Cross-Language Record Linkage by Exploiting Semantic Matching of Textual Metadata

Yuting SONG† Taisuke KIMURA† Biligsaikhan BATJARGAL‡ and Akira MAEDA†‡

† Graduate School of Information Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN

‡ Research Organization of Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN

†‡ College of Information Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 JAPAN

E-mail: †{gr0260ff, is0013hh}@ed.ritsumei.ac.jp, ‡biligsaikhan@gmail.com, †‡amaeda@is.ritsumei.ac.jp

Abstract Cross-language record linkage is a task of finding pairs of records that refer to the same entity across multiple databases in different languages. It is crucial to various research fields, such as federated search and data integration. The matching of textual values of metadata fields plays an important part in comparing record pairs. When matching textual values across languages, one problem is that the mismatches between semantically related translations of metadata values in source language and metadata values in target language, which refer to the same entity. For example, when comparing the records in Japanese (source language) and English (target language), the Japanese word "白雨" in metadata is translated into "rainfall". However, the corresponding word in English metadata is "storm", which is semantically related to "rainfall". As a consequence, the commonly used string-based matching cannot measure the relevance of semantically related words. In this paper, we propose a method for semantic matching of textual metadata, which is based on word embedding that can capture the semantic similarity relationships among words. The effectiveness of this method is evaluated on film related textual metadata in Japanese and English. Then, we use our method to link the identical Ukiyo-e prints between the databases in Japanese and English.

Keywords Cross-language record linkage, word embedding, semantic matching

1. Introduction

Record linkage is a process of matching records from several databases that refer to the same entities. It could be employed to integrate and combine the data from multiple sources, in order to improve data quality and to reduce costs and efforts in data acquisition [1].

In recent years, as the World Wide Web becomes widely matured in more and more countries, the information is being produced in variety of languages. Thus, the identical entities can exist in multiple databases in different languages. For example, the identical Ukiyo-e prints[1] are digitalized not only in Japanese digital museums with the metadata information in Japanese, but also in digital museums of foreign countries with metadata information in their native languages [2]. This situation poses new challenges to the task of classical record linkage, since it needs to link identical entities across languages boundaries.

In cross-lingual tasks, translation procedure is usually required to tackle the language barriers [14][20][21]. After the translating step, the record pairs are compared within the same language, which is similar to monolingual record linkage that has been studied for a long time.

In monolingual record linkage, the mismatches of metadata values are mainly due to the typographical

[1] The Ukiyo-e is a type of Japanese traditional woodblock printing, which is known as one of the popular arts of the Edo period (1603-1868).

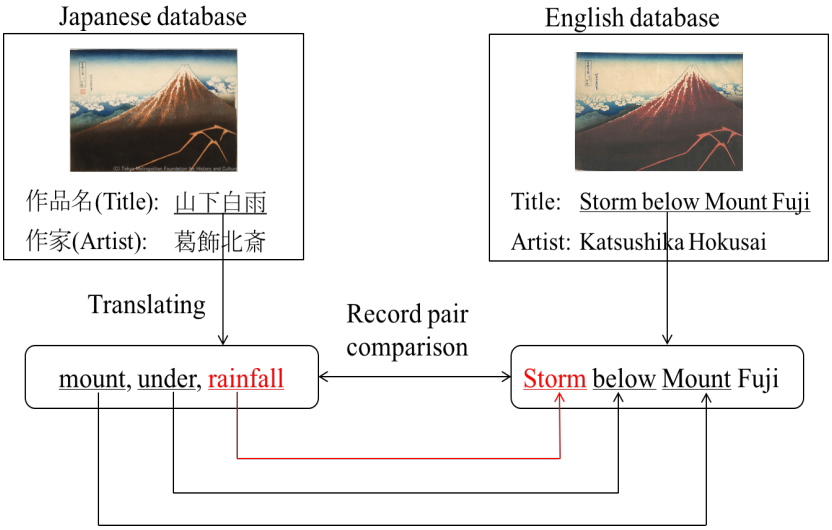

Figure 1. An example of mismatches between translated metadata values and metadata values in target language due to the use of different wordings to express the same meaning

