Latent Semantic Matching: Application to Cross-language Text Categorization without Alignment Information

Tsutomu Hirao and **Tomoharu Iwata** and **Masaaki Nagata**
NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan
{hirao.tsutomu,iwata.tomoharu,nagata.masaaki}@lab.ntt.co.jp

Abstract

Unsupervised object matching (UOM) is a promising approach to cross-language natural language processing such as bilingual lexicon acquisition, parallel corpus construction, and cross-language text categorization, because it does not require labor-intensive linguistic resources. However, UOM only finds one-to-one correspondences from data sets with the same number of instances in source and target domains, and this prevents us from applying UOM to real-world cross-language natural language processing tasks. To alleviate these limitations, we proposes *latent semantic matching*, which embeds objects in both source and target language domains into a shared latent topic space. We demonstrate the effectiveness of our method on cross-language text categorization. The results show that our method outperforms conventional unsupervised object matching methods.

1 Introduction

Unsupervised object matching is a method for finding one-to-one correspondences between objects across different domains without knowledge about the relation between the domains. Kernelized sorting (Novi et al., 2010) and canonical correlation analysis based methods (Haghighi et al., 2008; Tripathi et al., 2010) are two such examples of unsupervised object matching, which have been shown to be quite useful for cross-language natural language processing (NLP) tasks. One of the most important properties of the unsupervised object matching is that it does not require any linguistic resources which connects between the languages. This distinguishes it from other cross-language NLP methods such as machine translation based and projection based approaches (Dumais et al., 1996; Gliozzo and Strapparava, 2005; Platt et al., 2010), which we need bilingual dictionaries or parallel sentences.

When we apply unsupervised object matching methods to cross-language NLP tasks, there are two critical problems. The first is that they only find one-to-one matching. The second is they require the same size of source- and target-data. For example, the correct translation of a word is not always unique. French words '*maison*', '*appartment*' and '*domicile*' can be regarded as translation of an English word 'home'. In addition, English vocabulary size is not equal to that of French.

These discussions motivate us to introduce a shared space in which both source and target domain objects will reside. If we can obtain such a shared space, we can match objects within the space, because we can use standard distance metrics on this space. This will also enable us to use various kinds of non-strict matching. For example, k-nearest objects in the source domain will be retrieved for a query object in the target domain. In this paper, we propose a simple but effective method to find the shared space by assuming that two languages have common latent topics, which we call *latent semantic matching*. With latent semantic matching, we first find latent topics in two domains independently. Then, the topics in two domains are aligned by kernelized sorting, and objects are embedded in a shared latent topic space. Latent topic representations are successfully used in a wide range of NLP tasks, such as information retrieval and text classification, because they represent intrinsic information of documents (Deerwester et al., 1990). By matching latent topics, we can find relation between source and target domains, and additionally we can handle different numbers of objects in two domains.

We compared latent semantic matching with conventional unsupervised object matching meth-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 212–216,
Sofia, Bulgaria, August 4-9 2013.

ods on the task of cross-language text categorization, *i.e.* classifying target side unlabeled documents by label information obtained from source side documents. The results show that, with more source side documents, our method achieved the highest classification accuracy.

2 Related work

Many cross-language text processing methods have been proposed that require correspondences between source and target languages. For example, (Dumais et al., 1996) proposed cross-lingual latent semantic indexing, and (Platt et al., 2010) employed oriented principle component analysis and canonical correlation analysis (CCA). They concatenate the document pairs (source document and its translation) obtained from a document-level parallel corpus. They then apply multivariate analysis to acquire the translingual projection. There are extensions of latent Dirichlet allocation (LDA) (Blei et al., 2003) for cross-language analysis, such as multilingual topic models (Boyd-Graber and Blei, 2009), joint LDA (Jagadeesh and Daume III, 2010) and multilingual LDA (Xiaochuan et al., 2011). They require a bilingual dictionary or document-level parallel corpora.

Unsupervised object matching methods have been proposed recently (Novi et al., 2010; Haghighi et al., 2008; Yamada and Sugiyama, 2011). These methods are promising in terms of language portability because they do not require external language resources. (Novi et al., 2010) proposed kernelized sorting (KS); it finds one-to-one correspondences between objects in different domains by permuting a set to maximize the dependence between two sets. Here, the Hilbert-Schmidt independence criterion is used for measuring dependence. (Djuric et al., 2012) proposed convex kernelized sorting as an extension of KS. (Yamada and Sugiyama, 2011) proposed least-squares object matching which maximizes the squared-loss mutual information between matched pairs. (Haghighi et al., 2008) proposed another framework, matching CCA (MCCA), based on a probabilistic interpretation of CCA (Bach and Jordan, 2005). MCCA simultaneously finds latent variables that represent correspondences and latent features so that the latent features of corresponding examples exhibit the maximum correlation. However, these unsupervised object matching methods have limitations. They require that the source and target domains have the same data size, and they find one-to-one correspondences. There are critical weaknesses of these methods when we attempt to apply them to real world cross-language NLP applications.

