Towards an Integrated Corpus for the Evaluation of Named Entity Recognition and Object Consolidation

Knud Möller[1], Alexander Schutz[2] and Stefan Decker[1]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway
knud.moeller@deri.ie, stefan.decker@deri.ie
[2] Institut für Allgemeine Linguistik, Universität des Saarlandes, Saarbrücken
schutz@coli.uni-sb.de

1 Introduction

When faced with the task of incorporating legacy web data from existing HTML pages into the Semantic Web (SW), a widespread approach is to use Information Extraction (IE) and Named Entity Recognition (NER) techniques. Natural language texts are annotated automatically or semi-automatically, and thus formal data is extracted from the texts. While this allows to add new sets of data to the SW, the process cannot stop there. It is necessary to integrate the newly created formal data with existing formal data, i.e. to identify entities which are identical in both sets. To summarize, two main problems have to be tackled to allow the integration of information from unstructured data into the SW:

1. Find the set of entities E_D in a document (NER), and probably detect co-reference chains within the document.
2. Find matches between the elements of E_D and entities in a pre-existing knowledge base.

In order to evaluate any system trying to tackle both of these problems (e.g. KIM [1] or Semtag and Seeker [2]), conventional corpora are not suitable, since these are mostly tailord towards IE and NER only. These corpora can be used to evaluate a system's performance on an inner-document basis, i.e. how well it can detect entities in a document and probably chains of co-reference between them. However, what is needed is a means of evaluating a system with respect to how well it is able to match between the entities in a document and corresponding entities in a database. This problem falls into the area of Object Consolidation. We therefore propose a novel kind of corpus, which we will call an **Integrated Corpus for Named Entity Recognition and Object Consolidation**. The first incentive for proposing such a corpus came when we were looking for a way to evaluate the Geco project [3].

2 An Integrated Corpus

Our proposed integrated corpus consists of two interrelated parts:

– An annotated textual corpus C_1 for the evaluation of IE/NER components. This part of the corpus will be very similar to traditional corpora like MUC[3] or ACE[4] .

[3] MUC6 see http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T13, MUC7 see http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T02
[4] see http://www.ldc.upenn.edu/Projects/ACE

– A knowledge base (KB) C_2 containing objects corresponding to the entities mentioned in C_1.

These two parts are integrated by linking the annotated entities in C_1 to the corresponding objects in C_2, as Figure 2 illustrates.

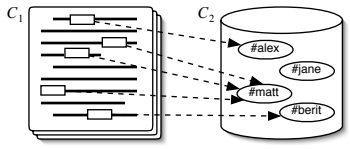

Fig. 1. The Integrated Corpus

For the first version of our corpus, we defined a set of 40 documents with approximately 100 words. These documents were excerpts from Wikipedia[5] biographies of various politicians, actors, scientists, bands, fictional and non-fictional characteres, etc. We compiled the corpus with the aim of including challenging problems for both the NER and the object consolidation task, such as different forms of the same name (e.g. "Bill Clinton", "Clinton", "Billy"), potentially ambiguous tokens (e.g. "Hope": location/verb) and pseudonyms (e.g. "Ringo Starr", "Richard Starkey"). The corpus was then annotated by one human annotator, currently only with respect to three different annotation types: PERSON, LOCATION and JOBTITLE.

In order to allow the integration of the textual corpus and the KB, the latter would have to contain the same entities as mentioned in the text. Of the 205 PERSON annotations in the textual corpus, 95 referred to individual entities. For each of these entities, we included a corresponding entity in the KB. Within the Geco project, we were working with FOAF[6] representations of people. For this reason, we chose to build a KB of `foaf:Person` instances.

Having completed both parts of the corpus, they had to be tied together. This was achieved by referencing the `Person` instances in the knowledge base from the annotations in the textual corpus. In FOAF, the assumption is made that each person can be uniquely identified by her email address. We therefore used email addresses (both real and made-up) as the referencing scheme. Once both parts of the corpus had been related in that way, the Integrated Corpus was complete.

3 Future Work

In this paper, we proposed a novel kind of evaluation corpus, which we called an **Integrated Corpus for the Evaluation of Named Entity Recognition and Object Consolidation**. It can be used for both the evaluation of NER systems and systems trying to solve object consolidation problems. We are aware of the fact that future versions of the textual part of our corpus will have to be extended in both size and depth. We will have to extend the size of the corpus, its scope and the number of annotation types. Another important task for a future version of our corpus is the development of a suitable kind of evaluation metrics. The conventional recall, precision and F-measure metrics could be applied individually to the textual part of the corpus and the linking between the annotations and the instances in the knowledge base. However, it would be desirable to provide a combined measure in order to rate the overall performance of a system with respect to our corpus.

[5] see `http://en.wikipedia.org`
[6] see `http://xmlns.org/foaf/0.1`

Bibliography

[1] Popov, B., Kiryakov, A., Manov, D., Ognyanoff, D., Goranov, M.: Kim - semantic annotation platform. Lecture Notes in Computer Science **124** (2003) 834–849

[2] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., Zien, J.Y.: Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In: Proceedings of the twelfth international conference on World Wide Web, ACM Press (2003) 178–186

[3] Möller, K.: Geco - using human language technology to enhance semantic web browsing. In: Proceedings of the Faculty of Engineering Research Day 2004, National University of Ireland, Galway. (2004)