

Image-Image Search for Comparable Corpora Construction

Yu Hong^{✉†} Liang Yao[†] Mengyi Liu[†] Tongtao Zhang[‡] Wenxuan Zhou[†] Jianmin Yao[†] Heng Ji[‡]

[†]School of Computer Science & Technology, Soochow University, China
{tianxianer, liangysky, mengyiliu22, chrisnotkris7}@gmail.com, jyao@suda.edu.cn

[‡]Rensselaer Polytechnic Institute, NY, USA
{zhangt13, jih}@rpi.edu

Abstract

We present a novel method of comparable corpora construction. Unlike the traditional methods which heavily rely on linguistic features, our method only takes image similarity into consideration. We use an image-image search engine to obtain similar images, together with the captions in source language and target language. On the basis, we utilize captions of similar images to construct sentence-level bilingual corpora. Experiments on 10,371 target captions show that our method achieves a precision of 0.85 in the top search results.

1 Introduction

We limit our discussion to the sentence-level comparable corpora. Each sample in the dataset is a pair of bilingual sentences whose constituents are translations of each other, mostly or in whole. Briefly, they contain semantically similar contents, although they are expressed in different languages. In order to make it easier to read, we name such a sample as a bilingual sentence pair. See an English-Chinese case as below (English translations are attached behind).

- 1) UN Secretary-General Ban Ki-moon appointed “Red” from the Angry Birds as Honorary Ambassador for Green.
- 2) 联合国秘书长潘基文 任命 “愤怒的小鸟” 中的 红色 小鸟 为 绿色荣誉大使.
United Nations appoint angry bird from red bird as Honorary Secretary-General Ambassador for green culture
Ban Ki-moon

Large-scale comparable corpora generally contain rich and diverse bilingual translation examples, such as phrase-level equivalents as well as aligned words. Therefore, so far, such corpora have been admitted to be extremely useful in training translation models. During the past decades, great effort has been made by researchers (Rauf et al 2009, Skadina et al 2012, Santanu et al 2014 and Ann et al 2014) to construct and expand the corpora. They fulfilled the goal mainly by using cross-language content similarity measurement techniques. Lexical information, topic model, knowledge base and domain-specific terminology have all been proven to be effective in the acquisition of document-level equivalents (Talvensaari et al 2007, Li et al 2010, Zhu et al 2013 and Hashemi et al 2014).



Figure 1: Similar images and their captions in English and Chinese news websites (In this case, we would like to believe that an English journalist and a Chinese peer both attended the ceremony and took the photos from different perspectives, and then released them in the domestic news stories)

✉ corresponding author

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organizers. Licence details: <http://creativecommons.org/licenses/by/4.0/>

Different from the previous work, we employ similar images as the bridge to retrieve the bilingual sentence pairs. We suppose that captions generally represent the semantic contents of images, so that if two images are visually similar, their captions are very likely to be semantically comparable. Figure 1 shows two real images which are respectively crawled from the English and Chinese news websites. Listed below the images are the captions which, as usual in the websites, are written in the native languages (Note that the captions have been exhibited in (1) and (2)). Not just the similar images, it can be found that the captions are comparable as well.

Accordingly, we collect the captions of similar images from websites, and specify them as the candidates of bilingual sentence pairs. We rank the candidates in the order of image similarity, and determine the most highly ranked ones as the reliable bilingual sentence pairs. In practice, we build an image-image search engine. The engine uses images as queries, and retrieves similar images based on consistency of image features of scale-invariant keypoints.

In this paper, we aim to independently evaluate the proposed method rather than a well-structured sophisticated system, and answer the question of whether it is possible to capture the bilingual caption pairs (sentence pairs) by image-image search, if they really exist. In reality, the system should additionally consist of the modules like crawling, webpage structure analysis and image indexing. For these modules, we only provide a brief introduction. Nevertheless, it is noteworthy that these techniques are undoubtedly important for mining large-scale comparable data from websites.

