Bypassed Alignment Graph for Learning Coordination in Japanese Sentences

Hideharu Okuma Kazuo Hara Masashi Shimbo Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
Ikoma, Nara 630-0192, Japan
{okuma.hideharu01,kazuo-h,shimbo,matsu}@is.naist.jp

Abstract

Past work on English coordination has focused on coordination scope disambiguation. In Japanese, detecting whether coordination exists in a sentence is also a problem, and the state-of-the-art alignment-based method specialized for scope disambiguation does not perform well on Japanese sentences. To take the detection of coordination into account, this paper introduces a 'bypass' to the alignment graph used by this method, so as to explicitly represent the non-existence of coordinate structures in a sentence. We also present an effective feature decomposition scheme based on the distance between words in conjuncts.

1 Introduction

Coordination remains one of the challenging problems in natural language processing. One key characteristic of coordination explored in the past is the structural and semantic symmetry of conjuncts (Chantree et al., 2005; Hogan, 2007; Resnik, 1999). Recently, Shimbo and Hara (2007) proposed to use a large number of features to model this symmetry, and optimize the feature weights with perceptron training. These features are assigned to the arcs of the *alignment graph* (or *edit graph*) originally developed for biological sequence alignment.

Coordinate structure analysis involves two related but different tasks:

1. Detect the presence of coordinate structure in a sentence (or a phrase).

2. Disambiguate the scope of coordinations in the sentences/phrases detected in Task 1.

The studies on English coordination listed above are concerned mainly with scope disambiguation, reflecting the fact that detecting the presence of coordinations in a sentence (Task 1) is straightforward in English. Indeed, nearly 100% precision and recall can be achieved in Task 1 simply by pattern matching with a small number of coordination markers such as "and," "or," and "as well as".

In Japanese, on the other hand, detecting coordination is non-trivial. Many of the coordination markers in Japanese are ambiguous and do not always indicate the presence of coordinations. Compare sentences (1) and (2) below:

(1) *rondon* *to* *pari* *ni* *itta*
(London) (and) (Paris) (to) (went)
(I went to London and Paris)

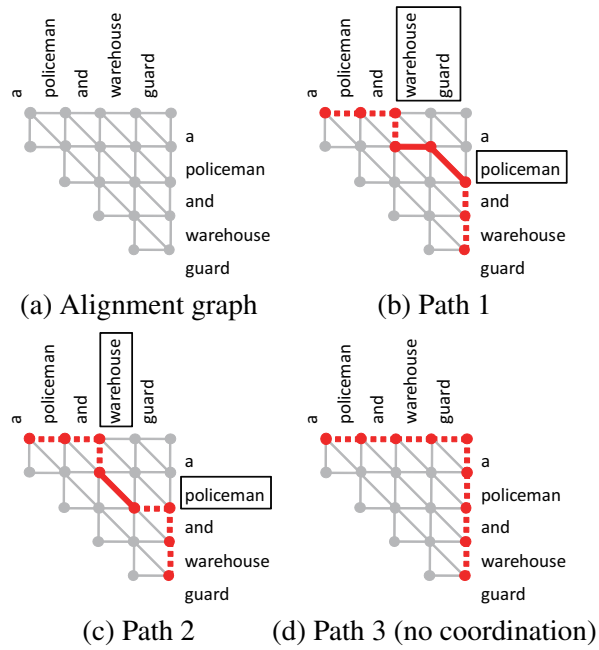
(2) *kanojo* *to* *pari* *ni* *itta*
(her) (with) (Paris) (to) (went)
(I went to Paris with her)

These sentences differ only in the first word. Both contain a particle *to*, which is one of the most frequent coordination markers in Japanese—but only the first sentence contains a coordinate structure. Pattern matching with particle *to* thus fails to filter out sentence (2).

Shimbo and Hara's model allows a sentence without coordinations to be represented as a normal path in the alignment graph, and in theory it can cope with Task 1 (detection). In practice, the representation is inadequate when a large number of training sentences do not contain coordinations, as demonstrated in the experiments of Section 4.

This paper presents simple yet effective modifications to the Shimbo-Hara model to take coordination detection into account, and solve Tasks 1 and 2 simultaneously.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 5–8,
Suntec, Singapore, 4 August 2009.



(a) Alignment graph (b) Path 1

(c) Path 2 (d) Path 3 (no coordination)

Figure 1: Alignment graph for "a policeman and warehouse guard" ((a)), and example paths representing different coordinate structure ((b)–(d)).

2 Alignment-based coordinate structure analysis

We first describe Shimbo and Hara's method upon which our improvements are made.

2.1 Triangular alignment graph

The basis of their method is a triangular *alignment graph*, illustrated in Figure 1(a). Kurohashi and Nagao (1994) used a similar data structure in their rule-based method. Given an input sentence, the rows and columns of its alignment graph are associated with the words in the sentence. Unlike the alignment graph used in biological sequence alignment, the graph is triangular because the same sentence is associated with rows and columns. Three types of arcs are present in the graph. A diagonal arc denotes coordination between the word above the arc and the one on the right; the horizontal and vertical arcs represent skipping of respective words.

