人工知能学会研究会資料
SIG-FIN-016-12

Comparison among multilingual financial words using the word2vec and clustering with news resources for automatic creation of financial dictionaries

Enda Liu[1], Tomoki Ito[1], and Kiyoshi Izumi

School of Engineering, The University of Tokyo

Abstract: Vector representation of words such as word2vec is an efficient method used in text mining. However, few papers are focusing on the multilingual studies. In this paper we present the comparative study on English and Japanese resources respectively, and then we try to investigate the possible relationship between the two vector models in two languages. We first extract two word2vec models by using news resources of ten years, and then we cluster them basing on their cosine similarity for both Japanese and English respectively. Second, we extract the words related to finance and then derive two dictionaries in two languages. Finally, we make a comparison between these two dictionaries and tempt to Sentiment estimation of a cluster of one language based on similar clusters of other language.

1 Introduction

Financial text mining is a very important part of aspect of the field of data mining and many studies have been done recent years basing on the machine learning and natural language processing. The prediction of stock price basing on the text mining of stock message board is one of the prevalent research topics, where a sentimental dictionary could be derived so that it becomes easy to identify the whether a word or a message contains positive or negative influences to one or multiple stock prices. However, most of the dictionaries are basing on the text resources in only single language and thereby the relationship of the positive-negative score of the words between multiple languages are seldom studied.

On the other hand, the sentimental resources are not balanced among languages. The amount and variety of the sentimental dictionaries in English is considered to be most since it is most commonly used, whereas other languages including Japanese are less professional, especially on some specific area, for instance, regarding to financial market. It becomes a meaningful and promising work to leverage English text resources and dictionaries in order to derive other dictionaries in other languages, such as in Japanese. Furthermore, a system might be developed for automatic creation of financial dictionaries with multi-lingual text resources basing on this.

In this paper, we make a comparison of the clustering results of two groups of identical words in Japanese and English, after implementing the word2vec [1] algorithm on both English and Japanese financial text resources respectively, attempting to excavate the relationship between them, which might become key factors for constructing the automatic creation system.

2 Framework of multilingual word clustering

2.1 Preprocessing the text resources

The preprocessing of the data consists of four parts: retrieving, cleaning, tagging and lemmatization of the original text resources.

In this study, we choose a stock message during the period of year 2010, from Stocktwits, a system that is able to automatically collect English information about stock on the Internet. Similarly, we retrieve Japanese financial message board online. Both of the raw data we obtained contain some unsolvable elements, such as special characters, http and email address, typo, and facial expression, and we therefore need to clean them into the original form.

Tagging and lemmatization are then conducted on both text resources, since we need to perform the vectorization of words by means of the Word2vec [1] which requires us to eliminate the possible morphologies in order to derive a reliable model. Tagger, also known as Part-Of-Speech tagger, assigning every element or token appeared in a sentence a label such as noun, verb, adjective, etc. For English text resources, we implement StandfordNLP [2] tool as tagger as well as NLTK [3] as lemmatizer. Lemmatizer is in charge of the transformation of plural nouns, comparative adjectives, paste tense verb and adverb to their base form. Similarly, we employ the MeCab [4] for analyzing Japanese resources during tokenization, tagging and lemmatizer. Furthermore, we remove unnecessary and meaningless semantic elements including determiner, such as “the”, punctuation marks, conjunction, and foreign word in order to train more

accurate word2vec model. The same scheme will also be adopted during the processing the Japanese version.

2.2 Deriving vector representation

Word2vec is a tool developed by Google based on deep learning which provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words [1]. In this study, we use this training the word2vec model with the preprocessed English and Japanese text resources respectively, with the dimensionality k of 200, which is considered to be reasonable figure during training.

2.3 Clustering for both languages

In this present stage of our experiments, we select 103 Japanese words that are not only commonly appeared in the stock message board but also considered to be important for prediction of stock price, as well as the 103 identical English words and phrases translated from the Japanese version. Basing on the vector representation generated from word2vec as discussed previously, we then conduct clustering for both Japanese and English word list using K-means with ten centroids (k=10). Here we use cosine similarity as the distance function when training the model, since the word2vec model is considered to have the property when we have two groups of words, in which the words have the similar relationship, and therefore the angles of vectors of words matter much.

