Part-of-Speech Tagging

Introduction to NLP, ACS 2010, Assignment 1 © Ted Briscoe (ejb@cl.cam.ac.uk) GS18

1 Task

Choose 2 sentences from each of the 4 sets below (8 total) and assign part-of-speech tags to each token of each sentence. Use the Penn Treebank tagset which is given inside the front cover of Jurafsky and Martin or on the web (e.g. http://www.comp.leeds.ac.uk/amalgam /tagsets/upenn.html). You should break up words into further tokens (clitics, morphemes) where this is appropriate to tag them, and assign the unique tag which is correct in context where a word is ambiguous. For instance the tagging of:

My aunt's can opener can open a drum

should look like this:

My/PRP$ aunt/NN 's/POS can/NN opener/NN can/MD open/VB a/DT drum/NN

Write up your answers and hand them in at the following Tuesday session. Include BRIEF notes on any difficulties or issues you had with specific assignments. It is more important to understand and be able to explain your reasoning than to get every tag right. Be prepared to discuss the difficult cases during the session. Please feel free to work on the task in groups, but the final selection of sentences and their tags should be your own.

2 Sentences

(1) a The old car broke down in the car park
  b At least two men broke in and stole my TV
  c The horses were broken in and ridden in two weeks
  d The horses were broken in and ridden in two weeks
  e Kim and Sandy both broke up with their partners

(2) a The horse which Kim sometimes rides is more bad tempered than mine
  b The horse as well as the rabbits which we wanted to eat have escaped
  c It was my aunt's car which we sold at auction last year in February
  d The only rabbit that I ever liked was eaten by my parents one summer
  e The veterans who I thought that we would meet at the reunion were dead

(3) a Natural disasters – storms, flooding, hurricanes – occur infrequently but cause devastation that strains resources to breaking point

b Letters delivered on time by old-fashioned means are increasingly rare, so it as well that that is not the only option available

c It won't rain but there might be snow on high ground if the temperature stays about the same over the next 24 hours

d The long and lonely road to redemption begins with self-reflection: the need to delve inwards to deconstruct layers of psychological obfuscation

e My wildest dream is to build a POS tagger which processes 10K words per second and uses only 1MB of RAM, but it may prove too hard

(4) a English also has many words of more or less unique function, including interjections (oh, ah), negatives (no, not), politeness markers (please, thank you), and the existential 'there' (there are horses but not unicorns) among others.

b Making these decisions requires sophisticated knowledge of syntax; tagging manuals (Santorini, 1990) give various heuristics that can help human coders make these decisions and that can also provide useful features for automatic taggers.

c The Penn Treebank tagset was culled from the original 87-tag tagset for the Brown Corpus. For example the original Brown and C5 tagsets include a separate tag for each of the different forms of the verbs *do* (e.g. C5 tag VDD for *did* and VDG tag for *doing*), *be* and *have*.

d The slightly simplified version of the Viterbi algorithm that we present takes as input a single HMM and a sequence of observed words $O = (o_1, o_2, ...o_T)$ and returns the most probable state/tag sequence $Q = (q_1, q_2, q_T)$ together with its probability.

e Thus the EM-trained "pure HMM" tagger is probably best suited to cases where no training data is available, for example, when tagging languages for which no data was previously hand-tagged.