Cross-lingual News Article Comparison Using Bi-graph Clustering and Siamese-LSTM

Enda Liu[1] Kiyoshi Izumi[1] * Kota Tsubouchi[2] Tatsuo Yamashita[2]

[1] School of Engineering, The University of Tokyo
[2] Yahoo Japan Corporation

Abstract: Calculating similarity score for monolingual text is a popular task since it could be used for various text mining system. However seldom research is focusing on multilingual text resources. On the other hand, machine learning based algorithms such as CBOW word embedding and clustering are widely used in extracting features of text. In this research, we develop and train a model that could calculate the similarity of the two finance news reports, by utilizing CBOW, spherical clustering, bi-graph extraction as well as the Siamese-LSTM deep learning model. In the end, we train the model by feeding news data that is closely related in the financial domain to help us to analyze the relationship among news reports written in different languages.

1 Introduction

Financial text mining has been widely used for financial analysis and one popular example is building financial prediction model with public database by means of machine learning algorithm and natural language processing [9].

However, there is only few studies about cross-lingual text mining and sometimes foreign news reports will also give impacts on local markets. Although there are some popular solutions for multilingual translation, they are neither designed nor specialized for financial domain. Hence, it becomes a needful work to construct a system that can establish relationships among multilingual financial text so that both developers and users are able to handle the foreign information promptly and accurately.

On the other hand, Vector representation of words such as skip-gram and CBOW[8] is efficient methods used in text mining. In this paper we present the cross-lingual study on English and Japanese text data respectively, investigating possible relationships between the two language models.

Finally, we also develop an piratical application basing on our research, which could calculate the similarity of the two cross-lingual finance news reports. Although calculating similarity score for monolingual text is a popular task since it could be used for various text mining system, seldom research is focusing on multilingual text. Our application is basing on the results of the extraction of bi-graph structures as well as the Siamese-LSTM deep learning model[5]. It performs well, proving the efficiency even though the training news data are closely related in the financial domain and also helps us to analyze the relationship among news reports written in different languages.

*連絡先：Dep. of Systems Innovation, School of Engineering, the University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan
E-mail: izumi@sys.t.u-tokyo.ac.jp

2 Framework of bi-graph structure extraction

The overview framework of the bi-graph structure extraction is shown in Figure 1. Generally, our framework includes five main blocks: data retrieving and preprocessing, word2vec modeling, dictionary extraction(vector assignment), and then clustering with spherical k-means and finally mapping for cross-lingual clusters in order to extract the bi-graph structures.

2.1 Data preparation and preprocessing

There are mainly three steps for data prepraration and preprocessing:

1. We first retrieve both the Japanese and English news reports related to finance and economics

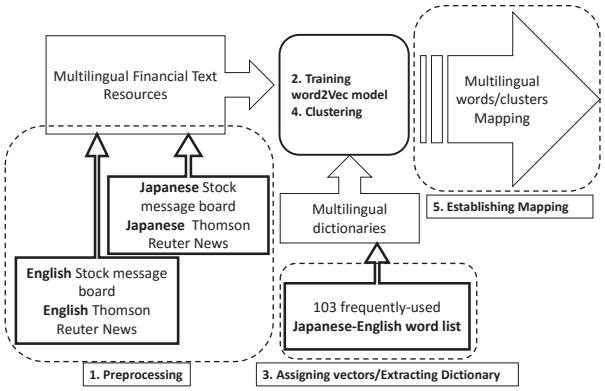

図 1: The data flow and framework for bi-graph clustering and structure extraction

from the database of Thomson Reuters [1] of year 2010.

2. Tokenization, Tagging and lemmatization are then conducted on both text data in order to acquire more precise word2vec model by only using small scale training data[8]. For English text, we implement StanfordNLP[6] as tagger as well as NLTK[7] as lemmatizer and for Japanese, we employ the MeCab[3] with the neologism dictionary[2].

3. Finally, we remove unnecessary and meaningless semantic elements including determiners, such as the word *the*, punctuations, conjunctions, and foreign words. Further more, we also remove unsolvable elements such as special characters, http and email address, typo and facial expression in order for better word2vec models through regular expression.

2.2 Word embedding

Google ' s word2vec is a prevalent tool basing on skip-gram or continuous bag-of-words architecture which provides an efficient way for word embedding[8]. In this research we perform two training for the Japanese news and the English news respectively in 200 dimension.