The rest of the paper is structured as follows. Section 2 overviews the related work. We present the methodology and detail the image-image search engine in section 3. Section 4 shows the experimental settings and results. We conclude the paper in Section 5.

2 Related Work

There has been a considerable amount of work done in acquiring bilingual comparable corpora. One of the most widely used methods is the bilingual dictionary based text retrieval approach. [Talvensaari et al \(2007\)](#) created Swedish-English comparable corpora based on Cross-Language Information Retrieval (CLIR). They extracted keywords from the documents in the source language, and translated them into the target language by using a bilingual dictionary. The translations were used as the query words to retrieve the document-level equivalents in the target language. [Bo et al \(2010\)](#) implemented a bidirectional CLIR by using English-French dictionary.

[Su et al \(2012\)](#) employed the Microsoft Bing Translator to produce pseudo equivalents. Their experiments show that the slightly weak translations can be used to construct comparable corpora. It was also illustrated that the performance of Statistical Machine Translation (SMT) trained on such corpora was better than using lexicon. [Su et al \(2012\)](#)'s work shows the possibility to utilize the pseudo equivalents and the boosting approach to iteratively improve SMT.

The recent work seeks to use topic model to improve CLIR. The key issue which is mainly considered in this case is to precisely calculate the similarity between the translations and the documents in the target language. [Preiss et al \(2012\)](#) transformed the topic models in the source language to the target language, and measured the similarity at the level of topic. [Zhu et al \(2013\)](#) utilized the bilingual LDA model and structural information in similarity measurement.

Besides, knowledge base like Wikipedia has been proven to be useful for the discovery of bilingual equivalents ([Ni et al., 2009](#), [Smith et al., 2010](#)). [Otero et al \(2010\)](#) used Wikipedia categories as the restriction to detect the equivalents within small-scale reliable candidates. [Skadinaa et al \(2012\)](#) proposed a method to merge the comparable corpora respectively obtained from news stories, Wikipedia articles and domain-specific documents.

3 Methodology

First of all, we present the methodological framework. Then we introduce the crucial part of the image-image search engine, i.e., SIFT based image similarity measurement. Finally, we list the preprocessing methods for collecting and processing raw data.

3.1 Cross-Media Information Retrieval

Our method can be regarded as a kind of Cross-Media Information Retrieval (CMIR) technique. The main framework of CMIR is closely similar to that of CLIR. The only difference between them is the

bridge used to link a text in the source language with the equivalents in the target language. For the former, the bridge is the image, while the latter the language (e.g., keyword and translation).

Figure 2 shows the framework of CMIR. We also provide that of CLIR for comparison. For our method, i.e., CMIR, we collect the texts which summarize the main contents in images, and map the texts to the images in a one-to-one way. On the basis, we search comparable texts by pair-wise image similarity measurement. By contrast, CLIR generally employs a slightly weak translator or bilingual dictionary to generate rough or partial translations (see Section 2). Such translations are used as queries by a text search engine to acquire higher-level equivalents, such as [Talvensaar](#) et al (2007) and [Bo et al](#) (2010)’s work, using the translations of keywords as the clues to detect document-level equivalents.



Figure 2: Frameworks of CMIR and CLIR. SDB is a source Data Bank (DB), while TDB a target DB.

To some extent, CMIR is easier to use than CLIR. The crucial issue for CMIR is only to improve the quality of the search results. CLIR needs to additionally consider the quality of the bilingual dictionaries or the performance of the weak translators.

In order to conduct CMIR, however, we need to ensure that there is indeed a correspondence between a pair of image and text. It means that the text sufficiently depicts the meanings of the image. To fulfil the requirement, we collect the images and their captions from the structure-fixed webpages, and use them to build the reliable data bank for CMIR.

