Intra-Sentential Subject Zero Anaphora Resolution using Multi-Column Convolutional Neural Network

Ryu Iida **Kentaro Torisawa** **Jong-Hoon Oh**
Canasai Kruengkrai **Julien Kloetzer**
National Institute of Information and Communications Technology
Kyoto 619-0289, Japan
{ryu.iida,torisawa,rovellia,canasai,julien}@nict.go.jp

Abstract

This paper proposes a method for intra-sentential subject zero anaphora resolution in Japanese. Our proposed method utilizes a Multi-column Convolutional Neural Network (MCNN) for predicting zero anaphoric relations. Motivated by Centering Theory and other previous works, we exploit as clues both the surface word sequence and the dependency tree of a target sentence in our MCNN. Even though the F-score of our method was lower than that of the state-of-the-art method, which achieved relatively high recall and low precision, our method achieved much higher precision (>0.8) in a wide range of recall levels. We believe such high precision is crucial for real-world NLP applications and thus our method is preferable to the state-of-the-art method.

1 Introduction

In such pro-drop languages as Japanese, Chinese and Italian, pronouns are frequently omitted in text. For example, the subject of *uketa* (suffered) is unrealized in the following Japanese example (1):

(1) *sono-houkokusho-wa* *seifu$_i$-ga*
the report-TOP government$_i$-SUBJ
jouyaku-o *teiketsushi* *(ϕ_i-ga)* *keizaitekini*
treaty-OBJ make it$_i$-SUBJ economically
higai-o *uke-ta* *koto-o* *shitekishi-ta*
damage-OBJ suffer-PAST COMP point out-PAST
The report pointed out that the government$_i$ agreed to a treaty and (it$_i$) suffered economically.

The omitted argument is called a *zero anaphor*, which is represented using ϕ. In example (1), zero anaphor ϕ_i refers to its antecedent, *seifu$_i$* (government). Such a reference phenomenon is called *zero anaphora*. Identifying zero anaphoric relations is an essential task in developing such accurate NLP applications as information extraction and machine translation for pro-drop languages. For example, in Japanese, 60% of subjects in newspaper articles are unrealized as zero anaphors (Iida et al., 2007).

This paper proposes a method for *intra-sentential subject zero anaphora resolution*, in which a zero anaphor and its antecedent appear in the same sentence and the zero anaphor must be a *subject* of a predicate, for Japanese. We target *subject* zero anaphors because they represent 85% of the intra-sentential zero anaphora in our data set (example (1) is such a case). Furthermore, this work focuses on intra-sentential zero anaphora because inter-sentential cases, in which a zero anaphor and its antecedent do not appear in the same sentence, are extremely difficult. The accuracy of the state-of-the-art method for resolving inter-sentential anaphora is low (Sasano and Kurohashi, 2011), and we believe the current technologies are not mature enough to deal with inter-sentential cases.

Our method *locally* predicts the likelihood of a zero anaphoric relation between every possible combination of potential zero anaphor and potential antecedent without considering the other (potential) zero anaphoric relations in the same sentence. The final determination of zero anaphoric relations for each zero anaphor in a given sentence is done in a *greedy way*; only the most likely candidate antecedent for each zero anaphor is selected as its antecedent as far as the likelihood score exceeds a

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1244–1254,
Austin, Texas, November 1-5, 2016.

given threshold. This approach contrasts with *global* optimization methods (Yoshikawa et al., 2011; Iida and Poesio, 2011; Ouchi et al., 2015), which have recently become popular. These methods use the constraints among possible zero anaphoric relations, such as "if a candidate antecedent is identified as the antecedent of a subject zero anaphor of a predicate, the candidate cannot be referred to by the object zero anaphor of the same predicate", and determine an optimal set of zero anaphoric relations in an entire sentence while satisfying such constraints, using such optimization techniques as sentence-wise global learning (Ouchi et al., 2015) and integer linear programming (Iida and Poesio, 2011).

Although the global optimization methods have outperformed the previous greedy-style methods, our contention is that greedy-style methods can still, in a certain sense, outperform the state-of-the-art global optimization methods. Ouchi et al. (2015)'s global optimization method achieved the state-of-the-art F-score for Japanese intra-sentential subject zero anaphora resolution, but its performance has not yet reached a level of practical use. In our setting, for example, it actually obtained a precision of only 0.61, and even after attempting to obtain more reliable zero anaphoric relations by several modifications, we could only achieve 0.80 precision at extremely low recall levels ($<$0.01). On the other hand, while our proposed greedy-style method obtained a lower F-score than Ouchi et al.'s method, it achieved much higher precision in a wide range of recall levels (e.g., around 0.8 precision at 0.25 in recall and around 0.7 precision at 0.4 in recall). We believe such high precision is crucial to real-world applications, even though the recall remains low, and thus our method is preferable to Ouchi et al.'s method in that sense.

