Semantic Similarity Measurements for Multi-lingual Short Texts Using Wikipedia

Tatsuya Nakamura, Masumi Shirakawa, Takahiro Hara and Shojiro Nishio
Department of Multimedia Engineering,
Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
Email: {nakamura.tatsuya, shirakawa.masumi, hara, nishio}@ist.osaka-u.ac.jp

***Abstract*—In this paper, we propose two methods to measure the semantic similarity for multi-lingual and short texts by using Wikipedia. In recent years, people around the world have been continuously generating information about their local area in their own languages on social networking services. Measuring the similarity between the texts is challenging because they are often short and written in various languages. Our methods solve this problem by incorporating inter-language links of Wikipedia into extended naive Bayes (ENB), a probabilistic method of semantic similarity measurements for short texts. The proposed methods represent a multi-lingual short text as a vector of the English version of Wikipedia articles (entities). We conducted an experiment on clustering of tweets written in four languages (English, Spanish, Japanese and Arabic). From the experimental results, we confirmed that our methods outperformed cross-lingual explicit semantic analysis (CL-ESA), which is a method to measure the similarity between texts written in two different languages. Moreover, our methods were competitive with ENB applied to texts that have been translated into English using Google Translate. Our methods enabled similarity measurements for multi-lingual short texts without the cost of machine translations.**

I. Introduction

One of the most important tasks of text analysis is to measure the similarity or dissimilarity between texts written in natural languages. It can be used to cluster [33] or classify [3] various topics of disorganized Web texts generated by people from manifold backgrounds. Recently, the immediacy and locality of information dissemination has been regarded as important, as many more people around the world have been continuously transmitting information about their local area in their own languages on social media. For example, the New York Times[1] and Asahi Shimbun Company[2] officially recognize the Twitter[3] accounts of local reporters, that are transmitting information in real time from their area. Because of the immediacy and locality, social media text is valuable information to be analyzed [12], [13]. Enabling similarity measurements between social media texts could improve the performance of many applications such as text clustering [1], text classification [32], and topic detection [2].

Measuring the similarity between such texts is difficult because they tend to be short and written in various languages, e.g. Twitter allows users to post messages called tweets up to 140 characters, and has supported 30 languages as of July 2012[4]. A short text consists of very limited number of terms, i.e., it contains little information that depicts its semantics. Therefore, statistical methods do not work well alone for short text similarity measurements. To measure the similarity between short texts, expanding their semantic information is required. Cross-lingual or multi-lingual[5] texts contain totally different terms to express the same meaning due to the difference of the languages. It is necessary to unify the language space to measure the similarity between cross-lingual or multi-lingual texts. A possible solution for unifying the language space is to use machine translations between languages. However, the cost for building machine translation models increases as the number of languages to be analyzed increases.

We considered using Wikipedia[6] to solve these problems. Wikipedia, a large-scale collaborative encyclopedia, has a more refined link structure and less noisy data than the Web, and it widely covers named entities, domain specific entities, and new entities. These features make Wikipedia an outstanding resource for text analysis [19]. In addition, Wikipedia supports over 280 language versions and has inter-language links, which are links between articles that refer to an equivalent entity but are written in different languages. Thanks to Wikimedia Foundation's contribution to provide the dump data online, it can easily be utilized. Wikipedia has been used as a knowledge source to accomplish similarity measurements on handling short texts [26] or cross-lingual texts [28]. However, to the best of our knowledge, situations that assume multi-lingual short texts as input have been ignored in research on similarity measurements using Wikipedia.

In this paper, we propose two methods to measure the similarity between short texts written in different languages using Wikipedia and Bayes' theorem. Our methods realize both the expansion of the semantic information and the unification of the language space at the same time by incorporating inter-language links into the extended naive Bayes (ENB) [26]. Given a short text written in any language as an input, our methods generate the vector of a certain language version of Wikipedia articles (entities) that are related to the input and add the vector to the input. The vector of Wikipedia entities for the input is used to measure the similarity using the metrics

[1] http://www.nytimes.com/twitter/

[2] http://www.asahi.com/twitter/

[3] https://twitter.com/

[4] https://blog.twitter.com/2012/twitter-now-catalan-and-ukrainian

[5] In this paper, we use "cross-lingual" for a set of two different languages and "multi-lingual" for a set of more than two different languages.

[6] http://www.wikipedia.org/

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.76

IEEE computer society

of the vector space model such as the cosine.

II. Related Work

Research on Wikipedia mining (knowledge extraction from Wikipedia) has attracted a great deal of attention since 2006 and its visibility as a study area has grown rapidly. The following sub sections explain related work on Wikipedia mining about semantic similarity measurements, short text analysis and multi-lingual information retrieval.

A. Semantic Similarity Measurements

Semantic similarity measurement represents one of the most active areas in Wikipedia mining research. The main reason for this is that similarity measurement is a fundamental technology for a variety of applications such as text clustering [9], [10], word sense disambiguation [18], and coreference resolution [24]. Sturbe et al. [31] employed several simple techniques previously developed for WordNet[7] [6] on Wikipedia. Given two articles (entities), their methods specifically compute the distance in the category structure or the degree of overlap in texts. They demonstrated the effectiveness of Wikipedia on standard datasets for similarity measurements (M&C, R&G, and 353-TC) and coreference resolution tasks [24]. Complete details on their work can be found in the literature [25].

Gabrilovich et al. [8] proposed explicit semantic analysis (ESA) to achieve highly accurate and inexpensive similarity measurements. ESA builds a weighted inverted vector that maps each term into a list of Wikipedia articles in which it appears, and computes the similarity between vectors generated from two terms or texts. Since ESA has the capability of computing relatedness between texts written in natural languages, it has been applied to NLP-related tasks [3], [4], [28], resulting in ESA becoming one of the most widely used methods for similarity measurements. Milne et al. [17] computed the relatedness between two articles using the degree of overlap in their incoming and outgoing links. Their method for incoming links is essentially similar to ESA because both methods compute the similarity between terms using articles in which the terms appear.

