Generative Modeling of Coordination by Factoring Parallelism and Selectional Preferences

Daisuke Kawahara and **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{dk, kuro}@i.kyoto-u.ac.jp

Abstract

We present a unified generative model of coordination that considers parallelism of conjuncts and selectional preferences. Parallelism of conjuncts, which frequently characterizes coordinate structures, is modeled as a synchronized generation process in the generative parser. Selectional preferences learned from a large web corpus provide an important clue for resolving the ambiguities of coordinate structures. Our experiments of Japanese dependency parsing indicate the effectiveness of our approach, particularly in the domains of newspapers and patents.

1 Introduction

Coordinate structures are a potential source of syntactic ambiguity in natural language. Although many methods have been proposed to resolve the ambiguities of coordinate structures, coordination disambiguation still remains a difficult problem for state-of-the-art parsers. Previous studies on coordination disambiguation used two kinds of clues:

- parallelism of conjuncts, and
- selectional preferences.

Syntactic, lexical and semantic parallelism of conjuncts is frequently observed in coordinate structures. For example, Dubey et al. (2005) empirically confirmed syntactic parallelism in coordinate structures. This clue was modeled by string matching, part-of-speech matching, number agreement, semantic similarities, and so forth (Agarwal and Boggess, 1992; Kurohashi and Nagao, 1994; Resnik, 1999; Chantree et al., 2005; Buyko and Hahn, 2008). For instance, consider the following example:

(1) eat Caesar salad and Italian pasta

We can observe lexical or semantic parallelism between *salad* and *pasta*, which can be automatically detected via a thesaurus or distributional similarity. In addition, syntactic parallelism can be observed; each conjunct has a modifier *Caesar* and *Italian*, respectively. These types of parallelism contribute to identifying the coordinate structure that conjoins *Caesar salad* and *Italian pasta*.

The other clue is selectional preferences, such as *eat* in the above example. Since *eat* is likely to have *salad* and *pasta* as its objects, it is plausible that *salad* and *pasta* are coordinated. Such selectional preferences of predicates are thought to support the construction of coordinate structures, and were used in Japanese dependency parsing by Kawahara and Kurohashi (2008). Selectional preferences of nouns (noun-noun modifications) were used by Resnik (1999), Nakov and Hearst (2005) and Kawahara and Kurohashi (2008). For example, let us see the following examples:

(2) a. mail and securities fraud
b. corn and peanut butter

In (2a), the coordination of *mail* and *securities* is guided by the estimation that *mail fraud* is a salient compound nominal phrase. In (2b), on the contrary, the coordinate structure that conjoins *corn* and *peanut butter* is led because *corn butter* is not a familiar concept.

Each clue has been empirically proven to be effective for coordination disambiguation. However, a unified approach that combines both clues has not been explored comprehensively. In this paper, we propose a unified framework for coordi-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 456–464,
Chiang Mai, Thailand, November 8 – 13, 2011.

nation disambiguation by incorporating both the clues into a generative parser. To capture syntactic parallelism of conjuncts, we formulate the generative process of pre-modifiers of conjuncts in a synchronized manner. In the above example, the generation process of *Caesar* from *salad* is synchronized with that of *Italian* from *pasta*. An interpretation of an unbalanced coordinate structure without synchronization (e.g., "Caesar salad and Italian") is penalized. Lexical parallelism, which is a tendency that some words, such as *salad* and *pasta*, are likely to be coordinated, is also modeled within the generative model.

In this paper, we focus on the Japanese language. A synchronization-based model of coordination disambiguation is integrated into a fully-lexicalized Japanese generative parser (Kawahara and Kurohashi, 2008). For the selectional preferences, we use case frames and statistics of noun-noun modifications that are automatically extracted from large raw corpora. Our method can resolve coordinate structures with parallelism on the basis of the synchronized generative model, and can also handle unlike coordinate structures using selectional preferences.

The remainder of this paper is organized as follows. Section 2 summarizes previous work related mainly to parsing models with coordination disambiguation. Section 3 briefly overviews the Japanese language and coordination ambiguity in Japanese. Section 4 illustrates our idea and describes our model in detail. Section 5 is devoted to our experiments. Finally, section 6 gives the conclusions.

