Word representations: a simple and general method for semi-supervised learning

GUANNAN LU

MARCH 17,2015

Outline

- Motivation
- Word representations
  - Distributional representations
  - Clustering-based representations
  - Distributed representations
- Supervised evaluation tasks
  - Chunking
  - Named entity recognition (NER)
- Experiments & Results
- Summary

Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming

Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
- A popular approach:
  - use unsupervised methods to induce word features

Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
- A popular approach:
  - use unsupervised methods to induce word features
    - clustering
    - word embeddings

Motivation

- Semi-supervised approaches can improve accuracy
- It can be tricky and time-consuming
- A popular approach:
  - use unsupervised methods to induce word features
    - clustering
    - word embeddings
- Questions:
  - Which features are good for what tasks?
  - Should we prefer certain word features?
  - Can we combine them?

Word Representations

- Word representation:
  - A mathematical object associated with each word, often a vector
- Word feature: each dimension's value
- Conventional representation
  - E.g. One-hot representation
  - Problems:
    - Data sparsity

Distributional representations

- Co-occurrence matrix F: $W \times C$
  - Each row F_w is initial representation of word w
  - Each column F_c is some context

Distributional representations

- Co-occurrence matrix F: $W \times C$
  - Each row F_w is initial representation of word w
  - Each column F_c is some context
- Function g: $f = g(F)$
  - Map F to f: $W \times d$ where $d << C$

Distributional representations

- Co-occurrence matrix F: $W \times C$
  - Each row F_w is initial representation of word w
  - Each column F_c is some context
- Function g: $f = g(F)$
  - Map F to f: $W \times d$ where $d << C$
- LSA: term-document matrix (Landauer et al., 1998)

Clustering-based representations

- Brown clustering (Brown et al., 1992)
  - A hierarchical clustering algorithm
  - A class-based bigram language model
  - Time complexity: $O(V*K^2)$
    - V is the size of the vocabulary, K is the number of clusters.
  - Limitations：
    - Only based on bigram statistics
    - not consider word usage

Distributed representations

- Not to be confused with distributional representations!

Distributed representations

- Not to be confused with distributional representations!
- also known as word embeddings
- dense, real-valued, low-dimensional
- Neural language models

Distributed representations

- Collobert and Weston embeddings (2008)
  - Neural language model
  - Discriminative and non-probabilistic
  - General architecture (e.g. SRL, NER, POS tagging)
- Differences on implementation
  - Not achieve the low log-rank
  - Corrupt the last word for each n-gram
  - Learning rates are separated

Distributed representation

- HLBL embeddings(2009)
  - Log-bilinear model
    - Predict the feature vector of the next word
  - Hierarchical structure (binary tree)
    - Represent each word as a leaf with a particular path
    - Calculate the product of the probability of each binary choice

Evaluation tasks

- Chunking: syntactic sequence labeling
  - CoNLL-2000 shared task
  - CRFsuite
  - Data
    - The Penn Treebank
    - 7936 sentences(training)
    - 1000 sentences (development)

Evaluation tasks

- NER: sequence prediction problem
  - The regularized averaged perceptron model (Ratinov and Roth, 2009)
  - CoNLL03 shared task
    - 204k words for training, 51k words for development, 46K words for testing
  - Out-of-domain dataset: MUC7 formal run (59K words)

Evaluation---Features

Chunking

- Word features: w_i for i in $\{-2,-1,0,+1,+2\}$, $w_i \wedge w_{i+1}$ for i in $\{-1,0\}$.
- Tag features: w_i for i in $\{-2,-1,0,+1,+2\}$, $t_i \wedge t_{i+1}$ for i in $\{-2,-1,0,+1\}$. $t_i \wedge t_{i+1} \wedge t_{i+2}$ for i in $\{-2,-1,0\}$.
- Embedding features [if applicable]: $e_i[d]$ for i in $\{-2,-1,0,+1,+2\}$, where d ranges over the dimensions of the embedding e_i.
- Brown features [if applicable]: $substr(b_i, 0, p)$ for i in $\{-2,-1,0,+1,+2\}$, where $substr$ takes the p-length prefix of the Brown cluster b_i.

NER

- Previous two predictions y_{i-1} and y_{i-2}
- Current word x_i
- x_i word type information: all-capitalized, is-capitalized, all-digits, alphanumeric, etc.
- Prefixes and suffixes of x_i, if the word contains hyphens, then the tokens between the hyphens
- Tokens in the window $c = (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2})$
- Capitalization pattern in the window c
- Conjunction of c and y_{i-1}.

