Image-Mediated Learning for Zero-Shot Cross-Lingual Document Retrieval

Ruka Funaki, Hideki Nakayama
Machine Perception Group
Graduate School of Information Science and Technology
The University of Tokyo
{funaki,nakayama}@nlab.ci.i.u-tokyo.ac.jp

Abstract

We propose an image-mediated learning approach for cross-lingual document retrieval where no or only a few parallel corpora are available. Using the images in image-text documents of each language as the hub, we derive a common semantic subspace bridging two languages by means of generalized canonical correlation analysis. For the purpose of evaluation, we create and release a new document dataset consisting of three types of data (English text, Japanese text, and images). Our approach substantially enhances retrieval accuracy in zero-shot and few-shot scenarios where text-to-text examples are scarce.

1 Introduction

Cross-lingual document retrieval (CLDR) is the task of finding relevant documents in one language given a query document in another language. While sufficiently large-scale corpora are critical for parallel corpus-based learning methods, manually creating corpora requires huge human effort and is unrealistic in many cases.

A straightforward approach is to crawl bilingual documents from the Web for use as training data. However, because most documents on the Web are written in one language, it is not always easy to collect a sufficient number of multilingual documents, especially those involving minor languages. Let us consider the multimedia information in documents. We can, for example, find abundant pairings of text and images, e.g., text with the ALT property of <IMG> tags in HTML, text with photos posted to social networking sites, and articles on Web news posted with images. Unlike text, an image is a universal representation; we can easily understand the semantic content of images regardless of our mother tongue. Motivated by this observation, we expect that we can learn the relation of two languages indirectly through images, even if we do not have sufficient bilingual text pairs (Figure 1).

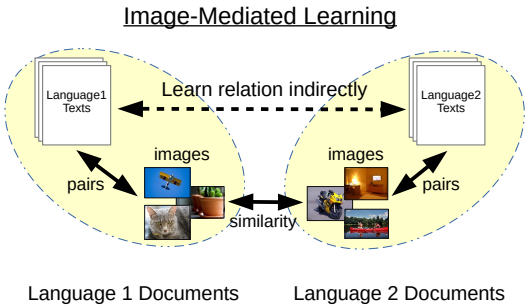

Figure 1: Concept of image-mediated learning. Our idea is to learn the relation between two languages indirectly by using images attached to text. If two documents written in different languages include images with similar image features, it is likely that the texts contained in the two documents are similar. Based on this idea, we seek the relation of texts written in different languages mediated by the similarity between images.

Generally, traditional image recognition techniques (or image features) are very poor compared with those in the natural language processing field. In recent years, however, deep learning has resulted in a breakthrough in visual recognition and dramatically improved image recognition accuracy in generic domains, which is rapidly approaching human recognition levels (Fang et al., 2015). We expect that these state-of-the-art image recognition technologies can effectively assist CLDR tasks.

We show that hub images enable zero-shot training of CLDR systems and improve retrieval accuracy given only a few parallel text samples.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 585–590,
Lisbon, Portugal, 17-21 September 2015.

2 Related Work

2.1 Multimodal Learning for CLDR

Multimodal learning, defined as a framework for machine learning using inputs from multiple media or sensors, has played a key role in various cross-modal applications. The most widely used standard method for multimodal learning is canonical correlation analysis (CCA) (Hotelling, 1936), which projects multimodal data into a shared representation. For example, CCA has been successfully used in image retrieval (tag to images) and image annotation (image to tags) (Hardoon et al., 2004; Rasiwasia et al., 2010; Gong et al., 2014). In the context of CLDR, each language's texts constitute one modality. CCA has also commonly been used for cross-lingual information retrieval (Vinokourov et al., 2002; Li and Shawe-Taylor, 2004; Udupa and Khapra, 2010). Whereas CCA can handle only two modalities, we need to consider relations between three modalities because we use images in addition to the two languages. Therefore, we focus on an extension of CCA, generalized canonical correlation analysis (GCCA), to handle more than two inputs (Kettenring, 1971).