[1]Official websites of Thomson Reuter: http://www.reuters.com/

[2]Neologism dictionary implementation on Mecab-ipadic: https://github.com/neologd/mecab-ipadic-neologd

2.3 Dictionary Extraction

In this research, we manually choose 103 typical Japanese words frequently appeared in the finance domain. We first translate those words into English, then assign vector representation basing on the word2vec model trained for both English and Japanese words.

In case of phrases that are not detected by the phrase detector, a trick applied here that we first retrieve the vector representation for each word appeared in the phrase and then conduct a summation on these vectors, considering the result as the vector representation for the whole phrase. This is proposed basing on the recent report[4] that word2vec could represent many linguistic regularities, suggesting that the vector addition and subtraction could still be able to represent the relative meaning of the phrase.

2.4 Spherical k-means clustering

We conduct spherical k-means clustering for both Japanese and English financial word list with centroids k to be 10. Concerning word and document vector clustering, it is reported the spherical k-means, which combine the k-means algorithm with cosine similarity as the distance function, could produce a satisfied clustering results[2]. We initialize with k-means++ seeding and overwrite the algorithm[1] with the help of Scikit learn platform[3].

A sample of clustering results for cross-lingual financial news reports of Thomson Reuters in both English and Japanese is shown in the Table 1 with the number of cluster centroid k=10. Here, we define the notation EN_i^m and JP_j^n, i,j,m,n $\in$ (1,2,3,...,k) referring to English cluster names and Japanese cluster names respectively where m and n indicate number of centroids we choose during clustering while i and j shows the cluster name in terms of number. Normally, as the the default size of cluster centroids of 10, the superscript m=10 and n=10 would be omitted.

This results indicate that for each cluster, the spherical kmeans approach indeed clusters words with similar meanings into the same cluster.

2.5 Mapping for cross-lingual clusters

The clustering results in Table 1 also shows that some Japanese words whose English translation are

[3]Scikit learn, official website: http://scikit-learn.org/stable/index.html

表 1: Example of English words clustering basing on Thomson Reuters news

EN Cluster	EN Word	JP Word	JP Cluster
EN_5	Improvement	改善	JP_5
EN_5	Jump	高騰	JP_1
EN_5	Fall	低下	JP_1
EN_5	Decline	減少	JP_9
EN_9	Decrease of profit	減益	JP_2
EN_9	Increase of income	増収	JP_2
EN_7	Concern	懸念	JP_{10}
EN_7	Risk	リスク	JP_{10}
EN_7	Aggravation	悪化	JP_{10}

clustered into the same cluster, for instance, EN_7, are also categorized into the cluster JP_{10}, indicating that there might be potential relationships between English clusters and Japanese clusters.

Here, we define the concept *common words* for cluster EN_i and JP_j as any translation pair where its English translation belongs to EN_i while its Japanese translation belongs to JP_j, denoted as $C_{(i,j)}$. According to the Table 1, for example, the Japanese-English pair *Decline* and *Genshou* is *common words* for cluster EN_5 and JP_9.

In order to identify the *cross-lingual cluster similarity* as the similarity among multilingual clusters quantitatively, we here define the *similarity between any two clusters* as:

$$sim(EN_i, JP_j) = \frac{2 \times size[C_{(i,j)}]}{size[EN_i] + size[JP_j]} \quad (1)$$

where $C_{(i,j)}$ as mentioned previously denotes the the common words of EN_i and JP_j, while the notation $size[A]$ refers to the size of a set A. In this equation, the similarity among clusters could reach 1 when $size[C_{(i,j)}] = size[EN_i] = size[JP_j]$ whereas it becomes 0 when there is no common words, that is $size[C_{(i,j)}] = 0$.

In reality, there might not be a one-to-one mapping for cross-lingual clustsers, instead it is also possible to 1-to-N mapping among them. Hence, we adjust the definition of the cluster similarity as:

$$sim(EN_i, JPF_p) = \frac{2 \times size[C_{(i,p)}]}{size[EN_i] + size[JPF_p]} \quad (2)$$

In this equation, considering full combinations of clusters from set JP_i j where $j \in (1, 2, .., k)$ and k=10,

表 2: Extended Japanese clusters with maximum cross-lingual similarity respecting to English clusters

English clusters	Most similar cluster	Similarity scores
EN_1	JP_6	0.4
EN_2	(JP_1+JP_2)	0.5
EN_3	JP_5	0.4
EN_4	JP_{10}	0.32
EN_5	JP_1	0.64
EN_6	(JP_9+JP_7+JP_2)	0.49
EN_7	(JP_{10}+JP_3)	0.30
EN_8	JP_4	0.38
EN_9	JP_3	0.4
EN_{10}	JP_4	0.43

we then could obtain a set, denoted as JPF_p where $p \in (1, 2, ..., Q)$. Here, Q is the number of the full combinations, normally calculated as:

$$Q = \sum_{n=1}^{k} \frac{k!}{n!(k-n)!} \quad (3)$$

To be more specific, the new sets after full combination for Japanese clusters, the JPF_p typically include:

