Natural Language Processing Exercise Sheet 1

Michaelmas Term 2016/17

2 Morphology and finite-state techniques [Lecture 2]

Exercise 2.1

What is a finite state transducer (FST), and what is it used for in computational linguistics? How does it differ from a finite state automaton?

Exercise 2.2

In German, the third person singular present inflection of weak verbs is generally formed by adding the 't' to the stem. Exceptions to this rule include verbs with stems that end in 't' or 'd' which are formed by adding 'et' instead of 't'. The following table gives some examples (^ is used as the affix marker in the underlying form).

stem	surface	underlying
kauf	kauft	kauf^t
arbeit	arbeitet	arbeit^t

Draw a finite state transducer (FST) that relates surface and underlying forms according to this pattern. (Only the inflected forms should be accepted by the transducer since the stems by themselves do not correspond to words.) Explain the notation that you use and outline how the FST could be used in morphological analysis and generation.

3 Prediction and part-of-speech tagging [Lecture 3]

Exercise 3.1

(a) Give an equation for finding the most probable sequence of part of speech (POS) tags that could be utilised by a stochastic POS tagger. You should assume a bigram model.

(b) Given the following training data, show the estimates that would be obtained for the probabilities in the equation you gave:

```
the_DT0 green_AJ0 bottle_NN1 leaked_VVD ._PUN
the_DT0 suppliers_NN2 bottle_VVB water_NN1 ._PUN
green_AJ0 water_NN1 suppliers_NN2 bottle_VVB ._PUN
```

(c) Explain what is meant by the terms smoothing and backoff in the context of stochastic POS tagging.

(d) One common source of errors in stochastic POS taggers is that nouns occurring immediately before other nouns (e.g. catamaran trailer) are often tagged as adjectives and, conversely, prenominal adjectives are often tagged as nouns (e.g. trial offer). Suggest possible reasons for this effect.

4 Context-free grammars and parsing [Lecture 4]

Exercise 4.1

Using the CFG given in the lecture notes (section 4.3):

(a) show the edges generated when parsing *they fish in rivers in December* with the simple chart parser in 4.7

(b) show the edges generated for this sentence if packing is used (as described in 4.9)

(c) show the edges generated for *they fish in rivers* if an active chart parser is used (as in 4.10)

Exercise 4.2

The following context-free grammar (CFG) accepts sequences of part-of-speech categories (e.g., Det N, Adj Adj N). With a lexicon, as shown, it can be used to parse some English noun phrases (NPs).

```
Start symbol:  NP
NP → Det N
NP → N
N → Adj N
N → N PP
PP → P NP
```

```
a, the:  Det
dog, dogs, house, houses, model, models:  N
brown, red, model:  Adj
in, under:  P
```

(a) Give a non-deterministic finite-state automaton (NDFSA) which accepts the same sequences of part-of-speech categories as this CFG. Explain the notation that you use.

(b) Give two examples of overgeneration that can be demonstrated with the lexicon shown, and explain how the CFG and NDFSA (and, if necessary, part-of-speech categories and lexicon) could be modified to prevent them.

(c) The CFG does not accept noun-noun compounds (e.g., the dog house, house dogs). Indicate how you could modify the original CFG and NDFSA to allow for them.

(d) Hand-constructed FSA have sometimes been used for part-of-speech tagging. Outline the possible practical and theoretical advantages and disadvantages of such an approach when compared to stochastic tagging using Hidden Markov Models.