Word Sense Disambiguation Based on Distance Metric Learning from Training Documents

Minoru Sasaki
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
Email: msasaki@mx.ibaraki.ac.jp

Hiroyuki Shinnou
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan
Email: shinnou@mx.ibaraki.ac.jp

***Abstract*—Word sense disambiguation task reduces to a classification problem based on supervised learning. However, even though Support Vector Machine (SVM) gives the distance from the data point to the separating hyperplane, SVM is difficult to measure the distance between labeled and unlabeled data points. In this paper, we propose a novel word sense disambiguation method based on a distance metric learning to find the most similar sentence. To evaluate the efficiency of the method of word sense disambiguation using the distance metric learning such as Neighborhood Component Analysis and Large Margin Nearest Neighbor, we make some experiments to compare with the result of the SVM classification. The results of the experiments show this method is effective for word sense disambiguation in comparison with SVM and one nearest neighbor. Moreover, the proposed method is effective for analyzing the relation between the input sentence and all senses of the target word if the target word has more than two senses.**

***Keywords*-word sense disambiguation, distance metric learning, similar example retrieval,**

I. INTRODUCTION

In natural language processing, acquisition of sense examples from examples that contain a given target word enables to construct an extensive data set of tagged examples to demonstrate a wide range of semantic analysis. For example, using the obtained data set, we can create a classifier that identifies its word sense by analyzing co-occurrence statistics of a target word. Also, we can construct a wide-coverage case frame dictionary automatically and construct thesaurus for each meaning of a polysemous word. To construct large-sized training data, language dictionary and thesaurus, it is increasingly important to further improve to select the most appropriate meaning of the ambiguous word.

If we have training data, word sense disambiguation (WSD) task reduces to a classification problem based on supervised learning. This approach is generally applicable to construct a classifier from a set of manually sense-tagged training data. Then, this classifier is used to identify the appropriate sense for new examples. A typical method for this approach is the classical bag-of-words (BOW) approach, where each document is represented as a feature vector counting the number of occurrences of different words as features. By using such features, we can easily adapt many existing supervised learning methods such as Support Vector Machine (SVM) [2] for the WSD task. However, even though SVM gives the distance from the data point to the separating hyperplane, SVM is difficult to measure the distance between labeled and unlabeled data points.

In this paper, to solve this problem, we propose a novel word sense disambiguation method based on a distance metric learning to find the most similar sentence. In general, when words are used with the same sense, they have similar context and co-occurrence features. To obtain feature vectors that are useful to discriminate among word sense efficiently, examples sharing the same sense are close to each other in the training data while examples from different senses are separated by a large distance by using the distance metric learning method.

In this method, we apply two distances metric learning approach. One approach is to find an optimal projection which maximizes the margin between data points from different classes such as Local Fisher Discriminant Analysis (LFDA)[7][9], Semi-Supervised Local Fisher Discriminant Analysis (SELF) [8]. Another alternative is to learn a distance metric such that data points in the same class are close to each other and those in different classes are separated by a large margin such as Neighborhood Component Analysis (NCA) and Large Margin Nearest Neighbor (LMNN). We present the results of experiments using these two approaches of the proposed method to evaluate the efficiency of word sense disambiguation.

The rest of this paper is organized as follows. Section 2 is devoted to the introduction of the related work in the literature. Section 3 describes distance metric learning method. Section 4 illustrates the proposed system. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

II. RELATED WORKS

This paper proposes a method based on a distance metric learning for WSD. In this section, some previous research

ISBN: 978-1-61208-240-0

using supervised approaches will be compared with ourproposed method.

k-nearest neighbor algorithm (k-NN) is one of the most well-known instance-based learning methods[1]. The k-NN classifies test data based on closest training examples in the feature space. One of the characteristics of this method is to calculate the similarity measure (e.g. cosine similarity) among instances. Therefore, this method can calculate a similarity measure between the new context and the training context, but do not consider the discriminative relations among the training data.

Support Vector Machines (SVM) has been shown to be the most successful and state-of-the-art approach for WSD[4][5]. This method learns a linear hyperplane that separates positive examples from negative examples from the training set. A test example is classified depending on the side of the hyperplane. Therefore, SVM have been successfully applied to a number of WSD problems. However, even though SVM gives the distance from the data point to the separating hyperplane, SVM is difficult to measure the distance between labeled and unlabeled examples. If the target word has more than two senses, This approach does not work so well.