In our proposed method, we use a Multi-column Convolutional Neural Network (MCNN) (Cireşan et al., 2012), which is a variant of a Convolutional Neural Network (CNN) (LeCun et al., 1998). An MCNN has several independent columns, each of which has its own convolutional and pooling layers. The outputs of all the columns are combined in the final layer to provide a final prediction. In this work, motivated by Centering Theory (Grosz et al., 1995) and other previous works, we exploit as distinct columns the word sequences obtained from the surface word sequence and the dependency tree of a target sentence in our MCNN. Although the existing works also exploited such word sequences, they used only particular types of information from them as features based on the researchers' linguistic insights. In contrast, we minimized such feature engineering due to using an MCNN.

The rest of this paper is organized as follows. In Section 2, we briefly overview previous work on zero anaphora resolution. In Section 3, we present the procedure of our zero anaphora resolution method and explain the column sets used in our MCNN architecture. We evaluate how effectively our method recognizes intra-sentential subject zero anaphora in Section 4 and summarize this work and discuss future directions in Section 5.

2 Related work

The typical zero anaphora resolution algorithms proposed so far have exploited the information of a predicate that potentially has a zero anaphor and its candidate antecedent in a supervised manner (Seki et al., 2002; Iida et al., 2003; Isozaki and Hirao, 2003; Iida et al., 2006; Taira et al., 2008; Sasano et al., 2008; Imamura et al., 2009; Hayashibe et al., 2011; Iida and Poesio, 2011; Sasano and Kurohashi, 2011; Yoshikawa et al., 2011). In addition, existing works have exploited the dependency path between a predicate and a candidate antecedent either by encoding such paths to the set of binary features of the words that appear in the path (Iida and Poesio, 2011) or by mining from the paths the sub-trees that effectively discriminate zero anaphoric relations (Iida et al., 2006). However, both methods just focus on the dependency paths between a predicate and a candidate antecedent without exploiting other structural fragments in the dependency tree representing a target sentence, whereas our method uses the text fragments that cover the entire dependency tree.

Another important clue was derived from discourse theories, such as Centering Theory (Grosz et al., 1995). In this theory, (zero) anaphoric phenomenon is explained based on the rules and principles regarding the recency and saliency of candidate antecedents. Okumura and Tamura (1996) developed a rule-based method based on the idea of Centering Theory. Iida et al. (2003) and Imamura et

al. (2009) used as features for machine learning the results of rule-based antecedent identification based on a variant of Centering Theory (Nariyama, 2002). However, we observed that actual anaphoric phenomena often do not obey Centering Theory. To robustly resolve zero anaphora, we need to explore additional clues that are represented in a target sentence (or text).

Recent work by Iida et al. (2015) newly introduced a sub-problem of zero anaphora resolution, *subject sharing recognition*, which is the task that judges whether two predicates have the same subject. In their method, a network of subject sharing predicates is created by their subject sharing recognizer, and then zero anaphora resolution is performed by propagating a subject to the unrealized subject positions through the path in the network. Even though the accuracy of subject sharing recognition exceeds that of zero anaphora resolution, the zero anaphoric relations identified using the results of subject sharing recognition are limited to those that can be reached by subject sharing relations. The recall of this method is not high.

Although most zero anaphora resolution methods independently identify a zero anaphoric relation for each predicate, some previous works optimized the global assignment of zero anaphoric relations in an entire sentence (or an entire text) while satisfying several constraints among zero anaphoric relations. For example, Iida and Poesio (2011) found the best assignment of subject zero anaphoric relations using integer linear programming. As mentioned in the Introduction, Ouchi et al. (2015) estimated the global score of all of the predicate-argument assignments in a sentence, which include the assignments of intra-sentential zero anaphoric relations, to find the best assignment using a hill-climbing technique. Their method has an advantage: it can exploit complicated relations (e.g., the combination of two potential zero anaphoric relations) as features to directly decide more than one predicate-argument relation simultaneously. We adopted Ouchi et al. (2015)'s method as a baseline in Section 4 because it achieved the state-of-the-art performance for intra-sentential zero anaphora resolution.

Collobert et al. (2011) proposed CNN architecture that can be applied to various NLP tasks, such as PoS tagging, chunking, named entity recognition and semantic role labeling. Following this work, CNNs have been utilized in such NLP tasks as document classification (Kalchbrenner et al., 2014; Kim, 2014; Johnson and Zhang, 2015), paraphrase (Hu et al., 2014; Yin and Schütze, 2015) and relation extraction (Liu et al., 2013; Zeng et al., 2014; dos Santos et al., 2015; Nguyen and Grishman, 2015). MCNNs were first introduced for image classification (Cireşan et al., 2012). In NLP tasks, they have been utilized for question-answering (Dong et al., 2015) and relation extraction (Zeng et al., 2015). Our MCNN architecture was inspired by a Siamese architecture (Chopra et al., 2005), which we extend to a multi-column network and replace its similarity measure with a softmax function at its top.