Graph-based methods [19], [22] derive related articles by computing reachability through links from a given article. Since they cope with a single term to generate a ranked list of related entities, they can easily be used for applications such as query expansion and ad matching. Ito et al. [11] proposed link co-occurrence analysis to speedily build an association thesaurus (defining relatedness between entities).

These studies demonstrated the effectiveness of Wikipedia as a knowledge source to compute the similarity or relatedness. However, it is not possible to measure the similarity between texts or words written in different languages because these studies assume the texts or words written in a single language as inputs. In this work, we propose methods that measure the similarity between texts written in different languages.

B. Short Text Analysis

A great deal of attention has recently been focused on analyzing short texts such as microblogs, search queries, ads and news feeds. Short texts vary from traditional documents in their brevity and sparsity, which makes statistical approaches less effective. Thus it is important to augment the semantics of short texts using external knowledge such as Wikipedia. Meij et al. [15] and Ferragine et al. [7] leveraged Wikipedia to achieve highly accurate entity disambiguation for short texts.

ESA [8] is also a powerful tool for the analysis of short texts. For example, Song et al. [27] illustrated the availability of ESA for short text clustering (as a comparative method), i.e., weighted lists of Wikipedia articles were used to compute the dissimilarity between texts. Sun et al. [32] utilized ESA to classify short texts with a support vector machine (SVM), which is a supervised machine learning technique. They used ESA for expanding the semantic information of short texts.

Shirakawa et al. [26] proposed extended naive Bayes (ENB), a probabilistic method of measuring the semantic similarity for real-world noisy short texts such as tweets. This method solved the compound problem including key term extraction, related entity finding, and aggregation of related entities using the Bayes' theorem to generate a vector of related entities from a short text. This vector is used to measure the similarity as well as ESA. They revealed that ESA is not suited for real-world noisy short texts because ESA's heuristic weighting mechanism highly depends on the rule of majority decision.

A problem with the above studies is that multi-lingual texts are not assumed as an input.

C. Multi-lingual Information Retrieval

Many researchers have tackled multi-lingual or cross-lingual information retrieval tasks using Wikipedia's inter-language links. Navigli et al. [20] pointed out that the performance on word or text analysis is influenced by the quality and quantity of a knowledge base written in a target language. They also built BabelNet, a very large and high quality multi-lingual semantic network, using Wikipedia, WordNet and machine translation. BabelRelate! [21] is a method for measuring the similarity between words (entities) using BabelNet. However, in BabelRelate!, text similarity measurements are not supported.

Sorg et al. [28] proposed cross-lingual explicit semantic analysis (CL-ESA) which measures the similarity between texts written in different languages. CL-ESA extends ESA [8] using inter-language links. It creates ESA vectors in a couple of languages and then limits the base of either one of the vectors to entities that have inter-language links to the other language. After that, it can measure the similarity between texts written in different languages by comparing the limited ESA vectors. The drawback of CL-ESA is that it is based on ESA which is not suited for noisy short texts as described in Section II-B. Moreover, the CL-ESA vectors tend to vary along with the language because CL-ESA only uses the text information of each language version of Wikipedia to create them.

Several studies have attempted to discovering new inter-language links. Sorg et al. [29] hypothesized that the equivalent

[7] http://wordnet.princeton.edu/

entities of different languages are linked to each other via the other entities that have the inter-language links, and proposed a method to discover new inter-language links using machine learning techniques. Penta et al. [23] also proposed a method to discover new inter-language links but, that did not use machine learning techniques, and realized better performance than Sorg's method. These methods are also capable of discovering the mistakes of inter-language links. In addition, Erdmann et al. [5] pointed out that there are a limited number of translation relationships that are covered with inter-language links and showed that the relationships can be expanded by analyzing inter-language links and redirect links. Our proposed methods in this paper are different from them in terms of aiming.

III. Proposed Method

In this section, we begin with challenges that should be considered when we try to measure the similarity between short texts written in any languages and solutions we employed in this work. We then introduce the outline of our two proposed methods. Next, we explain extended naive Bayes (ENB) [26], which is a probabilistic method of semantic similarity measurements between short texts. After that, we describe how to extend ENB using inter-language links.

A. Challenges and Solutions

Let us consider a task to measure the semantic similarity between multi-lingual and short texts; i.e., input is a pair of short texts written in any languages and output is the semantic similarity between them. In this task, there are two challenging problems.

The first problem is that the input texts are short. A short text consists of very limited number of terms, i.e., it contains little information that represents its semantics. To measure the similarity between short texts, expanding their semantic information is required. The extended naive Bayes (ENB) [26], which is a probabilistic method of semantic similarity measurements, is superior in distilling the most dominant topic as the vector of Wikipedia entities from a short text. In our proposed methods, we utilize ENB because it is suited for real-world noisy short texts (described in detail in Section III-C).

The second problem is that input texts are multi-lingual (written in many different languages). It is difficult to directly measure the similarity between multi-lingual texts because the language spaces of these texts are different. To measure the similarity between these texts, unifying their language spaces is necessary. A possible solution for this problem is to use machine translation between languages. However, the cost for building machine translation models becomes enormous as the number of languages to be analyzed increases. To address this problem, we extend ENB by incorporating inter-language links of Wikipedia instead of relying on the machine translation. Our methods unify the language space of input texts into a single language space; i.e., it maps any language versions of a Wikipedia entity to a certain language version of a Wikipedia entity using inter-language links. By this extension, our methods enable both the expansion of semantic information for short texts and unifying the language space for multi-lingual texts at the same time (see Section III-D).

