Word Representations: a Simple and General Method for Semi-Supervised Learning [Turian et al., 2012]

Massimo Innamorati

University of Edinburgh

February 26, 2016

Overview

Motivation

Observation

Semi-supervised NLP systems achieve higher accuracy than their supervised counterparts.

Motivation

Observation

Semi-supervised NLP systems achieve higher accuracy than their supervised counterparts.

Problem

Which features - or combination thereof - to use given a task?

Motivation

Observation

Semi-supervised NLP systems achieve higher accuracy than their supervised counterparts.

Problem

Which features - or combination thereof - to use given a task?

Focus

Clustering-based and distributed representations.
Sequence labelling tasks: NER and chunking.

Distributional Representations

Aim

Generate a cooccurence matrix F of size WxC.

Distributional Representations

Aim

Generate a cooccurence matrix F of size WxC.

Settings

Choose a context (window direction and size).

Distributional Representations

Aim

Generate a cooccurence matrix F of size WxC.

Settings

Choose a context (window direction and size).
Choose a count.

Distributional Representations

Aim

Generate a cooccurence matrix F of size WxC.

Settings

Choose a context (window direction and size).
Choose a count.
Choose a function g to reduce dimensionality of F_w.

Distributional Representations

Aim

Generate a cooccurence matrix F of size WxC.

Settings

Choose a context (window direction and size).
Choose a count.
Choose a function g to reduce dimensionality of F_w.

Previous Literature

[Salgren, 2006] Improves classification tasks (e.g. IR, WSD). Not known which settings ideal for NER & chunking.

Brown Clustering

Aim

Generate K hierarchical clusters based on bigram mutual information.

Brown Clustering

Aim

Generate K hierarchical clusters based on bigram mutual information.

Sample output:

```
cat 1101
dog 1100
city 1001
town 1011
```

Brown Clustering

Aim

Generate K hierarchical clusters based on bigram mutual information.

Sample output:

```
cat 1101
dog 1100
city 1001
town 1011
```

Pros & Cons

Hierarchy allows to choose among several levels.

Use of bigrams is restrictive.

Distributed Representations

Aim

Use a neural network to generate word vectors whose features capture latent semantic and syntactic properties.

Distributed Representations [Collobert & Weston, 2008]

Training

- *for* each epoch
  - Read an n-gram $x = (w_1, ..., w_n)$
  - Calculate $e(x) = e(w_1) \oplus .... \oplus e(w_n)$
  - Pick a *corrupted n-gram* $\tilde{x} = (w_1, ..., w_{n-q}, \tilde{w}_n)$ and calculate $e(\tilde{x})$
  - Get $s(x)$ by passing $e(x)$ through SLNN.
  - $L(x) = \max(0, 1 - s(x) + s(\tilde{x}))$

Distributed Representations
Hierarchical Log-Bilinear model [Mnih & Hinton, 2009]

Training

Given an n-gram, concatenate embeddings of $n-1$ first words.
Learn a linear model to predict the last word.

Aims

Hypothesis

It is a task-independent generalisation that supervised NLP systems can be improved by adding word representations as word vectors (thus turning them into semi-supervised systems).

Aims

Hypothesis

It is a task-independent generalisation that supervised NLP systems can be improved by adding word representations as word vectors (thus turning them into semi-supervised systems).

Method

Compare semi-supervised models obtained with off-the-shelf embeddings to previous ones with task-specific information, in particular [Ando & Zhang, 2005] and [Suzuki & Isozaki, 2008] for chunking and [Lin & Wu, 2009] for NER.

Chunking

Syntactic sequence labelling task, consisting of identifying phrases.

Chunking

Syntactic sequence labelling task, consisting of identifying phrases.

Method

Use publicly available CRFsuite chunker.
Add word embedding features learnt from RCV1 corpus (1.3M sentences).
Train on 8.9K sentences of WSJ newswire in Penn Treebank corpus.

NER

Classification task.

NER

Classification task.

Method

Use publicly available system by [Ratinov & Roth 2009].
Train on 14K sentences of Reuters newswire from CoNLL03 dataset.
Add word embedding features learnt from RCV1 corpus (1.3M sentences).
Test on Reuter + out-of-domain dataset MUC7 (with unseen NE types).

Induction of Word Representations

Brown

Models with 1000, 100, 320, and 3200 clusters.

