Unsupervised Part of Speech Tagging

Topics in Natural Language Processing

Nora Hollenstein

Combining Distributional and Morphological Information for Part of Speech Induction

A. Clark (2003)

Overview

- Unsupervised Part of Speech Induction
- Clark's approach
  - Morphological information
  - Frequency information
- Evaluation
  - Cross-linguistic evaluation
  - Infrequent words
- Summary

Unsupervised PoS Tagging

- No labeled training data
- Find word categories by analyzing raw text
- Part-of-speech Tagging as a clustering task, not a sequence labeling task
- Approaches to unsupervised tagging:
  - Clustering-based algorithms
  - Hidden Markov Models

Clark's approach

- Focus on infrequent words
- Focus on non-English languages
- Learn a deterministic clustering:
  - Ney-Essen algorithm
- Distributional information
  - Frequency prior
- Morphological information
  - Character-level information

Model

- Corpus of length *N*: $w_1,\ldots,w_n$
- Class membership function *g*

- Bigram model:

$$P(w_i \mid w_{i-1}) = P(w_i \mid g(w_i))\, P(g(w_i) \mid g(w_{i-1}))$$

Morphology & frequency

- Define class membership function *g*
- Morphological information: encoded in the characters (HMM)
- Add prior class probabilities α_i (MLE)
- Combine morphological information & frequency:

$$P(g) = \prod_{i=1}^{n} \prod_{g(w)=i} \alpha_i \, P_i(w)$$

Ney-Essen clustering

- Similar to k-means clustering algorithm

1. Define number of clusters c
2. Split corpus randomly into c clusters
3. For each word, move it to the class that would cause the largest increase in likelihood of a certain model.
4. Repeat until no word changes class anymore

Cross-linguistic evaluation

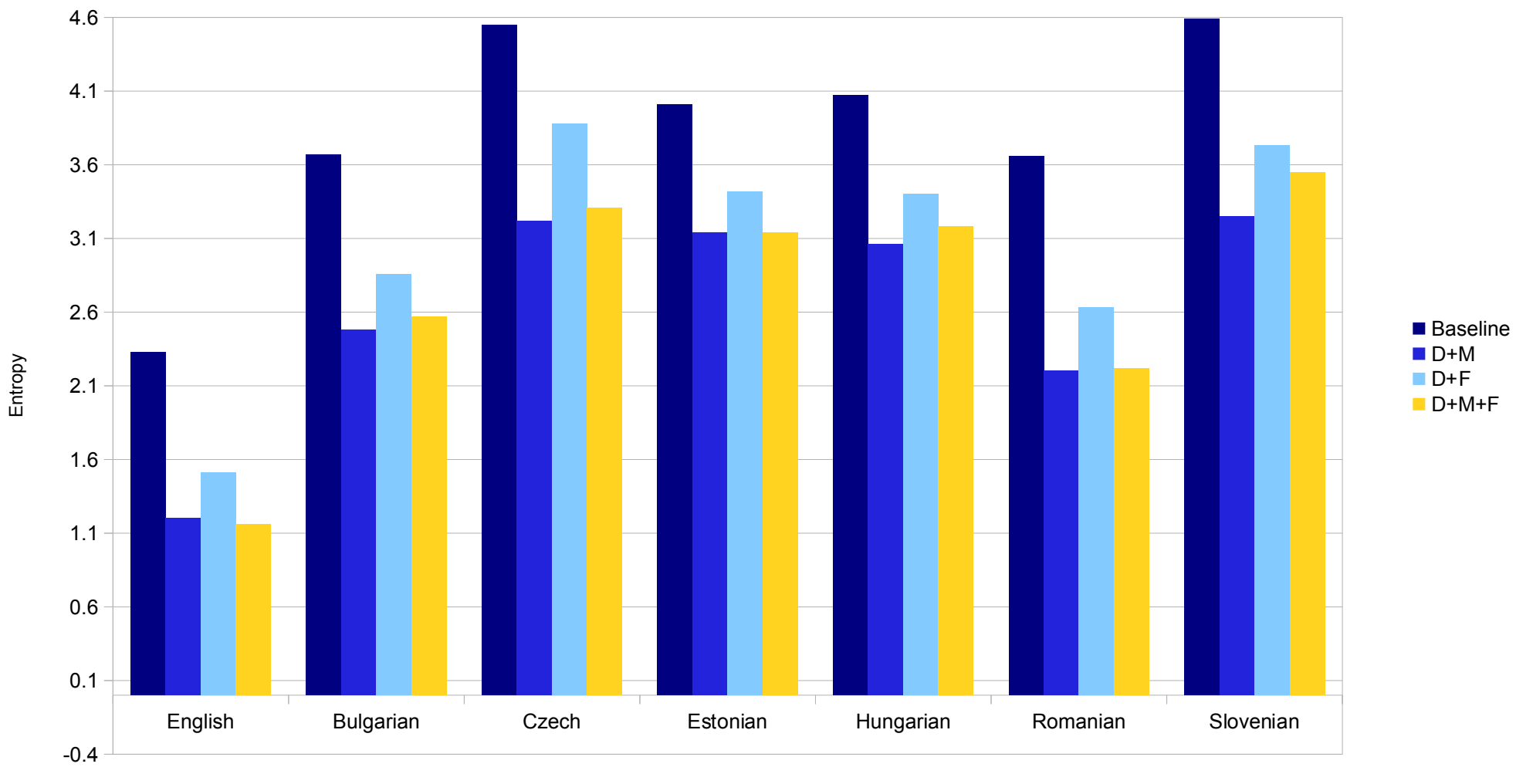
- English, Romanian, Czech, Slovenian, Bulgarian, Estonian and Hungarian
- MULTEXT-East parallel corpus
- Variation in tag sets between languages
- Small data sets: 90'000 – 120'000 tokens
- Model evaluated on conditional entropy $H(G|T)$:
  - Low entropy → mutual information between gold standard G and induced tags T is high