In practice, we collect the pairs of images and captions from both the news websites in the source language and that in the target language, respectively building source Data Bank (SDB) and target Data Bank (TDB). Given a caption C_s in SDB and the corresponding image I_s , we calculated the image similarity between I_s and all images $I_{t,s}$ in TDB. Then we rank all $I_{t,s}$ based on image similarity. Finally we select the captions $C_{t,s}$ of the most highly ranked $I_{t,s}$ as the equivalents of C_s . The pairs of C_s and C_t are used as the bilingual sentence pairs to construct the comparable corpora.

3.2 Image-Image Search

The image-image search engine uses each of the images in the SDB as a query. For every query, the engine goes through all the images in the TDB and measures their visually similarity to the query. The similarity will be used as the criterion to rank the search results. In this paper, we employ the Scale-Invariant Feature Transformation method (SIFT) for representing the images, creating scale-invariant keypoint-centered feature vector. On the basis, we calculate the image similarity by using the Euclidean distance of the keypoints.

SIFT is an image characterization method, which has been proven to be more effective than other methods in detecting the local details from different perspectives at different scales. This advantage causes precise image-to-image matching. Figure 3 shows the theory behind SIFT.

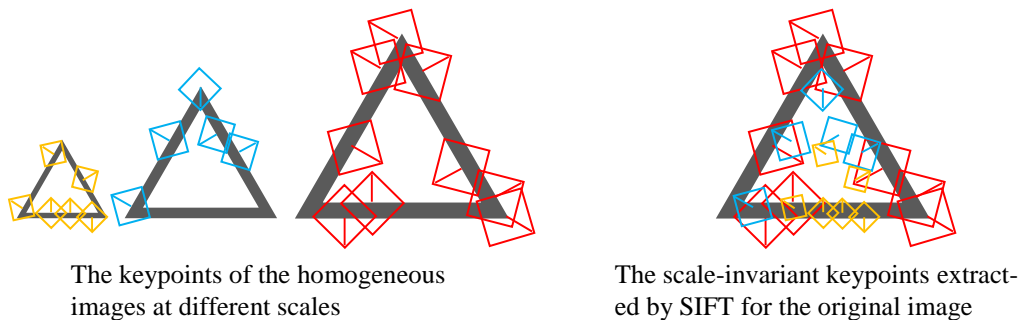


Figure 3: SIFT process (assume that the biggest triangle is the original image). The keypoints are denoted by the directed square marks (the direction is denoted by the line that radiates outward from the middle of the square marks)

First, SIFT zooms in and out on the original image, so as to obtain the homogeneous images at different scales (see the three triangles at the left side of Figure 3). Second, SIFT extracts keypoints respectively in the homogeneous, and merges them to generate a set of scale-invariant keypoints (see those points in the triangle at the right side of Figure 3). The feature space which is instantiated by those scale-invariant keypoints is scale-independent, and therefore extremely conducive to detecting visually similar images at different scales (Lowe et al 1999, Lowe et al 2004).

SIFT employs the most distinctive point in a small area as a key feature, i.e., the so-called keypoint. Due to the local processing in different areas, SIFT is not only able to obtain locally optimal features but maintain all the similar key features occurred in different parts of the image.

Following the state-of-the-art SIFT (Lowe et al 2004, Yan et al 2004 and Hakim et al 2006) method, we define a small area as the set of a sampling point and the adjacent points (neighbours). It is noteworthy that the area includes not just the neighbours in the original image but those in the homogeneous images at different scales. We use Gaussian function to fit the size of all the points in the area. On the basis, we use the difference of Gaussian function to determine the extreme point, and specify the point as the distinctive point in the area.































	#1st	#2nd	#3rd	#4th	#5th
$\theta=0.4$					
$\theta=0.5$					
$\theta=0.6$					
$\theta=0.7$					
$\theta=0.8$					
$\theta=0.9$					

Figure 4: Different versions of the top 5 image search results. They were respectively obtained when the threshold θ were finely turned from 0.6 to 0.9.

