Annotating Zero Anaphora for Question Answering

Yoshihiko Asao, Ryu Iida, Kentaro Torisawa
National Institute of Information and Communications Technology
Kyoto 619-0289, Japan
{asao, ryu.iida, torisawa}@nict.go.jp

Abstract

We constructed a large annotated dataset of zero pronouns that correspond to adjuncts marked by *-de* (translated to English as *in*, *at*, *by* or *with*) in Japanese. Adjunct zero anaphora resolution plays an important role in extracting information such as location and means from a text. To our knowledge, however, there have been no large-scale dataset covering them. In this paper, focusing on the application of zero anaphora resolution to question answering (QA), we proposed two annotation schemes. The first scheme was designed to efficiently collect zero anaphora instances that are useful in QA. Instead of directly annotating zero anaphora, annotators evaluated QA instances whose correctness hinges on zero anaphora resolution. Over 20,000 instances of zero anaphora were collected with this scheme. We trained a multi-column convolutional neural network with the annotated data, achieving an average precision of 0.519 in predicting the correctness of QA instances of the same type. In the second scheme, zero anaphora is annotated in a more direct manner. A model trained with the results of the second annotation scheme performed better than the first scheme in identifying zero anaphora for sentences randomly sampled from a corpus, suggesting a tradeoff between application-specific and general-purpose annotation schemes.

Keywords: zero anaphora, question answering, convolutional neural networks

1. Introduction

Zero anaphora refers to anaphora in which the anaphor has a phonetically null form. Zero anaphora resolution is an important sub-problem of many NLP tasks such as question answering (QA). To successfully apply machine learning algorithms to zero anaphora, it is crucial to construct a large and consistent annotated dataset.

In this paper, we focus on zero pronouns that correspond to non-obligatory adjuncts marked by postposition *-de* in Japanese, which mark the location, means, reason or manner of an event and can be translated to English prepositions such as *in*, *at*, *by* or *with*. Although the identification of adjunct zero pronouns plays an important role in extracting such information as location and means from a text, to our knowledge, there have been no attempts to create a large-scale dataset covering them.

While it is possible to exhaustively annotate zero anaphora in a corpus to achieve our goal, such an approach might be inefficient when we are interested only in instances that are useful in a specific application. Given this, we propose two different annotation schemes in this paper. In the first scheme, annotators annotate QA instances that potentially involve zero anaphora. In the second scheme, annotators directly annotate noun-predicate pairs with regard to whether they are in a zero anaphora relationship. We evaluated the performance of these two schemes using an existing neural network-based zero anaphora resolution method (Iida et al., 2016). Our experimental results show that the first scheme achieved better performance when it is used to train a module for a QA system, suggesting that the effectiveness of an annotation scheme depends on applications even in a relatively well studied task like zero anaphora resolution. Conversely, the model trained with annotation results of the second scheme achieved better performance in identifying zero anaphora for sentences randomly sampled from a corpus.

We collected 20,830 instances of *-de* zero anaphora with our first annotation scheme alone, while the Kyoto University Text Corpus (Kawahara et al., 2002), the largest existing resource that we are aware of, has only 333 instances of *-de* zero anaphora of the equivalent type. Table 1 shows a few illustrative examples of zero anaphora successfully collected in our work.[1]

レンタルバイクを借りて島をまわる。I rent a motorcycle$_i$ and *travel* the island [$^{\text{by}}$ $\emptyset_i$].
ソーラー発電が拡大すると環境はどう変化するのか。 How does our environment *change* [$^{\text{by}}$ $\emptyset_i$] if solar electricity$_i$ expands?
名古屋市立大学は、市民と科学者が喫茶店でコーヒーを飲みながら科学について話し合う「サイエンスカフェ」を名古屋市内で開催する。Nagoya City University will hold a 'science café' in the city of Nagoya$_i$, in which citizens and scientists *discuss* science [$^{\text{in}}$ $\emptyset_i$] while drinking coffee at a café.

