Semi-supervised learning for all-words WSD using self-learning and fine-tuning

Rui Cao **Jing Bai** **Wen Ma** **Hiroyuki Shinnou**
Ibaraki University, Department of Computer and Information Sciences
4-12-1 Nakanarusawa, Hitachi, Ibaraki JAPAN 316-8511
{18nd305g, 19nd301r, 19nd302, hiroyuki.shinnou.0828}
@vc.ibaraki.ac.jp

Abstract

In this paper, we propose a semi-supervised learning method using self-learning and fine-tuning for all-words word sense disambiguation (WSD). The all-words WSD can be regarded as a sequence labeling problem, so we use a bidirectional Long Short-term Memory (LSTM) to solve it. Furthermore, we propose the semi-supervised learning method to improve that LSTM model, where self-learning is essentially used. In general, self-learning is the method for a classification problem, not for a sequence labeling problem. To apply self-learning to an all-words WSD, the LSTM model is trained by not accumulating the loss from the low probability label. We also construct the model with additional labeled data and then fine-tune by using the original labeled data. As result, the precision has been improved from the precision of the model learned from only initial labeled data.

1 Introduction

In this paper, we propose a semi-supervised learning method using self-learning for all-words word sense disambiguation (WSD). WSD is a task to identify the sense of a polysemy word in a sentence, and hence is essential in semantic analysis. However, its use in an actual system is difficult because the general WSD is developed for limited target words only. Thus, an all-words WSD that provides senses to all polysemy words in a given sentence should be developed.

Normally, WSD can be solved through supervised learning. Thus, labeled training data, that are example sentences with sense tags, are required for each word of WSD. In an all-words WSD, a large number of words with sense tags are necessary because the target word is unlimited.

Thus, unsupervised learning should also be considered (Tanigaki et al., 2013; Komiya et al., 2015; Suzuki et al., 2018). However, a problem regarding accuracy exists in this case. Under such situation, the corpus with sense tags has been gradually prepared. Recently, the all-words WSD in a supervised learning framework has been attempted to address (Shinnou et al., 2017b; Shinnou et al., 2018). However, the currently available corpus with sense tags is limited and we cannot obtain a sufficient accuracy. Therefore, we attempt to develop an all-words WSD with high accuracy through semi-supervised learning.

Semi-supervised learning is a method used in training classifiers from a small amount of labeled data and large amount of unlabeled data. In the case of all-words WSD, the unlabeled data means a plain corpus. Because obtaining a large amount of plain corpus is easy, semi-supervised learning is promising approach for all-words WSD. Therefore, we propose the semi-supervised learning method to improve that LSTM model, where self-learning is essentially used. In general, self-learning is the method for a classification problem, not for a sequence labeling problem. To apply self-learning to the all-words WSD, the LSTM model is trained by not accumulating the loss from the low probability label. We construct the model with additional labeled data and then fine-tune by using the original labeled data. As result, the precision has been im-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 356-361, Hakodate, Japan, September 13-15, 2019

proved from the precision of the model learned from only initial labeled data.

2 Related Work

Many studies on semi-supervised learning for classifiers are already available. Co-training (Blum and Mitchell, 1998) and expectation-maximization (EM) (Nigam et al., 2000) algorithm are the popular and conventional methods. Co-training is a method utilized to improve classifier reciprocal by using two independent views. In the EM algorithm, a generation model $p(x;\theta)$ has been set and considered the label as potential variable to construct $p(z|x)$.

Based on this idea, the semi-supervised learning can be divided into two categories. The first one is employing a classifier trained by the original labeled data and then fine-tuning the classifier by data with a probability label. Self-learning (Abney, 2007) and label propagation (Zhu and Ghahramani, 2002) also belong to this category.