We model each keypoint by pixel-wise vectors in the keypoint-centered $16*16$ windows. The vector represents both the direction and the value of image gradient. Lowe et al (2004) detail the gradient measurement method.

In total, we extract keypoints for the image representation. Given two images, we calculate the similarity by the average Euclidean distance of the matching keypoints. For a keypoint x in the source image, we determine the matching point in the target image by the following steps: First, we acquire two most similar keypoints y and z in the target image. Assume that the similarity of (x, y) is smaller than (x, z) , second, we calculated the ratio r of the similarity $s(x, y)$ to $s(x, z)$. If r is bigger than a threshold θ , we determine that the keypoint z is the matching point of x ; otherwise there isn't any matching point of x in the target image. We set the threshold θ as 0.8.

A smaller value of r ($r < \theta$) will introduce many unqualified matching points in the image similarity calculation. It will reduce the precision of image search results. It means that most of the retrieved im-

ages are either dissimilar or unrelated. By contrast, a larger value causes few available matching points. It will influence the diversity of the search results. It means that most of the retrieved images are the same with each other or even extracted from the same provenances. Obviously, we would like to see that they derive from different `Medias` in different languages.

Figure 4 lists a series of images, which are the top 5 search results obtained by using different levels of θ . This group of search results are very representative in our experiments, able to reflect that the setting value 0.8 of θ is a reasonable boundary between correct and incorrect results. In particular, it can be found that such a threshold ensures the diversity of the correct results (Note that the query in this example is the left image in Figure 1).

3.3 Collecting and Processing Raw Data

We crawl the images and captions by using `crawler4j`¹, which is an open source toolkit specially developed for effectively crawling web data. On the basis, we use regular expressions to extract images and captions from the structured source files of the crawled web pages.

An optional preprocessing for an experimental system is to index images. It enables high-speed retrieval. We apply the locality-sensitive hashing (LSH) technique² for content-based image indexing.

4 Experiments

We conduct a pilot study for CMIR towards comparable corpora construction. The goal of this study is to verify whether image-image search is useful for the discovery of textual equivalents in TDB.

There is an important problem need to be solved firstly: *Reliability*. As mentioned in section 3.1, TDB is a data bank which contains a great number of images, along with the captions in the target language (named target captions for short). However, if we randomly select the captions in the source language (source captions) as the test samples for mining the bilingual captions, it is easy for us to encounter the problem that there is not a real equivalent in TDB. In the case, the experimental results are definitely unreliable. For example, the precision rate in the 5 highly-ranked target captions will always be 0. On the contrary however, if we added some ground-truth equivalents of the test samples to TDB, the experimental settings will be far from the real condition.

To solve the problem, we propose an automatic method of measuring comparability between source captions and target captions. Based on the measurement results, we collect pseudo ground-truth equivalents, and use them to enrich the test data. By this way, we can build an experimental environment similar to the real condition. We detail the method in section 4.4.

Besides, as usual, we show the corpus, traditional evaluation metrics and main experimental results one-by-one, which can be found in sections 4.1, 4.2 and 4.3. For the part of main result, we report the precision in top 5 highly-ranked target captions, as well as the ranking results, at four levels of comparability (parallel-level (abbr., `Par.`), comparable (`Com.`), pseudo-comparable (`Pse.`), and incomparable (`Inc.`)). In addition, we compare our method with the state-of-the-art CLIR method and the other image-image search engine.

4.1 Corpus

We crawled 42,633 images from Chinese news websites to initialize TDB. Each corresponds to a sole caption. The websites include China news, News of Sina and Xinhua Net (Chinese). In order to ensure sentence-level comparable corpora construction, we filtered the captions which are generated with multiple sentences or have a length of more than 20 Chinese words. Of course, the images of the captions were also filtered out of the TDB. Eventually, we obtained a TDB which contains 10,371 pairs of images and captions. As mentioned above, we didn't know whether there is an equivalent in the TDB for a source caption, and even if there does exist, we are blind to it (Black box).