Used clusters at depth 4, 6, 10, and 20.

Induction of Word Representations

Brown

Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.

Collobert & Weston

50 epochs.
Embeddings of dimensionality 25, 50, 100, or 200 learnt over 5-gram windows.

Induction of Word Representations

Brown

Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.

Collobert & Weston

50 epochs.
Embeddings of dimensionality 25, 50, 100, or 200 learnt over 5-gram windows.

HLBL

Embeddings of dimensionality 100 learnt over 5-gram windows.

Induction of Word Representations

Brown

Models with 1000, 100, 320, and 3200 clusters.
Used clusters at depth 4, 6, 10, and 20.

Collobert & Weston

50 epochs.
Embeddings of dimensionality 25, 50, 100, or 200 learnt over 5-gram windows.

HLBL

Embeddings of dimensionality 100 learnt over 5-gram windows.

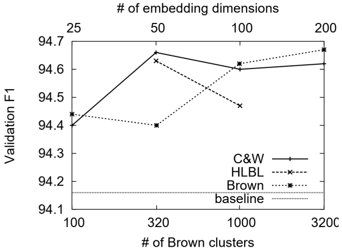
Scaling of Embeddings

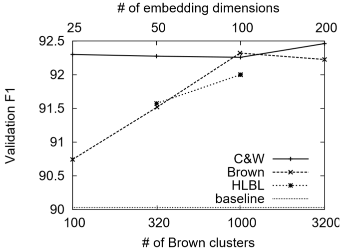
In all cases, the features are bounded by a scaling constant σ that sets their new standard deviation.

$$E \leftarrow \sigma \cdot / stddev(E) \tag{1}$$

Results

Influence of capacity of embeddings on chunking (top) and NER (bottom)





Results

Final results for chunking.

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**

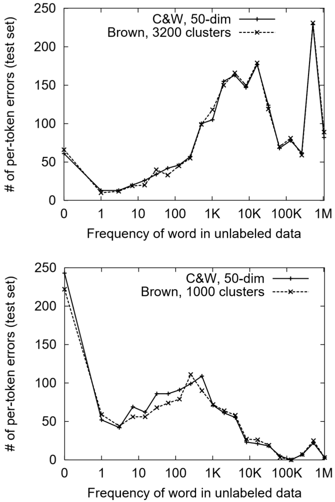
Results

Final results for NER.

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**	-

Results

Per-token errors given word frequency in chunking (top) and NER (bottom).



Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009].

Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009].

However, they are more general, and prove that task-agnostic embeddings can be used to improve supervised systems.

Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009].

However, they are more general, and prove that task-agnostic embeddings can be used to improve supervised systems.

It is also found that Brown embeddings are better for rare words, and a default method for scaling is presented.

Conclusion

These models do not outperform the state-of-the-art semi-supervised by [Ando & Zhang, 2005], [Suzuki & Isozaki, 2008], and [Lin & Wu, 2009].

However, they are more general, and prove that task-agnostic embeddings can be used to improve supervised systems.

It is also found that Brown embeddings are better for rare words, and a default method for scaling is presented.

Extending the embeddings to phrase representations may be useful.

References

Turian, J., Ratinov, L., and Bengio, Y. (2010)
Word representations: A simple and general method for semi-supervised learning
Proceedings of the 48th annual meeting of the association for computational linguistics. 12(3), 45 – 678.

Sahlgren, M. (2006)
The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces
Institutionen för lingvistik

Collobert, R., and Weston, J. (2008)
A unified architecture for natural language processing: Deep neural networks with multitask learning.
Proceedings of the 25th international conference on Machine learning.

Mnih, A., and Hinton, G. (2009)
A scalable hierarchical distributed language model.
Advances in neural information processing systems.

References

Ando, R. K., and Zhang, T. (2005).
A high-performance semi-supervised learning method for text chunking.
Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics.

Suzuki, J., and Isozaki, H. (2008).
Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data.
Association for Computational Linguistics (pp. 665-673).

Lin, D., and Wu, X. (2009).
Phrase clustering for discriminative learning. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th
International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1030-1038). Association for Computational Linguistics.

References

Ratinov, L., and Roth, D. (2009).
Design challenges and misconceptions in named entity recognition.
Proceedings of the Thirteenth Conference on Computational Natural Language Learning (pp. 147-155). Association for Computational Linguistics.

The End