Department of Computer Science
The University of Melbourne
COMP90042 WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2017)

Workshop exercises: Week 7

Discussion

1. What is **Information Extraction**? What might the "extracted" information look like?
   (a) What is **Named Entity Recognition** and why is it difficult? What might make it more difficult for persons rather than places, and *vice versa*?
   (b) What is the **BIO** trick, in a sequence labelling context? Why is it important?
   (c) What is **Relation Extraction**? How is it similar to NER, and how is it different?
   (d) Why are hand–written patterns generally inadequate for IE, and what other approaches can we take?

2. What is **Question Answering**, and how is it related to **Information Retrieval** and Information Extraction?
   (a) What is **semantic parsing**, and why might it be desirable for QA? Why might approaches like NER be more desirable?
   (b) What might be the main steps for answering a question for a QA system?

Programming

1. NLTK comes with a pre-trained named entity **chunker** `ne_chunk`, which takes a tagged sentence as input, and outputs a tree:

```
>>> print(nltk.ne_chunk(nltk.corpus.treebank.tagged_sents()[11]))
(S
  Dr./NNP
  (PERSON Talcott/NNP)
  led/VBD
  a/DT
  team/NN
  of/IN
  researchers/NNS
  from/IN
  the/DT
  (ORGANIZATION National/NNP Cancer/NNP Institute/NNP)
  ...
```

   Read up on how to traverse an `nltk.tree.Tree` object, and then convert the tree into a (flat) BIO-representation.

Catch-up

- What is **POS tagging**, and what are some common methods for applying it?
- What is a **Named Entity**? How is **ENAMEX** different to **TIMEX** and **NUMEX**?
- What is an **HMM** and what is it used for in language processing?
- What kinds of **relations** can exists between tokens ("words")? Constituents? Sentences? Documents?
- What is **parsing** and how is it different to **tagging**?

Get ahead

- Using the `WSTA_N4_hidden_markov_models` iPython notebook as a basis, train an HMM for NER (BIO) tags. Test a couple of sentences and consider where the output differs. What might be causing this to happen?
- Try using the `nltk.sem.extract_rels()` to extract a set of relations from the collection `nltk.corpus.ieer.parsed_docs()`, and then write a system that can answer simple questions like: "Where is [the] Bastille Opera?"