Toward Text Understanding: Integrating Relevance-tagged Corpus and Automatically Constructed Case Frames

Daisuke Kawahara, Ryohei Sasano, Sadao Kurohashi

Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, Japan
{kawahara, sasano, kuro}@kc.t.u-tokyo.ac.jp

Abstract

This paper proposes a wide-range anaphora resolution system toward text understanding. This system resolves zero, direct and indirect anaphors in Japanese texts by integrating two sorts of linguistic resources: a hand-annotated corpus with various relations and automatically constructed case frames. The corpus has relevance tags which consist of predicate-argument relations, relations between nouns and coreferences, and is utilized for learning parameters of the system and testing it. The case frames are indispensable knowledge both for detecting zero/indirect anaphors and estimating appropriate antecedents. Our preliminary experiments showed promising results.

1. Introduction

Text understanding is one of the ultimate goals of natural language processing. The first step for text understanding is to grasp various explicit/implicit relations in texts, such as syntactic relations, coreferences, and antecedents of indirect anaphora. Syntactic relation analysis, i.e. parsing, has achieved great success both in English and Japanese. Anaphora resolution, i.e. direct anaphora (coreference) resolution and indirect anaphora (bridging reference) resolution, in English is different from that in Japanese as shown in Table 1.

In English, direct anaphors consist mainly of pronouns and definite noun phrases, and has achieved some success by machine learning techniques based on linguistic clues, such as definiteness, number, and gender (Yang et al., 2003). On the other hand, indirect anaphora resolution is much more difficult, and a part of this phenomenon has been studied (Poesio et al., 2002).

In Japanese, both direct and indirect anaphora resolution are difficult. Direct anaphors are rarely expressed as pronouns, and become zero anaphors. This induces a big problem of detecting zero anaphors. To address this problem, elaborate knowledge for each verb is required. This observation applies to indirect anaphora resolution. That is, indirect anaphors are cast as zero anaphors of nouns, and can be detected by knowledge for each noun.

As for such knowledge, case frames can be employed. They describe what kinds of relations (case slots) each verb/noun has and what kinds of words can fill each case slot. The case frames can be utilized to detect zero/indirect anaphors and furthermore find their appropriate antecedents. In addition, a corpus in which many relations in texts are annotated is utilized for learning parameters of the system, testing and evaluating it.

This paper proposes a wide-range anaphora resolution system, which can resolve zero, direct and indirect anaphora in Japanese texts, based on the two kinds of resources: "Relevance-tagged corpus" and automatically constructed case frames. "Relevance-tagged corpus" is a handmade corpus with relevance tags that consist of predicate-argument relations, coreferences, and relations between nouns (Kawahara et al., 2002). The case frames, which are constructed from large corpora, describe relations between words and what kinds of words each word is related to (Kawahara and Kurohashi, 2002).

Table 1: Anaphora resolution in English and Japanese

	direct anaphora	indirect anaphora
E	ANT ··· V pronoun the NP	ANT ··· the NP
J	ANT ··· V ϕ	ANT ··· N (of ϕ)

2. Relevance-tagged corpus

"Relevance-tagged corpus" currently consists of about 5,000 sentences of 400 Japanese newspaper articles. Its annotation has three classes of relations: predicate-argument relations, coreferences, and relations between nouns.

2.1. Predicate-argument relations

In Japanese, postpositions function as case markers such as *ga* (nominative), *wo* (accusative), and *ni* (dative)[1]. To annotate predicate-argument relations, we give the predicate a tag that consists of an argument word and a case-marking relation (postposition itself).

For example, in Figure 1, *Ichiro* and *shimbun* 'newspaper' modify *yonde* 'read', and are arguments of *yonde*. The relation between *shimbun* and *yonde* is *wo* (accusative), which is indicated by the postposition following *shimbun*. Accordingly, the tag "*wo*:*shimbun*" is given to *yonde*.

In addition, *Ichiro* modifies *yonde*, but the relation between them is hidden by a topic marker (TM) *wa*. Since this *wa* functions as nominative, "*ga*:*Ichiro*" is given to *yonde*.

For *suteta* 'throw away', its nominative and accusative are zero anaphors. Since their antecedents are *Ichiro* and *shimbun*, respectively, the tags "*ga*:*Ichiro*" and "*wo*:*shimbun*" are given to *suteta*.

[1] In the examples of this paper, we use the abbreviations of the cases: nom (nominative), acc (accusative), dat (dative).

(1) *Ichiro-wa* *shimbun-wo* *yonde* *suteta*.
Ichiro-nom newspaper-acc read throw away

		read	throw away
		↑	↑
		ga:*Ichiro* *wo*:*shimbun*	*ga*:*Ichiro* *wo*:*shimbun*

(*Ichiro* read a newspaper and threw (it) away.)