- JP_1, JP_2, ... , JP_{10}
- (JP_1+JP_2) , (JP_1+JP_2) , ... , (JP_9+ JP_{10})
- (JP_1+JP_2+JP_3) , (JP_1+JP_2+JP_4) , ... , (JP_8+JP_9+ JP_{10}) ... ...
- (JP_1+JP_2+...+JP_{10})

Basing on the equation 2, we could derive the most similar combination of Japanese clusters in the regarding to each English cluster, as shown in the table 2.

3 An application: Cross-lingual Article Comparison

By utilizing the cross-lingual mapping results, we also develop an piratical application that could, when given a English news article, return the most similar Japanese news articles inside the database.

3.1 Framework of cross-lingual article comparison

There are mainly two part of this framework as shown in the Figure 2, three input layers and a Siamese LSTM module.

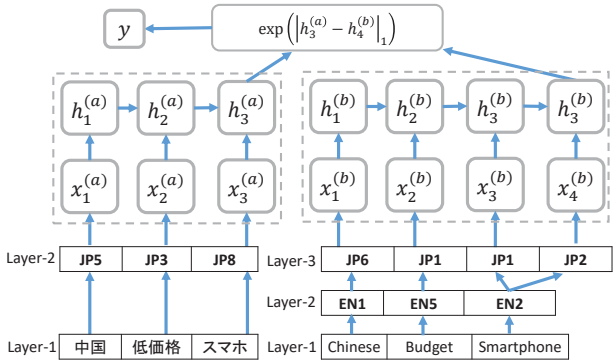

図 2: Framework of cross-lingual article comparison using two-LSTM modules with different embedding layer for the Japanese input and English input respectively.

3.1.1 Siamese LSTM

Siamese Long short-term Memory is a deep learning based machine learning model that could not only memorize the order information of input but also calculate the similarity between the given two text input [5]. This structure is normally applied for the monolingual text similarity task and has been proved efficiency, whereas it has not been applied in the cross-lingual situation so far. In our application, we adopts the same configuration which are claimed to be optimal.

3.1.2 Input layers

Differing from the monolingual case used in the Siamese LSTM module, there are either two or three layers appended for the text input in order to adapt for cross-lingual case, depending on whether it is Japanese text or English text.

1. Layer-1. The tokenized and normalized words are fed in term of fixed-length word2vec representation in the layer-1.
2. From layer-1 to layer-2, the cosine distance between each of the fed word and the centroid of each Japanese clusters will be calculated, and the the cluster with the closest distance is considered to be the representation of this word in the layer-2.
3. Layer-3 is only deployed in case of English input. It will map the cluster number in layer-2 from English domain to Japanese domain basing on the mapping results obtained in the bi-graph extraction section, so that the LSTM module could compare them equally.

3.2 The Data sets

Similar to the process conducted during the extraction of bi-graph structure extraction, we excavate data from Thomson Reuter News. With the help of existed news tags in the database, we extract the English and Japanese news pairs which are both relating to the same news events, which could be considered as similarity of 1. In contrast, any two cross-lingual news reports are defined to be dissimilar with a similar score of 0. In this experiment, we prepare 1000 cross-lingual pairs with similarity of 1 and 1000 pairs with similarity of 0 as training data.

3.3 Model training and evaluations

We perform the training 100 epochs times. In addition, instead of using mean-square-error as evaluation benchmark, we define a practical criteria, the @TOP-N indicator. Given a Japanese article, the comparison model would offer a list that lists all articles ordered from most the similar article to the least similar article. However there is only 1 correct answer with similarity score of 1, and therefore the @TOP-N indicator indicates whether the real similar article is identified among the top N similar articles suggested by the comparison system.

The results are shown in the table 3. As expected, the @TOP-N indicator for the training data is always slightly higher than testing. If we only focus on the testing data, there are about 33 articles are correctly paired within the TOP 10 range, which is not as good as our expectation. We consider this could be due to the high correlation among our test and training financial text.

