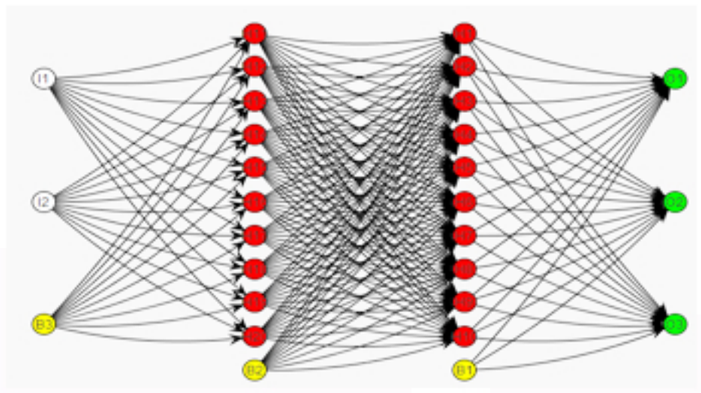
Named Entity Recognition and Classification using Word Vectors

By Rajkiran Veluri, Zia Ahmed for CS229



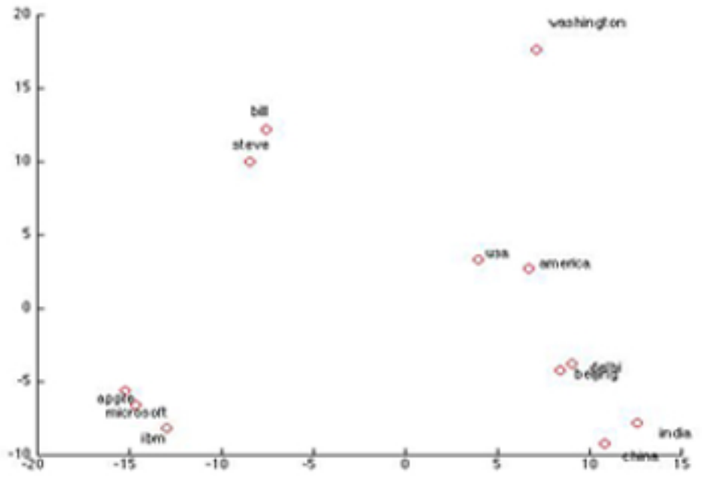


Fig: 1

MOTIVATION:

In this project we investigate the word2Vec model as proposed by Tomas Mikilov for determining word relationships and use the word vectors to implement a Named Entity Recognition System. Further we investigate clustering as a way to classify Word vectors and explore how K-means clustering can be used to improve the performance of the NER system.

MODELS:

Word Vectors

The method used to generate word vectors is the continuous bag-of-word model (CBOW) by Mikolov et al. (2013). It is a neural network model which tends to predict the target word based on the input window of context words surrounding the target word. The training process creates low-dimensional word vectors (each word is 200 dimensional) for each word in the training corpus. The word vectors which are contextually, syntactically and semantically similar tend to lie near each other in this low dimensional space, as shown in the PCA analysis of the few handpicked words from the vocabulary [Fig:1]. We use these word representations as features to build the NER system which is described next.

Named Entity Recognition

NER is a classification problem, where each input word is classified as being a Location, Person, Organization, Miscellaneous and Other (not any named entity). The algorithm uses a tokenized text to train a single-layer neural network model for named entity recognition with multiple classes. As the algorithms iterates through the dataset it learns both the classifier and the word representations. The training and the testing data for the NER algorithm is taken from CoNLL03 corpus. The data consists of sentences with one token per line and each token is associated with 5 possible labels: {O, LOC, MISC, ORG, PER} representing the classes defined above. The word vectors learned using the CBOW model were used to construct context windows that served as input features to the neural network.

The evaluation of the implemented algorithm was done using the CoNLL03 conlleval Perl script. The script evaluates the NER system's capability of identifying named entities. It gives a clear presentation of the performance of the system on various entity categories (person, location, organization, miscellaneous and other) based on the precision, recall and F1 measures.

Visualization with Principal Component Analysis and Preprocessing with K-means Clustering

We visualize the word vectors using PCA by reducing the 200 dimensional word vectors to 2 dimensions. This helps in gaining an intuition on how the vectors form clusters and thus lend semantic meaning from the context [surrounding window of words] in which they are situated. We use K-means as a pre-processing step with different number of clusters and see if they can be used for training NER labels and how they relate to the entities in our NER system. Then we run the NER system on the test data and see if the accuracy of the system is affected by this training procedure.

TUNING PARAMETERS:

The parameters of the system that were tuned for higher accuracy were:

- The regularization constant λ
- The learning rate α
- The context window size C
- The number of iterations (epochs)

The optimal values found for the tuning parameters are given in the following table:

Parameters	Optimal Value
Epoch	40
Learning Rate	0.075
Regularization	0.02
Context Window Size	5

Fig: 2

RESULTS:

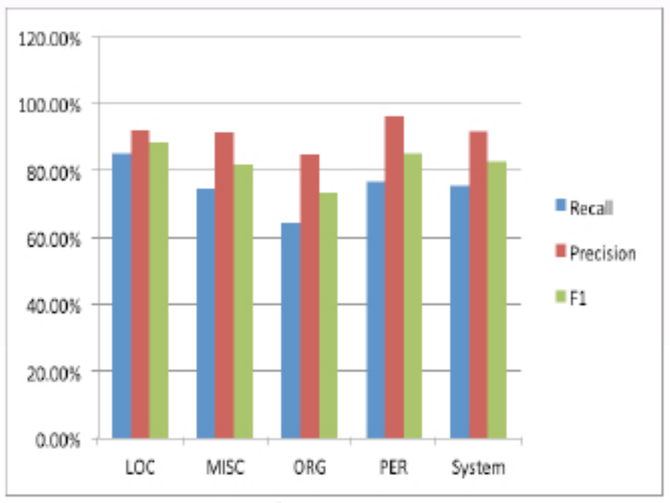


Fig: 3

	Recall	Precision	F1
LOC	85.20%	92.10%	88.52%
MISC	74.53%	91.39%	82.10%
ORG	64.58%	84.54%	73.22%
PER	76.53%	96.17%	85.23%
System	75.44%	91.73%	82.79%

Fig: 4

FURTHER WORK:

We are still working on our results on K-means to separate correlated entity clusters to achieve better accuracy. One thing we are trying to explore is if nested clustering (cluster of clusters) gives us more clearly separated entity classes[3]. We are also trying to see if Named Entity Recognition can be extended to create a system for identifying candidate answers in a Question Answering system.

REFERENCES AND ACKNOWLEDGEMENTS:

We would like Prof. Andrew Ng and our project TA Youssef Ahres for their support and guidance in enabling this project.

[1]) L. Ratinov and D. Roth, 2009, "Design Challenges and Misconceptions in Named Entity Recognition". Retrieved from https://aclweb.org/anthology/W/W09/W09-1119.pdf

[2] https://code.google.com/p/word2vec/

[3] S. K. Siencknik, *Adapting Word2Vec to Named Entity Recognition*, Proceedings of the 20th Nordic Conference of Computational Linguistics, 2015.

Data: http://mattmahoney.net/dc/text8.zip