B. Outline of Proposed Methods

We propose two methods to measure the semantic similarity for multi-lingual and short texts based on Wikipedia knowledge and the Bayes' theorem to address the challenges described in Section III-A. Our methods incorporate inter-language links of Wikipedia into ENB to solve the both problems of expanding semantic information for short texts and unifying the language space of semantic information at the same time. At this time, our methods unify the language space of semantic information into English because the English version of Wikipedia is the largest one among all languages and it contains many entities that have inter-language links from the other language versions of Wikipedia entities. Because there are two places to incorporate inter-language links into ENB, we propose two methods, SimpleMap and ProbMap.

Figure 1 shows an example of the outline of our two methods. At first, both methods extract key terms t that constitute input text T written in language L (L is Japanese in Figure 1). They next decide entities e_L of language L that each term t mentions at entity linking process. In Figure 1, "シャープ" (sharp) and "iPhone5" are key terms of the input text. From these terms, we are able to collect entities of language L, i.e., "シャープ" (Sharp Corporation), "シャープ_(記号)" (# of symbol), and "IPhone_5". SimpleMap then finds related entities c of language L for each entity e_L and aggregates c in a probabilistic scheme. After that, SimpleMap maps entity c of language L to equivalent entity c_{EN} of English. SimpleMap only uses English entities c that have inter-language links from language L as well as CL-ESA. On the other hand, ProbMap maps entities e_L to entities e_{EN} of English using inter-language links after entity linking. Here, entities e_L that do not have inter-language links are mapped to English entities via other entities having similar meanings. After that, ProbMap finds related English entities c for each English entity e_{EN} and aggregates c in a probabilistic scheme. Finally, both methods generate a vector of related English entities c to the input text as its semantic representation. The vector is used to measure the semantic similarity between multi-lingual short texts using cosine or other metrics.

The procedures described above are performed in a probabilistic manner. Namely, our methods acquire related entities and their probabilistic scores based on extended naive Bayes (ENB). The next and later sub sections explain the details of our methods, which are shown in Figure 1.

C. Extended Naive Bayes

Extended naive Bayes (ENB) [26] is a method to measure the semantic similarity for real-world noisy short texts based on the Bayes' theorem. It adds related Wikipedia entities to a short text as its semantic representation and uses the vector of the related entities for computing the semantic similarity between texts written in a single language. Adding related entities is generally a compound problem that involves the extraction of key terms, finding related entities for each key term, and the aggregation of related entities. ENB synthesizes these procedures using Bayesian inference techniques and achieves robust estimates of related Wikipedia entities for short texts that contain noise.

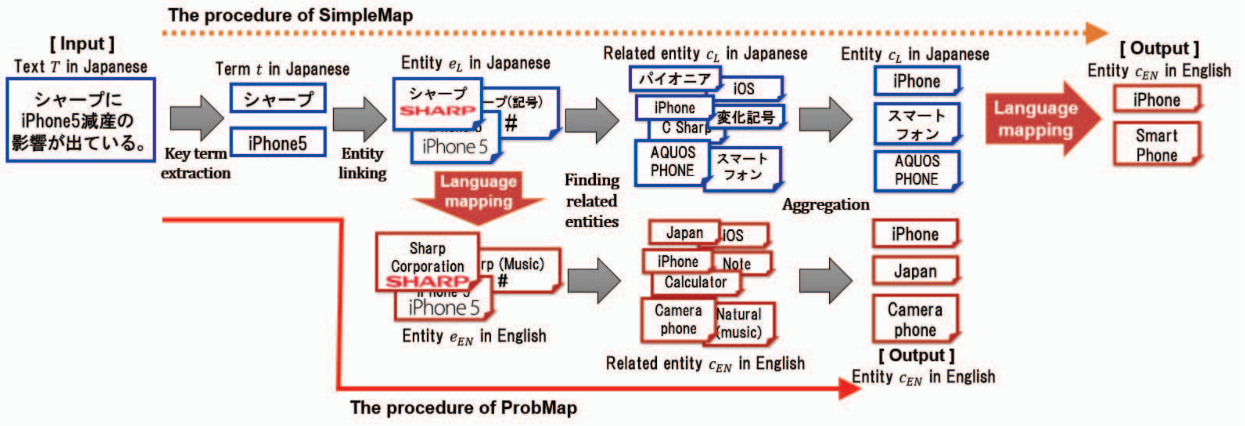


Fig. 1. Example of our methods to create a vector of English Wikipedia entities to measure similarity for multi-lingual short texts. The input assumes Japanese and the text means "The impact of cutting iPhone5 production has come to Sharp Corporation."

The details of ENB are as follows; based on the information that is obtainable from Wikipedia, it calculates $P(c|T)$, which is the probability that entity c is related to the key term set T contained in the input text, by using the following equation.

$$P(c|T) = \frac{\prod_{k=1}^{K} (P(t_k \in T)P(c|t_k) + (1 - P(t_k \in T))P(c))}{P(c)^{K-1}} \tag{1}$$

where t_k is the candidate of the key terms and K is the number of the candidates.