Experiment

- Unlabeled data
- RCV1 corpus (63 millions words in 3.3 million sentences)
- Preprocessing technique(Liang, 2005)
  - Remove all sentences that are less than 90% lowercase a-z.

Results

- Scaling of word embeddings

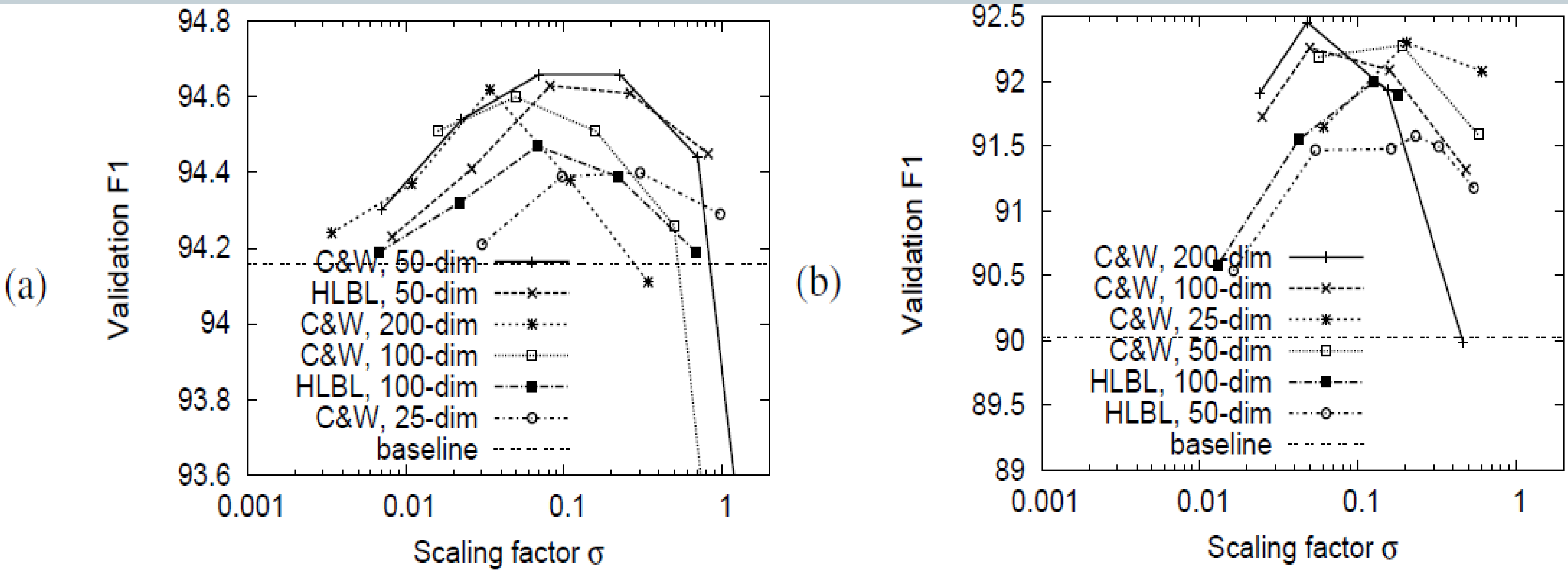


Figure 1: Effect as we vary the scaling factor σ (Equation 1) on the validation set F1. We experiment with Collobert and Weston (2008) and HLBL embeddings of various dimensionality. (a) Chunking results. (b) NER results.

