Introduction to Natural Language Processing: Practical 1

Prof. Ann Copestake, Dr. Stephen Clark

Michaelmas Term 2012

Hand-in of report: 16:00, 18th January 2013
(Note that the second report on practical 2, parsing, will be due the same day.)
The aim of this practical session is to evaluate and compare two publically available part-of-speech taggers of your choice.

Resources

Tagset: We suggest that you choose two taggers which use the Penn Treebank tagset.

Input sentences: A set of input sentences suitable for the evaluation can be found at the end of this handout (which is also on the module website as a PDF file). It is your responsibility to process these sentences so that they are in a suitable form for input to the taggers (see e.g. http://www.cis.upenn.edu/ treebank/tokenization.html but don't assume this will do the right thing!)

Possible taggers: Taggers which you might evaluate include the TnT tagger; the C&C maximum entropy tagger*; the Stanford POS tagger; CRFTagger. All of these can be found with a simple Google search. A website which lists some NLP tools is http://nlp.stanford.edu/links/statnlp.html.

*Note on C&C: the Windows version is not as reliable as the Linux or Mac versions. Dr Clark will not be able to help with Windows compilation problems.

What You Need To Do

- Download a few of the available taggers; compile them if necessary; and also read about their underlying probability models and search algorithms. Choose two for comparison. (It is your responsibility to get any taggers you choose to compile and run on the machines you are using.)

- Investigate what pre-processing — in particular tokenisation — is required for each tagger. A simple sed script for tokenisation consistent with the Penn Treebank is available here: http://www.cis.upenn.edu/~treebank/tokenizer.sed. You should investigate whether the tokenisation method you adopt is sufficient for the sentences in the input data, and, if not, extend it.
- Examine the output of the two taggers on the input sentences and see if the taggers make any errors. Can you make any generalisations about the errors each tagger makes? Given what you know about the probability models and search algorithms adopted by each tagger, can you explain why each tagger makes the mistakes it does?
- If you think that it will help your understanding or explanation of the errors made by the tagger you may add some new examples to those given.

Your Report

Your report should contain a concise summary and comparison of the errors made by each tagger, and any general conclusions that you have been able to draw regarding the performance of each tagger.

Your report should not be longer than 2,500 words and should include a word count. Your report should provide a pointer to a world-readable directory in your filespace that provides the complete output of the taggers that you ran and all your data as preprocessed by you.

Assessment

Your report will be graded out of 20 and will contribute 40% of the mark you receive for the module. Marks will be assigned for correctly identifying the errors made by each tagger, for insightful comparison, discussion of the issues of preprocessing, and for generalisations concerning the tagging models and types of errors observed.

Data

(1) The old car broke down in the car park.

(2) At least two men broke in and stole my TV.

(3) The horses were broken in and ridden in two weeks.

(4) Kim and Sandy both broke up with their partners.

(5) The horse which Kim sometimes rides is more bad tempered than mine.

(6) The horse as well as the rabbits which we wanted to eat have escaped.

(7) It was my aunt's car which we sold at auction last year in February.

(8) The only rabbit that I ever liked was eaten by my parents one summer.

(9) The veterans who I thought that we would meet at the reunion were dead.

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

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

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

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

(14) 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.

(15) 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.

(16) 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.

(17) 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*.

(18) 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.

(19) 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.

(20) Coming home from very lonely places, all of us go a little mad: whether from great personal success, or just an all-night drive, we are the sole survivors of a world no one else has ever seen.

(21) Skill without imagination is craftsmanship and gives us many useful objects such as wickerwork picnic baskets. Imagination without skill gives us modern art.

(22) An MoD spokesman said: "Surveys of Astute have now been completed and she will proceed to Faslane under her own power. She is being escorted by tugs and HMS Shoreham."

(23) But far fewer people fully understand how the Media Lab operates, fits into MIT, and encourages such a creative environment; about half of the anniversary celebration's program focused on simply defining what the Media Lab is.

(24) Instead of constantly worrying about funding, the faculty and students can focus on their project, with the exception of sponsors' weeks, when they have to convince companies to start or continue their support.

(25) The doctors are warning that the NHS cannot make the 20bn of savings by 2014 that ministers expect, while simultaneously undertaking a huge reorganisation that will see England's 152 primary care trusts (PCTs) abolished and consortiums of GPs assume responsibility for the commissioning of services for patients.