3 Proposed method

Our proposed method consists of the following four steps:

Step 1 Extract every pair of a predicate and a candidate antecedent, $\langle pred_i, cand_i \rangle$, that appears in a target sentence.

Step 2 Predict the probability of each pair using our MCNN.

Step 3 Rank in descending order all the pairs by their probabilities obtained in Step 2.

Step 4 Choose the top pair $\langle pred_i, cand_i \rangle$ in the ranked list and fill the zero anaphor position of predicate $pred_i$ by $cand_i$ if the position has not already been filled by another candidate. Remove $\langle pred_i, cand_i \rangle$ from the list and repeat this step as long as the score of the chosen pair exceeds a given threshold.

In Step 1, we extract set of pairs $\langle pred_i, cand_i \rangle$ in which candidate antecedent $cand_i$ is paired with predicate $pred_i$. Note that we extracted predicate $pred_i$, instead of a zero anaphor that is an unrealized subject of $pred_i$, because the (potential) zero anaphor of $pred_i$ is omitted in the text and cannot be extracted directly.

In Step 2, our MCNN gives a probability that indicates the likelihood of a zero anaphoric relation to judge for each pair whether $cand_i$ fills the blank subject position of $pred_i$ through zero anaphora and ranks all of the pairs by the probabilities in Step 3.

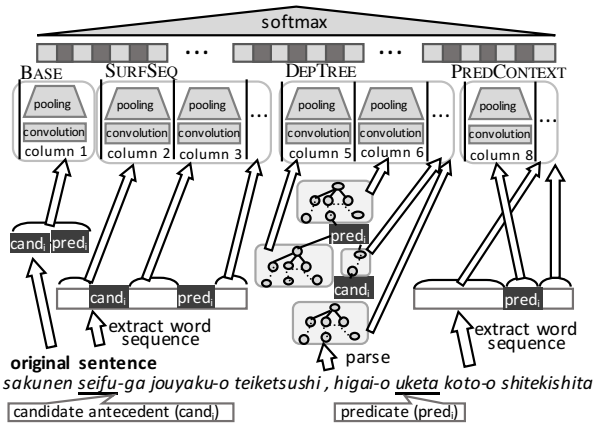

Figure 1: Our multi-column CNN architecture

Finally, in Step 4 we actually fill $cand_i$ in the blank subject positions of $pred_i$ in a greedy style in the order of the ranked list in Step 3, i.e., the zero anaphora resolution with a higher probability is done before that with a lower probability. If the subject position is already occupied by another candidate antecedent, candidate antecedents are no longer filled at that position.

3.1 Design of columns used in MCNN

In Step 2 of our method, we use a Multi-column Convolutional Neural Network (MCNN). Note that zero anaphoric phenomena can be divided into two different referential phenomena: *anaphoric* (i.e., an antecedent precedes its zero anaphor) and *cataphoric* (i.e., a zero anaphor precedes its antecedent) cases. To capture this difference, we divided the set of training instances into two subsets by the relative occurrence positions of a predicate and a candidate antecedent and respectively trained two independent MCNNs using each set.

Our MCNN simultaneously uses four column sets, as illustrated in Figure 1. In the following explanation for each column set, we assume that candidate antecedent $cand_i$ precedes predicate $pred_i$ in the surface order (for the opposite case, i.e., the cataphoric case, the positions of $cand_i$ and $pred_i$ are switched).

BASE The first column set consists of one column, which stores the word vectors of the *bunsetsu* phrases[1] including either $cand_i$ or $pred_i$. We call this column set the BASE column set.

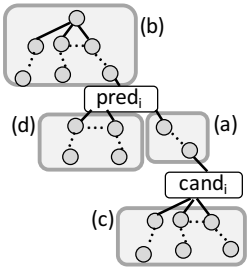

Figure 2: Columns (a, b, c, d) in DEPTREE column set

SURFSEQ The second column set consists of three columns, which store the word vectors of (a) the surface word sequence spanning from the beginning of the sentence to $cand_i$, (b) the sequence between $cand_i$ and $pred_i$, and (c) the remainder, i.e., from $pred_i$ to the end of the sentence. Note that $cand_i$ and $pred_i$ are not included in any column of this column set. We call this column set the SURFSEQ column set.

