Zero Pronoun Resolution Based on Automatically Constructed Case Frames and Structural Preference of Antecedents

Daisuke Kawahara and Sadao Kurohashi

University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, Japan
{kawahara,kuro}@kc.t.u-tokyo.ac.jp

Abstract. This paper describes a method to detect and resolve zero pronouns in Japanese text. We detect zero pronouns by case analysis based on automatically constructed case frames, and select their appropriate antecedents based on similarity to examples in the case frames. We also introduce structural preference of antecedents to precisely capture the tendency that a zero pronoun has its antecedent in its close position. Experimental results on 100 articles indicated that the precision and recall of zero pronoun detection is 87.1% and 74.8% respectively and the accuracy of antecedent estimation is 61.8%.

1 Introduction

Anaphora resolution is core technology to achieve a breakthrough in natural language applications, such as machine translation, text summarization, and question answering. To resolve anaphoric expressions, the following two clues can be considered:

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

As for the syntactic and semantic constraints, only the coarse constraints have been used so far. For instance, some previous researches used shallow semantic classes, such as human, organization, and object, and considered the agreement between the classes of an anaphor and its antecedent as the semantic constraints (e.g. [1–3]). The reason why only these coarse constraints have been used is that knowledge bases which provide precise selectional restriction have not been available. Recently, wide-coverage case frames have been constructed automatically from large corpora, and provide fine-grained selectional restriction [4]. We employ these case frames for the selectional restriction.

The case frames are also necessary to detect zero pronouns. A case frame has the information about case markers that a verb subcategorizes. By matching an input predicate-argument structure with a case frame, we can recognize case slots that have no correspondence with the input as zero pronouns. Previous work assumed perfect pre-detection of zero pronouns, or detected zero pronouns based

K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, pp. 12–21, 2005.

on hand-crafted sparse case frames [5]. On the other hand, we utilize the automatically constructed case frames, which enable accurate zero pronoun detection.

The second clue for anaphora resolution, i.e. distance tendency, has been attempted to capture by previous researches (e.g. [6, 3, 5]). They incorporated distance between an anaphor and its antecedent into a feature of machine learning techniques or a parameter of probabilistic models. The biggest problem with these approaches is that they do not consider structures in texts to measure distance, but rather just a flat distance, such as the number of words or sentences. To model the distance tendency precisely, we classify locational relations between zero pronouns and their possible antecedents by considering structures in texts, such as subordinate clauses, main clauses, and embedded sentences. We calculate how likely each location has antecedents using an annotated corpus, and acquire structural preference of antecedents.

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 propose a Japanese zero pronoun resolution system. We concentrate on zero pronouns, because they are much more popular than any other anaphoric expressions in Japanese. This system examines candidates in an increasing order of structural preference of antecedents, 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.

2 Zero Pronoun Resolution Based on Case Frames

We employ the automatically constructed case frames [4] for zero pronoun detection and selectional restriction that antecedents must agree. This section firstly outlines the method of constructing the case frames, and then describes the case analysis based on them and the zero pronoun detection using the case analysis results.

2.1 Automatic Construction of Case Frames

The biggest problem in automatic case frame construction 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 (e.g. "*tsumu*" (load/accumulate)), but by couples (e.g. "*nimotsu-wo tsumu*" (load baggage) and "*keiken-wo tsumu*" (accumulate experience)).

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 similar, but have separate case frames. To cope with this problem, the case frames are clustered using a similarity measure function. The similarity is calculated using a Japanese thesaurus, and its maximum score is 1.0. The details of this measure are described in [4].

Table 1. Case frame examples.

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

We constructed case frames by this procedure from newspaper articles of 20 years (about 21,000,000 sentences). The result consists of 23,000 predicates, and the average number of case frames for a predicate is 14.5. In Table 1, some examples of the resulting case frames are shown.

2.2 Zero Pronoun Resolution Based on the Case Frames

We build a zero pronoun resolution system that utilizes the case frames and the structural preference of antecedents, which is stated in Section 3. The outline of our algorithm is as follows.

1. Parse an input sentence using the Japanese parser, KNP.
2. Process each verb in the sentence from left to right by the following steps.
   2.1. Narrow case frames down to corresponding ones to the verb and its closest case component.
   2.2. Perform the following processes for each case frame of the target verb.
      i. Match each input case component with an appropriate case slot of the case frame. Regard case slots that have no correspondence as zero pronouns.
      ii. Estimate an antecedent of each zero pronoun.
   2.3. 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 above steps (2.1) and (2.2.i) in detail.

Narrowing Down Case Frames

As stated in Section 2.1, the closest case component plays an important role to determine the usage of a verb. In particular, when the closest case is "*wo*" or "*ni*", this trend is clear-cut. In addition, an expression whose nominative belongs to <agent> (e.g. "<agent> has accomplished"), does not have enough clue to decide its usage, namely a case frame. By considering these aspects, we impose the following conditions on narrowing down case frames.

...
(1) *Ishihara chiji-ga saisen-wo mezashite, chijisen-heno rikkouho-wo hyoumei-shita.*
governor reelection aim gov. election candidacy announce
(The governor Ishihara announced his candidacy for reelection to the governor.)
(2) *Jimintou-wa shiji-suru houshin-wo kettei-shitaga, Minsyutou-wa*
Liberal Democratic Party support policy decide Democratic Party
dokuji kouho-wo youritsu-suru koto-wo kentou-shiteiru.
original candidate support (that) examine
(The Liberal Democratic Party decided to support him, but the Democratic Party is examining to support its original candidate.)
...

Fig. 1. An example article.