3 Latent Semantic Matching

We propose latent semantic matching to find a shared latent space by assuming that two languages have common latent topics. Our method consists of following four steps: (1) for both source and target domains, we map the documents to a K-dimensional latent topic space independently, (2) we find the one-to-one correspondences between topics across source and target domains by unsupervised object matching, (3) we permute topics of the target side according to the correspondences, while fixing the topics of the source side, and (4) finally, we map documents in the source and target domains to a shared latent space by using permuted and fixed topics.

3.1 Topic Extraction as Dimension Reduction

Suppose that we have N documents in the source domain. $\boldsymbol{s}_n=(s_{ni})_{i=1}^{I}$ is the nth document represented as a multi-dimensional column vector in the domain, *i.e.* each document is represented as a bag-of-words vector. Here, each element of the vectors indicates the TF·IDF score of the corresponding word in the document. I is the size of the feature set, *i.e.,* the vocabulary size in the source domain. Also, we have M documents in the target domain. $\boldsymbol{t}_m=(t_{mj})_{j=1}^{J}$ is the mth document represented as a multi-dimensional vector. J is the vocabulary size in the target domain. Thus, the data set in the source domain is represented by an $I \times N$ matrix, $\boldsymbol{S}=(\boldsymbol{s}_1, \cdots, \boldsymbol{s}_N)$, the data set in the target is represented by a $J \times M$ matrix, $\boldsymbol{T}=(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_M)$.

We factorize these matrices using nonnegative matrix factorization (Lee and Seung, 2000) to find topics as follows:

$$\boldsymbol{S} \approx \boldsymbol{W}_S \boldsymbol{H}_S, \tag{1}$$

$$\boldsymbol{T} \approx \boldsymbol{W}_T \boldsymbol{H}_T. \tag{2}$$

$\boldsymbol{W}_S$ is an $I \times K$ matrix that represents a set of topics, *i.e.* each column vector denotes word weights for each topic. $\boldsymbol{H}_S$ is a $K \times N$ matrix that denotes a set of latent semantic representations of documents in the source domain, *i.e.* each row

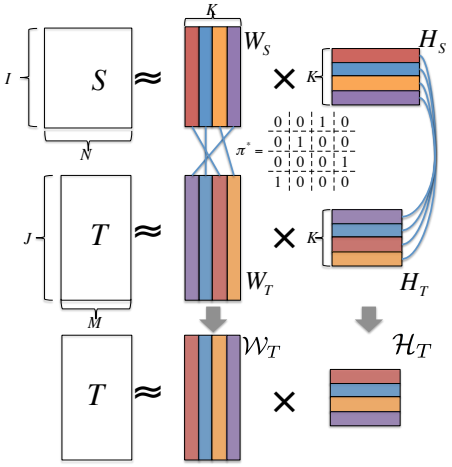

Figure 1: Topic alignments.

vector denotes an embedding of a document in the K-dimensional latent space. Similarly, $\boldsymbol{W}_T$ is an $I \times K$ matrix that represents a set of topics in the target domain, and $\boldsymbol{H}_T$ is a $K \times M$ matrix that denotes a set of latent semantic representations of target documents. K is less than I and J.

By factorizing the original matrices, we can independently map the documents in the source and target domains to the latent topic spaces whose dimensionality is K.

3.2 Finding Optimal Topic Alignments by Unsupervised Object Matching

To connect the different latent spaces, topics extracted from the source language must be aligned to one from the target language. This is reasonable because we can assume that both languages share the same latent concept.

However, we cannot quantify the similarity between the topics because we do not have any external language resources such as a dictionary. Therefore, we utilize unsupervised object matching method to find one-to-one correspondences between topics. In this paper, we employ kernelized sorting (KS) (Novi et al., 2010). KS finds the best one-to-one matching as follows:

$$\begin{aligned} \boldsymbol{\pi}^* &= \underset{\boldsymbol{\pi} \in \boldsymbol{\Pi}_K}{\arg\max}\, \mathrm{tr}(\overline{\mathcal{G}}_S \boldsymbol{\pi}^\top \overline{\mathcal{G}}_T \boldsymbol{\pi}), \\ & s.t. \quad \boldsymbol{\pi}\mathbf{1}_K = \mathbf{1}_K \text{ and } \boldsymbol{\pi}^\top \mathbf{1}_K = \mathbf{1}_K. \end{aligned} \quad (3)$$

Here, $\boldsymbol{\pi}$ is a $K \times K$ matrix that represents the one-to-one correspondence between topics, *i.e.* $\pi_{ij}=1$ indicates that the ith topic in the source language corresponds to the jth one of the target language.