III. Distance Metric Learning

Distance metric learning is to find a new distance measure for the input space of training data, while the pair of similar/dissimilar points preserves the distance relation among the training data pairs. In the Distance metric learning, there are two types of leaning approaches: dimensionalit y reduction and neighborhood optimization. In this section, we briefly explain two distance metric learning approaches.

A. Metric Learning with Dimensionality Reduction

This approach employs a linear transformation which assigns large weights to relevant dimensions and low weights to irrelevant dimensions. This is commonly used for data analysis such as noise removal, visualization and text mining and so on. The typical methods of its approach are 、 Local Fisher Discriminant Analysis (LFDA)[7][9] and Semi-Supervised Local Fisher Discriminant Analysis (SELF) [8]. LFDA finds an embedding transformation such that the between-class covariance is maximized and the within-class covariance is minimized, as shown in Figure 1.

This approach is efficient for representation of the relationship between data. However, problem arises when we apply this approach to predict new data. This method provides rotation of coordinate axes, not provide data points re-mapped to the original space, so that SVM generates a rotation of the hyperplane which is constructed in the original space. Therefore, there is little change in accuracy of performance compared to using the original feature space.

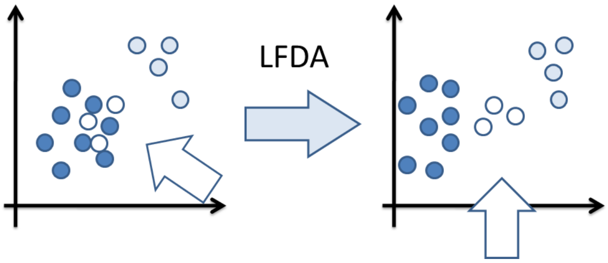

Figure 1. Local Fisher Discriminant Analysis

B. Metric Learning with Neighborhood Optimization

Alternative approach to distance metric learning is the method to learn a distance metric such that data points in the same class are close to each other and those in different classes are separated by a large margin. The two methods that implement this approach were developed, Neighborhood Component Analysis (NCA) [3] and Large Margin Nearest Neighbor (LMNN) [10].

1) NCA: NCA is a method for finding a linear transformation of training data such that the Mahalanobis distance between pairwise points is optimized in the transformed space. Given two data points x_i and x_j, the Mahalanobis distance between x_i and x_j is calculated by

$$d(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j)^T(\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

where $\mathbf{M} = \mathbf{A}^T\mathbf{A}$ is the distance metric that needs to be learned from the side information.

In this method, p_{ij} represents the probability of classifying the data point x_j to the data point x_i as neighbor as follows:

$$p_{ij} = \frac{\exp(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|^2)}{\sum_{k\neq i}\exp(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_k\|^2)} \tag{2}$$

. Then, the probability p_i is defined as the sum of the probability p_{ij} of classifying the data points x_j into the class c_i.

$$p_i = \sum_{j\in C_i} p_{ij}, \quad (C_i = \{j|c_i = c_j\}) \tag{3}$$

The optimization function $f(\mathbf{A})$ is defined as the sum of the probabilities of classifying each data point correctly. We maximize this objective function with respect to the linear transformation $f(\mathbf{A})$.

$$p_i = \sum_{j\in C_i} p_{ij}, \quad (C_i = \{j|c_i = c_j\}) \tag{4}$$

However, this objective function $f(\mathbf{A})$ is not convex, so there is a possibility of getting stuck in local minima.

2) LMNN: LMNN is a method for learning a distance metric such that data points in the same class are close to each other and those in different classes are separated by a large margin, as shown in Figure 2.

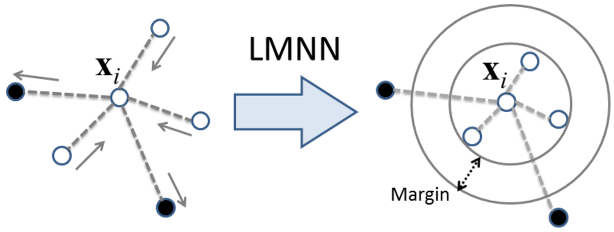

Figure 2. Large Margin Nearest Neighbor

In this method, the k neighbors of data $\mathbf{x}_i$ are the k nearest neighbors that share the same label y_i, and the matrix η is defined as $\eta_{ij} = 1$ if the input $\mathbf{x}_j$ is a target neighbor of input $\mathbf{x}_i$ and 0 otherwise. From these definitions, the cost function of LMNN is given by,

$$\varepsilon(\mathbf{A}) = \sum_{ij} \eta_{ij} ||\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j||^2 + c \sum_{ijl} \eta_{ij}(1-\eta_{il}) \\ [1 + ||\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j||^2 - ||\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_l||^2]_+, \tag{5}$$

where $[\cdot]_+$ denotes the positive part, for example, $[a]_+ = a$ if $a > 0$, and 0 otherwise, and c is some positive constant.