We built a SDB (i.e., source data bank) in the same way. The source captions in the SDB are collected from the English news websites like online CNN, BBC and Xinhua Net (English). It is a mini-sized data bank, containing only 52 pairs of source captions and images (See their topics in Table1).

1 <https://github.com/yasserg/crawler4j>

2 <https://github.com/embr/lsh>

Honestly, this SDB can only support an English-Chinese CMIR (or CLIR), in which the English captions and the images serve as the queries. In our experiments, we use the queries as the test data.

Russia military parade/10	Russia Putin/10	Obama depart/10	Brazil Olympic/7	Greek migrant/10	Putin birth/10	Michelle/6	Pluto/8
Vehicle Afghanistan/10	Israel bomb/10	Obama Cuba/10	Artistic Korea/5	Curry Warrior/10	Taj Mahal/10	Ankara/10	Nepal/9
Ecuador earthquake/10	Leo Oscar/10	Earthquake/10	Obama meet/10	Prime Russia/4	Xing Zhan/8	Xi talk/10	Wolf/10
Mexico explosion/10	Hindu fire/10	Leonardo/10	NASA image/7	Angry birds/10	Kim speak/8	Baghdad/6	
Miss South Africa/9	Mitsubishi/10	Kon tiki2/5	North Korea/9	Prime Italy/4	Diamond/10	Volcano/8	
Brussels damage/10	Seattle fire/7	Rocket/10	Putin Kerry/10	Trump/5	Mh370/10	Whale/7	
Pakistan floods/10	Satellite/10	River/10	Queen birth/10	Kobe/10	Castro/10	Protest/7	

Table 1: The topics of the pairs of target images and captions in the SDB, along with the numbers of the equivalents in the TDB (They are listed in the format `topic/number`)

Towards the images in the SDB, we collect similar images in the Chinese news websites, and use them and their captions as the ground-truth data. By this way, we collect at least 5 equivalents (similar image and comparable caption) for each sample in the SDB. In total, we collect 451 ground-truth equivalents. We added them to the TDB. From here on, the TDB is no longer a black box for us. The correct and incorrect equivalents are the prior knowledge for evaluating the CMIR and CLIR systems.

4.2 Evaluation Metrics

We conduct our CMIR process in TDB, with the aim to verify whether CMIR is able to seek out the comparable target captions in a large-scale data set. This is the kernel of the proposed corpora construction method. If it is promising, we can accomplish the corpora construction by continuous CMIR using a massive number of source captions as the queries. Therefore we focus on evaluating the CMIR in this paper, using the samples in the mini-sized SDB as the queries.

The basic evaluation metric is the Precision rate (P). We didn't consider the Recall (R) rate. It is because that the genuine requirements of comparable corpora construction are the noise-free high-quality equivalents, but not all. Not just the acquisition of qualified equivalents, $P@N$ also reflects the ability of a CMIR (or CLIR) system to filter incorrect equivalents out of the top n search results.

4.3 Main Results

We rank the retrieved target captions (candidate equivalents) by CMIR in terms of image similarity, and evaluate the performance in the top- n ($1 \leq n \leq 5$) highly ranked target captions. Table 2 shows the performance ($P@n$). Besides, we compare our method with [Talvensaaari et al \(2007\)](#)'s CLIR system. Note that the listed performance in the table is the Macro precision among the 52 test samples.

As shown in Table 2, CMIR achieves promising results. The precision in the top 5 search results is more than 60%. Besides, CMIR outperforms the state of the art CLIR, yielding nearly 2% performance gains at top 1 and in top 2.