$P(t_k \in T)$ [16] is the probability that term t_k appears in an article as an anchor text. This means the probability that term t_k belongs to key term set T given a text. The trie index is used to extract terms (candidates of the key term) from input texts. The trie index generally extracts all possible terms registered in knowledge bases. Hence, the longest match approach is adopted to detect appropriate terms. In this method, anchor texts and titles of the articles are used as the candidates of the key term. Given that $CountArticlesHavingAnchortexts(t_k)$ is the number of articles that contain term t as an anchor text and $CountArticlesHavingTerms(t_k)$ is the number of articles that contain term t, the probability is computed as

$$P(t_k \in T) \approx \frac{CountArticlesHavingAnchortexts(t_k)}{CountArticlesHavingTerms(t_k)}. \tag{2}$$

$P(c|t_k)$, which is the probability that entity c is related to term t_k, is computed from $P(e|t_k)$ and $P(c|e)$. $P(e|t_k)$ [18], which is the probability that term t_k is linked to entity e, is computed using the relationships between anchor texts and their destination articles. Given that $CountAnchortexts(t_k, e)$ is the number of times that the anchor text t is linked to entity e, the probability is calculated as

$$P(e|t_k) \approx \frac{CountAnchortexts(t_k, e)}{\sum_{e_i \in W} CountAnchortexts(t_k, e_i)} \tag{3}$$

where W denotes a set of all Wikipedia entities in a target language.

$P(c|e)$, which is the probability that entity c is related to entity e, is computed based on incoming and outgoing links of e. Given that $CountLinks(e, c)$ is the number of links (regardless of incoming or outgoing links) between two entities, e and c, the probability is computed as

$$P(c|e) \approx \frac{CountLinks(e, c)}{\sum_{c_j \in W} CountLinks(e, c_j)}. \tag{4}$$

By using $P(e|t_k)$ and $P(c|e)$, the probability $P(c|t_k)$ is computed as

$$P(c|t_k) = \sum_{e_i \in W} P(c|e_i)P(e_i|t_k). \tag{5}$$

$P(c)$, which is the prior probability of entity c, means the generality of c. The more general an entity is, the more likely it is to appear in texts. Given that $CountLinks(c)$ is the number of incoming and outgoing links that an entity c has, the prior probability can be computed as

$$P(c) \approx \frac{CountLinks(c)}{\sum_{c_j \in W} CountLinks(c_j)}. \tag{6}$$

A text is represented by using a vector of related entities ranked by $P(c|T)$. Similarity metrics such as cosine are used to measure the semantic similarity between vectors.

D. Multi-lingual ENB

In order to extend ENB for handling multilingual texts, we define two types of language mapping processes, which map entities in any languages to English entities. In the following, we introduce SimpleMap and ProbMap, which are semantic similarity measurements for multi-lingual short texts using different mapping processes.

1) SimpleMap: SimpleMap creates a vector of related entities using each language version of ENB and then converts the vector into English by only using related entities having the inter-language links to English. This mapping process is more straightforward than ProbMap and the same approach of CL-ESA [28].

SimpleMap calculates $P(c_L|T)$, the probability that entity c_L of language L is related to the text containing a set of terms T written in language L, using Equation (1). After that, SimpleMap maps entity c_L of language L to entity c_{EN} of English using inter-language links. Here, entity c_L without inter-language links to English is ignored. SimpleMap thus generates the vector of related English entities from the input.

2) ProbMap: In ProbMap, we define the language mapping process as a probability in order to incorporate the mapping process into ENB. Probabilistically defining the mapping process allows the multilingualization of ENB without degrading its robustness for short texts. In addition, this probabilistic approach enables us to utilize entities that do not have the inter-language link. In the following, we introduce the mapping probability using inter-language links of Wikipedia to multilingualize ENB.

a) Mapping Probability: Mapping probability $P(e_{EN}|e_L)$, which is the probability that entity e_L of language L is mapped to entity e_{EN} of English, is defined separately in two cases; whether e_L has the inter-language link to e_{EN} or not. When e_L has the inter-language link to e_{EN}, e_{EN} and e_L are assumed to be an equivalent entity. In this case, e_L can be mapped directly to e_{EN} (Figure 2 (a)). Mapping probability is defined by the following equation.

$$P(e_{EN}|e_L) = \begin{cases} 1 & , e_L \text{ has the inter-language link to } e_{EN} \\ 0 & , \text{otherwise} \end{cases} \tag{7}$$

Then, we consider the case in which entity e_L of language L does not have the inter-language link. It is not necessary to discover the English entity that is equivalent to e_L because ProbMap aims to obtain the entities that are related to input text. Therefore, ProbMap maps e_L to English entity e_{EN} via another related entity e'_L of language L that has the inter-language link to e_{EN}. In this case, it is important to find e_{EN} (i.e., e'_L) that is similar to e_L in terms of the topic. Hence we define the probability by using the similarity between e_L and e'_L, in order to map e_L to e_{EN} that has a meaning as similar as possible (Figure 2 (b)). Computing the similarity between all possible entities involves tremendous computational cost. To efficiently find similar entities, we only use the entity e'_L which has incoming links from e_L (i.e., outgoing links of e_L) and the inter-language link to English entity e_{EN}. The reason why we only use outgoing links of e_L to find alternative entities is based on the hypothesis [29] which is used to discover new inter-language links described in Section II-C. Specifically, equivalent entities between different languages are connected via other entities which have incoming links from them and the inter-language links between the languages. These intermediary entities are supposed to be topically similar to the original entity. Because the objective of ProbMap is to find related entities, we utilize these intermediary entities e'_L and e_{EN} as the target of mapping from e_L.

