An Empirical Investigation of Word Representations for Parsing the Web

NAIST

Sorami Hisamoto, Kevin Duh, Yuji Matsumoto
Nara Institute of Science and Technology

Overview

- Parsing is difficult for unrestricted web text (Accuracy: WSJ 90% → Web 80%).
- Word representation features obtained from large unlabeled data may combat data sparseness.
- We observed that word clusters/embeddings help most in the case of predicted part-of-speech (POS) tags.

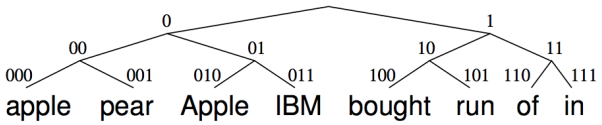
Dependency Parsing of Web Text

- Data: Google Web TreeBank from SANCL2012, containing 5 domains (*Answers, Emails, Newsgroups, Reviews, Weblogs*).
- Graph-based parser with arc-factored model.
- Extra word representations features are added on top of baseline features.

Brown Clustering

- Hierarchical clustering algorithm based on class-based bigram language model.
- It has been shown to improve accuracy. [Koo+ 2008]

An example of Brown clustering



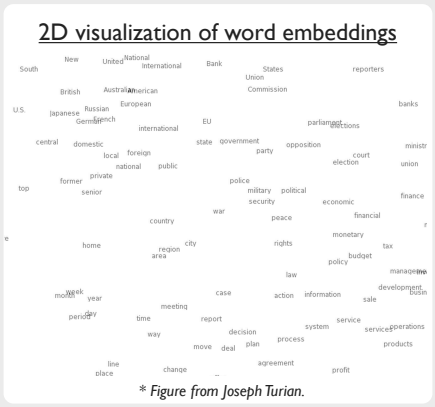
* Figure from Koo et al. 2008. "Simple Semi-supervised Dependency Parsing".

- We used short bit-string prefixes of the hierarchy, combined with word forms or POS tags, as features.

Collobert & Weston Embedding

- **Word Embedding**: word represented in a dense low dimensional real value vector form, often induced from a neural language model.
- It has been shown to improve accuracy of chunking & NER. [Turian+ 2010]
- We constructed features by clustering word embeddings:

2D visualization of word embeddings

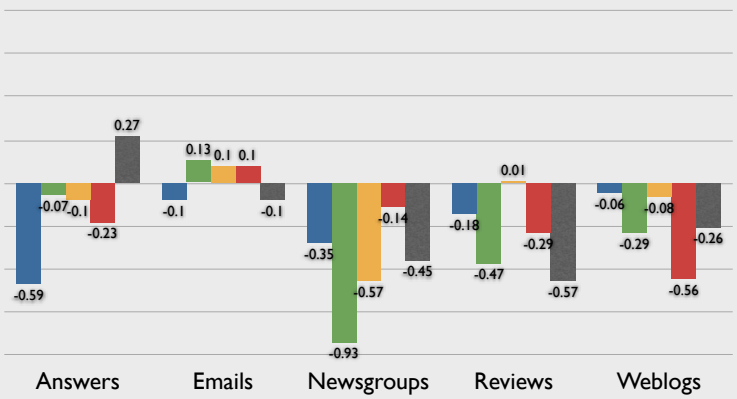


* Figure from Joseph Turian.

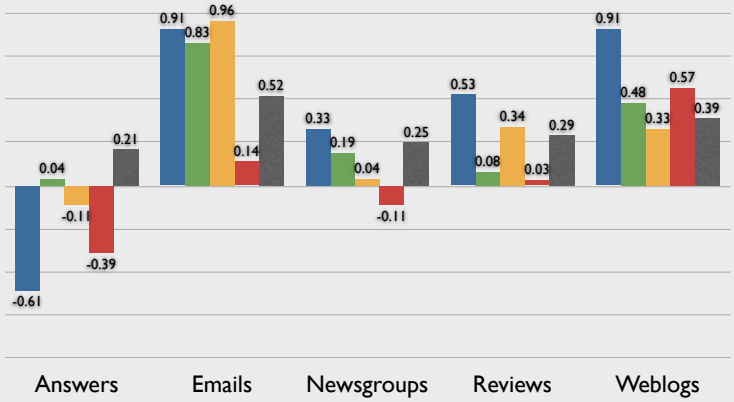
- We used repeated bisection algorithm to cluster embeddings, then use acquired cluster IDs as features, similar to Brown clustering.

Unlabeled Accuracy Relative to Baseline

Gold POS Tag Data Sets



Predicted POS Tag Data Sets



Brown (web), 50 clusters
C&W (web), 50 clusters ※
C&W (newswire), 50 clusters †
C&W (web), 1000 clusters ※
C&W (newswire), 1000 clusters †

† : Embeddings from [Turian+ 2010], trained on RCV1 (newswire) corpus. 1.3m sentences.
※ : Original embeddings induced using Google Web TreeBank's each domain, trained for 1 month. 30k to 2m sentences.

Discussion

- Extra features improved results on experiments with predicted POS tag data sets, but not with gold POS tag data sets.
- Brown clustering features outperforms word embedding features.

Future Work

- Induce word embeddings on in-domain data sets.
- Try different ways to construct features.