2 Related Work

Resnik (1999) and van Noord (2007) incorporated parallelism and selectional preferences into coordination disambiguation or parsing. Resnik (1999) integrated semantic similarities and noun-noun modifications into voting or decision trees to disambiguate the scope ambiguities of nominal compounds "n1 and n2 n3." He did not integrate this method into parsing, but applied it to an independent task. Van Noord (2007) proposed a MaxEnt model of Dutch parsing that incorporated selectional preferences learned from a large corpus. He used various features in the MaxEnt model including some features that capture parallelism. This indirect treatment of parallelism is different from our generative model that explicitly factors parallelism.

Several other studies have considered parallelism in parsing models. Charniak and Johnson (2005) incorporated some features of syntactic parallelism in coordinate structures into their MaxEnt reranking parser. Kübler et al. (2009) used a reranking parser with automatically detected scope possibilities to improve German parsing. As for a generative parser, Dubey et al. (2006) proposed an unlexicalized PCFG parser that modified PCFG probabilities to condition the existence of a coordinate structure. Hogan (2007) proposed a generative lexicalized parser that considered the symmetry of part-of-speech tags and phrase categories of conjuncts, which is more shallow information than our synchronization model. She also used cooccurrence statistics of conjunct heads, which are similar to our modeling of lexical parallelism, but her model did not use selectional preferences.

Kurohashi and Nagao (1994) proposed a rule-based method of Japanese dependency parsing that included coordination disambiguation. Their method first detects coordinate structures in a sentence using dynamic programming, and then determines the dependency structure of the sentence under the constraints of the detected coordinate structures. Shimbo and Hara (2007) and Hara et al. (2009) considered many features for coordination disambiguation and automatically optimized their weights, which were heuristically determined in Kurohashi and Nagao (1994), by using a discriminative learning model.

3 Japanese Grammar and Coordinate Structure

3.1 Japanese Grammar

Let us first briefly introduce Japanese grammar. The structure of a Japanese sentence can be described well by the dependency relation between *bunsetsu*s. A *bunsetsu* is a basic unit of dependency, consisting of one or more content words and the following zero or more function words. A *bunsetsu* corresponds to a base phrase in English and *eojeol* in Korean. The Japanese language is head-final, that is, a *bunsetsu* depends on another *bunsetsu* to its right (but not necessarily the adjacent *bunsetsu*).

For example, consider the following sentence:[1]

[1]In this paper, we use the following abbreviations: NOM (nominative), ACC (accusative), DAT (dative), ALL (alla-

(3) *ane-to gakkou-ni itta*
sister-CMI school-ALL went

(went to school with (my) sister)

This sentence consists of three *bunsetsu*s. The final *bunsetsu*, *itta*, is a predicate, and the other *bunsetsu*s, *ane-to* and *gakkou-ni*, are its arguments. Their endings, *to* and *ni*, are postpositions that function as case markers.

3.2 Coordinate Structure in Japanese

Coordinate structures in Japanese are roughly classified into two types. The first type is the nominal coordinate structure.

(4) *nagai enpitsu-**to** keshigomu-wo katta*
long pencil-CNJ eraser-ACC bought

(bought a long pencil and an eraser)

The other type is the predicative coordinate structure, in which two or more predicates form a coordinate structure.

(5) *kanojo-to kekkon-**shi** ie-wo katta*
she-CMI married-CNJ house-ACC bought

(married her and bought a house)

For both of these types, we can detect the possibility of a coordinate structure by looking for a *coordination key bunsetsu* that contains *to*, *-shi*, comma and so forth. That is to say, the left and right sides of a coordination key bunsetsu constitute possible pre- and post-conjuncts, and the key bunsetsu is located at the end of the pre-conjunct.

For the evaluation of our method, which is described in section 5, we use analyzed corpora that are annotated on the basis of the annotation criteria of the Kyoto University Text Corpus (Kurohashi and Nagao, 1998).[2] Under this annotation criteria, the last *bunsetsu* in a pre-conjunct depends on the last *bunsetsu* in a post-conjunct, as shown in the dependency trees of Figure 1.

4 Our Method

4.1 Idea

Consider, for example, the following sentence.