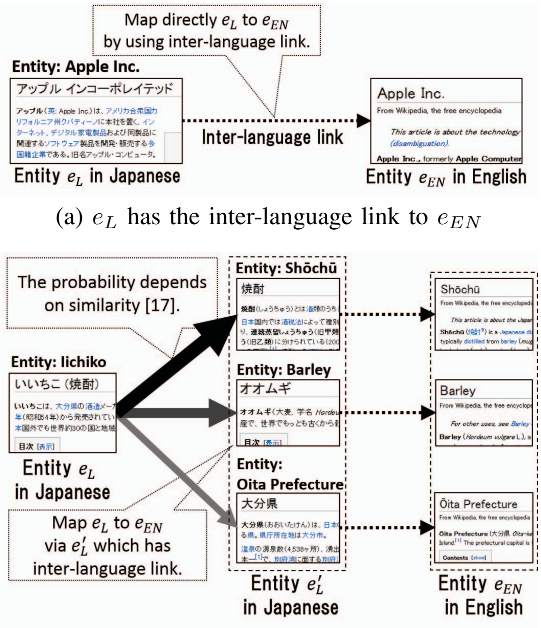
Given $Sim(e_L, e'_L)$ as the similarity between two entities e_L and e'_L, the probability that e_L is mapped to e_{EN} is computed as

$$P(e_{EN}|e_L) \approx \frac{Sim(e_L, e'_L)}{\sum_{e'_{L,i} \in R_L} Sim(e_L, e'_{L,i})} \tag{8}$$

where R_L donates a set of entities which have incoming links from e_L. Similarity $Sim(e_L, e'_L)$ is defined by the following equation [17].

$$Sim(e_L, e'_L) = \frac{\log(|W_L|) - \log(\min(|E_L|, |E'_L|))}{\log(\max(|E_L|, |E'_L|)) - \log(|E_L \cap E'_L|)} \tag{9}$$

where W_L donates a set of all entities on language L version of Wikipedia. E_L and E'_L are a set of entities that link to the entity e_L and e'_L respectively. This similarity is also used in the study of discovering new inter-language links [23]. It used this similarity as a criterion to determine which entity of a different language is equivalent to the input entity. In ProbMap, we employ this similarity as the likelihood that e_L is mapped to e_{EN} via e'_L.


Map directly e_L to e_{EN} by using inter-language link.
Entity: Apple Inc.
アップル インコーポレイテッド
Inter-language link
Apple Inc.
Entity e_L in Japanese
Entity e_{EN} in English
The probability depends on similarity [17].
Entity: Iichiko
いいちこ (焼酎)
Entity e_L in Japanese
Entity: Shōchū
焼酎
Entity: Barley
オオムギ
Entity: Oita Prefecture
大分県
Shōchū
Barley
Ōita Prefecture
Map e_L to e_{EN} via e'_L which has inter-language link.
Entity e'_L in Japanese
Entity e_{EN} in English

(a) e_L has the inter-language link to e_{EN}

(b) e_L does not have the inter-language link to e_{EN}

Fig. 2. Example of how to map the entity e_L of Japanese to the entity e_{EN} of English. Iichiko is one of distilled beverages made from barley and it is made in Oita Prefecture, Japan.

b) Incorporating Mapping Probability into ENB: Let us extend ENB using the mapping probability, which is defined in Section III-D2a, to handle multi-lingual short texts. In order to deal with different languages, SimpleMap and CL-ESA create a vector of related entities using each language version of Wikipedia and then converts it into a single language. However, they may generate largely different vectors from texts of the same meaning written in different languages. This is because the information of Wikipedia substantially differs along with the languages. ProbMap performs the unification of the language space before finding the related entities. That is, it finds related entities using the link structure of the English version of Wikipedia for input texts written in any languages. This enable us to reduce the influence of differences among different language versions of Wikipedia.

Using mapping probability $P(e_{EN}|e_L)$, Equation (5), which is the probability that entity c is related to term t, can be updated as the probability that English entity c_{EN} is related to term t of input language L.

$$P(c_{EN}|t) = \sum_{e_{EN} \in W_{EN}} P(c_{EN}|e_{EN}) \sum_{e_L \in W_L} P(e_{EN}|e_L) P(e_L|t) \tag{10}$$

Here, the probability $P(e_L|t)$ and the probability $P(c_{EN}|e_{EN})$ are calculated from the language L version and the English version of Wikipedia respectively. ProbMap then calculates $P(c_{EN}|T)$, the probability that English entity c_{EN} is related to the text containing a set of terms T written in language L, by using the following equation.

$$P(c_{EN}|T) = \frac{\prod_{k=1}^{K}\left(P(t_k \in T)P(c_{EN}|t_k) + (1 - P(t_k \in T))P(c_{EN})\right)}{P(c_{EN})^{K-1}} \qquad (11)$$

The probability $P(t_k \in T)$ and the probability $P(c_{EN})$ are calculated in the language L version and English version of Wikipedia respectively. Equation (11) is the same as Equation (1), i.e., ProbMap can obtain related entities in a unified language space of English within the framework of ENB.

IV. Evaluation

A. Setup

We carried out an experiment on clustering of tweets written in four languages (English, Spanish, Japanese and Arabic). In the same clustering algorithm, the performance of clustering depends on how the semantic distance (semantic similarity) is measured. Namely, the performance of semantic similarity measurements can be evaluated using clustering. We employed k-means clustering [14] as the clustering algorithm.

We utilized the $hashtags$ to create datasets (i.e., ground truth) for clustering task. $Hashtags$ are tags, such as $\#Google$ and $\#Boxing$, that Twitter users intentionally add to their tweets in order to clarify the topic of the tweet [13]. Hashtags are often used to create datasets for short texts clustering [26], [27]. In our experiment, we carefully selected independent, and unambiguous hashtags (topics) so that each cluster contained a maximum of appropriate tweets. Note that tweets still contain ambiguous terms.

Table I lists two datasets including four languages and their statistics that were used in the evaluation. The first dataset assumed information technology (IT) and the second dataset assumed sports (S). The procedure for constructing the dataset was as follows: 1) search tweets by using predefined hashtags and store those written in English, Spanish, Japanese and Arabic, 2) delete tweets that contain more than one predefined hashtag, 3) delete retweets (tweets starting with RT), 4) remove URLs in tweets, 5) remove hashtags at the end of tweets (to hide explicit clues for the topic) and the "#" of hashtags not at the end of tweets, and 6) delete tweets that consist of under four words written in English, Spanish and Arabic, and under 11 letters written in Japanese.

