Towards the Unsupervised Learning of Parts of Speech

Simon Fung
University of Alberta
AACL 2009

POS-Tagging

- Hand-tagging
- Transformational taggers
  - e.g. Brill tagger
- Supervised learning
  - e.g. HunPOS
- Unsupervised learning
  - e.g. Ravi & Knight (2009)
- Unsupervised POS induction

POS-Tagging

- Hand-tagging
- Transformational taggers
  - e.g. Brill tagger
- Supervised learning
  - e.g. HunPOS
- Unsupervised learning
  - e.g. Ravi & Knight (2009)
- Unsupervised POS induction

Questions

- Ideas about POS various and vague
- What determines POS?
  - context? what kind of context?
  - function words? morphology?
  - semantics?
- How well do words conform to POS in a language?
  - dense clusters?
  - how many rebels?

Goal

- improved unsupervised learning algorithm for POS
  - language-independent
  - incorporate morphology and syntactic context
  - semantics?
- currently: evaluate existing algorithms on non-Indo-European languages
  - so far: Lushootseed, Tagalog

Clark (2003)

- several algorithms developed
  - Clark (2003) least work to run
- both distributional & morphological info
- K-means clustering

Clark (2003)

- initialize k clusters
- maximize:

  P(word | prev. word) =

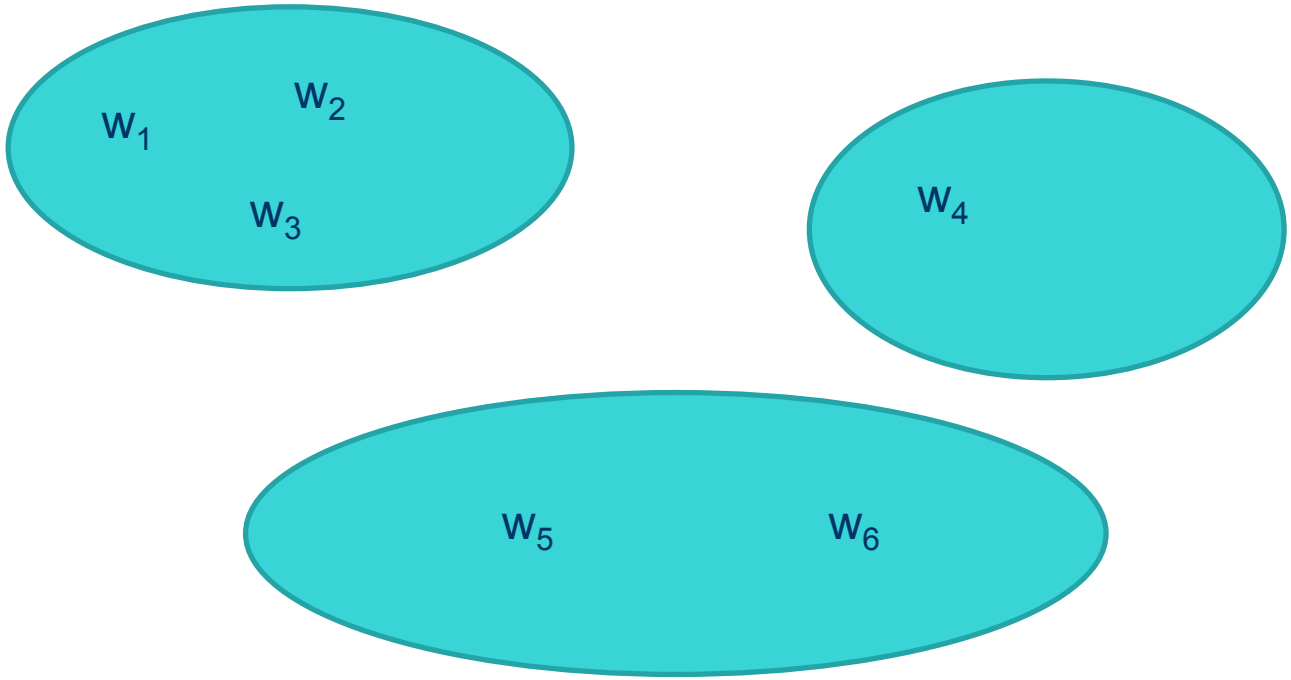
  P(word | category(word)) *

  P(category(word) | category(prev. word))

  dictionary

  grammar
- move each word to cluster that maximizes function

Clark (2003)



Corpora

- Lushootseed (Salishan)
  - 23,625 words
  - elicited stories (field work by Thom Hess)
- Tagalog (Austronesian)
  - 1,870,568 words
  - from Wikipedia
- both languages have disputed distinctions between nouns and verbs

Extracting from Wikipedia

- Parser available from PediaPress
  - Python library (mwlib)
  - writing your own parser not recommended
  - mwlib still not perfect, but final clean-up manageable (albeit tedious)
- text in other languages mixed in
  - e.g. English text in Tagalog articles
- advantage: free corpora available in different languages!

Evaluation

- Clark (2003) suggested 3 ways:
  - **manual evaluation**
  - **conditional entropy** of learned classes given pre-labeled POS
    - lower entropy = less surprise
  - **perplexity** of data based on bigram language model from learned classes
    - lower perplexity = less surprise

Evaluation

(see tables)

Evaluation

Perplexity per word \Clusters	8	32	64	128
Lushootseed	263.154	346.391	359.12	440.215
Tagalog	--	670.692	570.292	515.621

Clark (2003)

- drawbacks:
  - bigrams provide limited context
  - no difference between function and content words
    - syntactic vs. semantic clustering
  - limited morphological analysis
    - no recognition of morphological paradigms

Sketch of algorithm

- most frequent N words as function words
- contexts: one function word on each side
  - e.g. The second *sort* of information →

    the ___ *sort* of ___
- cluster content words by prob. distr. of contexts
  - can “see” layout of words in context space
  - cluster without specifying num. of clusters?
- then use content words to cluster function words
- morphological paradigms
  - can be as important than syntactic contexts

Issues

- one word, several POS (conversion)
  - e.g. swim, run, walk
- homography
  - e.g. bear, saw
- not easy, but oh, the glory . . .

References

- Clark, Alexander. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth Annual Meeting of the European Association for Computational Linguistics (EACL)*, pages 59-66, 2003.