Part-of-speech tagging revisited

A simple but useful form of linguistic analysis

Christopher Manning

Sources of information

- What are the main sources of information for POS tagging?
  - Knowledge of neighboring words
    - Bill   saw   that man yesterday
    - NNP NN   DT   NN   NN
    - VB   VB(D) IN   VB   NN
  - Knowledge of word probabilities
    - *man* is rarely used as a verb….
- The latter proves the most useful, but the former also helps

More and Better Features ➜ Feature-based tagger

- Can do surprisingly well just looking at a word by itself:
  - Word   the: the → DT
  - Lowercased word   Importantly: importantly → RB
  - Prefixes   unfathomable: un- → JJ
  - Suffixes   Importantly: -ly → RB
  - Capitalization   Meridian: CAP → NNP
  - Word shapes   35-year: d-x → JJ
- Then build a maxent (or whatever) model to predict tag
  - Maxent P(t|w):   93.7% overall / 82.6% unknown

Overview: POS Tagging Accuracies

- Rough accuracies:
  - Most freq tag: ~90% / ~50%

  - Trigram HMM: ~95% / ~55%
  - Maxent P(t|w):   93.7% / 82.6%
  - TnT (HMM++):   96.2% / 86.0%
  - MEMM tagger: 96.9% / 86.9%
  - Bidirectional dependencies: 97.2% / 90.0%
  - Upper bound: ~98% (human agreement)

Most errors on unknown words

How to improve supervised results?

- Build better features!

<div style="text-align:center">

                  RB

PRP VBD IN RB IN PRP   VBD .

They left   as soon as he   arrived .

</div>

  - We could fix this with a feature that looked at the next word

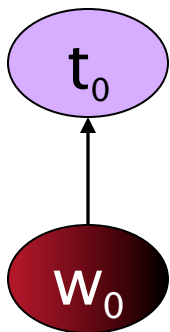<div style="text-align:center">

     JJ

NNP   NNS   VBD   VBN   .

Intrinsic flaws remained undetected .

</div>

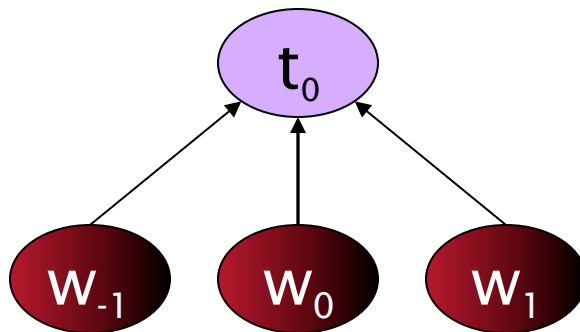 - We could fix this by linking capitalized words to their lowercase versions

Tagging Without Sequence Information

Baseline

Three Words

Model	Features	Token	Unknown	Sentence
Baseline	56,805	**93.69%**	82.61%	26.74%
3Words	239,767	**96.57%**	86.78%	48.27%

Using words only in a straight classifier works as well as a basic (HMM or discriminative) sequence model!!

Summary of POS Tagging

For tagging, the change from generative to discriminative model **does not by itself** result in great improvement

One profits from models for specifying dependence on **overlapping features of the observation** such as spelling, suffix analysis, etc.

An MEMM allows integration of rich features of the observations, but can suffer strongly from assuming independence from following observations; this effect can be relieved by adding dependence on following words

This additional power (of the MEMM ,CRF, Perceptron models) has been shown to result in improvements in accuracy

The **higher accuracy** of discriminative models comes at the price of **much slower training**

Part-of-speech tagging revisited

A simple but useful form of linguistic analysis

Christopher Manning