Results

- Capacity of word representations

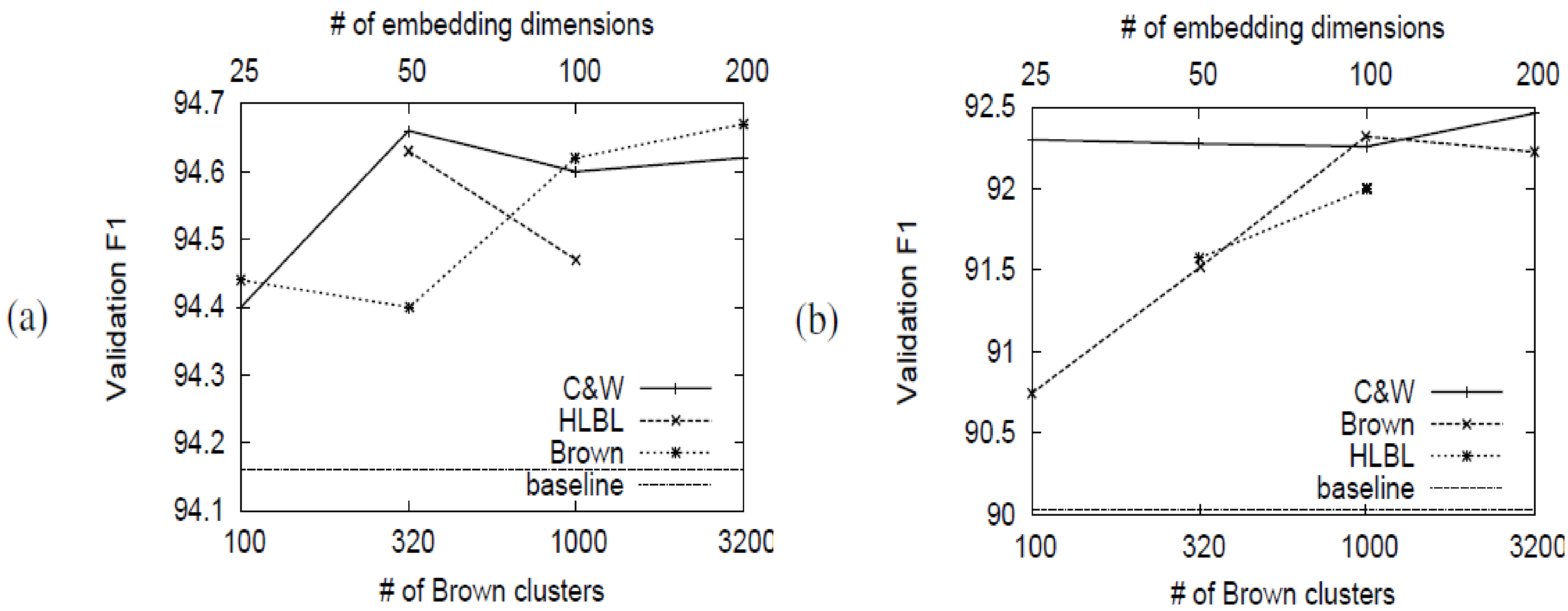


Figure 2: Effect as we vary the capacity of the word representations on the validation set F1. (a) Chunking results. (b) NER results.

Results

System	Dev	Test
Baseline	94.16	93.79
HLBL, 50-dim	94.63	94.00
C&W, 50-dim	94.66	94.10
Brown, 3200 clusters	**94.67**	**94.11**
Brown+HLBL, 37M	94.62	94.13
C&W+HLBL, 37M	94.68	94.25
Brown+C&W+HLBL, 37M	94.72	94.15
Brown+C&W, 37M	94.76	94.35
Ando and Zhang (2005), 15M	-	94.39
Suzuki and Isozaki (2008), 15M	-	94.67
Suzuki and Isozaki (2008), 1B	-	**95.15**