variations of string data, which can be measured by string-based approximate comparisons [7][8]. However, in cross-language record linkage, the mismatches between translated metadata values and metadata values in target language are not only due to the typographical variations of words but also the use of different wordings to express the same meaning. Figure. 1 gives an example of this type of mismatch. The word "白雨" in Japanese is translated into "rainfall" by a Japanese-English bilingual dictionary. However, the corresponding word in English title is "storm", which is translated by a human expert translator. Such mismatches cannot be measured by string-based comparison.

In this paper, we propose a method for cross-language record linkage, which employs the semantic matching of textual metadata. Our method is based on recently results in word embedding [9], which is dense vector representation of words. The learned word embedding can capture the semantic relationships of words, e.g. *vector*("Berlin") - *vector*("Germany") + *vector*("France") is close to *vector*("Pairs"). By using this property of word embedding, we measure the semantic relevance of textual metadata, which is based on matching of embedding of words in metadata.

The remainder of the paper is organized as follows. Section 2 outlines some related work of cross-language record linkage. Word embedding and the general process of cross-language record linkage are described in Section 3 and Section 4, respectively. Section 5 introduces our proposed method. Experimental setup and evaluation of the performance are presented in Section 6. Finally, the conclusion and future work follow in Section 7.

2. Related work

Cross-language entity linking [10][11] is related to our work to some extent, which aims to link the named entities in the texts in one language to a knowledge base in another language. In this task, much contextual information of named entities in texts and content of articles in knowledge bases can be employed. However, our work focuses on the record linkage where only the metadata values can be utilized, which are usually short texts, and sometimes in poor quality.

Cross-language knowledge linking [12][13] is another related task. Most methods are proposed using the structural information of data, such as interlink and outlink in the articles [12], to find the identical articles between knowledge bases in different languages. BabelNet [13] is a large multilingual lexical knowledge base built by combining Wikipedia and WordNet. However, our approach aims at linking the records in several databases in different languages that refer to the same real-world entity, not to find the identical lexicons or articles.

Our work is also related to cross-language ontology matching. With the development of the Linked Data[2], ontology matching is attracting interests of some researchers. Cross-language ontology matching is to find

[2] http://linkeddata.org/.

equivalent elements between two semantic data sources [14][15][16][17]. The difference between our goal and theirs is that our work focuses on general relational databases.

3. Word embedding

Word embedding, distributed representations for words, were firstly proposed by Rumelhart et al. [18] and have achieved impressive results in many natural language processing tasks [24], such as named entity recognition [25], word sense disambiguation [16]. Mikolov et al. [9] introduced two novel word embedding models, the skip-gram and continuous bag-of-words models, which are probably the most popular models and available in the word2vec toolkit [3]. These models learn word representations by employing a simple neural network architecture. Specifically, the skip-gram model is consisted of three layers, input, projection and output layers, to predict contextual words of the input word vector. The training objective is to learn word vector representations that are good at predicting its context in the same sentence [9]. Due to its simple architecture, the skip-gram model can be trained on a large amount of unstructured text data in a short time (billions of words in hours) using a conventional desktop computer.

The main advantage of learned word vector representations is that semantically similar words are close in the vector space. Moreover, the complex word relationships can be captured by performing simple algebraic operations on the word vectors. For example, *vector*("King") - *vector*("Man") + *vector*("Woman") is closest to the vector representation of the word "Queen" [27]. In this paper, although we utilize word2vec to learn word embedding, other word embedding models are also considerable [28][29].

4. Cross-language record linkage

Compared with the classical monolingual record linkage [3][4][5][6], the task of cross-language record linkage requires a translation procedure, since the records that will be matched are from databases in different languages.