Table 1: Examples of *-de* zero anaphora in our data

2. Related work

Anaphora or coreference has been annotated in several projects including Message Understanding Conference (MUC) (Hirschman and Chinchor, 1997), Automatic Context Extraction (ACE) (Doddington et al., 2004) and OntoNotes (Hovy et al., 2006), but zero anaphora is not annotated in their English corpora. The OntoNotes corpora for pro-drop languages like Chinese and Arabic contain coreference annotations for certain types of zero pronouns. They do not, however, include adjunct zero pronouns, which we deal with in this paper.

Another kind of resources that are relevant to our work is annotated corpora of semantic roles or frame elements such as PropBank (Palmer et al., 2005) and FrameNet

[1] We only deal with intra-sentential anaphora in this paper.

(Baker et al., 1998). In FrameNet, for example, frame elements that are not overtly encoded are annotated as Null Instantiation, some of which can be regarded as adjunct zero anaphors, although their antecedents are not annotated.

As for Japanese resources, zero anaphora was annotated in 5,000 sentences of the Kyoto University Text Corpus (Kawahara et al., 2002), including *-de* zero anaphora, although its size is small; it has only 333 instances of intra-sentential *-de* zero anaphora. Zero anaphora is also annotated for 20,000 sentences in the NAIST corpus (Iida et al., 2007), but only for *-ga* (nominative), *-o* (accusative) and *-ni* (dative).

While zero anaphora resolution has been recognized as an important task in pro-drop languages such as Chinese and Japanese, the task has been less prominent in languages like English, in which core arguments are usually realized as overt forms. However, adjuncts can be omitted in any language, and in such cases, they must be inferred from the context. For example, consider the following English sentence: *Shortly after her arrival in Tokyo, she began her career as a journalist*. Given this sentence, we can infer that she began her career as a journalist *in Tokyo*, but it is not a trivial task to identify this implicit relation. Thus, although the target language of our current work is Japanese, it has relevance to other types of languages as well.

3. *-de* zero anaphora of Japanese

In this paper, we focus on adjuncts marked by *-de* in Japanese. The postposition *-de* in Japanese marks the location, means, reason, or manner of an event or action denoted by the predicate it modifies; it can be translated into a number of English prepositions including *in*, *at*, *by* and *with* depending on the context. Examples of the usage of *-de* are shown below. In sentence (1), *-de* marks the location of the action, whereas in sentence (2), *-de* marks the means by which the action is performed.

(1) *kōen-de* *hashiru*
park-LOC run
'run in a park'

(2) *supūn-de* *taberu*
spoon-INSTR eat
'eat with a spoon'

In some cases, a *-de* phrase is not overt, but can be recovered from the context. Such cases constitute examples of *-de* zero anaphora:

(3) *Kare-wa* *gakkō$_i$-ni* *itte* $\emptyset_i^{de}$ *benkyōshita*
he-TOP school-to go.and studied
'He went to school$_i$ and studied $^{\text{at}}\emptyset_i$'

(4) *shizen-gengo-shori$_i$-o* *katsuyōshite*
natural-language-processing-ACC utilize.and
$\emptyset_i^{de}$ *iryō-o* *kaizendekiru*
medicine-ACC improve.can
'We can improve medicine $^{\text{with}}\emptyset_i$ utilizing natural language processing$_i$'

In sentence (3), the location of studying is not directly expressed, but can be inferred from the preceding part, *he went to school*. In the same vein, in sentence (4), *natural language processing* does not directly modify *improve*, but we can clearly see that it is the means of the action denoted by the predicate.

Solving *-de* zero anaphora is useful in a variety of applications including question answering. For example, given sentence (4), to build a QA system that can correctly answer to the question 'What can we improve medicine with?', the system must resolve *-de* zero anaphora. More generally, *-de* anaphora resolution plays a crucial role when we would like to extract information like location and means that must be inferred from context.[2]

4. Data construction

To generate datasets for annotation, we used WISDOM X (Mizuno et al., 2016), a question-answering system that our team has been developing.[3] The factoid QA module of WISDOM X accepts a question in Japanese and returns nouns as answers, as well as original sentences from the web corpus that support the answers. An example is below.

