Part II Information Retrieval Exercises

Stephen Clark
based on previous exercises by Simone Teufel

Michaelmas Term 2009

Lectures 5-8

Lecture 5 introduces the Information Extraction (IE) task, and describes the successful Mikheev system, which is based on a combination of high-precision regular expression rules and machine learning. Lecture 6 describes methods for *learning* patterns which can be used for IE, rather than relying on humans to create the rules by hand. Lecture 7 introduces the application of Question Answering (QA), and considers how to evaluate QA systems as well as describing a number of relatively sophisticated QA systems. Finally, Lecture 8 introduces the task of Summarisation, and describes some existing systems and evaluations.

1. Can you create any counterexamples to Mikheev's sure-fire rules on P.17 of the slides; i.e. examples that match the regular expression but which do not conform to the predicted semantic type?

2. What sorts of features does the machine learning algorithm of Mikheev use? Give some examples of useful internal and external features that the system could use in addition to those on p.18 of the slides. What are the advantages of the machine learning method over the rule-based system?

3. Experiment with the online demo of the C&C named entity system:
http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Demo
Use the "Try it for yourself" interface, under "C&C output for sentences". **Set the format option to GRs and the model option to Sentences.** For the input *Acme Inc. folded in December .*, the output should look like the following:

```
(ncmod _ Inc._1 Acme_0)
(dobj in_3 December_4)
(ncmod _ folded_2 in_3)
(ncsubj folded_2 Inc._1 _)
<c> Acme|Acme|NNP|I-NP|I-ORG|N/N Inc.|Inc.|NNP|I-NP|I-ORG|N
folded|fold|VBD|I-VP|O|S[dcl]\NP in|in|IN|I-PP|O|((S\NP)\(S\NP))/NP
December|December|NNP|I-NP|I-DAT|N .|.|.|O|O|.
```

The named entity information is in the `<c>` line, as the 5th field. So for the example, the words *Acme* and *Inc.* have been tagged as organisation (`I-ORG`), and *December* has been tagged as date (`I-DAT`). `O` indicates that the word is 'outside' of any named entity, i.e. not denoting one.

The named entity system is described in the following short paper:

```
Language Independent NER using a Maximum Entropy Tagger
James R. Curran and Stephen Clark
Proceedings of the Seventh Conference on Natural Language
Learning (CoNLL-03), pp.164-167, Edmonton, Canada, 2003
```

and is available here:

`http://www.cl.cam.ac.uk/~sc609/pubs.html`

Try and find some examples of persons, companies, organsiations and locations which the system performs well on, and some that it performs badly on. (You will need to separate out punctuation marks with a space – unless it is part of the word, eg *Inc.* – as in the previous example in which the final period has a space before it.) The system is entirely data-driven, not relying on hand-coded rules at all; i.e. all the knowledge that the system has is derived from the examples on which the system is 'trained'. Is the poor performance on some examples entirely down to the data on which the system was trained?

4. Explain the evaluation metrics MRR and Confidence-Weighted Score for QA. How is the Confidence-Weighted Score similar to Mean Average Precision used to evaluate document retrieval?

5. Explain how Kupiec et al. use a Naive Bayes classifier to perform sentence extraction. (The Kupiec et al. system was described in Lecture 8.) What are the simplifying assumptions that the Naive Bayes model makes? What are the weaknesses of sentence extraction as a method for producing summaries?

6. The majority of the techniques that we have seen for text processing do not rely on the system having much 'understanding' of the text. Should Information Retrieval be trying to use more sophisticated approaches using methods from Natural Language Processing and Artificial Intelligence? Will more sophisticated AI approaches be required for some tasks more than others? (You might consider this question in relation to each of the IR tasks considered in the course.)