The general process of cross-language record linkage is shown in Figure. 2. Firstly, the metadata values of a record, e.g. title, author, publisher of an image record, in source language are translated into target language. How to translate metadata values will be introduced in detail in Section 5.

Secondly, record pairs are compared by calculating the similarities between metadata values within the same language, which is similar to the monolingual record linkage. Since several metadata values are compared for each record pair, it results that numerical similarity values are calculated for that pair, which can be represented by a vector.

The next step is to classify the compared record pairs into three categories: matches, non-matches or possible matches. The possible matches will be further compared to classified into matches or non-matches. Finally, to evaluate the results of record linkage.

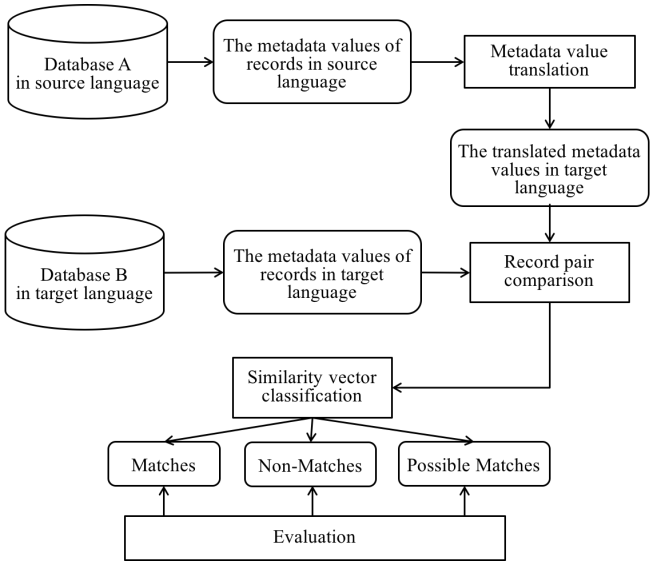

Figure 2. The general process of cross-language record linkage

5. Methodology

In this paper, we deal with the scenarios where textual metadata of records are descriptive metadata, which is used to describe a data resource for purposes such as discovery and identification, e.g. descriptive metadata for an image may include: title, creator, and description.

5.1 Translating textual metadata values

As mentioned above, translation process is required in order to compare the textual metadata within the same language. The two commonly used methods of translation are dictionary based method and machine translation based method [14][20][21][22][23].

Dictionary based method for translating textual metadata values is to find translations of each word in

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

metadata by using machine readable bilingual dictionaries. Due to its simplicity, many previous studies employed bilingual dictionaries to deal with cross-language problems [20][21][22]. However, bilingual dictionaries pose some problems. One problem is that most terms in dictionaries have more than one translations. In this case, word sense disambiguation is required. Another one is that general bilingual dictionaries only contain common words. As a consequence, the translations of rare proper names cannot be found by using bilingual dictionaries.

Machine translation based method is also used to solve cross-language problems [14][23], which is to find translations by employing a machine translation engine, such as well-known Google[4] and Bing[5] translators.

In the Section 6, we will show the experimental results of cross-language record linkage by translating the metadata values using both two methods explained above.

5.2 Semantic matching of textual metadata

Assume we are provided with a word embedding matrix $W \in R^{n \times d}$ for a finite size vocabulary of n words. The i^{th} row, $w_i \in R^d$, represents the embedding of the i^{th} word. The dimension of word embedding space is d.

