Cross-Language Microblog Retrieval using Latent Semantic Modeling

Archana Godavarthy
Department of Computer Engineering
Santa Clara University
500 El Camino Real
Santa Clara, CA, 95053
agodavarthy@gmail.com

Yi Fang
Department of Computer Engineering
Santa Clara University
500 El Camino Real
Santa Clara, CA, 95053
yfang@scu.edu

ABSTRACT

Microblogging has become one of the major tools of sharing real-time information for people around the world. Finding relevant information across different languages on microblogs is highly desirable especially for the large number of multilingual users. However, the characteristics of microblog content pose great challenges to the existing cross-language information retrieval approaches. In this paper, we address the task of retrieving relevant tweets given another tweet in a different language. We build parallel corpora for tweets in different languages by bridging them via shared hashtags. We propose a latent semantic approach to model the parallel corpora by mapping the parallel tweets to a low-dimensional shared semantic space. The relevance between tweets in different languages is measured in this shared latent space and the model is trained on a pairwise loss function. The preliminary experiments on a Twitter dataset demonstrate the effectiveness of the proposed approach.

Keywords

Cross Language Information Retrieval; Microblog Retrieval; Latent Semantic Modeling

1. INTRODUCTION

Microblogging platforms such as Twitter have emerged as a powerful source of real-time information sharing for people around the world. Its popularity is witnessed in different parts of the world such as Brazil, Japan, India, France and Turkey[1]. As a result, the content on Twitter is highly multilingual. One study[2] shows that only about 50% of the Twitter messages are in English and other popular languages on Twitter include Japanese, Portuguese, Malay and Spanish.

[1]http://www.forbes.com/sites/victorlipman20140524top-twitter-trends-what-countries-are-most-active-whos-most-popular

[2]http://techcrunch.com/2010/02/24/twitter-languages

ICTIR '16, September 12-16, 2016, Newark, DE, USA

DOI: http://dx.doi.org/10.1145/2970398.2970436

Table 1: An Example of parallel tweets sharing a common hashtag

HashTag: #Immigration
English Tweet: *Our overall #immigration policy is focus enfrcmnt resources more effectively on threats to pub safety & border security*
Spanish Tweet: *Por qué importa una reforma migratoria integral en EU: "US #immigration: Help wanted" http fb.me/81l8JGJWI , vía FT*
Google Translate [3] :*Why not give a comprehensive immigration reform in the US , "US #immigration: Help wanted" http fb.me81l8JGJWI via FT*

Microblogging services played a crucial role in quickly spreading the news about important international events such as Arab Springs. The news on microblogs is sometimes ahead of the major news media. It is very desirable to provide rapid cross-lingual data access for the huge number of multilingual users on microblogs. One solution is to deploy traditional cross-language information retrieval techniques (CLIR) such as machine translation. However, microblog content is quite different from the traditional textual data. For example, the tweets are short text with limitation of 140 characters. Some of the shortcomings of the tweets include informal usage of language, poor spelling and grammatical quality, and inclusion of symbols and diacritics. The existing CLIR approaches focus on regular text documents and may not be directly applicable to microblog data.

In this paper, we study cross-language microblog retrieval. We tackle a specific retrieval problem on microblogs: given a tweet in one language (e.g., English), retrieve a ranked list of relevant tweets in another language (e.g., Spanish). This retrieval task can also be used for recommending interesting tweets in different languages for multilingual users.

We build parallel corpora for tweets in different languages by bridging them via common hashtags. The intuition is that if two tweets share the same hashtag, they are semantically related. Due to the prevalence of hashtags on microblogs, we could find a large amount of parallel texts for tweets in different languages in an automatic manner. Table 1 contains an example of parallel tweets in English and in Spanish sharing a common hashtag. As we can see, the two tweets may not talk about exactly the same issue, but they are semantically related.

The traditional translation based approaches may not be

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

effective for cross-language microblog retrieval because tokens on microblogs are often irregular and noisy. Having a good translation on microblogs is a challenging research problem in itself [1]. In contrast, we model the latent semantics of the parallel tweets by mapping the original high-dimensional tweets to the low-dimensional semantic space and measuring their similarity in this shared semantic space. The mappings are learned from the parallel corpora in a pairwise manner. It is more reasonable than using pointwise loss function because the negative relevance judgments cannot be reliably obtained in this task. The model is trained by stochastic gradient descent (SGD). We conduct experiments on a dataset from Twitter for Spanish and English tweets. The preliminary results demonstrate the effectiveness of the proposed approach over the translation based approach.