(6) *houou-no kenkou-**to** tibet-no heiwa-wo*
pope-GEN health-CNJ tibet-GEN peace-ACC

inotta
prayed

(prayed (for) health of pope and peace of Tibet)

In this sentence, the coordination key "*to*" is a coordinate conjunction.[3] The coordinate structure in example (6) has four possible scopes. Among these, two structures are illustrated in Figure 1. In this figure, our parser generates the constituent words according to the arrows.

First, let us describe the effect of selectional preferences and lexical parallelism. In (a), two coordinated arguments, *kenkou* (health) and *heiwa* (peace), are generated from the verb *inotta* (prayed), and are eligible as accusative words of the verb *inotta* (prayed). *Kenkou* (health) is also generated from its coordinated head *heiwa* (peace). This generation is plausible because people often say this coordinated pair. In (b), the heads of conjuncts, *kenkou* (health) and *tibet*, are generated from the noun *heiwa* (peace). This is not appropriate because we are not referring to the nominal compound "*kenkou-no heiwa*" (peace of health). *Kenkou* (health) is also generated from its coordinated head *tibet*, but this generation has a low probability because this coordination is meaningless and rare.

These judgments are determined based on selectional preferences of predicates including nouns and lexical parallelism. As resources for considering these factors, we use automatically compiled case frames, and cooccurrences of noun-noun modifications and coordinated nouns.

Second, syntactic parallelism of conjuncts is also effective for coordination disambiguation. In (a), after the conjunct heads, *kenkou* (health) and *heiwa* (peace), are generated, the modifier in the pre-conjunct, *houou* (pope), is generated. In this generation, the generative probability of a genitive case from *kenkou* (health), $P(A(\text{GEN}) = \text{Y}|health)$, is considered. Note that $A(CASE)$=$\{\text{Y}, \text{N}\}$ is a binary function that returns Y if a case slot $CASE$ is filled with an ar-

tive), GEN (genitive), CMI (comitative), CNJ (conjunction) and TOP (topic marker).

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Kyoto%20University%20Text%20Corpus

[3] Note that the coordination key "*to*" can be used as a coordinate conjunction and also as a comitative case marker. The tasks of coordination disambiguation include the detection of coordinate conjunctions as well as the identification of coordination scopes. Both of these tasks are simultaneously carried out in our method.

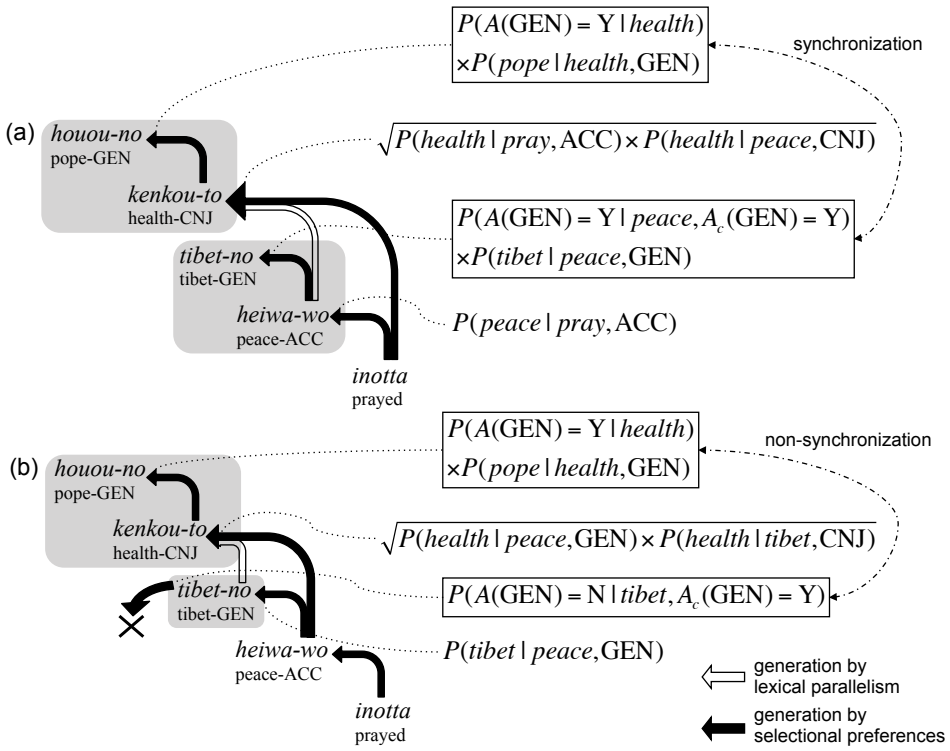

Figure 1: Two possible dependency and coordinate structures with some generative probabilities. The rounded rectangles represent conjuncts of coordinate structures.