Our method is to incorporate the semantic similarity between word pairs into the similarity of textual metadata. We use the cosine similarity metric to measure the word similarity. Specifically, the semantic similarity between word i and word j is formulated in (1).

$$word\_sim\,(i,j) = cosine(w_i\,,\, w_j) \qquad (1)$$

Note that in a bilingual dictionary, the corresponding translations of a word are sometimes phrases. For example, the English translation of Japanese word "時空" is "space and time". As mentioned in Section 3, the word embedding has the property that the complex word relationships can be captured by performing simple algebraic operations on the word vectors. For instance, *vector*("King") - *vector*("Man") + *vector*("Woman") is closest to the vector representation of the word "Queen". Utilizing this property of word embedding, we represent a phrase by a combination of embedded words (excluding the stop words) in that phrase. Recalling the example above, the phrase "space and time" is represented by *vector*("space") + *vector*("time"). The combined vector for a phrase can be regarded as a word. Therefore, the semantic similarity between a phrase and a word can also be measured by formula (1).

Our goal is to measure the semantic similarity of textual metadata. Intuitively, if a translated textual metadata contains more words that can match the words in metadata in target language, either exactly or semantically, it might have more possibility that they describe an identical entity. We represent textual metadata as a set of embedded words. The similarity between the translated textual metadata (M_{trans}) and metadata in target language (M_{target}) is formulated as the cumulative similarity of word pairs between M_{trans} and M_{target}. First, for each word i (w_i) in M_{trans}, we calculate the similarities between w_i and each word j (w_j) in M_{target}, the similarity between w_i and w_j is calculated by formula (1). Then, the maximum similarity between w_i and each word j (w_j) in M_{target} is regarded as the similarity contribution of w_i to similarity between M_{trans} and M_{target}, which is formulated in (2).

$$Contri(w_i) = max\big(word_{sim}(i,j)\big)\ \forall j \in \{1, \dots, n\} \qquad (2)$$

$Contri(w_i)$ represents the the similarity contribution of w_i in M_{trans} to similarity between M_{trans} and M_{target}. n represents the number of words in M_{target}.

Finally, we can define the similarity between M_{trans} and M_{target} as the cumulative similarity contribution of each w_i in M_{trans}, which is formulated in (3).

$$Sim\big(M_{trans}, M_{target}\big) = \sum_i^m Contri(w_i) \qquad (3)$$

$Sim\big(M_{trans}, M_{target}\big)$ represents the similarity between M_{trans} and M_{target}. m represents the number of words in M_{trans}.

Figure 3 illustrates our method to calculate the similarity between M_{trans} and M_{target}. In this example, we would like to link the identical film by using the film title. The film title "岸辺の旅" is translated into "The shore trip", which is translated metadata M_{trans}. First, stop words (e.g. "to", "the") are removed, leaving *shore*, *trip* in M_{trans} and *journey*, *shore* in M_{target}. The arrow from each word in M_{trans} to word in M_{target} are labeled with their contributions to the similarity between M_{trans} and M_{target}. The word *trip* in M_{trans} has semantically related to the word *journey* in M_{target}. This semantic relationship between words can be captured by word embedding. Consequently, the similarity between M_{trans} and M_{target} becomes higher than that without using word

[4] https://translate.google.com

[5] http://www.bing.com/translator

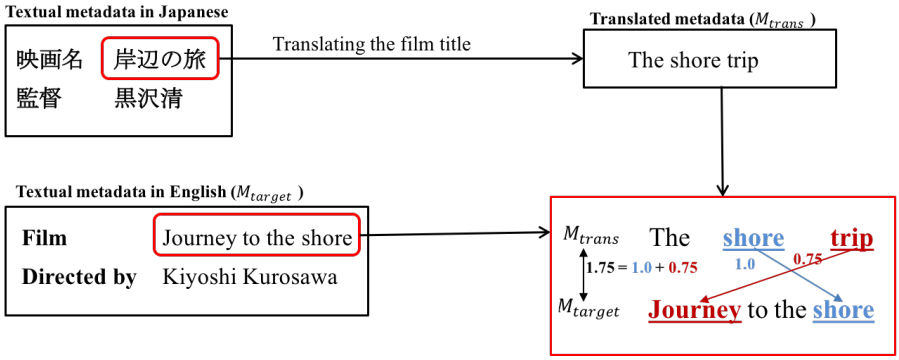

Figure 3. An example of semantic matching of textual metadata

embedding.