(5) **Question** *AI-de* *nani-ga* *jitsugensuru*
AI-INSTR what-NOM be.realized
'What will be realized by AI?'

Answer *kaji-robotto* 'housekeeping robot'

Sentence *AI-ga* *sarani* *hattensureba,*
AI-NOM further develop.if
***kaji-robotto**-ga* *jitsugensuru* *daroo.*
housekeeping-robot-NOM be.realized will
'If AI develops further, housekeeping robots will be brought into reality.'

Human annotators judge whether the answer is correct, or in other words, whether the sentence supports the answer given the question. In this example, we expect that the annotators' judgment is positive. When the judgment is positive, we can suppose that the sentence entails *housekeeping robots will be realized by AI*, although the *-de* phrase (translated as *by AI*) is not overtly expressed. This indicates that the original sentence involves zero anaphora, as shown in sentence (6).

(6) *AI$_i$-ga* *sarani* *hattensureba,*
AI-NOM further develop.if
***kaji-robotto**-ga* $\emptyset_i^{de}$ *jitsugensuru* *daroo.*
housekeeping-robot-NOM be.realized will
'If AI$_i$ develops further, housekeeping robots will be brought into reality $^{\text{by}}\emptyset_i$.'

Thus, this annotation task can be regarded as an indirect way to discover zero anaphora instances. With this scheme,

[2] Assuming zero pronouns for adjuncts may be non-conventional, but this is just one way of capturing covert semantic relationships between nouns and predicates in text. Instead of using the notion of adjunct zero anaphora, we can say that there are implicit semantic links between 'school' and 'study' in sentence (3) and between 'natural language processing' and 'improve' in sentence (4) that can be expressed by postposition *-de*. This alternative view does not affect our discussion.

[3] http://wisdom-nict.jp/

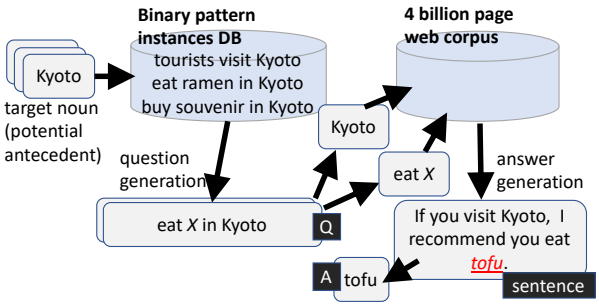

Figure 1: Construction of QA instances

we can focus on zero anaphora instances that are likely to be useful in QA, rather than zero anaphora in general. Another advantage of this annotation scheme is that annotators do not have to know what zero anaphora is, and therefore non-experts can more readily work on it.

In our approach, QA instances for annotation are automatically generated by combining the question suggestion module and the factoid QA module of WISDOM X. Figure 1 illustrates our workflow, which we also summarize below.

1. Choose a *-de candidate* (e.g., *Kyoto*)
2. Generate questions:
   (a) Extract binary pattern instances that contain the *-de* candidate (e.g., *eat ramen in Kyoto*)
   (b) Generate questions based on these binary pattern instances (e.g., *What is there to eat in Kyoto?*)
3. Generate answers:
   (a) Decompose the questions into a *-de* candidate and a unary pattern
   (b) Search for sentences with both the *-de* candidate and the unary pattern

In the first step, we choose a noun that is frequently used with *-de*, which we call the *-de candidate*. The following steps are designed to find a zero anaphor which has this *-de* candidate as the antecedent.

In the second step, we generate questions using the question suggestion module. The module works in the following way. First, it searches for binary pattern instances in our database that contain the *-de* candidate.[4] Here a *binary pattern instance* refers to a dependency tree fragment that consists of a predicate and two nouns that depend on it. For example, when the *-de* candidate is *Kyoto*, examples of binary pattern instances include ⟨*kankōkyaku-ga Kyōto-o otozureru*⟩ '*tourists visit Kyoto*', ⟨*Kyōto-de rāmen-o taberu*⟩ '*eat ramen in Kyoto*' and ⟨*Kyōto-de kaigi-ga hirakareru*⟩ '*a conference is held in Kyoto*'. While the *-de* candidate *Kyoto* may appear in a variety of syntactic positions, we only use questions in which it appears with *-de*. Next, questions are generated by replacing the non-*de*-candidate noun with an interrogative word. For example, the question *Kyōto-de nani-o taberu* (literally '*eat what in Kyoto*', i.e., '*what is there to eat in Kyoto?*') is generated based on such instances as ⟨*Kyōto-de rāmen-o taberu*⟩ '*eat ramen in Kyoto*'.