gument; otherwise, it returns N. Subsequently, the modifier in the post-conjunct, *tibet*, is generated. This generation includes the synchronous generation of a genitive case from *heiwa* (peace) with the probability $P(A(\text{GEN})=\text{Y}|peace, A_c(\text{GEN})=\text{Y})$, which is conditioned on the previously generated genitive case of the pre-conjunct. Since syntactic parallelism is preferred in coordinate structures, this probability has a larger value than other probabilities $P(A(\text{GEN})=\text{Y}|peace)$ without coordination and $P(A(\text{GEN})=\text{Y}|peace, A_c(\text{GEN})=\text{N})$ without synchronization.

In (b), $P(A(\text{GEN}=\text{N})|tibet, A_c(\text{GEN})=\text{Y})$ means that nothing is generated from *tibet*, whereas the head of the pre-conjunct has a genitive case. This probability has a small value because of non-synchronization (unbalanced coordinate structure).

4.2 Resources

As the resources of selectional preferences to support coordinate structures, we use automatically constructed case frames and cooccurrences of noun-noun modifications. As a parser for extracting these resources, we use the Japanese dependency parser, KNP,[4] which is also used as a base model in the following sections.

	CS	examples
yaku (1) (bake)	*ga*	I:18, person:15, craftsman:10, ···
	wo	bread:2484, meat:1521, cake:1283, ···
	de	oven:1630, frying pan:1311, ···
yaku (2) (have difficulty)	*ga*	teacher:3, government:3, person:3, ···
	wo	fingers:2950
	ni	attack:18, action:15, son:15, ···
yaku (3) (burn)	*ga*	maker:1, distributor:1
	wo	data:178, file:107, copy:9, ···
	ni	R:1583, CD:664, CDR:3, ···
⋮	⋮	⋮

Table 1: Acquired case frames of *yaku*. "CS" indicates case slots, such as *ga* (NOM), *wo* (ACC), *ni* (DAT) and *de* (LOC). Example words are expressed only in English due to space limitation. The number following each word denotes its frequency.

4.2.1 Automatically Constructed Case Frames

We employ automatically constructed case frames (Kawahara and Kurohashi, 2006). This section outlines the method for constructing the case frames.

A large corpus is automatically parsed, and case

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

frames are constructed from predicate-argument examples in the resulting parses. The problems of automatic case frame construction are syntactic and semantic ambiguities. That is to say, the parsing results inevitably contain errors, and verb senses are intrinsically ambiguous. To cope with these problems, case frames are gradually constructed from reliable predicate-argument examples.

First, predicate-argument examples that have no syntactic ambiguity are extracted, and they are disambiguated by a pair consisting of a verb and its closest case component. Such pairs are explicitly expressed on the surface of text, and are thought to play an important role in sentence meanings. For instance, examples are distinguished not by verbs (e.g., *yaku* (bake/broil/have difficulty)), but by pairs (e.g., *pan-wo yaku* (bake bread), *niku-wo yaku* (broil meat), and *te-wo yaku* (have difficulty)). Predicate-argument examples are aggregated in this way, and yield basic case frames.

Thereafter, the basic case frames are clustered to merge similar case frames. For example, since *pan-wo yaku* (bake bread) and *niku-wo yaku* (broil meat) are similar, they are clustered. The similarity is measured by using a distributional thesaurus based on the study described in Lin (1998).

By using this gradual procedure, we constructed case frames from a web corpus. The case frames were obtained from approximately 1.6 billion sentences extracted from the web. They consisted of 43,000 predicates, and the average number of case frames for a verb was 22.2. In Table 1, some examples of the resulting case frames of the verb *yaku* are listed.

4.2.2 Cooccurrences of Noun-noun Modifications

Adnominal nouns have selectional preferences to nouns, and thus this characteristic is useful for coordination disambiguation. We collect dependency relations between nouns, which have the form of "N_1-*no* N_2" (N_2 of N_1), from automatic parses of a large corpus. We performed this extraction using the web corpus of 1.6 billion sentences, and obtained 55.5 million unique dependency relations. We keep a cooccurrence frequency for each relation.