6. Experiments

In this section, we first evaluate our method on film related textual metadata from Japanese and English DBpedia[6]. Then, we use our method to link the identical Ukiyo-e prints between the databases in Japanese and English.

6.1 Experimental dataset

In order to evaluate the effectiveness of our proposed method, we construct a dataset that contains pairs of film titles from Japanese and English DBpedia. These pairs are the titles of article pairs that contain Japanese-English cross-lingual links with Wikipedia, which is extracted by DBpedia. The experimental dataset consists of 1,847 pairs of film titles. A small part of experimental dataset is shown in Table 1. Each row in the Table 1 represents a pair of film title that describes an identical film.

Table 1. A small part of experimental dataset

Film title in Japanese	Film title in English
怪獣総進撃	Destroy All Monsters
ゴジラ・エビラ・モスラ 南海の大決闘	Godzilla vs. the Sea Monster
星のオルフェウス	Metamorphoses
紀子の食卓	Noriko's Dinner Table
妖獣都市	Wicked City
南極物語	Antarctica
ゴジラの逆襲	Godzilla Raids Again

6.2 Experimental setup

In the experiments, word embedding is learned with word2vec by using the articles in English Wikipedia dump that contains more than 3 billion words.

In translation process, we translate the textual metadata in Japanese to English by using two methods, dictionary based method (*Dict*) and machine translation based method (*MT*), that are explained in Section 5.1. In dictionary based method, we use EDR[7] Japanese-English bilingual dictionary. In the machine translation based method, we use Microsoft translator API[8].

In the experiments, as the baseline method, we employ string based matching of textual metadata. This method is to link the records across languages by using the exact string matching of textual metadata. After translating the textual metadata into target language, the string based similarity between M_{trans} and M_{target} is measured by the percent of exactly matched words in M_{trans}, which is formulated in (4).

$$str\_sim(M_{trans}, M_{target}) = \frac{the\ number\ of\ string\ matched\ words\ in\ M_{trans}}{the\ total\ number\ of\ words\ in\ M_{trans}} \quad (4)$$

To evaluate the experimental results of cross-language record linkage, we utilize precision within Top-1, Top-5 and Top-10.

6.3 Experimental results

Table 2 show the performance of baseline method and our proposed method. According to the results, our method that employs semantic matching of textual metadata achieves better results than the comparison method of string based matching, especially in precision in Top-1 and precision within Top-5. By using machine translation based method, our method gains the precision of 0.52 in Top-1, which is the strictest evaluation metric.

[6] http://wiki.dbpedia.org

[7] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html

[8] https://www.microsoft.com/en-us/translator/translatorapi.aspx

Comparing two translation methods, machine translation based method achieves much better results than dictionary based method. This is because lots of words in Japanese film titles are not included in the bilingual dictionary.

The results of precision in Top-k (k=1, 5, 10) of the machine translation method is also shown in Table 2. The precision grows as k increases. In the precision within Top-10, our method achieves the precision of 0.64. Thus, if we do not want to find the exact cross-language links, our method can also provide cross-language link candidates.

Table 2. Performance of cross-language record linkage with two methods

	Top-1 precision		Top-5 precision		Top-10 precision	
	Dict	MT	Dict	MT	Dict	MT
String match	0.11	0.50	0.19	0.58	0.22	0.66
Semantic match	0.15	0.52	0.22	0.61	0.25	0.64

6.4 Linking the identical Ukiyo-e prints between the databases in Japanese and English

We also conducted experiments to verify the effectiveness of the proposed method on Ukiyo-e prints in Japanese and English.

Ukiyo-e is a type of Japanese traditional woodblock printing, which is known as one of the popular arts of the Edo period (1603-1868). These prints have been digitized and exhibited on the Internet in many libraries and museums with textual metadata in various languages [19]. Thus, it is suitable to evaluate our method with Ukiyo-e prints data.