2.2 Zero-Shot Learning for CLDR

Our core idea is to use another modality (image) as a hub to indirectly learn the relevance between two different languages. The work by Rupnik et al. is probably the closest to ours (Rupnik et al., 2012). In their study, they used a popular language (e.g., English) with enough bilingual documents shared with other languages as a hub to enhance CLDR for minor languages with few direct bilingual texts available. Nevertheless, this method assumes that parallel corpora of the hub and target languages exist and therefore, its application is limited to specific domains where manual translations are readily available, such as Wikipedia and news sites. Contrarily, because we use images as the hub, we can use documents closed with respect to each language for training. Considering that current generic Web documents are mostly closed with respect to one language, yet equipped with rich multimedia data, our setup is assumed to be more reasonable.

	Division	English	Images	Japanese
1	[train-E/I]	E_1	I_1	-
2	[train-I/J]	-	I_2	J_2
3	[train-E/J]	E_3	-	J_3
4	[test-E/J]	E_4	-	J_4

Table 1: Division of training and test data. Each division of training dataset is missing one of the three modalities.

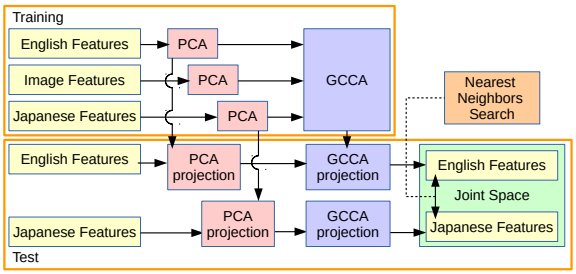

Figure 2: System overview

3 Our Approach

3.1 Overview of Image-Mediated Learning

We use the following notations for specifying each non-overlapping data division.

1. [train-E/I]: Training documents consisting of English text and images.
2. [train-I/J]: Training documents consisting of images and Japanese text.
3. [train-E/J]: Training documents consisting of English and Japanese text.
4. [test-E/J]: Test documents consisting of English and Japanese text.

We define IDs for each modality in each division as given in Table 1. For example, E_1 represents features of English text in the [train-E/I] division. Typical CLDR based on parallel corpora uses only [train-E/J] for training and [test-E/J] for evaluation. In the zero-shot learning scenario without any [train-E/J] data, we use only [train-E/I] and [train-I/J] for training. In the few-shot learning scenario, we also use a small number of [train-E/J] samples. We call this approach image-mediated learning.

An overview of our system is depicted in Figure 2. We compress features by principal component analysis (PCA) and train them by GCCA. For testing, we compress features by PCA, project features by GCCA, then, search the nearest neighbors from Japanese to English in the joint space.

3.2 Feature Extraction

A convolutional neural network (CNN) is one of the most successful deep learning methods for visual recognition. It is known that we can obtain very good image features by taking activation of hidden neurons in a network pre-trained by a sufficiently large dataset (Donahue et al., 2013). We apply the CNN model pre-trained using the ILSVRC2012 dataset (Russakovsky et al., 2015) provided by Caffe (Jia et al., 2014), a standard deep learning software package in the field of visual recognition.

As the text feature for both English and Japanese, we use the bag-of-words (BoW) representation and term frequency-inverse document frequency (TF-IDF) weighting. The MeCab (Kudo et al., 2004) library is used to divide Japanese text into words by morphological analysis. No preprocessing approaches like eliminating stop words and stemming, are used.

3.3 GCCA

GCCA is a generalization of CCA for any m modalities ($m = 3$ in our case). Although several slightly different versions of GCCA have been proposed (Carroll, 1968; Rupnik et al., 2012; Velden and Takane, 2012), we implement the simplest one (Kettenring, 1971) because GCCA itself is not the main focus of this study.

Let E, I, and J denote English, images, and Japanese, respectively. For feature vector $\mathbf{x}_k, \forall k \in \{E, I, J\}$, let $\mathbf{z}_k = (\mathbf{x}_k - \overline{\mathbf{x}}_k)\mathbf{h}_k$ denote its canonical variables. Σ_{ij} denotes a covariance matrix of modalities i and j where $i, j \in \{E, I, J\}$. Projection vectors $\mathbf{h}_k$ are computed such that they maximize the sum of correlations between each pair of modalities obtained by solving the following generalized eigenvalue problem:

$$\frac{1}{2}\begin{pmatrix} \mathbf{0} & \Sigma_{EI} & \Sigma_{EJ} \\ \Sigma_{IE} & \mathbf{0} & \Sigma_{IJ} \\ \Sigma_{JE} & \Sigma_{JI} & \mathbf{0} \end{pmatrix}\mathbf{h} = \rho\begin{pmatrix} \Sigma_{EE} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{II} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{JJ} \end{pmatrix}\mathbf{h} \quad (1)$$