	#P@1	#P@2	#P@3	#P@4	#P@5
CMIR (SIFT)	0.846	0.788	0.718	0.658	0.615
CLIR	0.827	0.769	0.756	0.745	0.703

Table 2: Main test results (Precision rates in top- n equivalents) for both CMIR and CLIR

It is easy to raise a question of whether the degree of comparability of source and target captions is proportionate to the similarity of their images. If it does, we can conclude that CMIR is conducive to the acquisition of high-quality equivalents. In order to answer the question, we verify the distributions of different levels of equivalents over the image-similarity based rankings. We consider four levels of equivalents, including *Par*, *Com*, *Pse*, and *Inc*. Note that the levels were manually annotated beforehand. Table 3 shows their definitions and a concrete example for each. A smaller sequence number in the ranking list implies a higher image similarity. For example, the image of the ranked 1st target caption (top equivalent) is most similar to the corresponding image of the source caption. Figure 5 shows the distributions for the rankings from 1 to 5.

Definition of Par.: Two sentences are the translation of each other or approximate translation with minor variations, which can be aligned on the word level. (see examples as below)

(English) *Italian Prime Minister Matteo Renzi meets with visiting Chinese Foreign Minister Wang Yi in Rome, Italy, on May 5, 2016.* /

(Chinese) 5月5日/May 5, 意大利/Italy 总理/premier 伦齐/Renzi 在/at 罗马/Rome 会见/meet with 中国/Chinese 外交部长/foreign minister 王毅/Yi Wang.

Definition of Com.: Two sentences in different languages depict the same event or topic, from very similar perspectives. One contains the translations of most constituents of the other. (see examples as below)

(English) *Flash floods in Pakistan and Kashmir Kill at Least 53.*

(Chinese) 巴基斯坦/Pakistan 爆发/break 洪灾/flood 和/and 山体/mountain 滑坡/landslides 至少/at least 53/53 死/dead 60/60 伤/injury.

Definition of Pse.: The sentences present the same event or topic from very different perspectives. They only contain several semantically equivalent words or phrases. (see examples as below)

(English) *people take photos of bodies of dead stranded sperm whales behind the dyke of Kaiser Wilhelm Koog.*

(Chinese) 8/8 头/number 抹香鲸/sperm whale 搁浅/stranded 德国/Germany 海滩/sea beach 起重机/crane 运输/transport 尸体/corpse.

Table 3: Definitions of the parallel-level (Par.), comparable (Com.) and pseudo-comparable (Pse.) equivalents, along with the examples. The rest cases are specified as incomparable (Inc.) sentences.

It can be found that most of the lower-level equivalents (Pse and Inc) were ranked at the bottom of the ranking list: more than 69% of Pse-level equivalents won the 3rd, 4th and 5th places, and 88% Inc won the same places (see the left diagram in Figure 5). On the contrary, most of the higher-level equivalents (Par and Com) were ranked at the top: more than 66% of Par-level equivalents won the 1st, 2nd and 3rd places, and 78% of Com won the same places. It illustrates that CMIR is able to distinguish the high-quality equivalents from the low-quality, and rank the former to the top of the ranking list. It helps a translator finely tune the proportions of comparable samples of different qualities in a bilingual corpora as requirement, e.g., noise-free smaller-sized corpora or large-scale noisy corpora (The former are reliable but provide less translation knowledge, the latter are just the opposite).

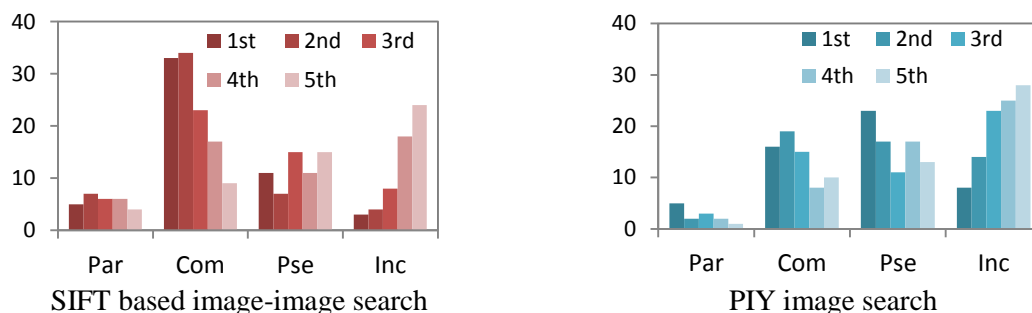


Figure 5: Distributions of different levels (Par, Com, Pse and Inc) of equivalents over the image-similarity based rankings. Exhibited in the left diagram are the distributions of the retrieved equivalents by our SIFT based image-image search engine, while the right by PIY image search. Each column in the histogram denotes the number of certain level of equivalents that arrive at the same ranking.