The titles of Ukiyo-e prints are used to link records. This dataset consists of 243 Japanese titles of Ukiyo-e prints from the Edo-Tokyo Museum[9] and 3,293 English titles from the Metropolitan Museum of Art[10], in which each Japanese title has at least one corresponding English title. A small part of Ukiyo-e titles is shown in Table 3. Each row in the Table 3 represents a pair of Ukiyo-e title that describes an identical Ukiyo-e print. Among the 243 Japanese titles, 143 titles are descriptive titles that contain at least one non-proper noun.

Table 3. A small part of experimental dataset of Ukiyo-e titles

Ukiyo-e title in Japanese	Ukiyo-e title in English
神奈川沖浪裏	Under the Wave off Kanagawa
蒲原 夜之雪	Evening Snow at Kanbara
従千住花街眺望ノ不二	Fuji Seen in the Distance from Senju Pleasure Quarter
藤川 棒鼻ノ図	Fujikawa, Scene at the Border
駿州江尻	Ejiri in Suruga Province
東海道程ヶ谷	Hodogaya on the Tōkaidō
東海道吉田	Yoshida on the Tōkaidō

Here we translate non-proper nouns of Japanese titles into English by using EDR Japanese-English bilingual dictionary. The proper nouns are transliterated by Hepburn Romanization system[11]. Both comparison method and our method that are mentioned in Section 6.2 are used in this experiment.

Table 4 shows the performance of two methods for cross-language record linkage using descriptive titles and all titles of Ukiyo-e prints. From the results, it can be seen that our method performs better than the string based matching method, especially for descriptive titles that contain one or more non-proper nouns. The reason is that descriptive titles contain one or more non-proper nouns, which are translated based on the meaning of words. However, Japanese proper nouns are transliterated based on the pronunciations, where our semantic matching method is not suitable.

Table 4. Results of cross-language record linkage on Ukiyo-e prints

	Top-1 precision of descriptive titles	Top-1 precision of all titles
String match	0.31	0.27
Semantic match	0.43	0.34

7. Conclusion

In this paper, we proposed a method that employs the distributed representations of words to measure semantic similarities of textual metadata for cross-language record linkage.

The preliminary experimental results have shown that this approach improves the precision of cross-language record linkage.

In the future work, we plan to improve the precision of cross-language record linkage by combining the word embedding to represent the metadata values. Besides, we

[9] http://digitalmuseum.rekibun.or.jp/app/selected/edo-tokyo

[10] http://www.metmuseum.org/

will also evaluate our method on the dataset in other languages.

References

[1] P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication", IEEE Transaction on Knowledge and Data Engineering, 24(9):1537-1555, 2012.

[2] T. Kimura, B. Batjargal, F. Kimura, and A. Maeda, "Finding the Same Artworks from Multiple Databases in Different Languages", In Conference Abstracts of Digital Humanities, 2015.

[3] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage", Journal of the American Statistical Association, 64 (328), pp. 1183-1210, 1969.

[4] S. Sarawagi and A. Bhamidipaty, "Interactive Deduplication using Active Learning", In Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 269-278, 2002.

[5] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg, "Adaptive Name Matching in Information Integration", IEEE Intelligent Systems, 18(5), pp. 16-23, 2003.

[6] A.K. Elmagarmid, P.G. Ipeirotis and V.S. Verykios, "Duplicate Record Detection: A Survey", IEEE Transactions on Knowledge and Data Engineering, 19(1), 1-16, 2007.

[7] G. A. Stephen, "String Searching Algorithms", World Scientific, 1994.

[8] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". In Proc. of the Section on Survey Research Methodology, pp. 354-359, 1990.

[9] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv preprint arXiv:1301.3781., 2013.

[10] P. McNamee, J. Mayfield, D. Lawrie, D. W. Oard and D. Doermann, "Cross-Language Entity Linking", In Proc. of the 5th International Joint Conference on Natural Language Processing, pp. 255-263, 2011.