2. RELATED WORK

Cross-language information retrieval (CLIR) is a well studied field in the IR community. The major IR evaluation platforms including CLEF[4], NTCIR[5], and TREC[6] have frequently organized the CLIR benchmark evaluations. Dumais et al. [2] is a classic work on general-purpose cross-language IR, which uses latent semantic features to retrieve relevant cross script documents. Our work differs from this work in utilizing the pairwise training of the latent semantic features, instead of pointwise training. Tang et al. [10] apply latent factor models for cross-language citation recommendations.

Microblog retrieval has recently attracted increased attention. TREC has hosted the Microblog track since 2011 [7] to study the retrieval tasks on microblogs. The characteristics of microblogs are investigated, but the TREC tasks have been exclusively focused on English content while a noticeable portion of tweets in the corpus are in other languages [9]. TweeMT[7] is a workshop and shared task on machine translation applied to tweets. Jehl et al. [4] propose probabilistic translation-based approach for Arabic tweets. They utilize the retrieved results to build parallel corpora between different languages. Hu et al.[3] use crowdsourcing methods to translate SMS messages between different languages. Ling et al. [5] build parallel corpora by identifying the presence of multiple languages in the same tweet. They also utilize retweets to identify parallel text. To the best of our knowledge, no prior work has utilized latent semantic modeling for cross-language microblog retrieval.

3. LATENT SEMANTIC MODELING

We utilize the parallel tweets in different languages by assuming that if two tweets in different languages share the same hashtag, they are semantically related. In this paper, we focus on English and Spanish tweets, but the model can be applied to any other languages. Without resorting to language translation, we model the latent semantics of the parallel tweets by projecting the original high-dimensional tweets into a low-dimensional semantic space. The projections are learned from the parallel corpora. The proposed approach is named as **C**ross-language **L**atent semantic **M**odel (CLM).

[4]http://www.clef-campaign.org/
[5]http://research.nii.ac.jp/ntcir/index-en.html
[6]http://trec.nist.gov/
[7]http://komunitatea.elhuyar.eus/tweetmt/

It is worth noting that the tweets that do not share the same hashtag could still be semantically related. Therefore, instead of doing pointwise relevance judgment (which would assume tweets that do not share the same hashtag are not relevant), we use a pairwise ranking approach by assuming that the tweets that share the same hashtag are more relevant than those that do not. This is a more reasonable assumption for our task than the pointwise relevance judgment.

Formally, let $\mathbf{e}_i \in R^n$ denote the i^{th} English tweet in the parallel corpus by using vector representation such as TF-IDF. Similarly, let $\mathbf{s}_j \in R^p$ be the vector of the j^{th} Spanish tweet. Given an English tweet $\mathbf{e}_i$, if the Spanish tweet $\mathbf{s}_j$ has a common hashtag with $\mathbf{e}_i$ while $\mathbf{s}_m$ does not, we have a pairwise triple $(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)$ indicating the Spanish tweet $\mathbf{s}_j^+$ is more relevant than $\mathbf{s}_m^-$. We then learn the mapping from TF-IDF feature space to the joint space with reduced dimensionality of K. $\phi_e : R^n \to R^K, \phi_s : R^p \to R^K$.

We apply linear mapping for both ϕ_e and ϕ_s as

$$\phi_e = \mathbf{e}_i^T \mathbf{W} \tag{1}$$

$$\phi_s = \mathbf{s}_j^T \mathbf{Q} \tag{2}$$

where $\mathbf{W}$ and $\mathbf{Q}$ are the projection matrices for English and Spanish tweets respectively. The relevance score $r(\mathbf{e}_i, \mathbf{s}_j)$ between the English and Spanish tweets can be measured by the inner product in the shared space:

$$r(\mathbf{e}_i, \mathbf{s}_j) = \phi_e \phi_s^T = \mathbf{e}_i^T \mathbf{W} (\mathbf{s}_j^T \mathbf{Q})^T = \mathbf{e}_i^T \mathbf{W} \mathbf{Q}^T \mathbf{s}_j \tag{3}$$

We sample $\mathbf{e}_i$, $\mathbf{s}_j^+$, and $\mathbf{s}_m^-$ to form the pairwise triples from the corpus based on the presence of common hashtags. Section 4.2 provides the details of our sampling strategy in the experiments. Given all the pairwise triples $(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)$, we define the pair-wise loss function as follows:

$$L = \sum_{e_i} \sum_{s_j^+, s_m^-} -\log \sigma\Big(h(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)\Big) \tag{4}$$

where $h(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-) = r(\mathbf{e}_i, \mathbf{s}_j^+) - r(\mathbf{e}_i, \mathbf{s}_m^-) = \mathbf{e}_i^T \mathbf{W} \mathbf{Q}^T (\mathbf{s}_j^+ - \mathbf{s}_m^-)$ and $\sigma(x)$ is the sigmoid function defined as: $\sigma(x) = \frac{1}{1+\exp(-x)}$.