	Overall Average
KS	0.252 ± 0.112
CKS	0.249 ± 0.033
LSOM	0.278 ± 0.086
LSM(300)	0.298 ± 0.077
LSM(600)	$\mathbf{0.359 \pm 0.062}$

Table 1: Average accuracy over all language pairs

$\boldsymbol{\Pi}_K$ indicates the set of all possible matrices storing one-to-one correspondences. $\mathcal{G}$ denotes the $K \times K$ kernel matrix obtained from topic proportion, $\mathcal{G}_{ij}=\mathcal{K}(\boldsymbol{W}_{i,:}^\top, \boldsymbol{W}_{:,j})$, and $\overline{\mathcal{G}}$ is the centered matrix of $\mathcal{G}$. $\mathcal{K}(,)$ is a kernel function. $\mathbf{1}_K$ is a K-dimensional column vector of all ones. $\boldsymbol{\pi}^*$ is obtained by iterative procedure.

According to $\boldsymbol{\pi}^*$, we obtain permuted matrices, $\mathcal{W}_T=\boldsymbol{W}_T\boldsymbol{\pi}^*$ and $\mathcal{H}_T=\boldsymbol{\pi}^{*\top}\boldsymbol{H}_T$, and the product of permuted matrices is the same with that of unpermuted matrices as follows:

$$\boldsymbol{T} \approx \boldsymbol{W}_T\boldsymbol{H}_T = \mathcal{W}_T\mathcal{H}_T. \quad (4)$$

Fig. 1 shows the topic alignment procedure.

Since documents from both domains are represented in a shared latent space, we can directly calculate the similarity between the nth document in the source domain and the mth document in the target domain based on $H_{T:,m}$ (mth column vector of H_T) and $\mathcal{H}_{S:,n}$ (nth column vector of $\mathcal{H}_S$).

4 Cross-language Text Categorization via Latent Semantic Matching

Cross-language text categorization is the task of exploiting labeled documents in the source language (e.g. English) to classify documents in the target language (e.g. French). Suppose we have training data set $\{\boldsymbol{s}_n, y_n\}_{n=1}^N$ in the source language domain. $y_n \in Y$ is the class label for the nth document. We can train a classifier in the K-dimensional latent space with data set $\{\boldsymbol{H}_{S:,n}^\top, y_n\}_{n=1}^N$. $H_{S:,n}^\top$ is the projected vector of $\boldsymbol{s}_n$. Also, the mth document in the target language domain $\boldsymbol{t}_m$ is projected into the latent space as $\mathcal{H}_{T:,m}^\top$. Here, the documents in both domains are projected into the same size latent space and the basis vectors of the spaces are aligned. Therefore, we can classify a document in the target domain $\boldsymbol{t}_m$ by a classifier trained with $\{H_{S:,n}^\top, y_n\}_{n=1}^N$.

	Books
English	Hack, Parent, tale, subversion, Interesting, centre, Paper, T., prejudice, Murphy
German	Lydia, Sebastian, Seelenbrecher, Patient, Fitzek, Patrick, Fiktion, Patientenakte, Realitt, Klinik
	Electronics
English	SD800, Angle, Digital, Optical, Silver, understnad, camra, 7.1MP, P3N, 10MP
German	*****, 550D, 600D, Objektiv, Canon, ablichten, Body, Werkzeug, Kamera, einliet
	Kitchen
English	Briel, Electra-Craft, Chamonix, machine, Due, crema, supervisor, technician, espresso, tamp
German	ESGE, Prierkopf, Zauberstab, Gummikupplung, Suppe/Sauce, Braun , Bolognese, prieren, Testsieger, Topf
	Music
English	Amy, Poison, Doherty, Schottin, Mid, Prince, Song, ausdrucksstark , Tempo, knocking
German	Norah, mini, 'Little, 'Rome, 'Come, Gardot, Lana, listenings , dreamlike, digipak
	Watch
English	watch, indicate, timex, HRM, month, icon, Timex, datum, troubleshooting, reasonable
German	Orient, Diver, Lnette, Leuchtpunkt, Zahlenringes, Handgelenksdurchmesser, Stoppsekunde, Uhrforum, Konsumbereiche, Schwingungen/Std

Table 2: Examples of aligned latent topics

5 Experimental Evaluation

5.1 Experimental Settings

We compared our method, latent semantic matching (LSM), with three unsupervised object matching methods: Kernelized Sorting (KS), Convex Kernelized Sorting (CKS), Least-Squares Object Matching (LSOM). We set the number of the latent topics K to 100 and employed the k-nearest neighbor method (k=10) as the classifier.