4.2.3 Cooccurrences of Coordinated Nouns

Some nouns are likely to be coordinated. We use this characteristic as lexical parallelism. We collect cooccurrences of coordinated nouns from automatic parses of a large corpus. We extracted 54.1 million unique noun pairs from the web corpus of 1.6 billion sentences.

4.3 Our Model

We employ the probabilistic generative model of dependency and case structure analysis (Kawahara and Kurohashi, 2008) as a base model. This base model resolves coordination ambiguities only on the basis of selectional preferences of predicates and nouns on which conjuncts depend. To capture syntactic parallelism, we integrate the synchronized generation process into the base model. Lexical parallelism is also factored within the generation of pre-conjuncts of coordinate structures.

Our model assigns a probability to each possible dependency structure, T, and case structure, L, of the input sentence, S, and outputs the dependency and case structure that have the highest probability. In other words, the model selects the dependency structure T_{best} and the case structure L_{best} that maximize the probability $P(T, L|S)$ or its equivalent, $P(T, L, S)$, as follows:

$$\begin{aligned}(T_{best}, L_{best}) &= \mathrm{argmax}_{(T,L)} P(T, L|S) \\ &= \mathrm{argmax}_{(T,L)} \frac{P(T, L, S)}{P(S)} \\ &= \mathrm{argmax}_{(T,L)} P(T, L, S). \quad (1)\end{aligned}$$

The last equation follows from the fact that $P(S)$ is constant.

In the model, a clause (or predicate-argument structure) is considered as a generation unit and the input sentence is generated from the root of the sentence. The probability $P(T, L, S)$ is defined as the product of the probabilities of generating clauses C_i as follows:

$$P(T, L, S) = \prod_{C_i \in S} P(C_i|C_h), \quad (2)$$

where C_h is the modifying clause of C_i. Since the Japanese language is head final, the main clause at the end of a sentence does not have a modifying head; we account for this by assuming $C_h = \mathrm{EOS}$ (End Of Sentence).

The probability $P(C_i|C_h)$ is defined in a manner similar to that in Kawahara and Kurohashi (2008). This probability is calculated as the product of generative probabilities of a case frame, its case slots and governed argument nouns. The differences between the probability in the above

study and that in our study are the generative probability of case slots and the generative probability of argument nouns. We describe these two probabilities in the following sections.

4.3.1 Generative Probability of Case Slot

In the base model, the generative probability of case slots is defined as follows:

$$P(A(s_j)=\{\mathrm{Y},\mathrm{N}\}|CF_l), \quad (3)$$

where CF_l is a case frame; s_j is a case slot of the case frame CF_l; and $A(s_j)$ is a binary function that returns Y if a case slot s_j is filled with an argument; otherwise, N.

In our model, if the target predicate or noun does not constitute a coordinate structure, we use the probability (3) for the case slot generation. If the target predicate or noun constitutes a coordinate structure and has a pre-conjunct, we use the following modified probability that depends on whether the same case slot of a pre-conjunct is filled.

$$P(A(s_j)=\{\mathrm{Y},\mathrm{N}\}|CF_l, A_c(s_j)=\{\mathrm{Y},\mathrm{N}\}), \quad (4)$$

where $A_c(s_j)$ represents the situation of the same case slot of the pre-conjunct.

In practice, to avoid the data sparseness problem, we interpolate this probability, which is conditioned on case frames, with the probability conditioned on predicates in the same manner as in Collins (1999).

4.3.2 Generative Probability of Argument Nouns

In the base model, the generative probability of argument nouns in a clause is defined as the product of the generative probability of an argument noun $P_{n_{jk}}$:

$$\prod_{s_j:A(s_j)=Y} \prod_{n_{jk}\in N_{s_j}} P_{n_{jk}}, \quad (5)$$

where N_{s_j} is a set of nouns including a noun filled in the case slot s_j and its coordinated nouns. The generative probability of an argument noun is given as follows:

$$P_{n_{jk}} = P(n_{jk}|CF_l, s_j). \quad (6)$$

In our model, the direct argument noun filled in the case slot s_j is generated with the above probability. The coordinated nouns, which have no direct dependency relation to the predicate, are generated with the following probability:

$$P'_{n_{jk}} = \sqrt{P(n_{jk}|CF_l, s_j) \times P(n_{jk}|n_{jh}, \mathsf{CNJ})}, \quad (7)$$

