semantic and syntactic subcategorization information, is the mapping from a citation form to an ontological concept.

The complete system is composed of many pre-existing components and has been tested using two weeks of news from English, Spanish, Russian, and Japanese newspapers. We first give an overview of the steps used to generate an event based cross-document summary.

2 Overview of Processing

The system uses a set of pre-existing modules. These are: automatic language/codeset recognition for a text (Ludovik et al., 1999), sentence based summarization (biased towards domain keywords) (Cowie et al.., 1998), part of speech tagging, noun phrase recognition, proper name recognition and classification (Cowie et al. 1993; Cowie, 1996), ontology based extraction, translation of the final filled template to English, and output generation. Additional document and template filters have been added at the front and back ends of the system to reduce the amount of text to be processed and to remove templates which are only sparsely filled.

For example, when a text is gathered in Spanish by the web spider it will be checked to see if any of the person names of interest occur in the document using a list of names in Spanish. If this is the case the document is then part-of-speech tagged and noun phrases and proper names are recognized. In the present system proper names are handled using a table lookup process, rather than a more complex (and accurate) pattern based method. The ontology based extraction fills out the slots to produce a completed template. This is then translated by looking up words in the lexicon and by transliterating, or translating, proper names. The completed template is then stored with references to the original document.

A set of templates are then used to produce one of a variety of reports either for all events or for a single event type. These can be sorted on the different slots in the template. A table is then produced using HTML containing links to each original document, to document summarization and translation tools, and the slot fillers from each template.

In the rest of this paper we focus on the configurable extraction method, the preliminary tests carried out on the system, and we close with a discussion on the improvements to resources and tools needed to make this a robust and useful technology.

3 Extraction

The three events used in the present system are "election", "travel", and "meeting". For each of these a template was defined containing slots whose content would seem likely to occur in newspaper articles. Each of these slots was then mapped to one or more ontology concepts to produce a "control template". The three events are currently defined as follows:

```
ELECTION
   {"ELECT", "ELECT"}
   {"PERSON-ELECTED", "NAME-
      HUMAN"}
   {"PLACE", "NAME-PLACE"}
   {"DATE", "TIME"}
   {"POSITION-ELECTED-TO","SOCIAL-
      ROLE"}

TRAVEL
   {"TRAVEL", "TRAVEL-EVENT"}
   {"PERSON-TRAVELLING", "NAME-
      HUMAN"}
   {"ROLE", "SOCIAL-ROLE"}
   {"TO-PLACE", "NAME-PLACE"}
   {"DATE", "TIME"}

MEETING
   {"MEETING", "MEETING"}
   {"PERSON1", "NAME-HUMAN"}
   {"PERSON1", "NAME-HUMAN"}
   {"PERSON3", "NAME-HUMAN"}
   {"ROLE1", "SOCIAL-ROLE"}
   {"ROLE2", "SOCIAL-ROLE"};'
   {"ROLE3", "SOCIAL-ROLE"}
   {"PLACE", "NAME-PLACE"}
   {"DATE", "TIME"}
```

The left hand label defines the name/role of the slot, the right hand defines one, or more, ontological concepts which should be found for any phrase in the text which is a potential filler for the slot. The method of template definition is completely generic, and should allow a user with a reasonable knowledge of the ontology to rapidly configure an extraction system for new simple event types.

To perform an extraction, after the phrase recognition step, each headword in a sentence is looked up in the lexicon and its associated concepts found. Each lexicon entry is then matched with the concepts in the control template slots. A match may also be found using ancestors of the concept found in the lexicon entry. Thus for the lexicon entry "Bishop", in English, the attached concept is "RELIGIOUS-ROLE", which is a kind of "SOCIAL-ROLE". The combination of lexical entries which has the highest match, and which contains the key concept for the event is chosen and a completed extraction template is produced.

Lexical subcategorization patterns, which will also help increase the accuracy of this selection process, have not been used yet.

The ontological lexicons for Japanese and Russian were created by joining a bi-lingual Source language to English lexicon with an English to ontology lexicon. This process adds a significant amount of artificial ambiguity to the final source language to ontology lexicon. Using correctly created lexicons for each language and syntactic knowledge for each lexical entry would allow the extraction process to operate more accurately.