where $\mathbf{h} = (\mathbf{h}_E^T, \mathbf{h}_I^T, \mathbf{h}_J^T)^T$. The canonical axises $\mathbf{h}$ are normalized such that $\frac{1}{3}\sum_{k\in\{E,I,J\}}^{3} \mathbf{h}_k^T \Sigma_{kk} \mathbf{h}_k = 1$. Additionally, we add a regularization term to the self covariance matrices to prevent over-fitting; that is, we set $\Sigma_{kk} \rightarrow \Sigma_{kk} + \alpha I$, where α is a parameter to avoid the singularity issue.

Despite our training datasets having only two of the three modalities as given in Table 1, we can handle this situation naturally by computing covariance matrices from the available data only. For example, in the few-shot learning scenario, we compute Σ_{EI} using E_1 and I_1, and Σ_{EE} using E_1 and E_3. In the zero-shot learning scenario, because [train-E/J] is not available, we compute Σ_{EE} using E_1 only and use a zero matrix for Σ_{EJ}.

3.4 Nearest Neighbor Search in Joint Space

We can find relevant documents in another language by computing the distances from the query documents using the coupled canonical subspaces. Having set Japanese as the query language, we retrieve documents written in English. Nearest neighbors are obtained as follows:

$$\hat{j} := \arg\min_j d(\mathbf{z}_J^i, \mathbf{z}_E^j), \quad (2)$$

where $\mathbf{z}_J^i$, $\mathbf{z}_E^j$ are projected feature vectors of the query and target documents, respectively, and $d(\cdot)$ is a distance function, which in our case, is the Euclidean distance.

4 Experiment

4.1 Pascal Sentence Dataset with Japanese Translation

The UIUC Pascal Sentence Dataset (Rashtchian et al., 2010) contains 1000 images, each of which is annotated with five English sentences describing its content. This dataset was originally created for the study of sentence generation from images, which is one of the current hot topics in computer vision. To establish a new benchmark dataset for image-mediated CLDR, we included a Japanese translation for each English sentence provided by professional translators[1], as shown in Figure 3. In this experiment, we bundled the five sentences attached to each image for use as one text document. Therefore, in our setup, each of the 1000 documents in the dataset consists of three items: an image, and the corresponding English and Japanese text.

[1]Dataset is available at:
http://www.nlab.ci.i.u-tokyo.ac.jp/dataset/pascal_sentence_jp/

English Texts	Images	Japanese Texts
- A family on a boat with a cross on a river - A happy couple with a young child wearing a life preserver sitting on a boat. - A man, a woman, and a child sit on boat with a large cross on it. - A man, women and small child sitting on top of a boat moving along the river. - Family of three sitting on deck, child wearing red vest, brush and shoes are seen in the foreground.		- 川で十字架のあるボートに乗っている家族。 - ボートに座っている、救命具を着た幼い子どもと幸せなカップル。 - 男性、女性と子どもが大きな十字架のあるボートに座っています。 - 川に沿って動いているボートの上部に座っている男性、女性と小さな子ども。 - ブラシと靴が前景に写されている、子どもが赤いベストを着て、デッキに座っている三人の家族。
- A black and white cow in a grassy field stares at the camera. - A black and white cow standing in a grassy field. - A black and white cow stands on grass against a partly cloudy blue sky. - a cow is gazing over the grass he is about to graze - Black and white cow standing in grassy field.		- 草地の黒と白の雌牛がカメラをじっと見ています。 - 草地に立っている黒と白の雌牛。 - 一部曇った青空を背に黒と白の雌牛が草地に立っています。 - 雌牛が食べようとしている草をじっと眺めています。 - 草地に立っている黒と白の雌牛。
⋮	⋮	⋮

Figure 3: Examples from the Pascal Sentence Dataset with Japanese translations: each image has about five sentences describing it from different perspectives.