The second one is mapping data to space.[1] Initially, mapping unlabeled data into space which can divide them well, then mapping labeled data to that space. Finally, the process identifies and constructs classifiers in that space. Generally, if the data can be mapped into a low-dimensional space,a small amount of labeled data is sufficient to estimate the boundaries between classes. hence, the semi-supervised learning can be approved. The multi-body theory (Rifai et al., 2011) and method using generation model (Cozman et al., 2003) belong to this category. Additionally, the semi-supervised learning method using deep generation model has a similar framework with the semi-supervised learning using the generation model. Thus, we consider the method of mapping the unlabeled data into space that can accurately divide them to be used by the network. (Kingma et al., 2014; Rasmus et al., 2015; Salimans et al., 2016)

The pre-trained method is a representative of the semi-supervised learning for a sequence labeling model (Peters et al., 2017; Qi et al., 2009). To training the identify vector as input, which can be recognition by a recognizer from the unlabeled data, and added it to the training and test data. The recent pre-training method used for a network-based language model, referred to as ELMo (Peters et al., 2018), also belongs to this type. BERT (Devlin et al., 2018) also belongs to the same framework which was developed from ELMo.

For the all-words WSD, some unsupervised learning using the topic model has been proposed (Boyd-Graber et al., 2007; Komiya et al., 2015). This should be easily extended to semi-supervised learning because a generative model has been established.

3 All-words WSD Based on Bidirectional LSTM

The all-words WSD can be regarded as a sequence labeling problem that provides labels (sense) to each word in the input word sequence. An LSTM is used when the sequence labeling problem handles a neural network and corresponds to the time series by learning from the hidden layer of time t and the state of input from $t-1$. It is also a model that addresses the time series data, Natural language processing can treat word sequence from words and sentences that are regarded as the time series data. Therefore, the word after time t which be paying attention is available, and then the data can be also analyzed from the reverse direction. The model in (Figure 1) is using forward direction and reverse direction LSTM while obtaining the output for time t. Hence, the model is referred to as bidirectional LSTM.

4 Bidirectional LSTM with Self-Learning

Self-leaning utilizes the current classifier to provide a label with the probability for the unlabeled data and considers the labels with high probabilities as correct labels. By adding the data to the labeled data (training data), the accuracy of the classifier is gradually increased. In self-learning of the sequence labeling model, the sequence labeling model receives unlabeled word strings as inputs and provides a label with probability for each word. Thus, the labels with high and low probabilities are mixed and the word sequence cannot be simply added to the training data. Therefore, self-learning for a sequence labeling model has two problems: (1) enhancing the training data and (2) using increased training data.

[1] generally contains a lower dimensional space than the original data.