(2) *Shikashi* *imouto-wa* *sore-wo* *yomi-takatta*.
but sister-TM it-acc want to read

but	sister-TM	it-acc	want to read
	↑	↑	↑
	no:*Ichiro*	=:*shimbun*	*ga*:*imouto* *wo*:*sore*

(But (his) sister wanted to read it.)

Figure 1: Tagging example

2.2. Relations between nouns

Not only predicates but also nouns have some intrinsic relations with other nouns in a text. When two nouns in a text are related to each other, a tag is given to the latter noun.

In the second sentence of Figure 1, since *imouto* 'sister' means "*Ichiro no imouto*" '*Ichiro's* sister', the tag "*no*:*Ichiro*" is given to *imouto*, though "*Ichiro no*" does not appear in the sentence. In this example, *imouto* requires intrinsic relations to other nouns. This is a so-called relational noun.

Not only relational nouns but also almost all of nouns have some intrinsic relations: *kuruma* 'car' and *handle*, *mado* 'window' and *curtain*. We also handle these relations.

2.3. Coreferences

When two nouns refer to the same entity, these two nouns are coreferential. To mark a coreference relation, "=" is used. A tag of this relation is given to the latter noun of two coreferential nouns.

In Figure 1, *sore* 'it' refers to *shimbun* 'newspaper', and the tag "=:*shimbun*" is given to *sore*.

3. Automatic case frame construction

To realize text understanding, world knowledge is indispensable. As to this knowledge, we exploit case frames, which describe relations between words and what kinds of words each word is related to. We construct the case frames for verbs and for nouns using the following two methods.

3.1. Verbal case frames

The biggest problem in the automatic construction of verbal case frames is verb sense ambiguity. Verbs which have different meanings should have different case frames, but it is hard to disambiguate verb senses precisely. To deal with this problem, predicate-argument examples which are collected from a large corpus are distinguished by coupling a verb and its closest case component. That is, examples are not distinguished by verbs such as *naru* 'make/become' and *tsumu* 'load/accumulate', but by couples such as "*tomodachi ni naru*" 'make a friend', "*byouki ni naru*" 'become sick', "*nimotsu wo tsumu*" 'load baggage', and "*keiken wo tsumu*" 'accumulate experience'.

Table 2: Verbal case frame examples

	CM	examples
youritsu (1) 'support'	*ga* *wo* *ni*	<agent>, group, party, ··· <agent>, candidate, applicant <agent>, district, election, ···
youritsu (2) 'support'	*ga* *wo* *ni*	<agent> <agent>, assemblyman, minister, ··· <agent>, candidate, successor, ···
⋮	⋮	⋮

Table 3: Nominal case frame examples

	case slot	examples
hyoujou 'expression'	[one] [feelings]	people, partner, ··· relief, margin, ···
hisashi (1)	[house, window]	parking, store, hall, ···
hisashi (2)	[cap]	cap, helmet, ···
hikidashi (1)	[desk, chest]	desk, chest, dresser, ···
hikidashi (2)	⟨other⟩	credit, fund, saving, ···
coach	[sport] ⟨belonging⟩	baseball, swimming, ··· team, club, ···

† *hisashi* means 'eaves/visor', and *hikidashi* means 'drawer'.

This process makes separate case frames which have almost the same meaning or usage. For example, "*nimotsu wo tsumu*" 'load baggage' and "*busshi wo tsumu*" 'load supply' are separate case frames. To merge these similar case frames and increase coverage of the case frame, the case frames are clustered.

To sum up, the procedure for the automatic construction of verbal case frames is as follows.

1. A large raw corpus is parsed by the Japanese parser, KNP, and reliable predicate-argument examples are extracted from the parse results.
2. The extracted examples are bundled according to the verb and its closest case component.
3. The case frames are clustered using a similarity measure function, resulting in the final case frames. The similarity is calculated using a Japanese thesaurus (Ikehara et al., 1997), and its maximum score is 1.0. The details of the similarity measure function are described in (Kawahara and Kurohashi, 2002).

We constructed verbal case frames by this procedure from newspaper articles of 25 years (about 25,000,000 sentences). The result consists of 23,000 predicates, and the average number of case frames for a predicate is 14.5. In Table 2, some examples of the resulting case frames are shown.

3.2. Nominal case frames

In the case of verbs, syntactic structures such as subject/object/PP in English or case markers such as *ga*, *wo*, *ni* in Japanese can be utilized as a strong clue to distinguish several obligatory cases and adjuncts (and adverbs), which makes it feasible to construct the case frames automatically like above.