	# of sents.	# of coord.	# of words in coord.
newspaper	1,000	630	14.7
patent	1,000	1,264	14.8
web	759	453	11.4

Table 2: Statistics of three test sets: the number of sentences, the number of coordinate structures and the average number of words that constitute a coordinate structure. Since a sentence can contain more than one coordinate structure, the number of coordinate structures in the patent set is larger than the number of sentences.

where n_{jh} is a head of n_{jk}, which constitutes a coordinate structure (designated as CNJ) with n_{jh}.

For instance, in Figure 1, the probability of generating *kenkou* (health) and *heiwa* (peace) from the verb *inoru* (pray) is written as follows:[5] [6]

$$P(\text{peace}|CF_{pray}, \mathsf{ACC}) \times \sqrt{P(\text{health}|CF_{pray}, \mathsf{ACC}) \times P(\text{health}|\text{peace}, \mathsf{CNJ})}.$$

This probability is estimated on the basis of the cooccurrence data of coordinated nouns described in section 4.2.3.

4.4 Practical Issue

The proposed model considers all the possible dependency structures including coordination ambiguities. To reduce this high computational cost, we introduced the CKY framework to the search (Eisner, 1996).

5 Experiments

5.1 Experimental Settings

We evaluated the dependency structures that were output by our proposed model. The necessary lexical resources for this parser, which include case frames, statistics of noun-noun modifications and coordinated nouns, and lexical parameters of our model, were acquired from automatic parses of 1.6 billion Japanese sentences crawled from the web (Kawahara and Kurohashi, 2006).

[5] In the probabilities in Figure 1, "pray" is used instead of "CF_{pray}" for simplicity.

[6] This probability can be intuitively understood from the approximation: $P(\text{peace}, \text{health}|\text{pray}) = P(\text{peace}|\text{pray}) \times \sqrt{P(\text{health}|\text{pray}, \text{peace})^2} \approx P(\text{peace}|\text{pray}) \times \sqrt{P(\text{health}|\text{pray}) \times P(\text{health}|\text{peace})}$.

		pref (baseline)	pref+parallelism	improve
newspaper	all	7,356/8,248 (89.2%)	7,398/8,248 (89.7%)	0.5%**
	coordination key	1,226/1,592 (77.0%)	1,251/1,592 (78.6%)	1.6%**
	coordination scope	2,291/2,631 (87.1%)	2,320/2,631 (88.2%)	1.1%**
patent	all	9,758/11,318 (86.2%)	9,852/11,318 (87.0%)	0.8%**
	coordination key	1,839/2,528 (72.7%)	1,887/2,528 (74.6%)	1.9%**
	coordination scope	3,776/4,573 (82.6%)	3,839/4,573 (83.9%)	1.3%**
web	all	4,563/5,114 (89.2%)	4,584/5,114 (89.6%)	0.4%**
	coordination key	893/1,125 (79.4%)	906/1,125 (80.5%)	1.1%*
	coordination scope	1,242/1,462 (85.0%)	1,257/1,462 (86.0%)	1.0%*

Table 3: Dependency accuracies of "pref" (baseline) and "pref+parallelism" (proposed) in the domains of newspapers, patents and web. ** and * represent statistically significant with $p < 0.01$ and with $p < 0.05$, respectively.

The parameters related to unlexical types were calculated from a training part of the Kyoto University Text Corpus. The Kyoto University Text Corpus is syntactically annotated in dependency formalism, and consists of 40K Japanese newspaper sentences. The training part is the remaining part excluding the test 1,000 sentences that are described below.

To evaluate the effectiveness of our model, our experiments were conducted using three test sets: newspaper set, patent set and web set. Table 2 lists some statistics of these test sets. As the newspaper set, we randomly extracted 1,000 sentences from the Kyoto University Text Corpus. The patent set consists of 1,000 sentences drawn from 2004's patent filings of the domain of "*Microbe/Ferment*." The web set consists of 759 sentences from the web, which are not included in the raw corpus of 1.6 billion sentences. This web set is the same as the test set used in previous studies. All the test sets follow the annotation criteria of the Kyoto University Text Corpus. As the input of our experiments, all the test sets were automatically segmented and tagged using the JUMAN morphological analyzer.[7]

We used the probabilistic generative model of dependency and case structure analysis (Kawahara and Kurohashi, 2008) as a baseline system for the purpose of comparison. This parser resolves coordination ambiguities based only on selectional preferences. We use the above-mentioned case frames in the baseline parser, which also requires automatically constructed case frames.