DEPTREE The third set consists of four columns. We extracted four partial dependency trees from the entire dependency tree of a target sentence: (a) the dependency path between $pred_i$ and $cand_i$, (b) the sub-trees that depend on $pred_i$, (c) the sub-trees on which $cand_i$ depends and (d) the remaining sub-trees, which are illustrated in Figure 2. Note that $cand_i$ and $pred_i$ are not included in the partial trees. Each column stores the word vectors of the word sequence in which the words in (the set of) the partial trees are ordered by their surface order. We call this set the DEPTREE column set.

PREDCONTEXT The fourth set consists of three columns, which store the word vectors of (a) the bunsetsu phrase including $pred_i$, (b) the surface word sequence that appears before (a) (from the beginning of the sentence) and (c) the sequence that appears after (a) (until the end of the sentence). We call this column set the PREDCONTEXT column set.

Among the four column sets, the SURFSEQ column set was designed to introduce the clues based

[1] A bunsetsu phrase is a Japanese base phrase that consists of at least one content word optionally followed by function words.

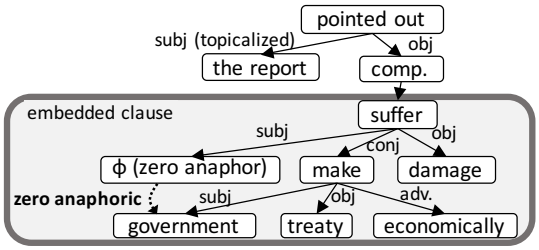

Figure 3: Dependency tree of example (1)

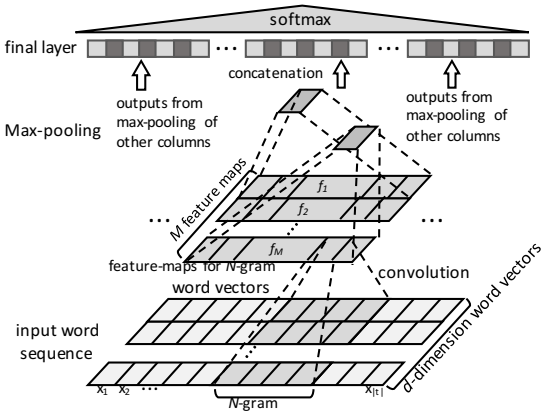

Figure 4: Column of our MCNN

on Centering Theory, in which the antecedent for a given zero anaphor can basically be identified by the *recency* and *saliency* properties of a candidate antecedent. More precisely, in the set of the most salient candidate antecedents, the most recent one is preferred. For example, suppose example (2) in which the predicate *increase* has a subject zero anaphor and its antecedent is *France*:

(2) *nihon-wa* *shoshikataisaku-ni*
Japan-TOP countermeasures to falling birth rate-IOBJ
shippaishi-taga, *furansu-wa* *sore-ni* *seikoushi*
fail-PAST/BUT France-TOP it-IOBJ succeed
(ϕ_i-ga) *shusseiritsu-o* *fuyashiteiru*
(it$_i$-SUBJ) birth rate-OBJ increase
Japan failed to develop countermeasures to its falling birth rate, but France$_i$ succeeded and (ϕ_i) increased its birth rate.

In this situation, there are two most salient candidate antecedents, *Japan* and *France*, because they are marked with topic marker *wa*, which basically indicates the highest degree of candidate saliency. In this case, *France* is selected as the antecedent because it appears more recently than *Japan*, and such recency can be estimated by consulting the surface word sequence between *France* and *increase*: no other salient candidates are included in the word sequence. Also, the other two types of word sequences (i.e., the sequence that spans from the beginning of the sentence to $cand_i$ and that spans from $pred_i$ to its end) are important for confirming whether a more salient candidate than $cand_i$ appears in each word sequence. If such a more salient candidate is found, it should be a stronger candidate of the antecedent.

The DEPTREE column set is introduced for capturing a different aspect of intra-sentential zero anaphora. In the explanation based on Centering Theory, the most salient candidate (e.g., the candidate marked with *wa* (topic marker)) is selected as an antecedent, but example (1) in Section 1 cannot be interpreted based on saliency and recency. In example (1), the *report* is the most salient candidate in the sentence because it is marked with topic marker *wa*, but the less salient candidate *government* becomes the antecedent of zero anaphor ϕ. Such a problem is often solved by introducing the dependency tree of a sentence. Figure 3 represents the dependency tree of example (1) in which the antecedent of ϕ_i appears in the embedded clause. In such a case, an antecedent probably exists among the most salient candidates in the embedded clause. To introduce such structural clues, we used the partial dependency trees as columns in the DEPTREE column set.