As we can see, this loss function maximizes the difference between the ranking scores $r(\mathbf{e}_i, \mathbf{s}_j^+)$ and $r(\mathbf{e}_i, \mathbf{s}_m^-)$ by using $\log \sigma(x)$ which is a monotonically increasing function. As a result, the tweet $\mathbf{s}_j^+$ becomes more relevant than the tweet $\mathbf{s}_m^-$. This optimization criterion is similar to Bayesian Personalized Ranking [8] that has demonstrated success in recommender systems with implicit feedback.

We learn the parameters $\mathbf{W}$ and $\mathbf{Q}$ by minimizing the loss function Eqn.(4) on the training data. The model is trained by Stochastic Gradient Descent (SGD). For each triple $(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)$ in the training data, the gradients of the loss function L with respect to the parameters $\mathbf{W}$ and $\mathbf{Q}$ are computed as follows:

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial \mathbf{W}} = \left(\sigma\Big(h(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)\Big) - 1\right) \mathbf{e}_i (\mathbf{s}_j^+ - \mathbf{s}_m^-)^T \mathbf{Q}$$

$$\frac{\partial L}{\partial \mathbf{Q}} = \frac{\partial L}{\partial h} \frac{\partial h}{\partial \mathbf{Q}} = \left(\sigma\Big(h(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)\Big) - 1\right) (\mathbf{s}_j^+ - \mathbf{s}_m^-) \mathbf{e}_i^T \mathbf{W}$$

For each triple, the parameters are updated as follows:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \alpha \frac{\partial L}{\partial \mathbf{W}}$$
$$\mathbf{Q}_{t+1} = \mathbf{Q}_t - \alpha \frac{\partial L}{\partial \mathbf{Q}} \quad (5)$$

where α is the learning rate. The converged values of $\mathbf{W}$ and $\mathbf{Q}$ are then used to rank all the candidate Spanish tweets $\mathbf{s}_j$ given any new English tweet $\mathbf{e}_i$ in the test data based on the descending order of ranking score $r(\mathbf{e}_i, \mathbf{s}_j)$ in Eqn.(3).

4. EXPERIMENTS

4.1 Data Collection

We collected English and Spanish tweets from Twitter streaming API via tweepy[8]. Between Nov. 11 to Dec. 5, 2015, we gathered about 1 million English tweets and 400,000 Spanish tweets for our experiments. We preprocess the tweets by removing the Retweets, URLs, usernames, punctuations, and stopwords. Since our approach is based on matching the tweets in different languages using hashtags, we filtered out the tweets that do not have any hashtag. Additionally, we pair each tweet with a single hashtag and remove the most & least frequent terms occurring in the corpus. After preprocessing, we obtained 167,203 English tweets with 35,305 unique tokens and 198,690 Spanish tweets with 30,531 tokens. We obtained 5,880 common hashtags shared by both English and Spanish tweets. We randomly form pairwise triples for each of the English tweet pairing with a positive Spanish tweet, $\mathbf{s}_j^+$ and a negative Spanish tweet, $\mathbf{s}_m^-$. We used 80% of these triples for the training and the remaining 20% for testing.

4.2 Baselines and Settings

To evaluate the performance of the proposed approach, we randomly picked 30 English tweets from the test set as queries. These 30 tweets cover a variety of topics in entertainment, sports, politics, technology, etc. Given each test English tweet $\mathbf{e}_i$, each of the baselines retrieves a ranked list of Spanish tweets from the whole corpus. Jaccard and BM25 are the baselines we used. They are Translation based, i.e. given an English tweet $\mathbf{e}_i$, we obtain its Spanish translation, $\mathbf{s}_t$ using Google Translate. Using $\mathbf{s}_t$, we apply different similarity based retrieval methods to retrieve the Spanish tweets that are most similar to $\mathbf{s}_t$.