Lexical Entry Example

```
elect  elect  V      LG
-np[PROPERTY-NAME agent]
-v
-np[PROPERTY-NAME beneficiary]
-pp_adjunct[PROPERTY-NAME inverse-social-role-relation CONSTRAINT social-object PREP to]
```

Ontology Entry Example

```
(MAKE-FRAME ELECT (IS-A (VALUE (COMMON VOTE)))
   (DEFINITION
   (VALUE (COMMON "to select for an office
      by voting")))
   (BENEFICIARY (SEM (COMMON
    HUMAN)))
   (INSTRUMENT (SEM (COMMON
    BALLOT-BOX))))
```

Two Examples of Extraction

The following examples are both produced by the extraction method operating on bracketed texts produced by part-of speech tagging and phrase recognition.

```
On Thursday April 16, Clinton began
his two day state visit in
Santiago, Chile to meet with
Chilean President, Eduardo Frei,
and then onto the Summit of the
Americas.

Slot=MEETING ; Filler = meet
Slot=PERSON1 ; Filler = Clinton
Slot=PERSON2; Filler = Eduardo Frei
Slot=ROLE1                  : EMPTY
Slot=ROLE2;Filler=Chilean President
Slot=PLACE ; Filler = Santiago
Slot=DATE ;Filler=Thursday April 16
```

Президент Борис Ельцин встретился с Президентом США Клинтоном 19 ноября 1987 г. в Москве.

```
SLOT = MEETING ; filler = встретился
SLOT = PERSON1  ; filler = Борис Ельцин
SLOT = PERSON2 => EMPTY
SLOT = ROLE1       ; filler = Президент
SLOT = ROLE2       ; filler = Президентом
SLOT = PLACE       ; filler = Москве
SLOT = DATE        ; filler = 19 ноября 1987
```

4 Testing

Two weeks of news stories were gathered from two newspapers in each of our four languages; English, Spanish, Russian, and Japanese. We then filtered this document collection and kept only those documents which mentioned specific surnames, for eighteen different people. This entailed generating lists of these names in all four languages, including morphological variants for Russian. This was intended to focus the extraction process to specific domains (business and politics principally). The extraction process was then run on the remaining set of documents and the resulting templates translated and used to generate the final tables of events.

Many of the entries are inaccurate. One of the principal causes is the lack of syntactic information to constrain the extraction process. Simple improvements could be made by adding constraints based on appositions, prepositions, particles and morphology. However, a significant number of entries do contain useful information and the ability to scan, in one language, the output from eight sources in four languages is obviously a useful one.

5 Problems

There are many problems associated with a system of this degree of complexity. Many are related to the quality and coverage of the resources available for processing. Techniques, for example, for proper name recognition and classification are well known. However, good quality name recognition software is only freely available at the present for English. Using general web resources it is often difficult to discover document creation dates, an important piece of information in a system of this type.

Co-reference resolution is not handled in this system at present. This is normally achieved in current information extraction systems by allowing merging of templates from adjacent sentences.

The availability of large scale onomastica (bilingual lists of proper names) is also crucial to the translation of extracted information. Work is currently underway to develop these resources for a variety of languages.

The problem of reference in general is a more interesting one. There are currently two Boris Berezovsky(s) appearing in the news. One is a pianist, the other the Russian politician. The question is, "How is it possible to let our end user appreciate which person a reference is being made to?". Perhaps some document classification system needs to be added to allow the automatic detection of document topics, which could be used to provide additional information in the interface, either for display or for filtering.

Conclusion

The current system demonstrates the feasibility of a knowledge based approach to information extraction. It appears that it is possible to generate meaningful documents from multi-language sources, although the initial amount of effort required to get reasonable coverage and robust performance is significant, particularly in the area of resource development.

Acknowledgements

We would like to acknowledge the contributions of many colleagues at CRL. We would also like to thank MITRE and Kyoto University for allowing us to use their English and Japanese part of speech taggers.

References

Cowie, J., L. Guthrie, T. Wakao, W. Jin, J. Pustejovsky and S. Waterman (1993) The Diderot Information Extraction System. In *Proceedings of the First Conference of the Pacific Association for Computational Linguistics, (PACLING 93)*. Vancouver, Canada.

Cowie, J. (1996) CRL's approach to MET (Multilingual Named Entity Recognition*)*. In *Proceedings of the Tipster Text II 24 Month Workshop*. Morgan Kaufman.

Cowie, J., E. Ludovik and H. Molina-Salgado (1998) Improving Robust Domain Independent Summarization. In *Proceedings of Natural Language Processing and Industrial Applications*. Moncton, Canada.

Harman, D.K., ed. (1995) NIST Special Publication: *The Fourth Text Retrieval Conference (TREC-4*), Systems Laboratory, NIST.

Ludovik, Y., R. Zacharski and J. Cowie (1999) Language Recognition for Mono- and Multilingual Documents. In *Proceedings of VEXTAL'99*, Venice, Italy, pp. 209-214.

Mahesh, K. and S. Nirenburg. (1995) A Situated Ontology for Practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, International Joint Conference on Artificial Intelligence. Montreal, Canada.

MUC6 (1995) *Proceedings of the Sixth Message Understanding Conference (MUC6)* Morgan Kaufman. San Mateo, California.