Anaphoricity determination, which is the task of judging whether a candidate anaphor has an antecedent, was established as a subtask of coreference resolution. This problem was basically solved by exploring the possible candidate antecedents for a given anaphor candidate in its search space, and the results were used for improving the overall performance of coreference resolution, especially in English (Ng, 2004; Wiseman et al., 2015). Inspired by such previous works, we designed the PRED-CONTEXT set to determine the anaphoricity of zero anaphors, i.e., to judge whether a zero anaphor candidate has its antecedent in a sentence, by consulting the surface word sequences before and after $pred_i$.

3.2 MCNN architecture

In our MCNN (Figure 4), we represent each word in text fragment t by d-dimensional embedding vec-

tor x_i and t by matrix $\mathbf{T} = [\mathrm{x}_1, \ldots, \mathrm{x}_{|t|}]$.[2] $\mathbf{T}$ is then wired to a set of M feature maps where each feature map is a vector. Each element O in the feature map is computed by a filter denoted by f_j ($1 \leq j \leq M$) from the N-gram word sequences in t for a fixed integer N, as $O = \mathrm{ReLU}(\mathbf{W}_{f_j} \bullet \mathrm{x}_{i:i+N-1} + b_{f_j})$, where $\bullet$ denotes element-wise multiplication followed by the summation of the resulting elements (i.e., a Frobenious inner product of $\mathbf{W}_{f_j}$ and $\mathrm{x}_{i:i+N-1}$) and $\mathrm{ReLU}(x) = \max(0, x)$. In other words, we construct a feature map by convolving a text fragment with a filter, which is parameterized by weight $\mathbf{W}_{f_j} \in \mathbb{R}^{d \times N}$ and bias $b_{f_j} \in \mathbb{R}$. Note that there can be several sets of feature maps where each set covers N-grams for different N. Note that the weight of the feature maps for each N-gram in each column set is shared.

As a whole, these feature maps are referred to as a *convolution layer*. The next layer is called a *pooling layer*. Here we use max-pooling (Scherer et al., 2010; Collobert et al., 2011), which simply selects the maximum value among the elements in the same feature map. Our assumption is that the maximum value indicates the existence of a strong clue, i.e., N-gram, for our final judgment. The selected maximum values from all the M feature maps are simply concatenated, and the resulting M-dimensional vector is given to our final layer.

The final layer has vectors coming from multiple feature maps in multiple columns. They are again simply concatenated and constitute a high dimensional feature vector. The final layer applies a linear softmax function to produce the class probabilities of the zero anaphoric labels: *true* and *false*. We use a mini-batch stochastic gradient descent (SGD) with the Adadelta update rule (Zeiler, 2012), apply random initialization within (-0.01, 0.01) for $\mathbf{W}_{f_j}$, and initialize the remaining parameters at zero.

4 Experiments

4.1 Revising annotation results

In our preliminary investigation of the intra-sentential zero anaphoric relations in the NAIST Text Corpus (Iida et al., 2007), since we found more annotation errors than we expected, we decided to revise the annotation results. In this revision, we additionally annotated the subject sharing relations, where two predicates have the same subject regardless whether the subject is realized or omitted, between pairs of predicates in our data set. Note that two predicates can have a subject sharing relation even if neither has a realized subject as far as a subject exists that can naturally fill the subject position of the two predicates. We used the annotated results of subject sharing relations to efficiently detect the annotation errors of intra-sentential zero anaphoric relations, as shown below.

Twenty-six human annotators directly annotated the subject sharing relations for pairs of predicates in a sentence. For this annotation, we automatically extracted from the NAIST Text Corpus all the pairs of predicates that appear in the same sentence and obtained 227,517 predicate pairs. For making the annotation results more reliable, each subject sharing relation was individually judged by three annotators, and the final label was decided by a majority vote. After that, further revisions of the subject sharing relations and the zero anaphoric relations were performed by focusing on the inconsistent annotations between the newly annotated subject sharing relations and the original predicate-argument relations in the NAIST Text Corpus. More precisely, we scrutinized the suspicious annotations such that a subject, which was determined through the annotated subject sharing relations, is not the same as a subject that was directly annotated in the NAIST Text Corpus. In this revision phase, both the subject sharing and zero anaphora relations for such suspicious instances were independently re-annotated by three annotators, and their final labels of both relations were determined by a majority of the their decisions.[3] As a result, 2,120 zero anaphoric instances were newly added to the corpus and 1,184 instances were removed from it for a total of 19,049 instances of intra-sentential subject zero anaphoric relations.[4]

[2] We use zero padding for dealing with text fragments of variable length (Kim, 2014).

[3] We are planning to release the annotated results and information on the data separation used in our evaluation from https://alaginrc.nict.go.jp/.