- Jaccard: The translated tweet $\mathbf{s}_t$ is used to rank all the candidate Spanish tweets based on Jaccard similarity [6].
- BM25: Similarly we compute the scores of the Spanish tweets using BM25 [6] for each query term $\mathbf{s}_{ti} \in \mathbf{s}_t$. For parameters in BM25, we use default values ($k = 1.5$, $b = 0.75$).
- CLM: our proposed model. For the given $\mathbf{e}_i$, we compute the relevance score as defined in Eqn.(3) using the converged $\mathbf{W}$ and $\mathbf{Q}$.
- Hybrid (CLM+BM25): We normalize the ranking scores of CLM and translation approaches respectively using the min-max normalization $(x - x_{min})/(x_{max} - x_{min})$, for feature scaling. The two scores are combined as follows:

$$\lambda \widehat{r}(\mathbf{e}_i, \mathbf{s}_j) + (1 - \lambda)\widehat{bm}(\mathbf{s}_t, \mathbf{s}_j) \quad (6)$$

where $\widehat{r}(\mathbf{e}_i, \mathbf{s}_j)$ and $\widehat{bm}(\mathbf{e}_i, \mathbf{s}_j)$ are the normalized scores from the respective approaches. λ is the combination weight which is set to be 0.5 in the experiments, giving equal weightage to both methods.

[8]http://www.tweepy.org

Table 2: Experimental results for different methods.

Method	P@5	P@10	AP@10	NDCG@10
Jaccard	0.060	0.110	0.157	0.074
BM25	0.080	0.100	0.170	0.075
CLM	**0.493**	**0.463**	0.536	**0.485**
Hybrid	**0.493**	0.460	**0.572**	0.479

We randomly form the pairwise triples $(\mathbf{e}_i, \mathbf{s}_j^+, \mathbf{s}_m^-)$ for training the CLM model by SGD. Specifically, we first randomly select an English tweet $\mathbf{e}_i$ with a hashtag $\mathbf{g}$, and then randomly sample a positive tweet $\mathbf{s}_j^+ \in \mathbf{I}_g^+$ from the pool of Spanish tweets $\mathbf{I}_g^+$ that have the same hashtag $\mathbf{g}$. For the negative tweet, we randomly sample $\mathbf{s}_m^-$ from the pool of Spanish tweets that excludes $\mathbf{I}_g^+$, i.e., $\mathbf{V} \setminus \mathbf{I}_g^+$, where $\mathbf{V}$ is the set of all Spanish tweets. Once a sufficient number of triples are sampled, we randomly shuffle them to avoid bias for certain hashtags. The model is then trained on these permuted instances by SGD.

The initial values of the parameters $\mathbf{W}$ and $\mathbf{Q}$ in the SGD algorithm are uniformly randomly sampled from $[0, 1]$ and the stopping criteria is when the relative change of the loss function in Eqn.(4) is less than 0.1%. The learning rate (α in Eqn.(5)) is set to be 0.1. The default dimension of the latent semantic space is $K = 500$ (we investigate the impact of K in Section 4.4).

4.3 Evaluation Metrics

Given the English tweets as queries, the retrieved Spanish tweets are manually judged based on binary relevance, i.e., relevant or not. We evaluate the results obtained from all the baselines mentioned above. We use Google Translate and Twitter to judge the relevance of the results. We used evaluation metrics like P@5, P@10, AP@10 and NDCG@10 which are widely used to evaluate the results of ranked lists [6]. To calculate NDCG@10, we assume that there are at least 10 relevant Spanish tweets for each query, so the ideal ranking has all of the top 10 results as relevant.

4.4 Results

Table 2 contains the experimental results for different methods. All the values are averaged over the 30 queries. As we can see, our proposed CLM model yields substantial improvement over the two baselines: Jaccard and BM25, which are Translation based methods. We found that the improvement is statistically significant at 0.99 confidence level by the paired t-test. As mentioned in Section 4.2, we based our Hybrid model on CLM and BM25. The improvement of the Hybrid model over CLM is not as significant as the improvement of CLM over the Translation methods. We can see that P@5 is high for CLM and Hybrid. NDCG is higher for the CLM model, as the Hybrid model is affected by the poor performance of BM25 for some queries. The AP@10 of the Hybrid is higher than CLM model, as the performance

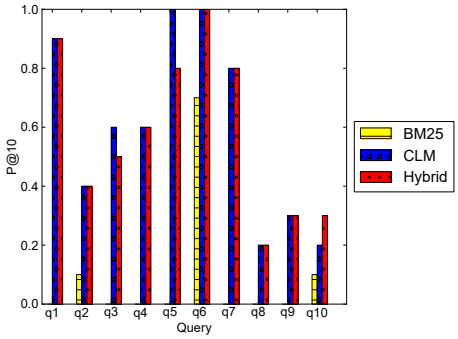

Figure 1: P@10 per query (we only show 10 out of the total 30 queries.)