Before clustering, our 103 word dictionaries are not normalized, consisted of various tense and plurals. Hence we implement the tagging and lemmatizing again, similar to the preprocessing section discussed previously, and then we remove the useless elements.

During the clustering the English word lists, in case of a word that is not included in our trained word2vec model, we remove it directly. In addition, the English word list also contains phrases which will never be involved in our word2vec model since we train the model word by word rather than by phrases. In case of this, a trick is adopted here that we first retrieve the vector representation for each word appeared in the phrases and then we use the summation results of these vectors as the vector representation for the whole phrase, although this method is still worth discussing more, regarding to its rationality.

3 Experiments Results

The piece of clustering results for both Japanese and English shows in the Table 1.

Table 1: Example of English words clustering

Cluster Name	Word	Corresponding Japanese Word	Cluster name of the Japanese word
EN_1	Consumption tax increase	消費増税	JP_5
EN_1	Tax increase	増税	JP_5
EN_8	Demand	需要	JP_1
EN_8	Growth	伸び	JP_3
EN_9	Improvement	改善	JP_8
EN_9	Contribution	貢献	JP_8
EN_9	Establishment	新設	JP_{10}

After clustering with centroid k=10 for both English and Japanese 103 common word lists, we assign each of the clustered group of words a name. For English clusters, we define the group name EN_j where j is from 1 to 10, whereas for the Japanese clusters, we have the group name JP_i where i is from 1 to 10. Next we compare their relationship, the probability whether one group is corresponding to another group. To be specific, Each English cluster have corresponded Japanese words with identical meanings, so we count the occurrence frequency of the corresponding clusters of these Japanese words. Table 2 demonstrates these relationships.

We could find that there should be some directly relationship between this clusters. For example, for EN_1 and EN_4, the words belonging to them are all corresponds to JP_5 and JP_4, indicating that it is highly possible to establish relations among those clusters.

Table 2:

Comparison of English clusters and Japanese clusters

Cluster Name [Total number of words in it]	The cluster name and its frequency of occurrence of the corresponding Japanese word with identical English meaning: cluster name [frequency of occurrence]
$EN_1[2]$	JP_5[2]
$EN_2[12]$	JP_5[4], JP_{10}[2], JP_4[2], JP_1[2], JP_7[1], JP_3[1]
$EN_3[2]$	JP_5[1], JP_4[1]
$EN_4[1]$	JP_4[1]
$EN_5[5]$	JP_7[3], JP_4[1], JP_2[1]
$EN_6[7]$	JP_7[2], JP_5[2], JP_4[1], JP_{10}[1], JP_8[1]
$EN_7[5]$	JP_7[2], JP_5[1], JP_4[1], JP_{10}[1]
$EN_8[16]$	JP_5[6], JP_1[4], JP_7[2], JP_4[2], JP_{10}[1], JP_8[1]
$EN_9[40]$	JP_7[13], JP_5[13], JP_8[5], JP_1[4], JP_4[3], JP_{10}[2], JP_9[1]

EN_{10}[5]	JP_6[2], JP_4[2], JP_5[1]

4 Future Works

The study discussed in this paper might be considered as the prior study and an attempt for the automatic creation system of financial dictionaries. In this experiment we derive the preliminary results that through vector representative model such as word2vec combining with machine learning algorithms like k-means, it could be concluded that there is a significant relationship between English and Japanese financial text resources, which is highly promising to obtain further expected results contributing to our final goal.

Here we propose several potential improvements for the future experiment:

1) Introduce the phrase2vector model that could offer not only single word but also a phrase a vector representation directly.
2) Alter the number of centroid for k=10, for instance, to k=5,15,20,25
3) Implement other vector representation schemes instead of only word2vec.
4) Establish mathematical similarity expressions in order to excavate more latent relationships among multilingual clusters.

References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems pp. 3111-3119, (2013)

[2] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, Proceedings of HLT-NAACL, pp. 252-259, (2003)

[3] NLTK official site, http://www.nltk.org/

[4] MeCab official site, http://mecab.sourceforge.net/