[4] After this revision, a small number of inconsistent annotated results have both a syntactically dependent subject and a subject zero anaphor because the revision was performed locally. There were 30 inconsistent instances in the testing set and 100 in the training and development sets. We only removed such instances from the testing set without changing the other

Type	#docs	#sentences	#zero anaphors (intra-sentential)
train	1,757	23,152	11,453
dev	586	7,526	3,691
test	586	7,705	3,875

Table 1: Statistics of our data set

4.2 Experimental settings

The documents in the corpus were divided into five subsets, three of which were used as a training data set, one as a development data set, and one as a testing data set. The statistics of our data set are summarized in Table 1. We evaluated the performance of our intra-sentential subject zero anaphora resolution method and three baseline methods described below using the revised annotated results in our data set.

We implemented our MCNN using Theano (Bastien et al., 2012). We pre-trained 300-dimensional word embedding vectors for 1,658,487 words[5] using Skip-gram with a negative-sampling algorithm (Mikolov et al., 2013)[6] on a set of all the sentences extracted from Wikipedia articles[7] (35,975,219 sentences). We removed from the training data all the words that only appeared once before training. In training, we treated them as unknown words and assigned them a random vector. To avoid overfitting, we applied early-stopping and dropout (Hinton et al., 2012) of 0.5 to the final layer. We used an SGD with mini-batches of 100 and a learning rate decay of 0.95. We ran ten epochs through all of the training data, where each epoch consisted of many mini-batch updates. We utilized 3-, 4- and 5-grams with 100 filters each and used the F-score of positive instances as our evaluation metric. The total number of the nodes in the final layers of our MCNN was 3,300: 11 columns $\times$ 3 N-gram $\times$ 100 filters.

Word segmentation, PoS tagging and dependency parsing of the sentences in the NAIST Text Corpus were performed by a Japanese morphological analyzer, MeCab[8] (Kudo et al., 2004), and a dependency parser, J.DepP[9] (Yoshinaga and Kitsuregawa, 2009).

4.3 Baselines

We compared our method with three baseline methods. The first baseline is a single-column convolutional neural network in which the column includes the entire surface word sequence of a sentence. To give the positions of $pred_i$ and $cand_i$ to the network, we concatenated to each word vector an additional 2-dimensional vector, where the first element is set to one if the corresponding word is $pred_i$, the second element is set to 1 if the corresponding word is $cand_i$, and otherwise they are set to 0. This baseline was adopted for estimating the impact of a multi-column network compared to a single-column one.

The remaining two baselines are Ouchi et al. (2015)'s global optimization method and Iida et al. (2015)'s method based on subject sharing recognition. Note that Ouchi's method outputs predicate-argument relations for three grammatical roles (subj, obj, iobj), but for this evaluation we used only the outputs related to intra-sentential subject zero anaphora resolution. As done in Ouchi et al. (2015), we averaged their performances across ten independent runs because the initial random assignment of the predicate-argument relations that was employed in their method changes the performance. Ouchi's method does not require any development data set, so we used both the development and training data sets for training their joint model. For training the subject sharing recognizer used in Iida's method, we used the annotated subject sharing relations in the training and development data sets. In these two baselines, we used the same morphological analyzer and dependency analyzer as for our method.

4.4 Results

Table 2 shows the results for each method. Their performances were evaluated by measuring recall, precision, F-score and average precision (Avg.P). To assess the effectiveness of each column set introduced in Section 3.1, we evaluated the performance of our method using every possible combination of column sets that includes at least the BASE column set. We also gave the precision-recall (PR)

two sets.

[5]Words occurring less than five times in all the sentences were ignored to train the word embedding vectors.

[6]We set the skip distance to 5 and the number of negative samples to 10.

[7]https://archive.org/details/jawiki-20150118

[8]http://taku910.github.io/mecab/

[9]http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

Method		#cols.	Recall	Precision	F-score	Avg.P
Ouchi et al. (ACL2015)		—	**0.539**	0.612	**0.573**	0.670
Iida et al. (EMNLP2015)		—	0.484	0.357	0.411	—
single column CNN (w/ position vec.)		1	0.365	0.524	0.430	0.540
MCNN	BASE	1	0.446	0.394	0.419	0.448
	BASE+SURFSEQ	4	0.458	0.597	0.518	0.679
	BASE+DEPTREE	5	0.339	0.688	0.454	0.690
	BASE+SURFSEQ+DEPTREE	8	0.417	0.695	0.521	0.730
	BASE+SURFSEQ+PREDCONTEXT	7	0.459	0.631	0.531	0.702
	BASE+DEPTREE+PREDCONTEXT	8	0.298	**0.728**	0.422	0.702
	BASE+SURFSEQ+DEPTREE+PREDCONTEXT (**Proposed**)	11	0.418	0.704	0.525	**0.732**

#cols. stands for the number of columns used in each MCNN.