On the other hand, in the case of nouns, obligatory cases of noun N_h appear, in most cases, in the single form of noun phrase "N_h of N_m" in English, or "N_m *no* N_h" in Japanese. This single form can express several obligatory cases, and furthermore optional cases, for example, "*rugby no coach*" (obligatory case concerning what sport), "*club no coach*" (obligatory case concerning which institution), and "*kyonen* 'last year' *no coach*" (optional case). Therefore, the key issue to construct nominal case frames is to analyze "N_h of N_m" or "N_m *no* N_h" phrases to distinguish obligatory case examples and others.

Nominal case frames are constructed from large corpora based on an accurate analysis of "N_m *no* N_h" phrases using an ordinary dictionary and a thesaurus (Kurohashi and Sakai, 1999). First, syntactically unambiguous noun phrases "N_m *no* N_h" are collected from the automatic parse results used for the verbal case frames. The extracted noun phrases are analyzed using two methods: dictionary-based analysis (DBA) and semantic feature-based analysis (SBA).

DBA utilizes an ordinary dictionary, because it has obligatory case information of nouns in its definition sentences. For example, "*rugby no coach*" can be interpreted by the definition of *coach* ("a person who teaches technique in some sport") as follows: the dictionary describes that the noun *coach* has an obligatory case of *sport*, and the phrase "*rugby no coach*" specifies that the *sport* is *rugby*. That is, the interpretation of the phrase can be regarded as matching *rugby* in the phrase to *some sport* in the *coach* definition.

Since diverse relations in "N_m *no* N_h" are handled by DBA, the remaining relations can be detected by SBA, that is, simple rules which check the semantic features (in the thesaurus (Ikehara et al., 1997)) of N_m and/or N_h. For example, a rule "N_m:ORGANIZATION, N_h:HUMAN $\rightarrow$ $\langle$belonging$\rangle$" analyzes a phrase "*team no coach*", and we can see that *team* has $\langle$belonging$\rangle$ relation to *coach*.

We constructed nominal case frames by this procedure from newspaper articles of 25 years. The result consists of 17,000 nouns, and the average number of case frames for a noun is 1.06. Some examples of the resulting case frames are shown in Table 3. In this table, "$[\cdots]$" denotes an analysis result by DBA, and "$\langle\cdots\rangle$" denotes an analysis result by SBA.

4. Anaphora resolution system

We build a Japanese anaphora resolution system using "Relevance-tagged corpus" and the case frames. This system simultaneously resolves various anaphora, such as zero, direct, and indirect anaphora. So far, previous researches have tackled each resolution task independently. However, these anaphora should be solved together, because various kinds of relations are related interactively.

For the anaphora resolution, the following two clues can be considered:

- Anaphors and their context have syntactic and semantic constraints to their antecedents.
- Anaphors are likely to have their antecedents in their close position.

As for the first clue, we employ the automatically constructed case frames, which provide wide-coverage and fine-grained selectional restriction.

The second clue, namely the distance tendency, has been tried to capture by previous researches. However, they used only flat distance, such as the number of words or sentences. To model the distance tendency more precisely, we classify locational relations between anaphors and their possible antecedents by considering structures in texts, such as subordinate/main clauses and embedded sentences. Using "Relevance-tagged corpus", we calculate how likely each location has antecedents, and acquire the order of antecedent location preference (Kawahara and Kurohashi, 2004).

In addition to these two devices, we exploit a machine learning technique to consider various features related to the determination of an antecedent, including syntactic constraints, and make a Japanese anaphora resolution system. This system examines candidates in the order of antecedent location preference, and selects as its antecedent the first candidate which is labeled as positive by a machine learner and satisfies the selectional restriction based on the case frames.

The outline of our algorithm is as follows.

1. Parse an input sentence using the Japanese parser, KNP.
2. Process each verb and noun in the sentence from left to right by the following steps.
   - 2.1. Perform the following processes for each case frame of the target verb/noun.
     - i. Match a word which have syntactic relation to the target word with an appropriate case slot of the case frame. Regard case slots that have no correspondence as zero/indirect anaphors.
     - ii. Estimate an antecedent of each anaphor detected.
   - 2.2. Select a case frame which has the highest total score, and output the analysis result for the case frame.

The rest of this section describes the steps (i) and (ii) in detail.

4.1. Matching syntactically related elements with case slots

A word that have syntactic relation to the target word is matched with an appropriate case slot in the case frame.

If the target word is a verb, its syntactically related words are its case components. They are matched against the case frame according to their case markers (Kurohashi and Nagao, 1994).