In the third step, the generated questions are input into our QA system, WISDOM X. While WISDOM X uses multiple methods to find answers, for our present work, we only use answers obtained via the method that we describe here. WISDOM X decomposes the question into the *-de* candidate and the *unary pattern* that corresponds to the question minus the *-de* candidate. Here a *unary pattern* refers to a dependency tree fragment that consists of a predicate and a slot for a noun, such as ⟨*X-ni iku*⟩ '*go to X*' or ⟨*X-ga aku*⟩ '*X opens*'. In our '*eat what in Kyoto*' example, the *-de* candidate is *Kyoto* and the unary pattern is *'eat X'*. The system then searches for sentences that contain *both* the unary pattern and *-de* candidate. We will obtain, for example, the following sentence from the corpus: '*If you visit Kyoto, I recommend you eat tofu*'. Here, the noun *tofu* fills *X* of the *'eat X'* pattern; therefore, *tofu* is presented as an answer along with the original sentence, which the answer is based on. As is the case in this example, we only use sentences in which the *-de* candidate ('*Kyoto*' in this example) and the target predicate ('*eat*') are *not* in a dependency relationship. We expect sentences collected with this method to have a higher-than-average chance of involving *-de* zero anaphora. This is because our procedure guarantees that the *-de* candidate is a noun that frequently occurs in the *-de* position. Note that, because WISDOM X does not search for contexts beyond a single sentence, our target is limited to intra-sentential anaphora.

We created two datasets for annotation, **QAAnnot** and **AllNouns**, based on the QA instances generated by the procedure above.

QAAnnot Annotators directly evaluate QA instances that potentially involve zero anaphora. This task can be simultaneously interpreted as both a QA evaluation task and a zero anaphora annotation task. For this task, we obtained 100,000 QA instances in the following manner. First, we generated questions for 10,000 nouns randomly sampled from the nouns that most frequently appear in the *-de* position in the TSUBAKI corpus (Shinzato et al., 2008) of 600 million web pages. Next, we randomly sampled questions according to the frequency distribution of the predicates; these questions were then input into WISDOM X until we obtained 100,000 QA instances. Finally, human annotators judge each QA instance for its correctness.

AllNouns While **QAAnnot** may be optimized for QA, its special annotation scheme may have a negative impact on performance when it is used to train a model for identifying *-de* zero anaphora in general. To investigate this, we created the second dataset called **AllNouns**; for this dataset, we obtained 10,000 QA instances using the same procedure as **QAAnnot**. Unlike **QAAnnot**, however, annotators do not see questions or answers, but instead judge whether each noun in the original sentence is in a *-de* anaphora relationship with the target predicate. On average there were

[4] Our binary pattern database is based on the four-billion-page web corpus that we constructed, which is also used by the factoid QA module.

	#sentences	#predicates	#pairs	#-*de* zero anaphora		#annotators	approx. person days	Fleiss' κ
QAAnnot	100,000	100,000	100,000	20,830	(20.8%)	24	500	0.556
AllNouns	10,000	10,000	70,971	7,749	(10.9%)	8	100	0.539
General	1,433	3,790	17,392	1,126	(6.5%)	4	20	0.495
KTC	5,127	14,987	90,731	333	(0.4%)	–	–	–

KTC: Kyoto University Text Corpus (Kawaraha et al. 2002)

Table 2: Annotation results

seven nouns to annotate per sentence. More than one noun can be simultaneously in a *-de* anaphora relation with the same target predicate.