Considering the important influence of image-image search to translation-oriented CMIR, we conduct an additional experiment to evaluate different image search engines. We employ an open-source engine, named PIY³, which is a well-developed and easy to use. PIY calculates image similarity by using 3D colour histogram. Figure 5 shows the performance of the PIY based CMIR (see the right diagram). It can be found that PIY has the same advantage with our CMIR method, capable of raising high-quality equivalents in the search results. Nevertheless, PIY achieved a worse precision in top-5 search results. The macro-average precision is 0.5, far below the performance of our method.

³ <http://www.pyimagesearch.com/2014/12/08/adding-web-interface-image-search-engine-flask/>

4.4 Collaborative Evaluation

We propose an automatic method to measure the comparability between a source caption and the target (a candidate equivalent). The method can be used to evaluate the results of CMIR without knowing the ground truth. It measures the comparability by using the following features:

- Content similarity (f_c) is calculated by the Cosine measure between TFIDF based VSM models of source caption and target caption (Only content words are considered in the calculation).
- Co-occurrence of entities (f_e) is calculated by the joint co-occurrence rates of entity mentions in source caption and target caption.
- Length ratio (f_l) is the difference of the length of the captions. If they have the same length, f_l is equal to 1, otherwise a smaller value (divide the length of the shorter by that of the longer)

On the basis, we measure the comparability by combing the features by the linear weighted sum method: $C = \alpha \cdot f_c + \beta \cdot f_e + \gamma \cdot f_l$, where the parameters α , β and γ are empirically set as 0.8, 0.15 and 0.05. Further, we divide the ground-truth samples (i.e., bilingual caption pairs) into three classes in terms of the prior level of comparability, i.e., `Par`, `Com` and `Pse`. For each class, we calculate the average C . Table 3 shows the calculation results. It can be found that the average C -measure of the classes of the ground truth closely fit the manual ratings of the classes (the rating score 3 corresponds to the `Par`-level, 2 to `Com` and 1 to `Pse`). The Pearson factor between the ratings and the C scores is high up to 0.99. It illustrates that the trend of gradient descent of C is similar to that of manual ratings.

	#Par.	#Com.	#Pse.	Pearson
Rating	3.0	2.0	1.0	0.993
C-measure	0.646	0.496	0.397	

Table 3: Comparability C-measure and Pearson parameter

Accordingly, we use C -measure to ensure the reliability of the evaluation process when there is lack of known ground-truth equivalents. We set $\mathfrak{G}(C, \varepsilon)$ as a linear function of the deviation ε from the average C -measure of certain level of equivalents: $\mathfrak{G}(C, \varepsilon) = C - a \cdot \varepsilon$. We estimate the optimal factor a in the training data by maximizing the precision. We use $\mathfrak{G}(C, \varepsilon)$ as the criteria to determine whether a target caption is a qualified equivalent for a certain level of comparability. For example, if $C(x, y) > \mathfrak{G}(C, \varepsilon)$, x is comparable to y (at `Par`-level, `Com` or `Pse`). The qualified equivalents will be used as the ground-truth data to evaluate the performance of the CMIR systems.

An instantiated $\mathfrak{G}(C, \varepsilon)$ enables an experiment on large-scale test data (source captions as queries) and rich ground-truth data. The test result, therefore, will be more reliable than the current case. Active learning can be applied for enhancing the evaluation process.