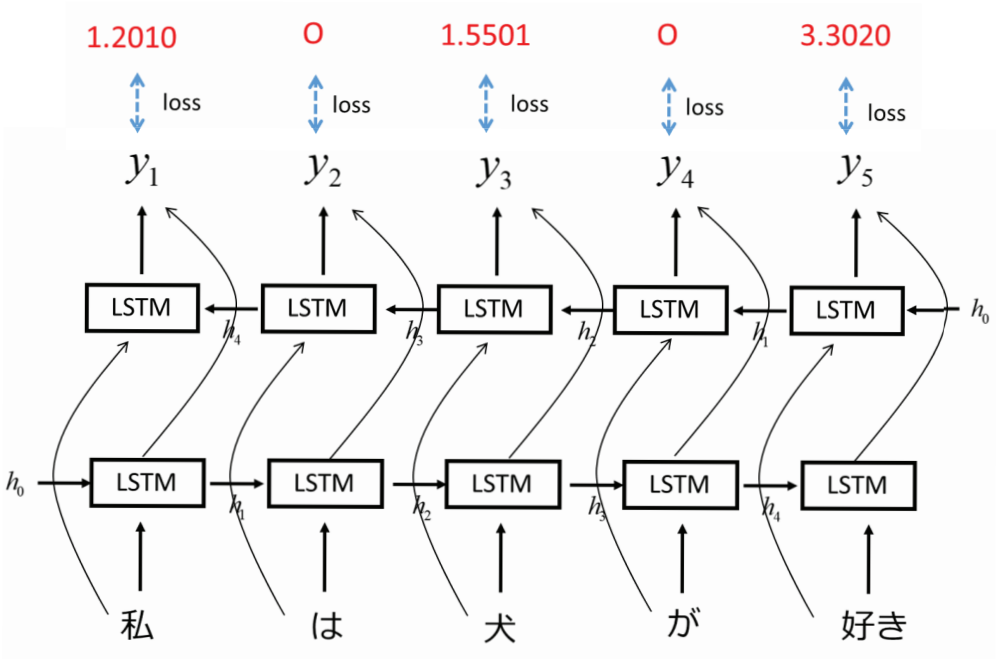

Figure 1: Learning of a bidirectional LSTM

4.1 Avoiding learning from low probability labels

For the first problem mentioned in the previous section, we do not learn from a label with low probability (confidence degree). Thus, the sequence labeling model provides labels with probabilities for each unlabeled word and then adds the word list to the training data regardless of the probability. Performing this process using LSTM is easy. For each word in LSTM, $loss_i$ is obtained from the difference between the output value and label of the word w_i, thereby accumulating the loss. When the processing is completed up to the end of the sentence, the network parameters are updated based on the accumulated $\sum_i loss_i$. If a label with low confidence degree exists. then $loss_i = 0$ is acceptable.

4.2 Using supplemental labeled data

For the second problem described in the previous section, the following three approaches are considered. In this case, the training data with label are assumed to be D. and the labeled data with probability obtained through self-learning are assumed to be A.

In this study, we attempt the following three approaches and then determine the most effective approach.

(a) Using $D \cup A$ in training the bidirectional LSTM model

(b) Using D in training the bidirectional LSTM model and A to fine-tune the model

(c) Using A in training the bidirectional LSTM model and D to fine-tune the model

5 Experiment

In this study, the sense ID in the Word List by Semantic Principles (WLSP) provided by National Institute for Japanese Language and Linguistics is regarded as sense. the Japanese sense dataset, Balanced Corpus of Contemporary Written Japanese (BCCWJ) tagged with WLSP, has been released from National Institute for Japanese Language and Linguistics (NINJAL) (Kato et al., 2017). We utilize it as a sense-tagged corpus for Japanese all-words WSD. Approximately 10% of this data is used as test data T, whereas the rest are labeled training data D. Regarding the number of sentences, D has 12,482 sentences and T has 1,498 sentences. Moreover, unlabeled data U are used in self-learning with regard

to the label. We used 100,000 sentences that are randomly extracted from the Mainichi Shimbun from 1993 to 1999.

Two layers were used as a bidirectional LSTM model. To convert the words into distributed representations we used the nwjec2vec (Shinnou et al., 2017a), which is exiting Japanese distributed expression data without learning.

Then, we utilized D in training the bidirectional LSTM model and evaluated it using T, where T was divided into 36,263 words by using the system. Considering that division of 2212 words was different from the correct answer data, the remaining 34,051 words (sense) were used as the evaluation subject. Meanwhile, 18,522 words are polysemy. The correct answer rate of these 18,522 words was determined as the correct answer rate of the all-words WSD. Figure 2 show the results. Moreover, the abscissa represents the number of epochs during the learning of the bidirectional LSTM, whereas the ordinate represents the correct answer data as described previously. The correct answer rate of the model was obtained after 18 epoch with the best value of 0.799. Because the system in (Shinnou et al., 2018) was used, the correct answer rate of the model constructed after 20 epochs where the value of 0.796 the base correct answer rate is 0.796.