For, KS, CKS and LSOM, we find the one-to-one correspondence between documents in the source language and documents in the target language. Then, we assign class labels of the target documents according to the correspondence.

In order to build a corpus with various language pairs for evaluation, we crawled product reviews from Amazon U.S., German, France and Japan with five categories: 'Books', 'Electronics', 'Music', 'Kitchen', 'Watch'. The corpus is neither sentence level parallel nor comparable. For each category, we randomly select 60 documents as the test data (M=300) for all methods and 60 documents as the training data (N=300) for KS, CKS, LSOM and LSM(300). We also compared latent semantic matching with 120 training documents for each category (N=600), and called this method LSM(600). Note that since KS, CKS and LSOM require that the data sizes are the same for source and target domains, they cannot use training data more than test data. To avoid local optimum solutions of NMF, we executed our methods with 100 different initialization values and chose the solution that achieved the best objective function of KS.

5.2 Results and Discussion

Table 1 shows average accuracies with standard division over all language pairs. From the table, classification accuracy of all methods significantly outperformed random classifier (accuracy=0.2). The results showed the effectiveness of both unsupervised object matching and latent semantic matching. When comparing LSM(300) with KS, CKS and LSOM, LSM(300) obtained better results than these unsupervised object matching methods. The result supports the effectiveness of the latent topic matching. Moreover, LSM(600) achieved the highest accuracy. There are large differences between LSM(600) and the others. This result implies not only the effectiveness of the latent topic matching but also increasing the number of source side documents (labeled training data) contributes to improving classification accuracy. This is natural in terms of supervised learning but only our method can deal with source side documents that are larger in number.

Table 2 shows examples of latent topics in English and German extracted and aligned by LSM(600). We can see that some author names, words related to camera, and cooking equipment appear in 'Books', 'Electronics' and 'Kitchen' topics, respectively. Similarity, there are some artists' names in 'Music' and watch brands in 'Watch'.

6 Conclusion

As an extension of unsupervised object matching, this paper proposed latent semantic matching that considers the shared latent space between two language domains. To generate such a space, topics of the target space are permuted by exploiting unsupervised object matching. We can measure distances between objects by standard metrics, which enable us retrieving k-nearest objects in the source domain for a query object in the target domain. This is a significant advantage over conventional unsupervised object matching methods. We used Amazon review corpus to demonstrate the effectiveness of our method on cross-language text categorization. The results showed that our method outperformed conventional object matching methods with the same number of training samples. Moreover, our method achieved even higher performance by utilizing more documents in the source domain.

Acknowledgements

The authors would like to thank Nemanja Djuric for providing code for Convex Kernelized Sorting and the three anonymous reviewers for thoughtful suggestions.

References

Francis Bach and Michael Jordan. 2005. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3(Jan.):993–1022.

Jordan Boyd-Graber and David Blei. 2009. Multilingual topic model for unaligned text. In *Proc. of the 25th UAI*, pages 75–82.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Nemanja Djuric, Mihajlo Grbovic, and Slobodan Vucetic. 2012. Convex kernelized sorting. In *Proc. of the 26th AAAI*, pages 893–899.

Susan Dumais, Lanauer Thomas, and Michael Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Proc. of the Workshop on Cross-Linguistic Information Retieval in SIGIR*, pages 16–23.

Alfio Gliozzo and Carlo Strapparava. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 9–16.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proc. of ACL-08: HLT*, pages 771–779.

Jagarlamudi Jagadeesh and Hal Daume III. 2010. Extracting multilingual topics from unaligned corpora. In *Proc of the 32nd ECIR*, pages 444–456.

Daniel Lee and Sebastian Seung. 2000. Algorithm for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562.

Quadrianto Novi, Smola Alexander, Song Le, and Tuytelaars Tinne. 2010. Kernelized sorting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(10):1809–1821.

Jhon Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representation from discriminative projections. In *Proc. of the 2010 Conference on EMNLP*, pages 251–261.

Abhishek Tripathi, Arto Klami, and Sami Virpioja. 2010. Bilingual sentence matching using kernel CCA. In *Proc. of the 2010 IEEE International Workshop on MLSP*, pages 130–135.

Ni Xiaochuan, Sun Lian-Tao, Hu Jian, and Chen Zheng. 2011. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proc. of the 4th WSDM*, pages 375–384.

Makoto Yamada and Masashi Sugiyama. 2011. Cross-domain object matching with model selection. In *Proc. of the 14th AISTATS*, pages 807–815.