Advanced Science and Technology Letters
Vol.120 (AIA 2015), pp.784-788
http://dx.doi.org/10.14257/astl.2015.120.154

Comparison of NER Performance Using Word Embedding

Miran Seok[1,2], Hye-Jeong Song[1,2], Chan-Young Park[1,2], Jong-Dae Kim[1,2], Yu-seop Kim[1,2,1]

[1]Department of Convergence Software, Hallym University, Korea
[2]Bio-IT Research Center, Hallym University, Korea
smr4880@hanmail.net, {hjsong, cypark, kimjd, yskim01}@hallym.ac.kr

Abstract. Recent studies in NER use the supervised machine learning. This study used CRF as a learning algorithm, and applied word embedding to feature for NER training. Word embedding is helpful in many learning algorithms of NLP, indicating that words in a sentence are mapped by a real vector in a low-dimension space. As a result of comparing the performance of multiple techniques for word embedding to NER, it was found that CCA (85.96%) in Test A and Word2Vec (80.72%) in Test B exhibited the best performance.

Keywords: Natural Language Processing, Named Entity Recognition, Word Embedding

1 Introduction

Recently, there has been great interest in using unlabeled data to learn word representation that may be used as a feature in supervised classifiers for a natural language processing task. Word embedding is a feature learning technique in NLP, indicating that words in a sentence are mapped by a real vector in a low-dimension space. This study uses word embedding as a feature for learning named entity recognition and CRF [2-6] as a learning algorithm. As a result of comparing the performance of multiple techniques (GloVe, Word2Vec, and CCA) to NER in order to deduct word embedding, CCA in Test A and Word2Vec in Test B showed an improvement in performance of approximately 1.8% and 3.6%, respectively, compared to the baseline.

[1] He is a corresponding author

ISSN: 2287-1233 ASTL

2 Word Embedding

2.1 Summation

Named entity refers to phrases that include the names of persons, organizations, and locations, and NER classifies all words of a document in predefined categories. NER, generally covered as a sequence prediction issue, is a highly important phase in extracting and managing intelligence information.

2.2 Data

This study used the 2003 shared task corpus (English) of CoNLL for NER learning.

```
U.N.        NNP  I-NP  I-ORG
official    NN   I-NP  O
Ekeus       NNP  I-NP  I-PER
heads       VBZ  I-VP  O
for         IN   I-PP  O
Baghdad     NNP  I-NP  I-LOC
.           .    O     O
```

Fig. 1. Example of CoNLL-2003 shared task data file.

The first item of each line is the word, the second item is the part of speech, the third item is the syntactic chunk tag, and the fourth item is the named entity tag. The O tag refers to tokens that do not belong to a certain named entity.

3 Word Embedding

Word embedding is mapping words with vectors of real numbers, helping make better performance in natural language processing by grouping similar words. Word embedding can capture various dimensions of meanings and phrase information relevant to the potential features of words within the vector.

3.1 Global Vector

GloVe is an unsupervised learning algorithm to obtain vector representations of words. Learning is carried out in global word-word co-occurrence statistics counted from the corpus. The GloVe model learns items that are not 0 in the global word-word co-occurrence matrix. This matrix shows how often the words co-occur in the given corpus.

3.2 Word2Vec

Word2Vec is a linguistic model based on a neural network that learns the embedding of each word in the corpus. Word2Vec provides skip-gram architecture and continuous bag-of-words (CBOW) architecture. This study established word embedding using the skip-gram model of Word2Vec. Skip-gram predicts the neighboring words or context when a single word is given.

3.3 Canonical Correlation Analysis (CCA)

Canonical correlation analysis is a statistical method used to investigate the relationship between two or more variable sets. CCA can be calculated as representations similar to covariance matrices. The algorithm for the two representations is based on singular value decomposition (SVD). Also, CCA can be used for dimensional reduction, and provides the correlation between two sets at the d dimension and the relevant projection vector.

4 Conditional Random Field

Conditional random fields (CRFs) represent a probabilistic framework for labeling and segmenting sequential data [3]. We used the CRF model for the NER task. The input is a sentence consisting of a sequence of words $x_1, \cdots, x_n$; tags $y_1, \cdots, y_n$ suitable for each word (token) are allocated.

$$P(y|x) = \frac{1}{Z(x)} \prod_{i=1}^{x} exp \sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, i) \tag{1}$$

In equation (1), $f_k(y_{i-1}, y_i, x, i)$ is feature function, and λ_k is weight learned about the feature. Z(x) is a normalization function. This study used the average perceptron algorithm for CRF learning.

5 Feature Representation

This study used two feature sets of baseline and word embedding for NER.

5.1 Baseline features

The first feature set consists of unigram and bigram of the words, parts of speech, and the unigram and bigram of the named entity tag.

The second feature set is applied to rare words that appear fewer than five times in the training data. Prefix and suffix features are less than four characters in length, and isDigit, isUpper and isHyphen features encode other morphological characteristics.

5.2 Word embedding features

This study used dimension values of word embedding as a new feature. Embedding was established using GloVe, Word2Vec, and CCA, consisting of 44,532 words that appeared at least 101 times in the data.

6 Experiments and Results

In this study, NER was learned using the average perceptron algorithm of CRFsuite, and the 2003 shared task corpus (English) of CoNLL was used for training and testing.

The Reuters Corpus Volume 1 was used to create word embedding used as a feature. The dimension of word embedding used in this study is 50, and window_size for context information was set as four words in front and four words in back, based on the current word.

The experiment was performed in the case where the baseline feature was used and the case where GloVe, Word2Vec, and CCA embeddings were added as features to the baseline.

Table 1. Experimental F1-score results.

	Test A	Test B
Baseline	84.12%	77.06%
Baseline +GloVe	85.18%	79.48%
Baseline +Word2Vec	85.89%	**80.72%**
Baseline +CCA	**85.96%**	80.68%

7 Conclusion

When embedding was established using various word embedding methods and used as a feature for NER, CCA and Word2Vec was obtained highest F1 score in each test. Moreover, the word expression ability of word embedding could be found through the nearest neighbors.

In addition to NER, word embedding can be applied to multiple NLP tasks such as syntactic parsing and sentiment analysis to obtain improved results in the future.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the

Ministry of Science, ICT and future Planning (2015R1A2A2A01007333) and by the Ministry of Education, Science and Technology (2010-0010612).

References

1. Qiu, L., Cao, Y., Nie, Z., and Rui, Y.: Learning Word Representation Considering Proximity and Ambiguity. Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
2. McCallum, A., and Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons. In: Conference on Computational Natural Language Learning (2003)
3. Song, Y., Kim, E., Lee, G.G., and Yi, B.K.: POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004. In: International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (2004)
4. Konkol, M., and Konopík, M.: Crf-based czech named entity recognizer and consolidation of czech ner research. Text, Speech, and Dialogue, Springer Berlin Heidelberg (2013)
5. Xu, Z., Qian, X., Zhang, Y., and Zhou, Y.: CRF-based Hybrid Model for Word Segmentation, NER and even POS Tagging. IJCNLP (2008)
6. Ekbal, A., and Bandyopadhyay, S.: Voted NER system using appropriate unlabeled data. In: 2009 Named Entities Workshop: Shared Task on Transliteration. Association for Computational Linguistics (2009)