If the target word is a noun, its syntactically related words are not always case components, but are obligatory or optional elements. To distinguish them, a similarity threshold is employed. That is, a syntactically related word whose similarity to a case slot exceeds a threshold is considered as an obligatory element, namely a case component, and can be assigned to the case slot. The case component is assigned to the most similar case slot among the case slots in the case frame.

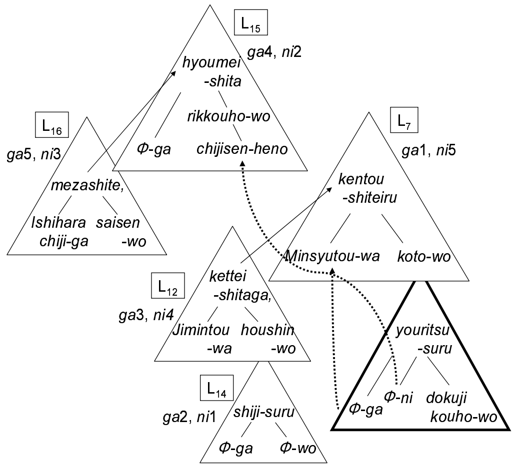

Figure 2: Analysis Example

The result of the above matching process tells if the zero/indirect anaphors exist. That is, vacant case slots in the case frame, which have no correspondence with the input case components, mean zero/indirect anaphors.

For example, in the case of *youritsu* 'support' in Figure 2, *wo* case slot has a corresponding case component, but *ga* and *ni* case slots are vacant. Accordingly, two zero anaphors are identified in *ga* and *ni* case of *youritsu*.

4.2. Antecedent estimation

We estimate antecedents of zero, direct and indirect anaphors based on examples in the case frames and the classifier. We examine possible antecedents in order of the antecedent location preference, and label them positive/negative using the binary classifier. If a possible antecedent is classified as positive and its similarity to examples in its case slot exceeds a threshold, it is determined as the antecedent. At this moment, the procedure finishes, and further candidates are not tested.

For example, *youritsu* 'support' in Figure 2 has zero anaphors in *ga* and *ni*. The ordered possible antecedents for *ga* are L_7:*Minsyutou*, L_{14}:*Jimintou*(ϕ *ga*), L_{14}:"*Ishihara chiji*"(ϕ *wo*), $\cdots$. The first candidate *Minsyutou* (similarity:0.73), which is labeled as positive by the classifier, and whose similarity to the case frame examples exceeds the threshold (0.60), is determined as the antecedent.

5. Experimental results

We conducted two experiments to evaluate the zero anaphora resolution and the indirect anaphora resolution.

5.1. Experimental result of zero anaphora resolution

We ran an experiment on 100 newspaper articles in "Relevance-tagged corpus" to evaluate the zero anaphora resolution. The antecedent location preference and the classifier are learned from 279 newspaper articles. Table 4 shows the experimental result.

5.2. Experimental result of indirect anaphora resolution

We ran an experiment on 10 newspaper articles in "Relevance-tagged corpus" to evaluate the indirect anaphora resolution. The experimental setting is same as the zero anaphora resolution. Table 5 shows the experimental result.

Table 4: Experimental result of zero anaphora resolution

precision	recall	F
515/924 (0.557)	515/1087 (0.474)	0.512

Table 5: Experimental result of indirect anaphora resolution

precision	recall	F
25/45 (0.556)	25/41 (0.610)	0.581

6. Conclusion

We have proposed a anaphora resolution system that resolves zero, direct, and indirect anaphora in Japanese texts. For zero anaphora resolution, the precision and recall were 55.7% and 47.4%. For indirect anaphora resolution, the precision and recall were 55.6% and 61.0%. Major errors are caused by context sensitivity of obligatory cases, multiple candidates with the same semantic feature, and word sense ambiguity in example matching. We plan to investigate resolution errors further to improve the accuracy.

7. References

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentarou Ogura, and Yoshifumi Oyama Yoshihiko Hayashi (eds.), 1997. *Japanese Lexicon*. Iwanami Publishing.

Kawahara, Daisuke and Sadao Kurohashi, 2002. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*.

Kawahara, Daisuke and Sadao Kurohashi, 2004. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*.

Kawahara, Daisuke, Sadao Kurohashi, and Kôiti Hasida, 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Kurohashi, Sadao and Makoto Nagao, 1994. A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. In *IEICE Transactions on Information and Systems*, volume E77-D No.2.

Kurohashi, Sadao and Yasuyuki Sakai, 1999. Semantic analysis of Japanese noun phrases: A new approach to dictionary-based understanding. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira, 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Yang, Xiaofeng, Guodong Zhou, Jian Su, and Chew Lim Tan, 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.