We compared our methods (SimpleMap and ProbMap) with CL-ESA [28] and original ENB [26] with Google-Translate (Translate). In each method, we changed the maximum number k of non-zero dimensions (related entities) of the vector obtained with each method, namely $k \in \{10, 20, 50, 100, 200, 500, 1000, 2000\}$. We used normalized mutual information (NMI) [30] as the metric to evaluate the performance. NMI expresses scores based on information theory and is regarded as one of the most reliable metrics for clustering. NMI scores are between 0 and 1, and larger scores indicate better results. We conducted k-means clustering 20 times with random initial clusters and recorded the average score for each method.

B. Results

Figure 3 shows the results of tweet clustering in all languages (maximum NMI scores are described in figures) and Figure 4 shows the results of each language. The horizontal axis shows the number of non-zero dimensions (related entities).

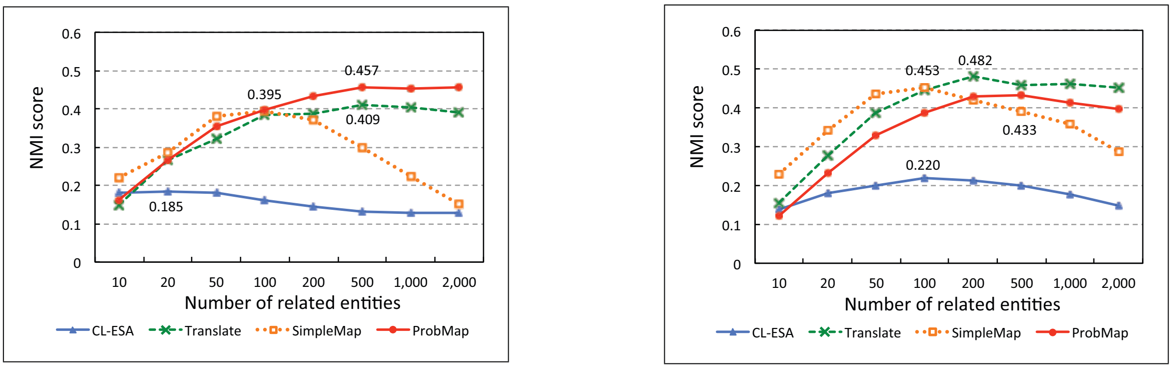
ENB-based methods (Translate and our methods) apparently achieved good performance for both datasets in comparison with the CL-ESA. According to Figure 4, CL-ESA achieved good performance in English and Spanish that are European languages and the majority of the datasets. However, it did not work well in Japanese and Arabic that are the minority of the datasets. This is because CL-ESA is not good at absorbing the difference of information between different language versions of Wikipedia. CL-ESA highly depends on the contents of Wikipedia which are subject to the different cultures of each language. As a result, related entities obtained with CL-ESA are different between European languages and Japanese or Arabic. It is ineffective especially when we consider multi-lingual tasks which deal with more than two different languages. On the other hand, ENB-based methods are good at absorbing the difference of information between different language versions of Wikipedia because they utilize the link structure to find related entities. ENB-based methods are able to finely find the related articles of short texts even if the texts are multi-lingual and short.

Of the ENB-based methods, ProbMap outperformed SimpleMap and Translate in the IT dataset. We applied t-test to compare the best performance between ProbMap and SimpleMap or Translate, and they were significantly different with the p-value < 0.01. SimpleMap tended to be affected by languages because it only used the link structure of each language version of Wikipedia. In Translate, the link structure of only English version of Wikipedia was used to find related entities for input texts in all languages to absorb the difference of the languages. However, the performance depends on the accuracy of machine translation. ProbMap extracted the key terms in the language space of the input text and found related entities using the link structure of the English version of Wikipedia only. ProbMap is able to effectively absorb the difference of information (e.g. text contents and the link structure) among different language versions of Wikipedia.

In the S dataset, Translate outperformed SimpleMap and ProbMap. We applied t-tests to compare the best performance between SimpleMap or ProbMap and Translate. There is not a significant difference with the p-value < 0.01 between SimpleMap and Translate, however, there is a significant difference with the p-value < 0.01 between ProbMap and Translate. The reason why Translate achieved the highest performance in the S dataset is that Translate was able to correctly extract key terms from texts written in Arabic. The S dataset contains many domain-specific proper nouns such as athlete names and team names. These nouns tend to be undefined on small size of Wikipedia such as the Arabic version. In social media, moreover, these proper nouns were often expressed casually. Because SimpleMap and ProbMap only used anchor texts and titles of the entities as the candidates of key terms in each language version of Wikipedia, they are hard to be extracted. In fact, Translate extracted about 11 candidates per Arabic tweet translated into English, while SimpleMap and our method

TABLE I. Two datasets for evaluation and their statistics.

Dataset	Information technology (IT)	Sports (S)
Hashtag / The number of tweets (En / Es / Ja / Ar)	#iPhone / 1.881 (405/445/524/507) #Google / 1,993 (440/545/467/541) #Python / 645 (468/112/26/39) #Microsoft / 1,864 (465/486/452/461) #Firefox / 946 (349/341/54/202)	#F1 / 2,070 (507/536/504/523) #Golf / 1,669 (432/515/514/208) #MLB / 1,499 (497/507/495/0) #NBA / 1,922 (549/532/345/496) #SerieA / 1,406 (415/406/42/543) #Boxing / 999 (532/313/80/74)
Total number of tweets	7,329 (2,127/1,929/1,523/1,750)	9,565 (2,932/2,809/1,980/1,844)
Average number of words per tweet[8]	En:12.6 / Es:12.7 / Ja:55.3 / Ar:12.1	En:13.6 / Es:14.7 / Ja:45.9 / Ar:12.3
Average number of candidate of key terms per tweet	En:10.5 / Es:10.0 / Ja:14.2 / Ar 7.4	En:11.3 / Es:11.7 / Ja:13.7 / Ar:8.0



Information technology (IT) dataset.