4.2 Evaluation

We randomly sampled data from the dataset for each division in Table 1 without any overlap; we ignored the modality of each document that was not available in each data division (e.g., Japanese text in [train-E/I]). We ran experiments with varying sample sizes for [train-E/I] and [train-I/J], that is, 100, 200, 300, and 400. Furthermore, we gradually increased the number of [train-E/J] samples from 0 to 100 to emulate the few-shot learning scenario. The size of the test data [test-E/J] was fixed at 100. Following this setup, we performed image-mediated CLDR based on GCCA, and compared the results with those obtained by standard CLDR using only [train-E/J] data with CCA. We evaluated the performance with respect to the top-1 Japanese to English retrieval accuracy in the test data. Given that we used 100 test samples, the chance rate was 1%. For each run, we conducted 50 trials randomly replacing data and used the average score. All features were compressed into 100 dimensions via PCA and α was set to 0.01.

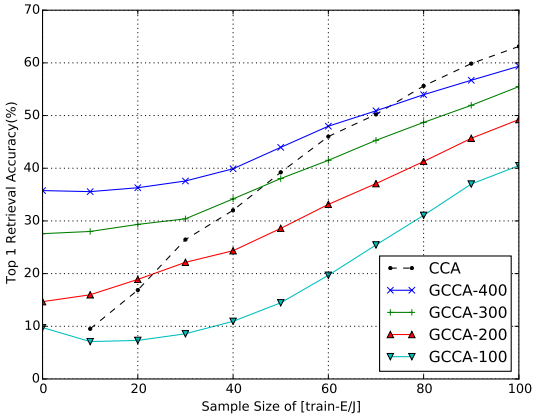

Figure 4: Retrieval accuracy varying the number of [train-E/J] data. Each colored line shows the performance of our method with a different sample size of [train-E/I] and [train-I/J] data (e.g., GCCA-400 denotes respective 400 samples of [train-E/I] and [train-I/J] for GCCA). We used image features extracted from GoogLeNet and text features represented as bags-of-words.

The experimental results, illustrated in Figure 4, clearly show that better accuracy is obtained with a greater number of text-image data in both zero-shot and few-shot scenarios. We can expect even better zero-shot accuracy with more text-image data, although, we cannot increase [train-E/I] and [train-I/J] more than 400 each in the current setup because of the restricted dataset size. We summarized results in zero-shot scenario in Table 2 in several cases. Although both GCCA and CCA show improved performance as the sample size of [train-E/J] increases, not surprisingly, GCCA is gradually overtaken by CCA when we have enough samples to learn the relevance between English and Japanese texts directly. However, accuracies of image-mediated learning in the cases when [train-E/J] is scarce are higher than CCA baseline. Hence, we confirmed that the image-mediated model is also effective in the few-shot learning scenario.

Model	Accuracy(%)
GCCA-400(BoW)	37.4 ± 3.8
GCCA-300(BoW)	27.6 ± 3.4
GCCA-200(BoW)	14.7 ± 3.2
GCCA-100(BoW)	9.8 ± 2.3
GCCA-400(TF-IDF)	42.0 ± 4.6
GCCA-300(TF-IDF)	31.6 ± 4.3
GCCA-200(TF-IDF)	17.2 ± 2.6
GCCA-100(TF-IDF)	11.8 ± 2.7

Table 2: Accuracy of zero-shot learning. Image features are extracted from GoogLeNet and both the bag-of-words (BoW) and the TF-IDF model are used as text features.

Feature	Accuracy(%)
CNN(GoogLeNet), BoW	37.4 ± 3.8
CNN(VGG), BoW	31.3 ± 3.5
CNN(CaffeNet), BoW	25.1 ± 3.4
FisherVector, BoW	10.8 ± 2.7
CNN(GoogLeNet), TF-IDF	42.0 ± 4.6
CNN(VGG), TF-IDF	37.8 ± 2.9
CNN(CaffeNet), TF-IDF	29.7 ± 4.4
FisherVector, TF-IDF	12.6 ± 2.7

Table 3: Accuracy of zero-shot learning in multiple image features. The sample size of both [train-E/I] and [train-I/J] is 400. Both the bag-of-words (BoW) and the TF-IDF model are used as text features.

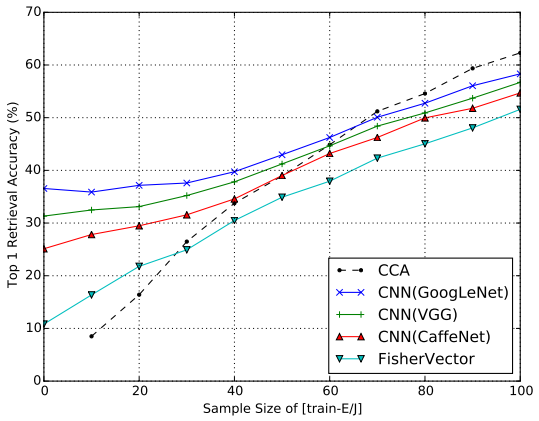

Figure 5: Retrieval accuracy using different image features in image-mediated CLDR. The sample size of both [train-E/I] and [train-I/J] is 400. Text features are based on the bag-of-words model.