We also created a small third dataset for evaluation, which we refer to as **General**. For this dataset, sentences were randomly sampled from the four-billion-page web corpus. For each noun-predicate pair that was identified as not being in a dependency relationship, annotators annotated whether it was in a *-de* anaphora relationship or not. In order to restrict our data to Japanese body texts, we only used sentences that (i) have at least two postpositions and (ii) end with the Japanese full stop (。).

To identify nouns and predicates to annotate, we used the morphological analyzer MeCab (Kudo et al., 2004), as well as the dependency parser J.DepP (Yoshinaga and Kitsuregawa, 2009).

5. Annotation results

In this section, we describe annotation results obtained from the annotation tasks described above. Table 2 summarizes the sizes of our datasets as well as our annotation results.

For each dataset, three annotators independently evaluate each instance, and the final judgment was determined by a majority vote. An annotator was sometimes replaced by another person after completing a set of 1,000 instances. The total numbers of annotators participated in our work, as well as the time required to create each dataset, are included in Table 2.

While 20.8% of the instances were judged to be positive in **QAAnnot**, the percentage was only 10.9% in **AllNouns** and 6.5% in **General**. This difference is expected given the way each dataset was built; **QAAnnot** has the highest proportion of positive instances because it is designed to annotate only likely candidates of zero anaphora. **AllNouns** has a smaller proportion of positive instances because not only the original *-de* candidates but also the other nouns are annotated. The third dataset, **General**, has the lowest proportion of positive instances, as it consists of random sentences in the corpus.

Table 2 also shows a comparison with the Kyoto University Text Corpus (Kawahara et al., 2002). In the Kyoto University Text Corpus, the number of noun-predicate pairs that are targets of zero anaphora annotation based on our criteria is 90,731. Among them, only 333 pairs (0.3%) are annotated as *-de* zero anaphora. The number of *-de* zero anaphora that we collected is substantially larger, in terms of both absolute numbers and the frequency relative to the size of the corpus.

6. Experiments

	train	valid	devel	test
QAAnnot	62,527	20,805	10,401	10,409
QAAnnot-Small	44,603	14,753	-	-
AllNouns	44,603	14,753	7,231	7,253
General	-	-	-	18,265

Table 3: Dataset sizes for our experiments

To see how our annotation results could be generalized to new data, we employed the multi-column convolutional neural network (MCNN) (Cireşan et al., 2012). An MCNN is a variant of a convolutional neural network and has multiple independent columns, each of which has its convolutional and pooling layers. MCNNs have recently been successfully used to model subject zero anaphora in Japanese (Iida et al., 2016). In this work, Iida et al. extracted eleven distinct column inputs from the target predicate (*pred*), the *-de* candidate (*cand*) and their context. The column inputs consist of (a) the word sequence of *cand* and *pred*, (b) the surface word sequences before *cand*, between *cand* and *pred* and after *pred*, (c) the four word sequences extracted from the dependency tree, and (d) *pred* and the word sequences before and after *pred*. We used the same eleven column inputs for our experiments; our architecture is illustrated in Figure 2. More details on the definition of each column are given in (Iida et al., 2016).

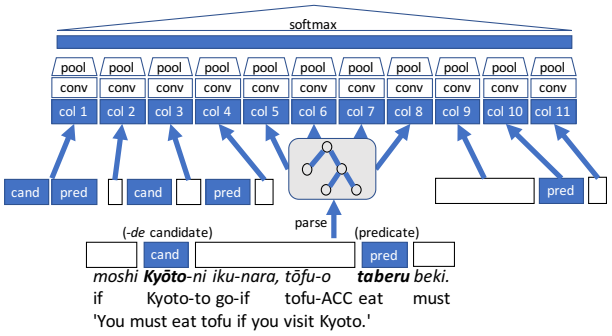

Figure 2: Our multi-column convolutional network architecture

Our MCNN was implemented with Theano (Bastien et al., 2012). We used 300-dimensional word embedding vectors pre-trained with Wikipedia articles using Skip-gram with a negative-sampling algorithm (Mikolov et al., 2013). We treated all the words that only appeared once as unknown words and assigned them a random vector. We used an SGD with mini-batches of

