A Preliminary Evaluation of Word Representations for Named-Entity Recognition

Joseph Turian

joint work with Lev Ratinov, Yoshua Bengio, and Dan Roth

NIPS GRLL Workshop, 20091211
Natural language processing

- Running example, sequence labeling:

\[
\text{Pr dist over labels} \rightarrow m \rightarrow (3*n) \times m \rightarrow 3*n
\]

word -1, word 0, word +1
Natural language processing

- Words, words, words
- Words, words, words
- Words, words, words
How do we handle words?

• Not very well
“One-hot” word representation

- $n = |vocabulary|$, e.g. 50K for PTB2

$Pr$ dist over labels

- $3^*|V| \times m$
- $3^*|V|$ elements
- word -1, word 0, word +1
One-hot word representation

- 85% of vocab words occur as only 10% of corpus tokens
- Bad estimate of Pr(label|rare word)
Approach

- Manual feature engineering
Approach

• Manual feature engineering
Approach

• Induce word reprs over large corpus, unsupervised
• Use word reprs as word features for supervised task
Less sparse word reprs?

• Class-based (clustering) word reprs
• Distributed word reprs
Less sparse word reprs?

• Class-based (clustering) word reprs
• Distributed word reprs
Class-based word repr

• $|C|$ classes, hard clustering

$m$

$(|V|+|C|) \times m$

$|V|+|C|$
Class-based word repr

• Hard vs. soft clustering
• Hierarchical vs. flat clustering
Less sparse word reprs?

- Class-based (clustering) word reprs
  - Brown (hard, hierarchical) clustering
  - HMM (soft, flat) clustering
- Distributed word reprs
Less sparse word reprs?

- Class-based (clustering) word reprs
  - Brown (hard, hierarchical) clustering
  - HMM (soft, flat) clustering
- Distributed word reprs
Brown clustering

• Hard, hierarchical class-based LM
• Brown et al. (1992)
• Greedy technique for maximizing bigram mutual information
• Merge words by contextual similarity
Brown clustering

cluster(chairman) = `0010'
2-prefix(cluster(chairman)) = `00'

(image from Terry Koo)
Brown clustering

- Hard, hierarchical class-based LM
- 1000 classes
- Use prefixes = 4, 6, 10, 20
Less sparse word reprs?

- Class-based (clustering) word reprs
  - Brown (hard, hierarchical) clustering
  - HMM (soft, flat) clustering
- Distributed word reprs
HMM approach

- Soft, flat class-based repr
- Multinomial distribution over hidden states = word representation
- 80 hidden states
- Huang and Yates (2009)
- No results with HMM approach yet
Less sparse word reprs?

- Class-based (clustering) word reprs
- Distributed word reprs
Distributed word repr

- k- (low) dimensional, dense representation
- “word embedding” matrix $E$ of size $|V| \times k$
Sequence labeling w/ embeddings

“word embedding” matrix $E$ of size $|V| \times k$

$|V| \times k$, tied weights

word -1, word 0, word +1
Less sparse word reprs?

- Class-based (clustering) word reprs
- Distributed word reprs
  - Collobert + Weston (2008)
  - HLBL embeddings (Mnih + Hinton, 2007)
Less sparse word reprs?

- Class-based (clustering) word reprs
- Distributed word reprs
  - Collobert + Weston (2008)
  - HLBL embeddings (Mnih + Hinton, 2007)
Collobert + Weston 2008

\[
\text{score} > \mu + \overline{\text{score}}
\]

50*5

w1 w2 w3 w4 w5

1

100

w1 w2 w3 w4 \overline{w5}
50-dim embeddings: Collobert + Weston (2008)
t-SNE vis by van der Maaten + Hinton (2008)
Less sparse word reprs?

- Class-based (clustering) word reprs
- Distributed word reprs
  - Collobert + Weston (2008)
  - HLBL embeddings (Mnih + Hinton, 2007)
Log bilinear Language Model (LBL)

Linear prediction of w5

\[
\text{Pr} = \frac{\exp (\text{predict} \cdot \text{target})}{Z}
\]
HLBL

- HLBL = hierarchical (fast) training of LBL
- Mnih + Hinton (2009)
Approach

• Induce word reprs over large corpus, unsupervised
  – Brown: 3 days
  – HLBL: 1 week, 100 epochs
  – C&W: 4 weeks, 20 epochs

• Use word reprs as word features for supervised task
Unsupervised corpus

- RCV1 newswire
- 40M tokens (vocab = all 270K types)
Supervised Task

- Named entity recognition (NER)
- Averaged perceptron (linear classifier)
- Based upon Ratinov + Roth (2009)
Embeddings: Normalization

![Graph showing F1 on tuning data vs normalization constant for HLBL100, HLBL50, C&W50.]
Clean unsupervised corpus?

Hypothesis: Use curriculum learning? (Bengio et al, 2009)
Results (late-breaking)
Hypothesis

• If you have a linear supervised model (e.g. Perceptron):
  – Prefer reprs that are sparse + high-dim (e.g. Brown)
  – Too difficult to estimate Prs given dense + low-dim (embeddings)
• Especially true when supervised task involves many rare words (e.g. NER)
Difficulties with word embeddings

- No stopping criterion during unsup training
- More active features (slower sup training)
- Hyperparameters
  - Learning rate for model
  - (optional) Learning rate for embeddings
  - Normalization constant
- vs. Brown clusters, no hyperparams
Summary

• Accuracy on NER w/ linear sup classifier
  – Brown > embeddings
  – HLBL 100-dim > HLBL 50-dim
  – HLBL 50-dim \approx C&W 50-dim
• Unsup training speed
  – HLBL 20x faster to train than C&W
• Brown has 0 hyperparam
• Brown faster sup training vs embeddings
• Sparse high-dim > dense low-dim?