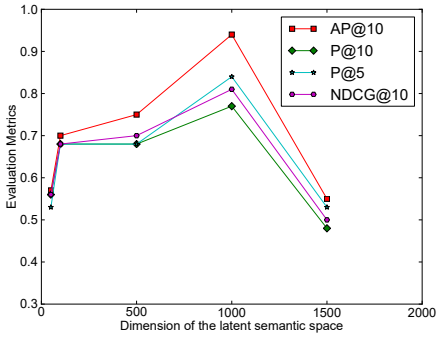

Figure 2: Performance of *CLM* with different dimensionality (K) of the latent semantic space.

of some of the queries improved by utilizing the relevant retrievals by BM25.

To gain further insight into the proposed model, we investigate the performance of different methods using P@10 metric on the per query basis. Due to lack of space we only show 10 out of the total 30 queries. As we can see, for most of the queries, CLM outperforms Translation methods with a large margin. We find there exist several reasons. First of all, many tokens in the tweets are not translated correctly in the Translation methods. This is a challenging issue because words on microblogs are often quite noisy. Secondly, the Translation methods rely on the match of the words in the tweets. If two tweets do not have overlap of tokens, the similarity by translation would be zero. Our proposed CLM model tackles these challenges by modeling the latent semantic space instead. The relations between tweets in different languages are automatically learned via the presence of common hashtags. Consequently, no translation is needed and fuzzy match is possible.

For many queries the performance improvement of the Hybrid model comes directly from the CLM. The CLM model retrieved tweets that do not have exact terms in the query. For example, for query q5 with text: "2025 facebook?", having the same text after Spanish transaltion, the CLM model retrieved tweets like "snapchat redessociales" and "elimina botón redessociales feedly" with the Google translation of "Snapchat social networks" and "delete button social networks feedly" respectively. Even though the term "facebook" does not appear in the Spanish tweets, our model learned the latent semantic to determine these as relevant tweets for social networks like facebook.

However there are few cases where Hybrid utilizes the high performance of both CLM and BM25. For example consider query q10 with query text: "urbana christmasgift", with its corresponding translation "christmasgifts Urbana". BM25 retrieves a relevant tweet: "urbana calzado" (urban footwear in English) which was not retrieved by CLM. In a couple of cases the Hybrid performs worse than CLM, owing to the poor retrievals of BM25. These observations indicate further improvement could be achieved if we use adaptive weights (e.g., high $1-\lambda$ for popular topic queries) to combine Translation and CLM instead of the fixed weight. We will explore this idea in future work.

Figure 2 shows the impact of the dimensionality of the latent semantic space on the CLM model, with $K = 50$, 100, 500, 1000, 1500, respectively. We can see that the model performance peaks at $K = 1000$ in all the metrics. As K is further increased to 1500, the performance quickly deteriorates, probably due to model overfitting.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate the task of cross-language microblog retrieval. We propose a pairwise latent semantic learning approach that models the low-dimensional shared semantic space between tweets in different languages. The experimental results demonstrate the effectiveness of the proposed approach over the traditional translation baseline. This work is an initial step towards a promising research direction. In future work, we plan to conduct more extensive experiments to validate the proposed approach, including different losses and different learning-to-rank approaches. We also plan to expand our evaluation by including the tweets without hashtags.

6. REFERENCES

[1] I. Alegria, N. Aranberri, C. España Bonet, P. Gamallo, H. Gonçalo Oliveira, E. Martínez Garcia, I. S. V. Roncal, A. Toral, and A. Zubiaga. Overview of tweetmt: a shared task on machine translation of tweets. 2015.

[2] S. T. Dumais, T. A. Letsche, M. L. Littman, and T. K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI*, 1997.

[3] C. Hu, P. Resnik, Y. Kronrod, V. Eidelman, O. Buzek, and B. B. Bederson. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using haitian creole emergency sms messages. In *SMT*, 2011.

[4] L. Jehl, F. Hieber, and S. Riezler. Twitter translation using translation-based cross-lingual retrieval. In *SMT*, 2012.

[5] W. Ling, G. Xiang, C. Dyer, A. W. Black, and I. Trancoso. Microblogs as parallel corpora. In *ACL*, 2013.

[6] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*. 2008.

[7] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the trec-2011 microblog track. In *TREC*, 2011.

[8] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009.

[9] I. Soboroff, I. Ounis, C. Macdonald, and J. Lin. Overview of the trec-2012 microblog track. In *TREC*, 2012.

[10] X. Tang, X. Wan, and X. Zhang. Cross-language context-aware citation recommendation in scientific articles. In *SIGIR*. ACM, 2014.