test data	training data	R	P	F	avg. P
QAAnnot	**QAAnnot**	0.263	0.629	0.371	**0.519**
	QAAnnot-Small	0.199	**0.677**	0.307	0.510
	AllNouns	**0.335**	0.523	**0.408**	0.452
AllNouns	**QAAnnot**	0.198	0.563	0.293	0.402
	QAAnnot-Small	0.124	**0.589**	0.204	0.369
	AllNouns	**0.288**	0.569	**0.382**	**0.451**
General	**QAAnnot**	0.125	0.379	0.188	0.218
	QAAnnot-Small	0.081	**0.414**	0.135	0.197
	AllNouns	**0.186**	0.413	**0.257**	**0.269**

Table 4: Performance comparison for different combinations of training and test data

100 and a learning rate decay of 0.95. We used 3-, 4- and 5-grams with 100 filters each. Average precision was used as our evaluation metric. Tuning embeddings in training was turned off, as we found no performance improvements. **QAAnnot** and **AllNouns** were divided into training, validation, development and test data as summarized in Table 3, such that noun-predicate pairs from the same sentence were included in the same bin.[5] For **General**, all instances were used as test data. Further, because **QAAnnot** is larger than **AllNouns**, we could not easily determine whether the performance differences between these two datasets were due to differences in annotation methods or size. To avoid this complication, we constructed **QAAnnot-Small** by randomly sampling **QAAnnot** such that the sizes of training and validation data exactly matched those of **AllNouns**; note that **QAAnnot-Small** was only used for training.

Table 4 summarizes our results. Regardless of the type of training data, the average precision was highest when the test data was **QAAnnot**, and lowest when the test data was **General**. This is probably because **QAAnnot** has the largest proportion of positive instances, whereas **General** has the smallest.

The model trained with **QAAnnot** outperformed that of **AllNouns** when the test data was **QAAnnot**, whereas the model trained with **AllNouns** outperformed that of **QAAnnot** when it was **AllNouns** or **General**, in terms of average precision. As expected, the model trained with **QAAnnot-Small** performed worse than that of **QAAnnot** for all test data, but not to the extent that the gap between the two annotation schemes was reversed, showing that the performance differences between **QAAnnot** and **AllNouns** were not caused by their size differences. Our results suggest that **QAAnnot** has an advantage when our goal is to improve a module of a QA system, while **AllNouns** is better for identifying *-de* zero anaphora in randomly sampled sentences.

Figure 3, 4 and 5 show the precision-recall curves corresponding to different training data when **QAAnnot**, **AllNouns** and **General** were used as test data respectively. To improve QA, we would be able to set a threshold such that only relatively reliable answers are returned.

[5]The total number of instances does not exactly match the total number of annotated noun-predicate pairs because sometimes the same noun occurs more than once in a sentence. In these cases, annotations were made only once without distinguishing different occurrences of the same noun, but each occurrence of the noun was used as a different instance for experiments.

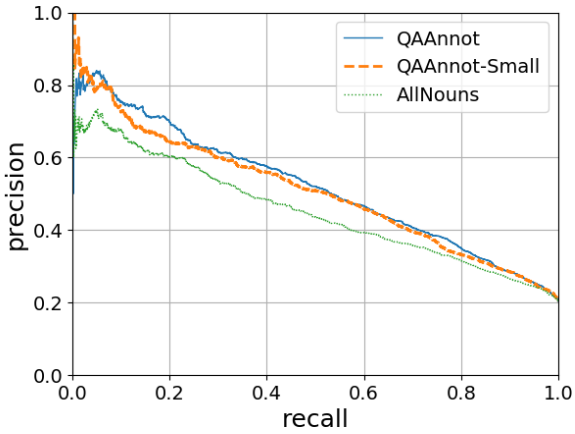

Figure 3: The precision-recall curves for experiments in which **QAAnnot** was used as test data