Coordinate structure in a sentence is represented by a complete path starting from the top-left (initial) node and arriving at the bottom-right (terminal) node in its alignment graph. Each arc in this path is labeled either *Inside* or *Outside* depending on whether its span is part of coordination or not; i.e., the horizontal and vertical spans of an *Inside* segment determine the scope of two conjuncts. Figure 1(b)–(d) depicts example paths. *Inside* and *Outside* arcs are depicted by solid and dotted lines, respectively. Figure 1(b) shows a path for coordination between "policeman" (vertical span of the *Inside* segment) and "warehouse guard" (horizontal span). Figure 1(c) is for "policeman" and "warehouse." Non-existence of coordinations in a sentence is represented by the *Outside*-only path along the top and the rightmost borders of the graph (Figure 1(d)).

With this encoding of coordinations as paths, coordinate structure analysis can be reduced to finding the highest scoring path in the graph, where the score of an arc is given by a measure of how much two words are likely to be coordinated. The goal is to build a measure that assigns the highest score to paths denoting the correct coordinate structure. Shimbo and Hara defined this measure as a linear function of many features associated to arcs, and used perceptron training to optimize the weight coefficients for these features from corpora.

2.2 Features

For the description of features used in our adaptation of the Shimbo-Hara model to Japanese, see (Okuma et al., 2009). In this model, all features are defined as indicator functions asking whether one or more attributes (e.g., surface form, part-of-speech) take specific values at the neighbor of an arc. One example of a feature assigned to a diagonal arc at row i and column j of the alignment graph is

$$f = \begin{cases} 1 & \text{if } \mathit{POS}[i] = \text{Noun},\ \mathit{POS}[j] = \text{Adjective}, \\ & \text{and the label of the arc is } \textit{Inside}, \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathit{POS}[i]$ denotes the part-of-speech of the ith word in a sentence.

3 Improvements

We introduce two modifications to improve the performance of Shimbo and Hara's model in Japanese coordinate structure analysis.

3.1 Bypassed alignment graphs

In their model, a path for a sentence with no coordination is represented as a series of *Outside* arcs as we saw in Figure 1(d). However, *Outside* arcs also appear in partial paths between two coordinations, as illustrated in Figure 2. Thus, two differ-

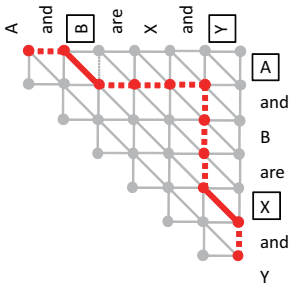


Figure 2: Original alignment graph for sentence with two coordinations. Notice that *Outside* (dotted) arcs connect two coordinations

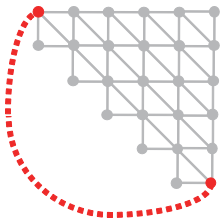


Figure 3: alignment graph with a "bypass"

ent roles are given to *Outside* arcs in the original Shimbo-Hara model.

We identify this to be a cause of their model not performing well for Japanese, and propose to augment the original alignment graph with a "bypass" devoted to explicitly indicate that no coordination exists in a sentence; i.e., we add a special path directly connecting the initial node and the terminal node of an alignment graph. See Figure 3 for illustration of a bypass.

In the new model, if the score of the path through the bypass is higher than that of any paths in the original alignment graph, the input sentence is deemed not containing coordinations.

We assign to the bypass two types of features capturing the characteristics of a whole sentence; i.e., indicator functions of sentence length, and of the existence of individual particles in a sentence. The weight of these features, which eventually determines the score of the bypass, is tuned by perceptron just like the weights of other features.

3.2 Making features dependent on the distance between conjuncts

Coordinations of different type (e.g., nominal and verbal) have different relevant features, as well as different average conjunct length (e.g., nominal coordinations are shorter).

This observation leads us to our second modification: to make all features dependent on their occurring positions in the alignment graph. To be precise, for each individual feature in the original model, a new feature is introduced which depends on whether the Manhattan distance d in the alignment graph between the position of the feature occurrence and the nearest diagonal exceeds a fixed threshold[1] θ. For instance, if a feature f is an indicator function of condition X, a new feature f' is introduced such that

$$f' = \begin{cases} 1, & \text{if } d \leq \theta \text{ and condition } X \text{ holds,} \\ 0, & \text{otherwise.} \end{cases}$$

Accordingly, different weights are learned and associated to two features f and f'. Notice that the Manhattan distance to the nearest diagonal is equal to the distance between word pairs to which the feature is assigned, which in turn is a rough estimate of the length of conjuncts.

This distance-based decomposition of features allows different feature weights to be learned for coordinations with conjuncts shorter than or equal to θ, and those which are longer.

