Machine Learning for Language Learning and Processing

Sharon Goldwater, Frank Keller, Mirella Lapata,
Victor Lavrenko, Mark Steedman

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

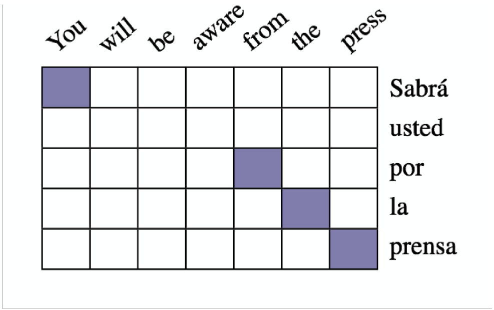
October 15, 2008

1. Machine Learning and NLP
    - Latent Variables
    - Multi-class and Structured Variables
    - Discrete, Sparse Data
    - Other Problems
2. Research Interests
    - Parsing
    - Language Acquisition
    - Language Generation
    - Information Retrieval

Latent Variables

Natural language processing (NLP) problems typically involve *inferring latent (non-observed) variables.*

- given a bilingual text, infer an alignment;
- given a string of words, infer a parse tree.



Multi-class and Structured Variables

The learning targets in NLP often are *multi-class,* e.g., in part of speech tagging:

- standard POS tag sets for English have around 60 classes; more elaborate ones around 150 (CLAWS6);
- morphological annotation often increases the size of the tag set (e.g., Bulgarian, around 680 tags).

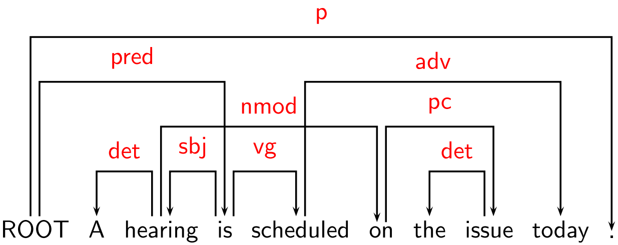
Everything	in	the	sale	will	have	been	used	in	films	.
PNI	PRP	AT0	NN1	VM0	VHI	VBN	VVN	PRP	NN2	PUN

Multi-class and Structured Variables

NLP tasks are often *sequencing tasks,* rather than simple classification:

PNI → PRP → AT0 → NN1 → VM0 → VHI → VBN → VVN . . .

Many NLP tasks are not classification but often involve *hierarchical structure,* e.g., in parsing:

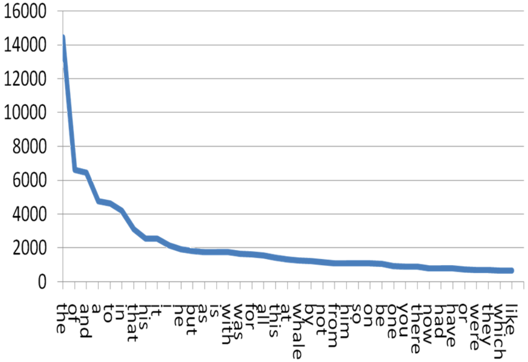


Most structures in the Penn Treebank are unique.

Discrete, Sparse Data

Linguistic data is different from standard ML data (speech, vision):

- typically *discrete* (characters, words, texts);
- follows a *Zipfian* distribution.



Discrete, Sparse Data

The Zipfian distribution leads to ubiquitous *data sparseness:*

- standard maximum likelihood estimation doesn't work well for linguistic data;
- a large number of smoothing techniques have been developed to deal with this problem;
- most of them are ad hoc; Bayesian methods are a principled alternative.

$$G \sim \mathrm{PY}(d, \theta, G_0)$$

Other Problems

- NLP typically uses *pipeline models;* errors propagate;
- models often highly *domain-dependent* (models for broadcast news will not work well for biomedical text, etc.);
- there is no single *error function* to optimize; evaluation metrics differ from task to task (BLEU for MT, ROUGE for summarization, PARSEVAL for parsing).

POS:	NNP	NNP	VBD	TO	NNP	NN
Chunk:	[- NP -]		[- VP -]	[-PP-]	[- NP -]	[-NP-]
NER1:	[- Per -]			[- O -]	[-Org-]	[- 0 -]
NER2:	[- Per -]		[- O -]	[- O -]	[- O -]	[- O -]
NER3:	[- Per -]		[- O -]	[- O -]	[- Org -]	
NER4:	[- Per -]		[- O -]	[- O -]	[-Org-]	[- O -]

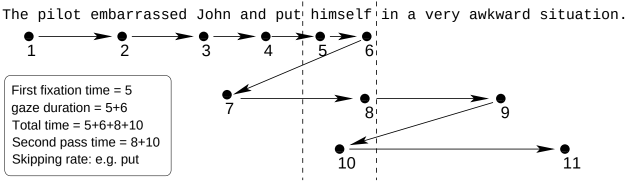
1. Machine Learning and NLP
   - Latent Variables
   - Multi-class and Structured Variables
   - Discrete, Sparse Data
   - Other Problems

2. Research Interests
   - Parsing
   - Language Acquisition
   - Language Generation
   - Information Retrieval

Parsing (Keller, Steedman)

Current focus of research in probabilistic parsing:

- models for more expressive syntactic representations (CCG, TAG, dependency grammar);
- semi-supervised induction of grammars and parsing models;
- cognitive modeling:
  - incrementality;
  - limited parallelism, limited memory;
  - evaluation against behavioral data.



Language Acquisition (Goldwater, Steedman)

Research focuses on Bayesian models for improving unsupervised NLP and understanding of human language acquisition:

- What constraints/biases are needed for effective generalization?
- How can different sources of information be successfully combined?

ML methods and problems:

- infinite models, esp. those for sequences/hierarchies;
- incremental, memory-limited inference methods;
- joint inference of different kinds of linguistic information (e.g., morphology and syntax).

Language Generation (Lapata)

Research focuses on data-driven models for language generation:

- fluent and coherent text that resembles human writing;
- general modeling framework for different input types (time series data, pictures, logical forms).

ML methods and problems:

- mathematical programming for sentence compression and summarization;
- latent variable models for image caption generation;
- models have to integrate conflicting constraints and varying linguistic representations.

Language Generation (Lapata)

Example: image captioning beyond keywords.

troop, Israel, force, ceasefire, soldiers

Thousands of Israeli troops are in Lebanon as the ceasefire begins.

Information Retrieval (Lavrenko)

Universal search:

- learn to relate relevant text/images/products/DB records;
- data: high-dimensional, extremely sparse, but dimensionality reduction is a bad idea;
- targets: focused information needs, not broad categories;
- semi-supervised: lots of unlabeled data, few judgments.

Learning to rank:

- partial preferences → ranking function;
- objective: non-smooth, can be very expensive to evaluate.

Novelty detection:

- example: identify first reports of events in the news;
- supervised task, but hard to learn anything from labels;
- best approaches unsupervised, performance very low.