4.3 Effect of Image Features

We also verified the effect of the performance of image features in our framework (see Figure 5 and Table 3). CNN has improved dramatically over the last few years, and many new powerful pre-trained networks are currently available. We compared three different features extracted from GoogLeNet (Szegedy et al., 2014), VGG 16 layers (Chatfield et al., 2014), and CaffeNet (Jia et al., 2014; Krizhevsky et al., 2012). Additionally, we tested the Fisher Vector (Perronnin et al., 2010), which was the standard hand-crafted image feature before deep learning. We extracted features from the pool5/7x7_s1 layer in GoogLeNet, fc6 layer in VGG, and fc6 layer in CaffeNet. For the Fisher Vector, following the standard implementation, we compressed SIFT descriptors (Lowe, 1999) into 64 dimensions by PCA, and used a Gaussian mixture model with 64 components. We used four spatial grids for the final feature extraction. Overall, the order of performance of features corresponds to that known in the image classification domain (Russakovsky et al., 2015). This result indicates that when more powerful image features are used, better performance can be achieved in image-mediated CLDR.

5 Conclusion

We proposed an image-mediated learning approach to realize zero-shot or few-shot CLDR. For evaluation, we created and released a new dataset consisting of Japanese, English, and image triplets, based on the widely used Pascal Sentence Dataset. We showed that state-of-the-art CNN-based image features can substantially improve zero-shot CLDR performance. Considering that image features have continued to improve rapidly since the deep learning breakthrough and the universality of images in Web documents, this approach could become even more important in the future.

Acknowledgments

This work was supported by JST CREST, JSPS KAKENHI Grant Number 26730085. We thank the three anonymous reviewers for their helpful comments.

References

Douglas J Carroll. 1968. Generalization of canonical correlation analysis to three or more sets of vari-

ables. In *Proceedings of the 76th Annual Convention of the American Psychological Association*, volume 3, pages 227–228.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *Proceedings of the British Machine Vision Conference*, pages 1–11.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the International Conference on Machine Learning*, pages 647–655.

Hao Fang, Saurabh Gupta, Forrest Iandola, K. Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A Multiview Embedding Space for Modeling Internet Images, Tags, and their Semantics. *International Journal of Computer Vision*, 106(2):210–233.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical Correlation Analysis: an Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664.

Harold Hotelling. 1936. Relations between Two Sets of Variants. *Biometrika*, 28:321–377.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe : Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678.

Jon Robers Kettenring. 1971. Canonical Analysis of Several Sets of Variables. *Biometrika*, 58(3):433–451.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Yaoyong Li and John Shawe-Taylor. 2004. Using KCCA for Japanese-English Cross-language Information Retrieval and Classification. In *Learning Methods for Text Understanding and Mining Workshop*, pages 117–133.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the Fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision*, pages 143–156.

Cyrus Rashtchian, Peter Young, Micah Hodosh, Julia Hockenmaier, and North Goodwin Ave. 2010. Collecting Image Annotations Using Amazon ' s Mechanical Turk. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A New Approach to Cross-modal Multimedia Retrieval. *Proceedings of the International Conference on Multimedia*, pages 251–260.

Jan Rupnik, Andrej Muhič, and Primo Škraba. 2012. Cross-Lingual Document Retrieval through Hub Languages. In *Neural Information Processing Systems Workshop*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*.

Christian Szegedy, Scott Reed, Pierre Sermanet, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions. *CoRR abs/1409.4842*.

Raghavendra Udupa and Mitesh M Khapra. 2010. Improving the Multilingual User Experience of Wikipedia Using Cross-Language Name Search. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 492–500.

Michel Velden and Yoshio Takane. 2012. Generalized Canonical Correlation Analysis with Missing Values. *Computational Statistics*, 27(3):551–571.

Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. 2002. Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. *Advances in Neural Information Processing Systems*, pages 1473–1480.