Statistical and Learning Methods in Natural Language Processing

Ido Dagan
dagan@macs.biu.ac.il

Shuly Wintner
shuly@cs.haifa.ac.il

Spring 2004

Practicalities

Office hours: Sunday 15:00-16:00, Jacobs 43. Phone: (828)8180.

Times: Sunday 10:00-14:00.

Place: Education Building 3502.

Prerequisites: Computational Linguistics.

Textbook: Nothing mandatory. Some of the material can be found in **Foundations of Statistical Natural Language Processing**, by Chris Manning and Hinrich Schuetze.

Grading: The final grade will be based on a final project (approximately 60%) and a short exam (approximately 40%).

Organization according to application

- Morphology
  - Segmentation
  - Part of speech tagging, Morphological disambiguation
  - Lexical acquisition
- Syntax
  - parsing
  - shallow parsing
  - attachment
- Semantics
  - word-sense disambiguation
  - categorization

Segmentation

Problem: Given a word w, find a sequence of morphemes $m_1, \ldots, m_k$ such that $w = m_1 \cdots m_k$.

Example: im|possible, in|credible, ir|regular, ir|resistable, in|finite, in|dependent, ...

ink, imply, Iran,...

Example: resist|able, comfort|able, ed|ible, incred|ible, imposs|ible, ...

table, stable, ...

More complex cases: segmenting sentences to words in Asian languages.

Where can statistical methods help?

Part of speech tagging

Problem: Given a text where each word is associated with all its possible parts of speech, determine the most likely POS for the word with respect to its context.

Example:

who	PRON(int), PRON(rel)
can	AUX, V(inf), N(sg)
it	EXPLETIVE, PRON(3sg)
be	V(inf)
?	PUNC

Part of speech tagging

Problem: Given a text where each word is associated with all its possible parts of speech, determine the most likely POS for the word with respect to its context.

Example:

who	PRON(int), PRON(rel)
can	AUX, V(inf), N(sg)
it	EXPLETIVE, PRON(3sg)
be	V(inf)
?	PUNC

Where can statistical methods help?

Morphological disambiguation

A generalization of Part-of-speech Tagging

Problem: Given a text where each word is associated with all its possible morphological analyses, determine the most likely analysis for the word with respect to its context.

Example:

BIWM XMI$I HCLIXH $W@RT BLBW$ AZRXI LHIKNS
LMS&DH K$HIA LWB$T XGWRT DMH $DIMTH XGWRT NPC

Analysis

Where can statistical methods help?

Lexical acquisition

Problem: Given a morphological analyzer based on a partial lexicon, and a corpus of texts, expand the lexicon automatically.

Where can statistical methods help?

Parsing

Problem: Given a grammar and a sentence, generate all the structures that are induced by the grammar on the sentence.

Observation: not all structures are equally likely.

Where can statistical methods help?

Shallow parsing

Problem: Given a sentence, segment it into phrases such that no two phrases overlap.

Example (from http://pi0657.kub.nl/cgi-bin/tstchunk/demo.pl):

```
[NP Identifying NP] [NP the NP] [NP root NP]
{PNP [Prep of Prep] [NP a given word NP] PNP}
{PNP [Prep in Prep] [NP a NP] PNP}
[NP Semitic NP] [NP language NP]
[VP is VP] [NP an important task NP] ,/,
{PNP [Prep in Prep] [NP some cases NP] PNP}
[NP a NP] [NP crucial part NP]
{PNP [Prep of Prep] [NP morphological analysis NP] PNP}
```

Attachment

Problem: Given an ambiguous syntactic structure, determine which of the candidate structures is most likely.

Example:

The teacher [wrote [three equations] [on the board]]

The author [wrote [three novels [on the civil war]]]

Word-sense disambiguation

Problem: Given a text in which each word is associated with several senses, determine the correct sense in the context of each of the words.

Example:

The dictionary entry of brilliant

Some usage examples with various senses of brilliant

Where can statistical methods help?

Text categorization

Problem: Given a document and a (hierarchical) classification of "topics", determine which topics are addressed by the document.

Example: classify Internet pages to Yahoo!'s tree of topics.

Where can statistical methods help?