[11] J. Mayfield, D. Lawrie, P. McNamee and D. W. Oard, "Building a Cross-Language Entity Linking Collection in Twenty-One Languages", In Proc. of the Cross Language Evaluate Forum, pp. 3-13, 2011.

[12] Z. Wang, J. Li, Z. Wang and J. Tang, "Cross-Lingual Knowledge Linking across Wiki Knowledge Bases", In Proc. of the 21st international conference on World Wide Web, pp. 459-468, 2012.

[13] R. Navigli and S. P. Ponzetto, "Babelnet: Building a Very Large Multilingual Semantic Network", In Proc. of ACL, pp. 216–225, 2010.

[14] B. Fu, R. Brennan, and D. O'Sullivan, "Cross-Lingual Ontology Mapping – an Investigation of the Impact of Machine Translation", In Proc. of Asian Semantic Web Conference, pp. 1–15, 2009.

[15] J. Li, J. Tang, Y. Li and Q. Luo, "Rimom: A Dynamic Multistrategy Ontology Alignment Framework", IEEE Transactions on Knowledge and Data Engineering, 21(8), pp.1218–1232, 2009.

[16] J. Tang, J. Li, B. Liang, X. Huang, Y. Li, and K. Wang, "Using Bayesian Decision for Ontology Mapping", Web Semantics: Science, Services and Agents on the World Wide Web, 4(4), pp.243–262, 2006.

[17] X. Zhang, Q. Zhong, F. Shi, J. Li, and J. Tang, "Rimom Results for OAEI 2009", In Proc. of the 4th International Conference on Ontology Matching, pp.208-215, 2009.

[18] D.E. Rumelhnrt, G.E. Hinton, R.J. Williams, "Learning Representations by Back-Propagating Errors", Nature, 323, pp.533-536, 1986.

[19] B. Batjargal, T. Kuyama, F. Kimura and A. Maeda, "Identifying the Same Records across Multiple Ukiyo-e Image Databases using Textual Data in Different Languages", In Proc. of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 193-196, 2014.

[20] L. Ballesteros and C. Bruce, "Dictionary Methods for Cross-Lingual Information Retrieval", In Proc. of International Conference on Database and Expert Systems Applications, pp. 791-801, 1996.

[21] G.A. Levow, D.W. Oard and P. Resnik, "Dictionary-based Techniques for Cross-Language Information Retrieval", Information Processing & Management, 41(3), pp.523-547, 2005.

[22] N. Ehsan, F.W. Tompa and A. Shakery, "Using a Dictionary and N-gram Alignment to Improve Fine-Grained Cross-Language Plagiarism Detection", In Proc. of the 2016 ACM Symposium on Document Engineering, pp. 59-68, 2016.

[23] A. Barrón-Cedeño, P. Gupta and P. Rosso, "Methods for Cross-Language Plagiarism Detection", Knowledge-Based Systems, 50, pp.211-217, 2013.

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural Language Processing (Almost) from Scratch", Journal of Machine Learning Research, 12, pp. 2493–2537, 2011.

[25] P. S. Dhillon, D. Foster and L. Ungar, "Multi-view Learning of Word Embeddings via CCA", In Advances in Neural Information Processing Systems, 2011.

[26] X. Chen, Z. Liu and M. Sun, "A Unified Model for Word Sense Representation and Disambiguation", In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1025–1035, 2014.

[27] T. Mikolov, S. W. Yih and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", In Proc. of the Conference of the North American Chapter of the Association for Computational Liguistics, pp.746-751, 2013.

[28] J. Turian, L. Ratinov, Y. Bengio, and J. Turian, "Word Representations: A Simple and General Method for Semi-supervised Learning", In Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384–394, 2010.

[29] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation", In Proc. of EMNLP, 14, pp. 1532–43, 2014.