4 Experimental setup

We applied our improved model and Shimbo and Hara's original model to the EDR corpus (EDR, 1995). We also ran the Kurohashi-Nagao parser (KNP) 2.0[2], a widely-used Japanese dependency parser to which Kurohashi and Nagao's (1994) rule-based coordination analysis method is built in. For comparison with KNP, we focus on *bunsetsu*-level coordinations. A bunsetsu is a chunk formed by a content word followed by zero or more non-content words like particles.

4.1 Dataset

The Encyclopedia section of the EDR corpus was used for evaluation. In this corpus, each sentence is segmented into words and is accompanied by a syntactic dependency tree, and a semantic frame representing semantic relations among words.

A coordination is indicated by a specific relation of type "and" in the semantic frame. The scope of conjuncts (where a conjunct may be a word, or a series of words) can be obtained by combining this information with that of the syntactic tree. The detail of this procedure can be found in (Okuma et al., 2009).

[1] We use $\theta = 5$ in the experiments of Section 4.

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

Table 1: Accuracy of coordination scopes and end of conjuncts, averaged over five-fold cross validation. The numbers in brackets are the improvements (in points) relative to the Shimbo-Hara (SH) method.

	Scope of coordinations			End of conjuncts		
Method	Precision	Recall	F1 measure	Precision	Recall	F1 measure
KNP	n/a	n/a	n/a	58.8	65.3	61.9 (−2.6)
Shimbo and Hara's method (SH; baseline)	53.7	49.8	51.6 (±0.0)	67.0	62.1	64.5 (±0.0)
SH + distance-based feature decomposition	55.3	52.1	53.6 (+2.0)	68.3	64.3	66.2 (+1.7)
SH + distance-based feature decomposition + bypass	55.0	57.6	56.3 (+4.7)	66.8	69.9	68.3 (+3.8)

Of 10,072 sentences in the Encyclopedia section, 5,880 sentences contain coordinations. We excluded 1,791 sentences in which nested coordinations occur, as these cannot be processed with Shimbo and Hara's method (with or without our improvements).

We then applied Japanese morphological analyzer JUMAN 5.1 to segment each sentence into words and annotate them with parts-of-speech, and KNP with option '-bnst' to transform the series of words into a bunsetsu series. With this processing, each word-level coordination pair is also translated into a bunsetsu pair, unless the word-level pair is concatenated into a single bunsetsu (sub-bunsetsu coordination). Removing sub-bunsetsu coordinations and obvious annotation errors left us with 3,257 sentences with bunsetsu-level coordinations. Combined with the 4,192 sentences not containing coordinations, this amounts to 7,449 sentences used for our evaluation.

4.2 Evaluation metrics

KNP outputs dependency structures in Kyoto Corpus format (Kurohashi et al., 2000) which specifies the end of coordinating conjuncts (bunsetsu sequences) but not their beginning.

Hence two evaluation criteria were employed: (i) correctness of coordination scopes[3] (for comparison with Shimbo-Hara), and (ii) correctness of the end of conjuncts (for comparison with KNP). We report precision, recall and F1 measure, with the main performance index being F1 measure.

5 Results

Table 1 summarizes the experimental results. Even Shimbo and Hara's original method (SH) outperformed KNP. KNP tends to output too many coordinations, yielding a high recall but low precision. By contrast, SH outputs a smaller number of coordinations; this yields a high precision but a low recall.

The distance-based feature decomposition of Section 3.2 gave +2.0 points improvement over the original SH in terms of F1 measure in coordination scope detection. Adding bypasses to alignment graphs further improved the performance, making a total of +4.7 points in F1 over SH; recall significantly improved, with precision remaining mostly intact. Finally, the improved model (SH + decomposition + bypass) achieved an F1 measure +6.4 points higher than that of KNP in terms of end-of-conjunct identification.

[3] A coordination scope is deemed correct only if the bracketing of constituent conjuncts are all correct.

References

F. Chantree, A. Kilgarriff, A. de Roeck, and A. Willis. 2005. Disambiguating coordinations using word distribution information. In *Proc. 5th RANLP*.

EDR, 1995. *The EDR dictionary*. NICT. http://www2.nict.go.jp/r/r312/EDR/index.html.

D. Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proc. 45th ACL*, pages 680–687.

S. Kurohashi and M. Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Comput. Linguist.*, 20:507–534.

S. Kurohashi, Y. Igura, and M. Sakaguchi, 2000. *Annotation manual for a morphologically and sytactically tagged corpus, Ver. 1.8*. Kyoto Univ. In Japanese. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0/doc/syn_guideline.pdf.

H. Okuma, M. Shimbo, K. Hara, and Y. Matsumoto. 2009. Bypassed alignment graph for learning coordination in Japanese sentences: supplementary materials. Tech. report, Grad. School of Information Science, Nara Inst. Science and Technology. http://isw3.naist.jp/IS/TechReport/report-list.html#2009.

P. Resnik. 1999. Semantic similarity in a taxonomy. *J. Artif. Intel. Res.*, 11:95–130.

M. Shimbo and K. Hara. 2007. A discriminative learning model for coordinate conjunctions. In *Proc. 2007 EMNLP/CoNLL*, pages 610–619.