Table 2: Results of intra-sentential subject zero anaphora resolution

curves of our method using the four column sets (BASE+SURFSEQ+DEPTREE+PREDCONTEXT), the single column baseline, and Ouchi's method in Figure 5 to investigate the behavior of each method at a high precision level.[10] The PR-curves of our method and the single-column baseline were plotted just by altering the threshold parameters in Step 4 of our method (See Section 3). In contrast, the PR-curve of Ouchi's method cannot be easily plotted because it gives a score to each sentence, not to each zero anaphoric relation. For plotting the PR-curve, we used the normalized global score of a sentence as the score of any zero anaphoric relations in the sentence.[11] Note that the recall of their PR-curve reached just 0.539, shown in Table 2, because we could not estimate the scores of the zero anaphoric relations that were not outputted by their method. The PR-curves of the other methods also fail to reach 1.0 in recall. This is because the zero anaphoric relations are exclusive; a zero anaphor does not refer to more than one antecedent. If a method provides an incorrect zero anaphoric relation, a correct relation for the same zero anaphor will never be provided in its output. Also, note that the average precision of each method was calculated by averaging the precisions at the available recall levels for each method.

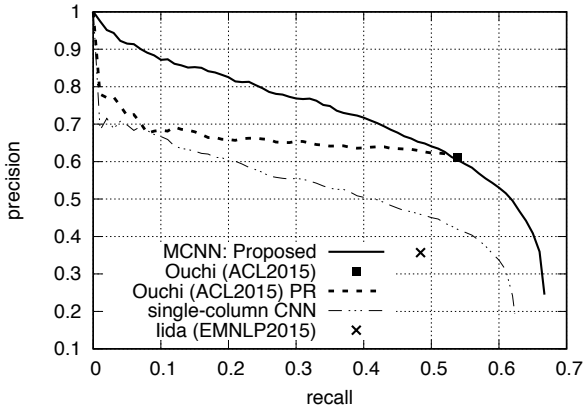

Figure 5: PR-curves of each method

The results in Table 2 show that our method using all the column sets achieved the best average precision among the combination of column sets that include at least the BASE column set. This suggests that all of the clues introduced by our four column sets are effective for performance improvement. Table 2 also demonstrates that our method using all the column sets obtained better average precision than the strongest baseline, Ouchi's method, in spite of an unfavorable condition for it.[12] The results also show that our method with all of the column sets achieved a better F-score than Iida's method and the single-column baseline. However, it achieved a lower F-score than Ouchi's method. This was caused by the choice of different recall levels for computing the F-score. In contrast, the PR-

[10] The PR-curve of Iida et al. (2015)'s method was not plotted because it does not provide the score of each zero anaphoric relation.

[11] The global score provided by Ouchi's method becomes greater based on the number of predicate-argument pairs in a sentence. To control this, we normalized the original global score by the sum of the frequencies of the single or double predicate-argument pairs because the feature functions were applied to such pairs in their method. This achieved the best performance among the normalization schemes we have tried so far.

[12] When calculating the average precision of each method, the relatively low values in precision at high recall levels (i.e., from 0.54 to 0.67) were used in our method but not in Ouchi's method, as seen in Figure 5.

Set	Method	Recall	Precision	F-score	Avg.P
Anaphoric	single-column CNN (w/ position vec.)	0.445	0.525	0.481	0.341
	MCNN (BASE)	**0.591**	0.330	0.424	0.367
	MCNN (BASE+SURFSEQ)	0.555	0.566	0.560	0.565
	MCNN (BASE+DEPTREE)	0.389	0.615	0.476	0.518
	MCNN (BASE+SURFSEQ+DEPTREE)	0.503	0.660	**0.571**	0.599
	MCNN (BASE+SURFSEQ+PREDCONTEXT)	0.535	0.611	0.570	0.581
	MCNN (BASE+DEPTREE+PREDCONTEXT)	0.330	**0.699**	0.449	0.528
	MCNN (**Proposed**)	0.492	0.673	0.569	**0.602**
Cataphoric	single-column CNN (w/ position vec.)	0.163	0.293	0.209	0.163
	MCNN (BASE)	0.171	0.130	0.148	0.099
	MCNN (BASE+SURFSEQ)	0.202	0.417	0.272	0.257
	MCNN (BASE+DEPTREE)	**0.268**	0.438	0.332	0.329
	MCNN (BASE+SURFSEQ+DEPTREE)	0.195	**0.525**	0.285	0.330
	MCNN (BASE+SURFSEQ+PREDCONTEXT)	0.258	0.406	0.316	0.276
	MCNN (BASE+DEPTREE+PREDCONTEXT)	0.240	0.488	0.322	0.341
	MCNN (**Proposed**)	0.251	0.522	**0.339**	**0.337**

Table 3: Results of instance-wise evaluation for anaphoric and cataphoric sets

curves for these two methods in Figure 5 show that our method obtained higher precision than Ouchi's method at all recall levels. Particularly, it got high precision in a wide range of recall levels (e.g., around 0.8 in precision at 0.25 in recall and around 0.7 in precision at 0.4 in recall), while the precision obtained by Ouchi's method at 0.25 in recall was just around 0.65. We believe this difference becomes crucial when using the outputs of each method for developing accurate real-world NLP applications.