IV. WSD Method Based on Distance Metric Learning

In this section, we will describe the details of the WSD classifier using distance metric learning mentioned in the previous section.

A. Feature Extraction

At the first step, our method extracts a set of features; nouns and verbs that have co-occurred with the target word by morphological analysis from each sentence in the training and test data. Then, each feature set is represented as a vector by counting co-occurrence frequencies of the words. The set of word co-occurrence vectors forms a matrix for each target word.

B. Classification Model Construction

For the obtained this matrix, classification model is constructed by using distance metric learning method. The experiments in this paper use two learning methods such as NCA and LMNN to transform the data points. For the transformed data set using the NCA, we find optimal dividing hyperplane that will correctly classify the data points of the training data by using SVM. For the transformed data set using the LMNN, we apply one-nearest neighbor method in order to classify a new data point.

When the classification model is obtained by training data, we predict one sense for each test example using this model. When a new sentence including the target word is given, the sense of the target word is classified to the most plausible sense based on the obtained classification model. To employ the SVM for distinguishing more than two senses, we use one-versus-rest binary classification approach for each sense. To employ the LMNN, we use the one-nearest neighbor (1-NN) classification rule to classify a test data set. The 1-NN method classifies a new sentence into the class of the nearest of the training data. Therefore, even if the target word has many senses, there is no need to repeat the classification process.

V. Experiments

To evaluate the efficiency of the method of word sense disambiguation using the distance metric learning such as NCA and LMNN, we make some experiments to compare with the result of the SVM classification. In this section, we describe an outline of the experiments.

A. Data

We used the Semeval-2010 Japanese WSD task data set, which includes 50 target words comprising 22 nouns, 23 verbs, and 5 adjectives [6]. In this data set, there are 50 training and 50 test instances for each target word.

B. Evaluation Method

To evaluate the results of the methods using NCA and LMNN for the test data, we compare their performances with the results of simple SVM and 1-NN training. We obtain the total number of correct prediction of each target word using three methods: SVM, 1-NN, NCA+SVM and LMNN+1-NN. Moreover, we also obtain precision value of each method over all the examples to analyze the average performance of systems.

VI. Experimental Results

A. Classification Performance

Table I and Table II show the results of the experiments of applying four methods. The proposed method using distance metric learning shows higher precision than the traditional one-nearest neighbor method. The distance metric learning provides an effective semantic relation between word senses so that this approach is effective for word sense disambiguation.

When NCA is applied to distance metric learning, the accuracy is increased on 9 words, decreased on ten words and the same on 31 words in comparison with SVM. Totally, NCA is not improved compared with SVM, because objective function of NCA tends to converge into a local optimum. To use the NCA for word sense disambiguation, further improvements are required for the prospective practical use. Examples of improvements include the use of a large data set, the use of other feature extraction methods or finding the optimal number of dimensions of projection etc.

When we use LFDA, we can not solve the generalized eigenvalue problem, since the co-occurrence matrix is very sparse. Hence, we apply SELF to their experiments instead of LFDA. The accuracy is increased on 1 word and the same

Table I
EXPERIMENTAL RESULTS(1/2)

word	1-NN	SVM	SELF+ SVM	NCA+ SVM	LMNN+ 1NN
現場 (genba)	30	39	39	37	29
場所 (basyo)	48	48	48	48	48
取る (toru)	13	13	13	13	14
乗る (noru)	27	25	25	20	27
会う (au)	28	33	33	33	33
前 (mae)	24	31	31	29	27
子供 (kodomo)	26	18	18	21	26
関係 (kankei)	39	39	39	39	39
教える (oshieru)	15	9	9	9	13
勧める (susumeru)	20	16	16	16	27
社会 syakai)	40	43	43	43	42
する (suru)	18	21	21	23	20
電話 (denwa)	31	28	28	35	33
やる (yaru)	46	47	47	47	47
意味 (imi)	26	27	27	23	26
あげる (ageru)	15	18	18	18	17
出す (dasu)	18	14	14	17	26
生きる (ikiru)	47	47	47	47	47
経済 (keizai)	47	49	49	49	49
良い (yoi)	24	12	12	15	23
他 (hoka)	50	50	50	50	50
開く (hiraku)	45	45	45	45	45
もの (mono)	44	44	44	44	44
強い (tuyoi)	43	46	46	46	45
求める (motomeru)	39	38	38	38	39