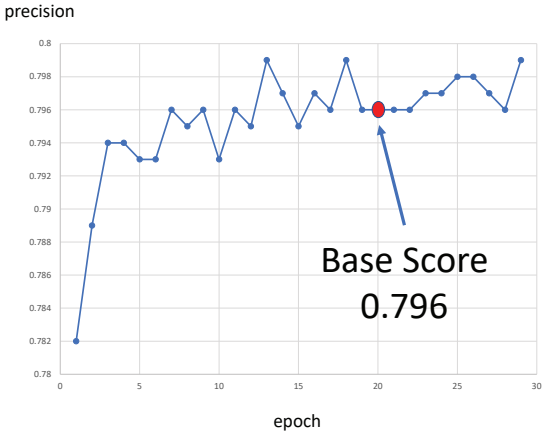

Figure 2: Using only D in training the model

Then, the model constructed after 20 epochs to the given U label with probability was used the label whose probability is less than 0.8 was replaced with the label of -1 to construct a supplemental version of the labeled data A.

(a) Using $D \cup A$ in training the model

We used $D \cup A$ as the new data to train the bidirectional LSTM model and then employed T to evaluate it. Figure 3 show the training results. The correct answer rate in this method was increased to 0.798.

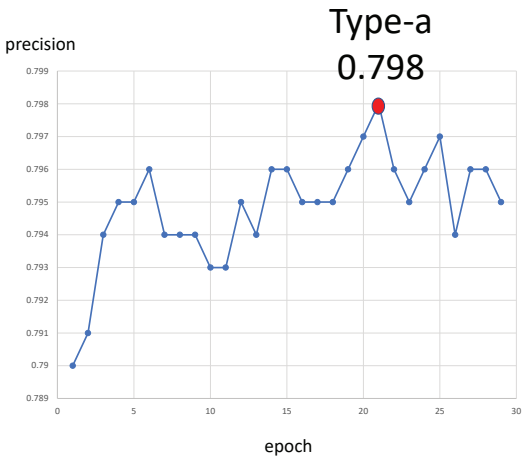

Figure 3: Using $D \cup A$ in training the model

(b) Fine-tuning ($D \rightarrow A$)

We first used D to train the bidirectional LSTM model, then A to fine-tune it, and finally T to evaluate it.Figure 4 shows the training results. In this case, the correct rate was reduced to 0.794.

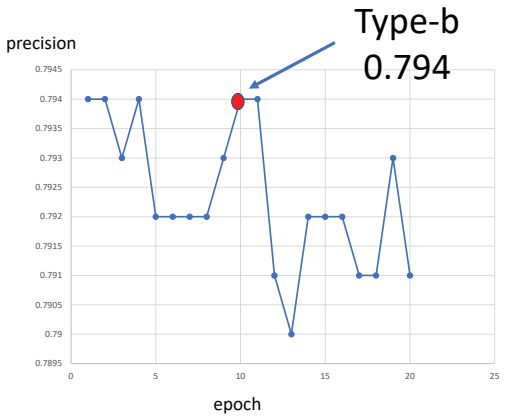

Figure 4: Fine-tuning ($D \rightarrow A$)

(c) Fine-tuning ($A \rightarrow D$)

We employed A to train the bidirectional LSTM model. D to fine-tune it. and T to evaluate it. Figure 5 shows the training result. In this case, the correct rate was increased to 0.799.

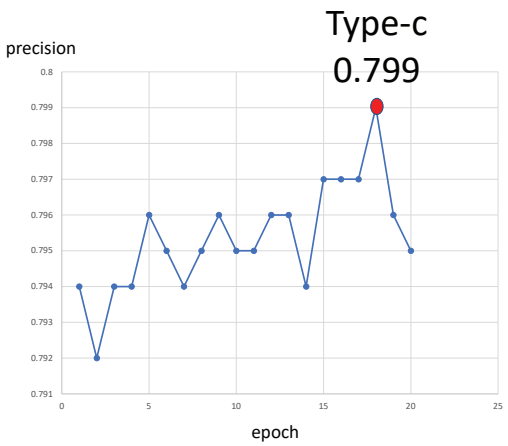

Figure 5: Fine-tuning ($A \rightarrow D$)

6 Discussion

About how to use the enhanced data, In (c) approach which creates the model based on enhanced and fine-tuning it with original labeled data. As shown in Figure 5, the correct answer rate is increased gradually, which is higher than of the base sequence labeling model. Therefore, semi-supervised learning method through self-learning can be considered to be a promising method.