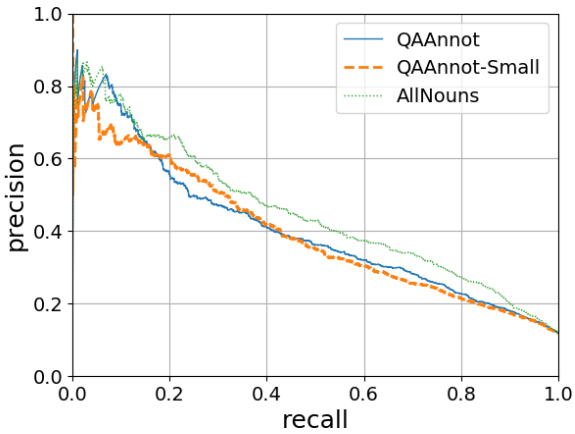

Figure 4: The precision-recall curves for experiments in which **AllNouns** was used as test data

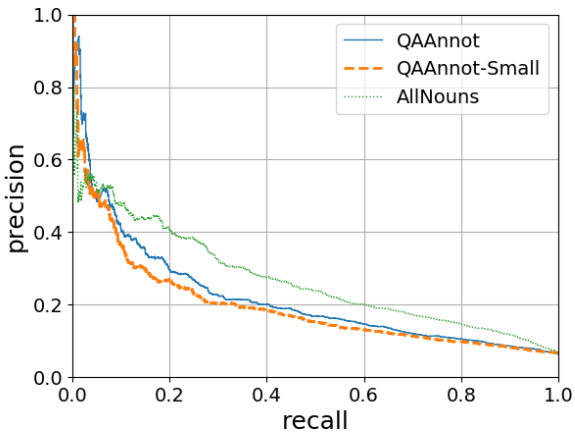

Figure 5: The precision-recall curves for experiments in which **General** was used as test data

To the best of our knowledge, this work is the first attempt to identify *-de* zero anaphora on a large scale, and thus a direct performance comparison with existing work is not possible. For reference, Ouchi et al. (2017) report zero anaphora resolution experiments for *-ga* (nominative), *-o* (accusative) and *-ni* (dative); their best F-measures were 50.65%, 35.07% and 9.83% respectively. Our best F-measure on **General** is 25.7%, which surpass the best score for dative, but not for nominative or accusative. This suggests that *-de* zero anaphora is more difficult to solve than *-ga* and *-o* zero anaphora.

7. Conclusions

In this paper, we described our work in constructing a large annotated dataset of zero pronouns that correspond to adjuncts marked by *-de* in Japanese. The contrast between **QAAnnot** and **AllNouns** can be equated to the contrast between an end-to-end approach and a component-based approach. While **QAAnnot** performs better for QA, it does not generalize well to sentences randomly sampled from a corpus, suggesting that there is a tradeoff between application-specific and general-purpose annotation methods.

8. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley Framenet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., and Bengio, Y. (2012). Theano: new features and speed improvements. In *In Proceedings of the NIPS 2012 Workshop: Deep Learning and Unsupervised Feature Learning.*

Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *International Conference of Pattern Recognition*, pages 3642–3649.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, volume 2, pages 837–840.

Hirschman, L. and Chinchor, N. (1997). MUC-7 coreference task definition.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139. Association for Computational Linguistics.

Iida, R., Torisawa, K., Oh, J.-H., Kruengkrai, C., and Kloetzer, J. (2016). Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-2016)*, pages 1244–1254.

Kawahara, D., Kurohashi, S., and Hashida, K. (2002). Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 2008–2013.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 230–237.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (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.

Mizuno, J., Tanaka, M., Ohtake, K., Oh, J.-H., Kloetzer, J., Hashimoto, C., and Torisawa, K. (2016). WISDOM X, DISAANA and D-SUMM: Large-scale NLP systems for analyzing textual big data. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016) (Demo Track)*.

Ouchi, H., Shindo, H., and Matsumoto, Y. (2017). Neural modeling of multi-predicate interactions for Japanese predicate argument structure analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1600.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C., and Kurohashi, S. (2008). TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Processings of the 3rd International Joint Conference on Natural Language Processing*, pages 189–196.

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