In addition to an evaluation that used all of the test instances, we also investigated how our method performed differently for anaphoric and cataphoric cases. In this evaluation, we first divided our data set into anaphoric and cataphoric sets by the relative position of the candidate antecedent and evaluated the performance by measuring the recall, precision, F-score and average precision for each set. This evaluation was done instance-wise, where we took into account each pair of a predicate and its candidate antecedent as a classification target, while in the previous evaluation the performance was measured for the set of zero anaphors in the test set. Thus, the figures in Table 2 and Table 3 are not comparable. Note that we only compared our method with the baseline using a single-column convolutional neural network because the other baselines are not able to output the score of each instance for measuring their average precision.

The results in Table 3 show that our MCNN-based method achieved better average precision than the single-column CNN baseline except the method that uses only the BASE column set for the cataphoric case. The results also demonstrate that each column set consistently contributes to improving the average precision for both the anaphoric and cataphoric cases. However, Table 3 shows that the average precision for the cataphoric set remains low. As one future direction for further improvement, we need to explore clues for identifying cataphoric relations more accurately.

5 Conclusion

This paper proposed an accurate method for intra-sentential subject zero anaphora resolution using a Multi-column Convolutional Neural Network (MCNN). As clues, our MCNN exploits both the surface word sequence and the dependency tree of a target sentence. Our experimental results show that the proposed method achieved better precision than the strong baselines in a wide range of recall levels.

As future work, we plan to use our MCNN architecture for inter-sentential zero anaphora resolution and develop highly accurate NLP applications using our intra-sentential subject zero anaphora resolution method.

Acknowledgement

We thank Hiroki Ouchi for providing his predicate-argument analyzer that was proposed in Ouchi et al. (2015).

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. In *Proceedings of the NIPS 2012 Workshop: Deep Learning and Unsupervised Feature Learning*.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of Computer Vision and Pattern Recognition Conference*, pages 539–546.

Dan Claudiu Cireşan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition*, pages 3642–3649.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 260–269.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 626–634.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Yuta Hayashibe, Mamoru Komachi, and Yuji Matsumoto. 2011. Japanese predicate argument structure analysis exploiting argument position and type. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 201–209.

Geoffrey E . Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs /1207.0580.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Processings of Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 2042–2050.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 804–813.

Ryu Iida, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 2003 EACL Workshop on The Computational Treatment of Anaphora*, pages 23–30.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Processings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the ACL Workshop: 'Linguistic Annotation Workshop'*, pages 132–139.

Ryu Iida, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, and Julien Kloetzer. 2015. Intra-sentential zero anaphora resolution using subject sharing recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2179–2189.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 85–88.

Hideki Isozaki and Tsutomu Hirao. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 184–191.

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.

Chunyang Liu, Wenbo Sun, Wenhan Chao, and Wanxiang Che. 2013. Convolution neural network for relation extraction. In *Proceedings of the 9th International Conference of Advanced Data Mining and Applications*, pages 231–242.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119.

Shigeko Nariyama. 2002. Grammar for ellipsis resolution in Japanese. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 135–145.

Vincent Ng. 2004. Learning noun phrase anaphoricity to improve conference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pages 151–158.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48.

Manabu Okumura and Kouji Tamura. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. In *Proceedings of The 16th International Conference on Computational Linguistics*, pages 871–876.

Hiroki Ouchi, Hiroyuki Shindo, Kevin Duh, and Yuji Matsumoto. 2015. Joint case argument identification for Japanese predicate argument structure analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 961–970.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 769–776.

Dominik Scherer, Andreas Müller, and Sven Behnke. 2010. Evaluation of pooling operations in convolutional architectures for object recognition. In *Proceedings of the 20th International Conference on Artificial Neural Networks*, pages 92–101.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Processings of the 19th International Conference on Computational Linguistics*, pages 1–7.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 523–532.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1416–1426.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.

Katsumasa Yoshikawa, Masayuki Asahara, and Yuji Matsumoto. 2011. Jointly extracting Japanese predicate-argument relation with Markov Logic. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1125–1133.

Naoki Yoshinaga and Masaru Kitsuregawa. 2009. Polynomial to linear: Efficient classification with conjunctive features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1542–1551.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. In *arXiv:1212.5701 (Dec 27, 2012)*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 2335–2344.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.