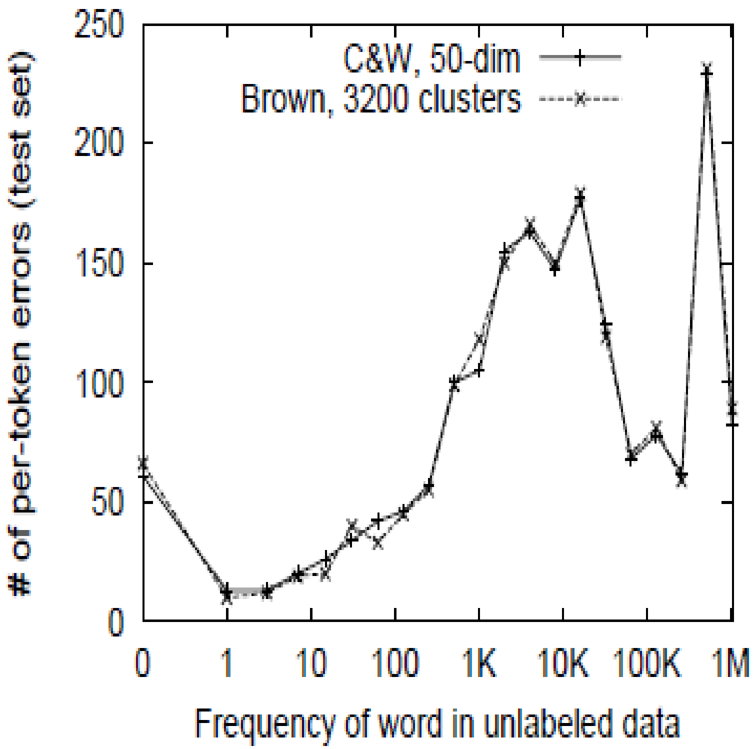
System	Dev	Test	MUC7
Baseline	90.03	84.39	67.48
Baseline+Nonlocal	91.91	86.52	71.80
HLBL 100-dim	92.00	88.13	75.25
Gazetteers	92.09	87.36	77.76
C&W 50-dim	92.27	87.93	75.74
Brown, 1000 clusters	92.32	**88.52**	**78.84**
C&W 200-dim	**92.46**	87.96	75.51
C&W+HLBL	92.52	88.56	78.64
Brown+HLBL	92.56	88.93	77.85
Brown+C&W	92.79	89.31	80.13
HLBL+Gaz	92.91	89.35	79.29
C&W+Gaz	92.98	88.88	81.44
Brown+Gaz	**93.25**	**89.41**	**82.71**
Lin and Wu (2009), 3.4B	-	88.44	-
Ando and Zhang (2005), 27M	93.15	89.31	-
Suzuki and Isozaki (2008), 37M	93.66	89.36	-
Suzuki and Isozaki (2008), 1B	**94.48**	89.92	-
All (Brown+C&W+HLBL+Gaz), 37M	93.17	90.04	82.50
All+Nonlocal, 37M	93.95	90.36	84.15
Lin and Wu (2009), 700B	-	**90.90**	-

Chunking

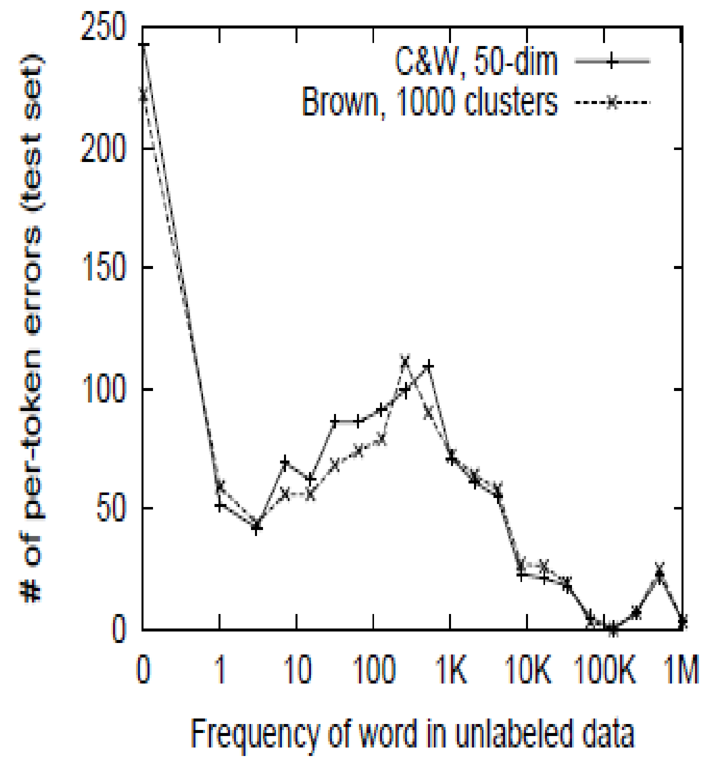
NER

Results

(a)



(b)



Chunking

NER

Summary

- Word features
  - in an unsupervised, task-inspecific, and model-agnostic manner
- The disadvantage
  - Accuracy might be lower than a task-specific semi-supervised method
- The contributions
  - The first work to compare different word representations
  - Combining different word representations can improve accuracy further
- Future work
  - Induce phrase representations
  - Apply to other supervised NLP systems

References

- Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18, 467–479.
- Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *ICML*.
- Landauer, T. K., Foltz, P.W., & Laham, D. (1998).An introduction to latent semantic analysis. *Discourse Processes*, 259–284.
- Liang, P. (2005). Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology
- Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. *NIPS* (pp. 1081–1088).
- Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. *CoNLL*.
- Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394). Association for Computational Linguistics.

Q&A

Any questions?

Thank you!