[7]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

5.2 Evaluation

We evaluated the dependency structures analyzed by the proposed model and the baseline model. The dependency structures obtained were evaluated with regard to unlabeled dependency accuracy — the proportion of correct dependencies out of all dependencies.

Table 3 lists the dependency accuracies. In this table, "pref" represents the baseline model, which is the probabilistic parser of dependency and case structure with only selectional preferences, and "pref+parallelism" represents our proposed model. "all" represents the overall dependency accuracies. The proposed model significantly outperformed the baseline system in all the sets (McNemar's test; $p < 0.01$).

In Table 3, the dependency accuracies are further classified into *coordination key* and *coordination scope*. Coordination key means the dependency accuracy of coordination key *bunsetsus*, which possibly lead coordinate structures. Coordination scope means the dependency accuracy of *bunsetsus* inside coordinate structures of the manual annotation.

5.3 Discussions

In the newspaper and patent sets, in particular, the accuracies of both coordination key and coordination scope were improved by 1.1% to 1.9%. These improvements were conduced by the consideration of syntactic and lexical parallelism. In the web set, the accuracies of coordination related dependencies were less improved than those of the newspaper and patent sets.

Figure 2 shows improved analyses; here, the dotted lines represent the analysis performed using the baseline "pref," and the solid lines rep-

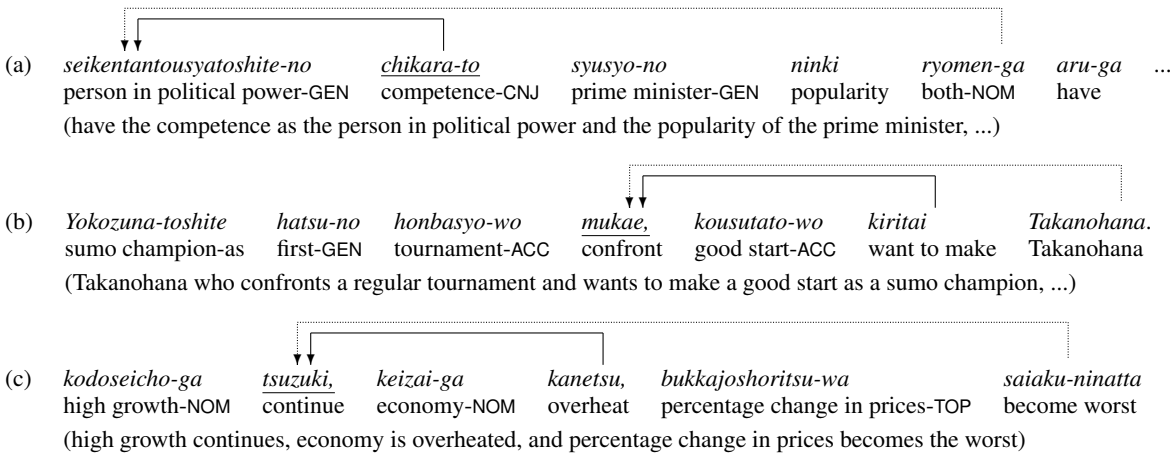

Figure 2: Improved examples. The dotted lines represent the results of "pref," and the solid lines, which are correct dependencies, represent the analysis of "pref+parallelism." The underlined *bunsetsu*s represent coordination key *bunsetsu*s.

resent the analysis performed using the proposed method, "pref+parallelism." These sentences are incorrectly analyzed by the baseline but correctly analyzed by the proposed method. For example, in sentence (a), the head of *seikentantousyatoshite-no* (person in political power-GEN) was correctly judged as *chikara-to* (competence-CNJ). This is because the two genitive (GEN) *bunsetsu*s were synchronously generated to prefer syntactic parallelism.