Sports (S) dataset.

Fig. 3. Results of tweet clustering in all four languages.

only extracted 8 candidates per Arabic tweet. Except Arabic result shown in Figure 4, SimpleMap and ProbMap carried out the same or better performance than Translate. By extending the extraction of key term, our methods will achieve better performance in Arabic.

SimpleMap achieved higher score than ProbMap in the S dataset. There, however, is not a significant difference with p-value < 0.01. Due to the S dataset having many entities that do not have inter-language links, the topics of the texts became ambiguous in mapping process. For example, the entities that do not have inter-language links were mapped via other 20 entities on average in the Japanese version of Wikipedia. Excessive indirect mapping can deteriorate the performance. When the topics are language-specific, this tends to happen. In such a case, only using the entities that have inter-language links is considered to be better than ProbMap.

V. Conclusions

We proposed two methods of semantic similarity measurements for multi-lingual and short texts by using Wikipedia and the Bayes' theorem. To measure the similarity between such texts is a challenging task because it is necessary to solve both expanding the semantic information for short texts and unifying the language space for multi-lingual texts at the same time. To address this problem, we incorporated inter-language links of Wikipedia into extended naive Bayes (ENB), which is a probabilistic method of semantic similarity measurements for short texts.

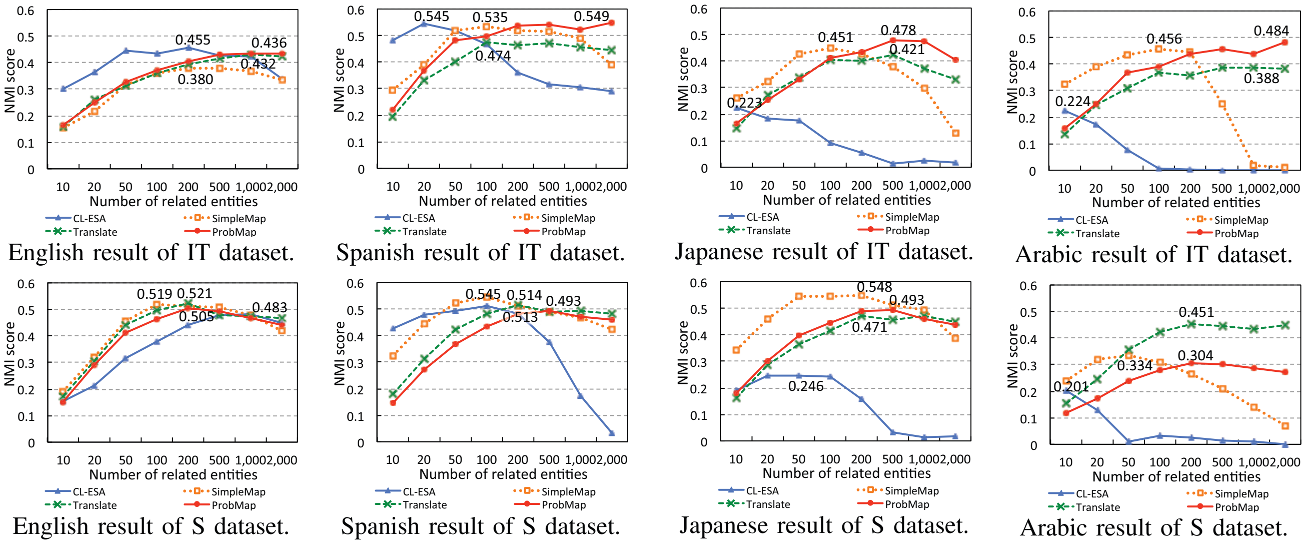
We conducted an experiment on clustering of tweets written in four languages including English, Spanish, Japanese and Arabic. From the experimental result, we confirmed that our methods outperformed cross-lingual explicit semantic analysis (CL-ESA), which is a method to measure the similarity between texts written in two different languages. CL-ESA is highly subject to the difference of each language version of Wikipedia because it utilizes the text contents of the articles. Our methods were able to absorb the difference because it utilizes the link structure of the English version of Wikipedia to find related entities for the texts written in any languages. Moreover, our methods were competitive with ENB applied to texts that have been translated into English using Google Translate. Our methods can perform similarity measurements for multi-lingual short texts using only Wikipedia, i.e., it does not require the cost of machine translations.

For the future work, we plan to combine SimpleMap and ProbMap. As the performance of the two methods depends on the contents of the input texts, some combination of SimpleMap and ProbMap may provide better performance compared to each method alone. We also plan to improve the extraction of key terms. Our method uses only anchor texts and titles of articles of Wikipedia as the candidates of key terms. In social media, however, the key terms are often casually expressed, making their extraction difficult especially if the size of their associated language version of Wikipedia is small. We will explore how to extract these terms independent of the size of associated language version of Wikipedia.

Acknowledgment

This work was supported in part by CPS-IIP Project (Integrated Platforms for Cyber-Physical Systems to Accelerate Implementation of Efficient Social Systems (FY2012 - FY2016)) in the research promotion program for national level challenges "research and development for the realization of next-generation IT platform" by the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

[8] In Japanese, it shows the average number of letters per tweet.



English result of IT dataset.

Spanish result of IT dataset.

