Some terms and resources

LING 8570 M. Covington Spring 2011

Things we do to natural-language texts:

Tokenization – breaking the text into words (tokens)

`"What is this?"` → `[what, is, this, '?']`

This is necessary for all natural language processing.

Units like *'s* and *n't* are usually treated as separate tokens: *doesn't* → *does n't*

Tokenization requires design decisions about things like that.

Correct tokenization of things like *Do you owe us $145,678.90?* is hard.

Stemming – removing suffixes to make related words look alike

The vastness of the canyon surprised us. → *the vast of the canyon surpris us*

This is a shallow method and does not aim for correctness. Information is lost.

Search engines do it. Language understanding systems don't.

It is done by just chopping off endings that might be suffixes, without a dictionary.

Tagging – labeling the parts of speech (nouns, verbs, etc.)

`What is this?` → `What/WP is/VBZ this/PRP ?/.`

Non-trivial because many words can be more than one.

Uses a dictionary and a collection of heuristics. Typically 95%-97% accurate.

Almost always based on Penn Treebank (see below).

Lemmatization – finding the dictionary form of each word

surprises → *surprise* *went* → *go* *children* → *child*

Uses morphological analysis and tagging. Aims to be correct.

Usually does not undo derivational morphology, just inflection.

Not a necessary step – parsing can proceed without it.

Parsing – recovering syntactic structure

Recovering the syntactic tree.

Uses a dictionary and a collection of grammar rules.

May or may not require tagging to have been done first.

Many different algorithms are possible, some of them unification-based.

Semantic interpretation – building a representation of meaning

Many methods, all of them poorly understood.

Often done concurrently with parsing.

Unification-based algorithms are a great help.

The Penn Treebank

A large collection of samples of English (about 7 million words tagged, 3 million words parsed) which was tagged and parsed by trained human beings at the University of Pennsylvania.

It includes 3 main corpora:

- A collection of articles from *The Wall Street Journal*
- The Brown University Corpus, a million-word collection of English journalism and literature from the mid-20th-century
- The "Switchboard" corpus (of telephone conversations, annotated for dysfluency)

The University of Georgia has a license for the whole data set, which is kept in \\AIHV\NLP. (It is not free; do not take copies elsewhere for non-UGA work.)

About 10% of it is distributed free with the (Python) Natural Language Toolkit.

About half of that, in turn, I have adapted for easy use in Prolog and will be distributing it to you.

The Penn Treebank introduces its own "tagset" (set of labels for parts of speech), which are different from those normally used in linguistics. For example, noun is NN rather than N.

Its tagging makes very fine distinctions. For example, a verb with no ending is VBP if the context makes it third-person plural (*They sing*) and VB otherwise (such as *to sing*).

The tagging in the Penn Treebank is not perfectly accurate. Taggers trained on it never achieve more than about 97% accuracy and this is probably the level of consistency in the Treebank itself. There is less consistency between corpora than within them. The Switchboard corpus *uses a different tagset* and even when you try to convert it to be equivalent, the match is not perfect.

The parsing in the Penn Treebank follows a grammar that is extensively documented.