Table II
EXPERIMENTAL RESULTS(2/2)

word	1-NN	SVM	SELF+ SVM	NCA+ SVM	LMNN+ 1NN
技術 (gijutu)	39	42	42	42	41
与える (ataeru)	21	29	29	28	25
市場 (shijou)	14	35	35	34	20
立つ (tatu)	18	26	26	22	16
手 (te)	41	39	39	39	40
考える (kangaeru)	49	49	49	49	49
見える (mieru)	19	26	26	23	23
一 (ichi)	45	46	46	46	46
入れる (ireru)	28	36	36	36	34
場合 (baai)	42	43	43	43	45
早い (hayai)	31	26	26	27	28
出る (deru)	22	30	30	30	28
入る (hairu)	20	25	25	26	34
はじめ (hajime)	38	30	30	33	44
情報 (jouhou)	39	40	42	37	32
大きい (ookii)	45	47	47	47	47
見る (miru)	39	40	40	40	40
可能 (kanou)	23	28	28	28	30
持つ (motu)	30	34	34	34	29
時間 (jikan)	43	44	44	42	44
文化 (bunka)	46	49	49	49	49
始める (hajimeru)	39	39	39	40	39
認める (mitomeru)	39	35	35	35	39
相手 (aite)	41	41	41	41	40
高い (takai)	26	43	43	43	43
precision	0.6544	0.6888	0.6896	0.6876	0.6964

on 49 words in comparison with SVM so that the experimental results of SVM and SELF are almost the same. LFDA obtains the optimal subspace that maximizes between-class and minimizes the within-class variance. However, this subspace is obtained by rotating and scaling the original coordinate space. Therefore, SVM produces the hyperplane equal to the transformation of it in the original space into the subspace obtained by LFDA.

When LMNN is applied to distance metric learning, precision of LMNN is slightly improved from 98.9% to 69.6% in comparison with SVM. It is possible to build a classification model that can perform better than NCA and SELF. Unlike NCA, we can obtain a global optimum solution by using LMNN so that we consider that LMNN is effective for word sense disambiguation.

B. *Efficiency of Distance Metric Learning*

In traditional SVM classification, an additional process is required for extensive analysis on the relation between the new data and the training data. However, in the proposed method, we can perform such analysis easily. In contrast to SVM, we can retrieve the most similar sentence using one nearest neighbor for the input sentence.

To employ the SVM for classifying more than two senses, we solve multi-class classification problems by considering the standard one versus rest strategy. If the target word has more than two senses, it is difficult to compare the distance between the test data and its nearest neighbor. The LMNN method employs one nearest neighbor rule and can calculate the distance to its nearest neighbor for each sense. Therefore, the proposed method is effective for analyzing the relation between the input sentence and all senses of the target word. Also, this method is effective for identifying uncommon word senses of target words.

VII. CONCLUSION

In this paper, we propose a novel word sense disambiguation method based on a distance metric learning to find the most similar sentence. To evaluate the efficiency of the method of word sense disambiguation using the distance metric learning such as NCA and LMNN, we make some experiments to compare with the result of the SVM classification. The results of the experiments show this method is effective for word sense disambiguation in comparison with SVM and one nearest neighbor. Moreover, the proposed method is effective for analyzing the relation between the input sentence and all senses of the target word if the target word has more than two senses.

Further work would be required to consider more effective re-mapping method of the training data to improve the performance of word sense disambiguation.

REFERENCES

[1] E. Agirre, O. Lopez, and D. Martínez, “Exploring feature spaces with svd and unlabeled data for word sense disambiguation,” in *In Proceedings of the Conference on Recent*

Advances on Natural Language Processing (RANLP' 05), Borovets, Bulgary, 2005.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood Component Analysis," in *Proceedings of Advances of Neural Information Processing*, 2004.

[4] R. Izquierdo, A. Suárez, and G. Rigau, "An empirical study on class-based word sense disambiguation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '09, 2009, pp. 389–397.

[5] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 10:1–10:69, Feb. 2009.

[6] M. Okumura, K. Shirai, K. Komiya, and H. Yokono, "Semeval-2010 task: Japanese wsd," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, ser. SemEval '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 69–74.

[7] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 905–912.

[8] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local fisher discriminant analysis for dimensionality reduction," *Machine Learning*, vol. 78, pp. 35–61, January 2010.

[9] M. Sugiyama and S. Roweis, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[10] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, June 2009.