Japanese result of IT dataset.

Arabic result of IT dataset.

English result of S dataset.

Spanish result of S dataset.

Japanese result of S dataset.

Arabic result of S dataset.

Fig. 4. Results of tweet clustering in each language.

References

[1] S. Banerjee, K. Ramanathan, and A. Gupta, “Clustering Short Texts Using Wikipedia,” in *SIGIR*, July 2007, pp. 787–788.

[2] H. Becker, M. Naaman, and L. Gravano, “Learning Similarity Metrics for Event Identification in Social Media,” in *WSDM*, Feb. 2010, pp. 291–300.

[3] M. W. Chang, L. Ratinov, D. Roth, and V. Srikumar, “Importance of Semantic Representation: Dataless Classification,” in *AAAI*, vol. 2, July 2008, pp. 830–835.

[4] O. Egozi, E. Gabrilovich, and S. Markovitch, “Concept-based Feature Generation and Selection for Information Retrieval,” in *AAAI*, vol. 2, July 2008, pp. 1132–1137.

[5] M. Erdmann, K. Nakayama, T. Hara, and S. Nishio, “Improving the Extraction of Bilingual Terminology From Wikipedia,” *ACM Trans. Multimedia Computing, Communications and Applications*, vol. 5, no. 4, pp. 1–17, Oct. 2009.

[6] C. Fellbaum, *WordNet: An Electronic Lexical Database*. The MIT Press, May 1998.

[7] P. Ferragina and U. Scaiella, “TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities),” in *CIKM*, Oct. 2010, pp. 1625–1628.

[8] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis,” in *JCAI*, July 2007, pp. 1606–1611.

[9] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen, “Enhancing Text Clustering by Leveraging Wikipedia Semantics,” in *SIGIR*, 2008, pp. 179–186.

[10] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, “Exploiting Wikipedia as External Knowledge for Document Clustering,” in *KDD*, 2009, pp. 389–396.

[11] M. Ito, K. Nakayama, T. Hara, and S. Nishio, “Association Thesaurus Construction Methods based on Link Co-occurrence Analysis for Wikipedia,” in *CIKM*, Oct. 2008, pp. 817–826.

[12] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?” in *WWW*, Apr. 2010, pp. 591–600.

[13] D. Laniado and P. Mika, “Making Sense of Twitter,” in *ISWC*, Nov. 2010, pp. 470–485.

[14] J. B. MacQueen, “Some Methods for Classification and Analysis of MultiVariate Observations,” in *Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[15] E. Meij, W. Weerkamp, and M. de Rijke, “Adding Semantics to Microblog Posts,” in *WSDM*, Feb. 2012, pp. 563–572.

[16] R. Mihalcea and A. Csomai, “Wikify!: Linking Documents to Encyclopedic Knowledge,” in *CIKM*, Oct. 2007, pp. 233–242.

[17] D. Milne and I. H. Witten, “An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links,” in *AAAI*, July 2008, pp. 25–30.

[18] D. Milne and I. H. Witten, “Learning to Link with Wikipedia,” in *CIKM*, Oct. 2008, pp. 509–518.

[19] K. Nakayama, T. Hara, and S. Nishio, “Wikipedia Mining for An Association Web Thesaurus Construction,” in *WISE*, Dec. 2007, pp. 322–334.

[20] R. Navigli and S. P. Ponzetto, “BabelNet: Building a Very Large Multilingual Semantic Network,” in *ACL*, July 2010, pp. 216–225.

[21] R. Navigli and S. P. Ponzetto, “BabelRelate! A Joint Multilingual Approach to Computing Semantic Relatedness,” in *AAAI*, July 2012, pp. 22–26.

[22] Y. Ollivier and P. Senellart, “Finding Related Pages using Green Measures: an Illustration with Wikipedia,” in *AAAI*, July 2007, pp. 1427–1433.

[23] A. Penta, G. Quercini, R. Chantal, and N. Shadbolt, “Discovering Cross-language Links in Wikipedia Through Semantic Relatedness,” in *ECAI*, Aug. 2012, pp. 27–31.

[24] S. P. Ponzetto and M. Strube, “Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution,” in *HLT-NAACL*, June 2006, pp. 192–199.

[25] S. P. Ponzetto and M. Strube, “Knowledge Derived from Wikipedia for Computing Semantic Relatedness,” *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 181–212, Oct. 2007.

[26] M. Shirakawa, K. Nakayama, T. Hara, and S. Nishio, “Probabilistic Semantic Similarity Measurements for Noisy Short Texts Using Wikipedia Entities,” in *CIKM*, Oct./Nov. 2013.

[27] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, “Short Text Conceptualization using a Probabilistic Knowledgebase,” in *IJCAI*, July 2011, pp. 2330–2336.

[28] P. Sorg and P. Cimiano, “Cross-lingual Information Retrieval with Explicit Semantic Analysis,” in *CLEF*, Sept. 2008.

[29] P. Sorg and P. Cimiano, “Enriching the Crosslingual Link Structure of Wikipedia – A Classification-based Approach,” in *AAAI Workshop*, June 2008.

[30] A. Strehl and J. Ghosh, “Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions,” *Journal of Machine Learning Research*, vol. 3, pp. 583–617, Dec. 2002.

[31] M. Strube and S. P. Ponzetto, “WikiRelate! Computing Semantic Relatedness using Wikipedia,” in *AAAI*, July 2006, pp. 1419–1424.

[32] X. Sun, H. Wang, and Y. Yu, “Towards Effective Short Text Deep Classification,” in *SIGIR*, July 2011, pp. 1143–1144.

[33] O. Zamir and O. Etzioni, “Web Document Clustering: A Feasibility Demonstration,” in *SIGIR*, 1998, pp. 46–54.