Cross-linguistic evaluation

Language	Hapaxes	Tags	*H(G)*	*H(G|T)*
English	4600	136	3.37	0.16
Bulgarian	9836	116	3.62	0.10
Czech	12048	956	4.41	0.21
Estonian	11643	404	3.92	0.14
Hungarian	13485	400	3.43	0.04
Romanian	8088	581	4.02	0.10
Slovene	10939	1033	4.34	0.20

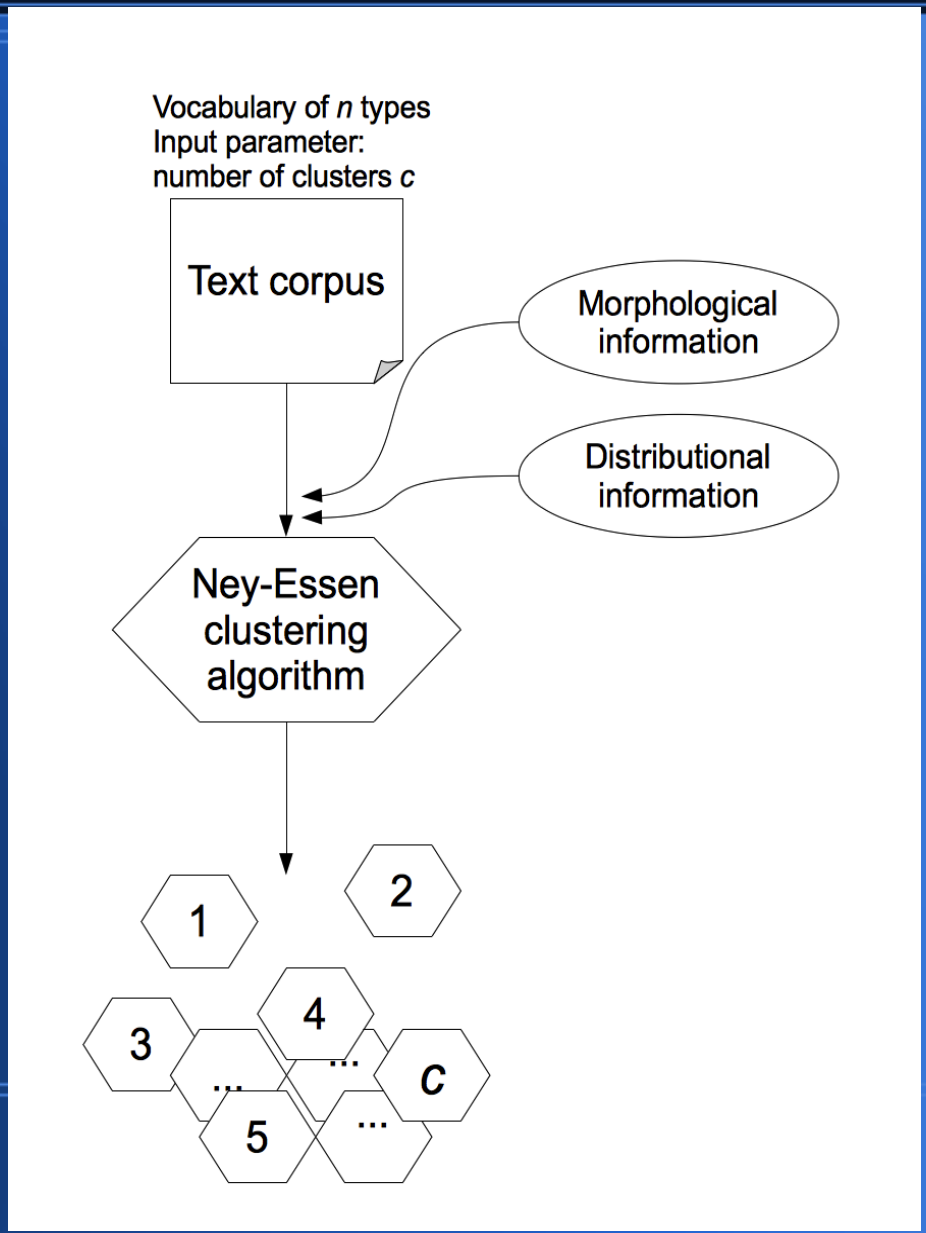
Table adapted from Clark (2003)

Evaluation


Entropy
4.6
4.1
3.6
3.1
2.6
2.1
1.6
1.1
0.6
0.1
-0.4
English
Bulgarian
Czech
Estonian
Hungarian
Romanian
Slovenian
Baseline
D+M
D+F
D+M+F

Summary

- Unsupervised PoS Induction including morphological and distributional cues
- Tagging as a clustering task
- Clark's algorithm works well for many languages
- Is still one of the best available algorithms (Christodoulopoulos et al. (2010)



References

- Clark, A. "Combining distributional and morphological information for part of speech induction." *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*-Volume 1. Association for Computational Linguistics, 2003.

- Christodoulopoulos, C., S. Goldwater, and M. Steedman. "Two Decades of Unsupervised POS induction: How far have we come?." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.

Questions?