However, the correct answer rate has a minimal improvement. Thus, self-learning was not effective in this experiment. Particularly in the self-learning of the discriminator, because information that can acquire new knowledge in the enhanced training data does not exist, using the semi-supervised learning is assumed to be ineffective. In the case of sequence labeling problem, we anticipated that the outcome would be good for the diversity label combination. However, this experiment did not work well.

The effect may be caused by modifying the amount of data (100,000 sentences in this experiment) or the parameter of the threshold (0.8 in this experiment) with the pseudo-label, which is regarded as the appropriate label. Therefore, we will examine these appropriate values in the future.

In addition, adjusting the amount of loss for every word in the learning process for the LSTM model may be effective. In this experiment, we set the weights to 0 when the probability based on the confidence degree is less than 0.8, and the others were set to 1. It is considered if the set weights as probability based on the confidence degree will get more appropriate for processing self-learning processing. The question of this point also will be investigated as the future problem.

7 Conclusion

In this paper, we proposed a semi-supervised learning method using self-learning for all-words word WSD. The all-words WSD is regarded as a sequence labeling problem, so we used a bidirectional LSTM to solve it. To improve that LSTM model, we attempts semi-supervised learning for it, where self-learning is essentially used. In general self-learning is for a classification problem, not for a sequence labeling problem. To apply self-learning to our problem, the LSTM model is trained by not accumulating the loss from the low probability label. We also proposed a method to train the model with additional labeled data and then to fine-tune by using the original labeled data. As result, the precision has been improved from the precision of the model learned from only initial labeled data. This improvement is just small. Hence, our proposed method is a little effective. In the future, we will try the loss from the probability based on the confidence degree.

References

Steven Abney. 2007. *Semisupervised learning for computational linguistics*. CRC Press.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A Topic Model for Word Sense Disambiguation. In *EMNLP-CoNLL-2017*.

Fabio G Cozman, Ira Cohen, and Marcelo C Cirelo. 2003. Semi-supervised learning of mixture models. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 99–106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. 2017. Annotation of 'word list by semantic principles' information on 'balanced corpus of contemporary written japanese'. In *Processing of NLP 2017*, pages 306–309 (In Japanese).

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani. 2015. Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation. In *PACLIC-29*, pages 35–43.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3):103–134.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Yanjun Qi, Pavel Kuksa, Ronan Collobert, Kunihiko Sadamasa, Koray Kavukcuoglu, and Jason Weston. 2009. Semi-Supervised Sequence Labeling with Self-Learned Features. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 428–437. IEEE.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554.

Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. 2011. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pages 2294–2302.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

Hiroyuki Shinnou, Masayuki Asahara, Kanako Komiya, and Minoru Sasaki. 2017a. nwjc2vec: Word Embedding Data Constructed from NINJAL Web Japanese Corpus (in Japanese). *Natural Language Processing*, 24(5):705–720.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. 2017b. Japanese all-words WSD system using the Kyoto Text Analysis ToolKit. In *PACLIC-31*, pages 392–399.

Hiroyuki Shinnou, Rui Suzuki, and Kanako Komiya. 2018. All-words WSD assigning Bunruigoi ID by using Bidirectional LSTM (in Japanese). In *NINJAL Language Resources Workshop, P2-04-E*.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2018. All-words Word Sense Disambiguation Using Concept Embeddings. In *LREC-2018*.

Koichi Tanigaki, Mitsuteru Shiba, Tatsuji Munaka, and Yoshinori Sagisaka. 2013. Density Maximization in Context-Sense Metric Space for All-words WSD. In *ACL-51*, volume 1, pages 884–893.

Xiaojin Zhu and Zoubin Ghahramani. 2002. *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107, Carnegie Mellon University.