表 3: The TOP-N score of the cross-lingual LSTM model

	Training Data	Testing Data
@TOP-10	58/1000)	33/1000
@TOP-5	32/1000	20/1000
@TOP-1	4/1000	6/1000

In the previous step for calculating the top-N score, we have obtained the ranking positions of the correct counterpart within the whole database for each of documents. Figure 3 shows this ranking results for training documents and test documents in red and in blue respectively. The more dots close to the top line are (rank = 1), the better the training results are. Though there are large amounts documents still ouside the top-10 range, most of them are approaching to the top-1 line compared with the case of the random distribution indicating the effectiveness of our model.

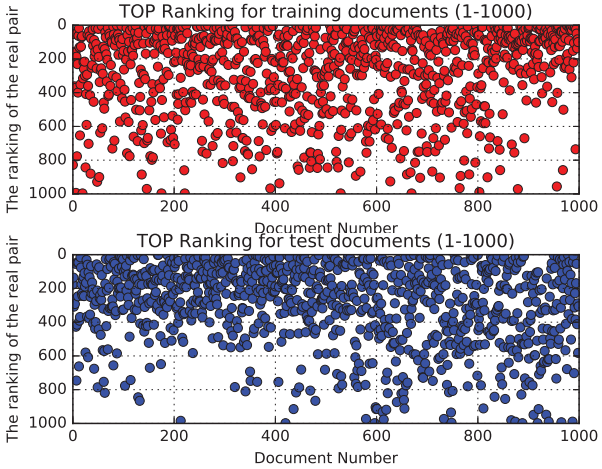

図 3: TOP ranking for training documents (red, above) and test documents (blue, below) respectively

Furthermore, when investigating the learning process of our model, we obtain the learning curve as shown in Figure 4, where the TOP-N (N=1,5,10) accuracy are increasing along with training epoches and become largest TOP accuracy at the 100 times training.

4 Conclusion

This paper give a cross-lingual solution for articles similarity comparison by using extraction of bi-graph structures basing on word2vec, spherical clustering

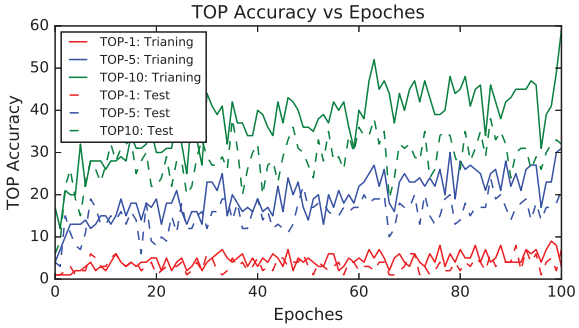

図 4: Learning curve in terms of TOP-1,5,10 accurary regarding to learning epoches

and Siamese LSTM with multilingual news reports from Thomson Reuters. We evaluate our model with finance news data and prove the effectiveness of our model, though the accuracy still needs to be improved in the future work. As the final goal, this kind of model is expected to assist the financial analyst in their prediction model.

参考文献

[1] BANERJEE, Arindam, et al. Clustering on the unit hypersphere using von Mises-Fisher distributions. In: *Journal of Machine Learning Research*. 2005. pp. 1345-1382.

[2] DHILLON, Inderjit S.; GUAN, Yuqiang; KOGAN, J. Refining clusters in high-dimensional text data. In: *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications at the Second SIAM International Conference on Data Mining*. 2002. pp. 71-82.

[3] KUDO, Taku; YAMAMOTO, Kaoru; MATSUMOTO, Yuji. Applying Conditional Random Fields to Japanese Morphological Analysis. In: *EMNLP*. 2004. pp. 230-237.

[4] MIKOLOV, Tomas; YIH, Wen-tau; ZWEIG, Geoffrey. Linguistic Regularities in Continuous Space Word Representations. In: *Hlt-naacl*. 2013. pp. 746-751.

[5] MUELLER, Jonas; THYAGARAJAN, Aditya. Siamese Recurrent Architectures for Learning Sentence Similarity. In: *AAAI*. 2016. pp. 2786-2792.

[6] TOUTANOVA, Kristina, et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003. pp. 173-180.

[7] TAGHVA, Kazem; ELKHOURY, Rania; COOMBS, Jeffrey. Arabic stemming without a root dictionary. In: *Information Technology: Coding and Computing, 2005*. ITCC 2005. International Conference on. IEEE, 2005. pp. 152-157.

[8] MIKOLOV, Tomas, et al. *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013.

[9] VU, Tien-Thanh, et al. An experiment in integrating sentiment features for tech stock prediction in Twitter. In: *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*. The COLING 2012 Organizing Committee, Mumbai, India, 2012. pp. 2338.