5 Conclusion

In this paper, we propose a CMIR method to obtain bilingual sentence pairs, with the aim to construct sentence-level comparable corpora. The CMIR applies SIFT algorithm for image similarity measurement. On the basis, it detects the captions of similar images in source data and target data, as use them as search results. Experiments show that CMIR is promising in acquiring the comparable captions.

In the future, we will focus on the implement of a CMIR-based corpora constructor. The first difficulty for us is to determine the source captions that indeed have at least one equivalent in TDB. Obviously, the CMIR results for other source captions all are incorrect. If add them to the corpora, the quality of the data set will be reduced largely. The resolution is to use burst measurement method to detect break news, and use the captions and images in the news stories as the source data. It may work well because that break news would be reported widely around the world. There should be always some topic-related stories occurred in the news websites in the target language. This largely increases the probability that target data contain the desired equivalents.

Another crucial issue is to predict the numbers of the target captions which will be added to the corpora. A possible solution is to measure the textual comparability for a massive number of highly ranked target captions, and use $\mathfrak{G}(C, \varepsilon)$ as the threshold to filter out the `Inc`-level samples. However this method will negatively influence efficiency. Nevertheless, this problem may raise an interest in the joint model of textual comparability and image similarity, as well as collaboration methods.

Acknowledgements

This research is supported by the U.S. DARPA DEFT Program (No. FA8750-13-2-0041), ARL NS-CTA (No. W911NF-09-2-0053), NSF CA-REER Award (IIS-1523198), and the National Natural Science Foundation of China, No.61672368, No.61373097, No.61672367, No.61272259. The authors would like to thank the anonymous reviewers for their insightful comments and suggestions. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. and CHN Governments. The U.S. and CHN Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. Yu Hong, Professor Associate in Soochow University, is the corresponding author of the paper, whose email address is tianxianer@gmail.com.

Reference

- Abdel Hakim, Alaa Elvalser, and Aly A. Farag. 2006. CSIFT: A SIFT descriptor with color invariant characteristics. *Conference on Computer Vision and Pattern Recognition*, pages 1345-1354.
- Abdul Rauf, Sadaf, and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23.
- Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652.
- Degen Huang, Lian Zhao, Lishuang Li, and Haitao Yu. 2010. Mining large-scale comparable corpora from chinese-english news collections. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 472–480.
- Fangzhong Su and Bogdan Babych. 2012. Development and application of a cross-language document comparability metric. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3956–3962.
- Hashemi, Homa B, and Azadeh Shakery. 2014. Mining a Persian–English comparable corpus for cross-language information retrieval. *Information Processing & Management*, pages 384-398.
- Irvine Ann, Chris Callison Burch. 2014. Using comparable corpora to adapt mt models to new domains. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 437–444.
- Inguna Skadiņa, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su. 2012. Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 438-445.
- Judita Preiss. 2012. Identifying comparable corpora using lda. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 558–562.
- Ke, Yan, and Rahul Sukthankar. 2004. PCA-SIFT: A more distinctive representation for local image descriptors. *Conference on Computer Vision and Pattern Recognition*, pages 456-466.
- Lowe David. 1999. Object recognition from local scale-invariant features. *The proceedings of the 7th IEEE International Conference on Computer vision*, pages 1150-1157.
- Morin, Emmanuel, and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, 49th Annual Meeting of the Association for Computational Linguistics*, pages 27–34.
- Pablo Gamallo Otero and Isaac González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 21–25.

- Pal Santanu, Partha Pakray, and Sudip Kumar Naskar. 2014. Automatic building and using parallel resources for SMT from comparable corpora. In Proceedings of 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 48–57.
- Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
- Tuomas Talvensaaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems*, 25(1):4.
- Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In Proceedings of the 18th International Conference on World Wide Web, pages 1155–1156.
- Zede Zhu, Miao Li, Lei Chen, and Zhenxin Yang. 2013. Building comparable corpora based on bilingual lda model. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 278–282.