The proposed model did not largely outperform the baseline in the web set. One of the reasons of this result was due to weak parallelism in the web set. We found that coordinate structures in the newspaper and patent sets tend to have greater syntactic parallelism than those in the web set. The average number of words that constitute a coordinate structure in the newspaper set was 14.7 and that in the patent set was 14.8, whereas that in the web set was 11.4, as shown in Table 2. Therefore, it was hard to show substantial improvement by considering such weak parallelism of coordinate structures in the web set.

In order to compare our results with a discriminative dependency parser, we input the patent set and the web set into an SVM-based Japanese dependency parser, CaboCha (Kudo and Matsumoto, 2002),[8] which was trained using the Kyoto University Text Corpus.[9] Its dependency accuracies were 86.3% (9,770/11,320) for the patent set and 88.7% (4,534/5,114) for the web set, which are lower than those of our proposed model. This low performance can be attributed to the lack of sufficient consideration of both parallelism and selectional preferences, as mentioned in Sassano (2004). Another cause of the low performance is the out-of-domain training corpus. This SVM-based parser was trained on a newspaper corpus, while the test sets were obtained from patent filings and the web because tagged corpora of these domains that are large enough to train a supervised parser are not available. In other words, our proposed model achieved a good performance on the patent set without using in-domain corpora.

6 Conclusion

In this paper, we have proposed a unified generative model of coordination that simultaneously considers parallelism and selectional preferences. Syntactic parallelism is modeled by the synchronized generation process of pre-modifiers of conjuncts, and lexical parallelism was factored within the generation of pre-conjuncts. Selectional preferences are acquired from large raw corpora as case frames and statistics of noun-noun modifications. The experimental results indicate the effectiveness of our model, particularly in the domains of newspapers and patents. The acquired case frames can be obtained from a non-profit organization and our analysis system will be freely available at our web site. Our future research involves incorporating ellipsis resolution to develop an integrated model for syntactic, case, and ellipsis analyses.

[8] http://chasen.org/~taku/software/cabocha/

[9] We did not input the newspaper set into CaboCha, because it is included in the training corpus used in CaboCha.

References

Rajeev Agarwal and Lois Boggess. 1992. A simple but useful approach to conjunct identification. In *Proceedings of ACL1992*, pages 15–21.

Ekaterina Buyko and Udo Hahn. 2008. Are morpho-syntactic features more predictive for the resolution of noun phrase coordination ambiguity than lexico-semantic similarity scores? In *Proceedings of COLING2008*, pages 89–96.

Francis Chantree, Adam Kilgarriff, Anne de Roeck, and Alistair Wills. 2005. Disambiguating coordinations using word distribution information. In *Proceedings of RANLP2005*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL2005*, pages 173–180.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Amit Dubey, Patrick Sturt, and Frank Keller. 2005. Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 827–834.

Amit Dubey, Frank Keller, and Patrick Sturt. 2006. Integrating syntactic priming into an incremental probabilistic parser, with an application to psycholinguistic modeling. In *Proceedings of COLING-ACL2006*, pages 417–424.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of COLING-96*, pages 340–345.

Kazuo Hara, Masashi Shimbo, Hideharu Okuma, and Yuji Matsumoto. 2009. Coordinate structure analysis with global structural constraints and alignment-based local features. In *Proceedings of ACL-IJCNLP2009*, pages 967–975.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proceedings of ACL2007*, pages 680–687.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Proceedings of LREC2006*.

Daisuke Kawahara and Sadao Kurohashi. 2008. Coordination disambiguation without any similarities. In *Proceedings of COLING2008*, pages 425–432.

Sandra Kübler, Erhard Hinrichs, Wolfgang Maier, and Eva Klett. 2009. Parsing coordinations. In *Proceedings of EACL2009*, pages 406–414.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of CoNLL2002*, pages 29–35.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534.

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of LREC1998*, pages 719–724.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*, pages 768–774.

Preslav Nakov and Marti Hearst. 2005. Using the web as an implicit training set: Application to structural ambiguity resolution. In *Proceedings of HLT-EMNLP2005*, pages 835–842.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

Manabu Sassano. 2004. Linear-time dependency analysis for Japanese. In *Proceedings of COLING2004*, pages 8–14.

Masashi Shimbo and Kazuo Hara. 2007. A discriminative learning model for coordinate conjunctions. In *Proceedings of EMNLP-CoNLL2007*, pages 610–619.

Gertjan van Noord. 2007. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of IWPT2007*, pages 1–10.