- The closest case component exists, and must immediately precede its verb.
- The closest case component and the closest case meet one of the following conditions:
  - The closest case is "*wo*" or "*ni*".
  - The closest case component does not belong to the semantic marker <agent>.
- A case frame with the closest case exists, and the similarity between the closest case component and examples in the closest case exceeds a threshold.

We choose the case frames whose similarity is the highest. If the above conditions are not satisfied, case frames are not narrowed down, and the subsequent processes are performed for each case frame of the target verb. The similarity used in this process is defined as the best similarity between the closest case component and examples in the case slot.

Let us consider "*youritsu*" (support) in the second sentence of Fig.1. "*youritsu*" has the case frames shown in Table 1. The input expression "*kouho-wo youritsu*" (support a candidate) satisfies the above two conditions, and the case frame "*youritsu* (1)" meets the last condition. Accordingly, this case frame is selected.

Matching Input Case Components with Case Slots in the Case Frame

We match case components of the target verb with case slots in the case frame. When a case component has a case marker, it must be assigned to the case slot with the same case marker. When a case component is a topic marked phrase or a clausal modifiee, which does not have a case marker, it can be assigned to some pre-defined case slots described in [7].

The result of case analysis tells if the zero pronouns exist. That is, vacant case slots in the case frame, which have no correspondence with the input case components, mean zero pronouns. In this paper, we concentrate on three case slots: "*ga*", "*wo*", and "*ni*".

In the case of "*youritsu*" (support) in Fig.1 and the selected case frame "*youritsu* (1)", "*wo*" case slot has a corresponding case component, but "*ga*" and "*ni*" case slots are vacant. Accordingly, two zero pronouns are identified in "*ga*" and "*ni*" case of "*youritsu*".

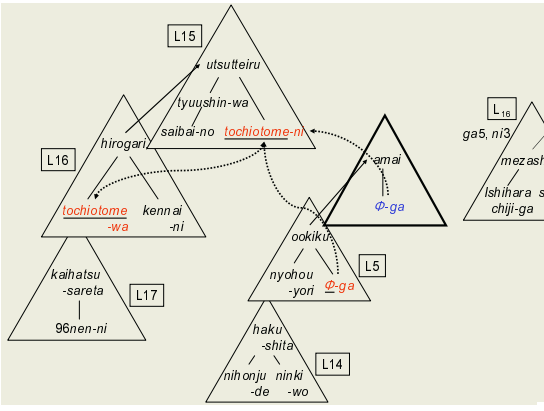


Fig. 2. How to handle antecedents.

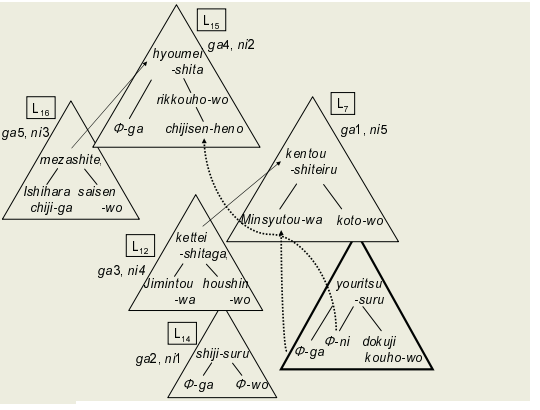


Fig. 3. Location classes in case of V_z = *youritsu*.

The procedure for estimating antecedents of detected zero pronouns is described in Section 5.

3 Learning Structural Preference of Antecedents

According to the selectional restriction of the case frames, possible antecedents are restricted to the eligible ones, but more than one possible antecedent still remain in many cases. To narrow down possible antecedents further, we exploit the distance tendency that zero pronouns are likely to have their antecedents in their close position. Previous researches measured the closeness by flat distance, such as the number of words or sentences between zero pronouns and their antecedents, and did not consider structures in texts. To model the distance tendency precisely, we classify locational relations between zero pronouns and their possible antecedents by considering the structures in texts, such as subordinate clauses, main clauses, and embedded sentences. We call the classification of the locational relations **location classes**, and calculate how likely each location class has antecedents based on an annotated corpus. Ordering these likelihoods yields the structural preference of antecedents, which is exploited in our zero pronoun resolution system.

This section describes how to handle antecedents in a training corpus, and then introduces the location classes. Finally, we illustrate how to calculate the structural preference of antecedents, namely the ordering of the location classes.

3.1 Handling Antecedents in the Relevance-Tagged Corpus

We learn the ordering of location classes using "Relevance-tagged Corpus" [8]. This corpus consists of Japanese newspaper articles, and has several types of relevance tags, such as predicate-argument relations, relations between nouns, and coreferences.

Table 2. Location classes of antecedents.

the sentence under consideration	
L_1 : case components of "parent predicate of V_z"	MC
L_2 : case components of "parent predicate of V_z"	
L_3 : case components of "parent predicate of V_z"	MC, P
L_4 : case components of "parent predicate of V_z"	P
L_5 : case components of "child predicate of V_z"	
L_6 : case components of "child predicate of V_z"	P
L_7 : case components of "parent predicate of parent noun phrase of V_z"	MC
L_8 : case components of "parent predicate of parent noun phrase of V_z"	
L_9 : case components of "parent predicate of parent predicate of V_z"	MC
L_{10}: case components of "parent predicate of parent predicate of V_z"	
L_{11}: case components of "predicate of main clause"	MC
L_{12}: case components of "predicate of subordinate clause depending on main clause"	
L_{13}: other noun phrases following V_z	
L_{14}: other noun phrases preceding V_z	
1 sentence before	
L_{15}: case components of "predicate of main clause"	MC
L_{16}: case components of "predicate of subordinate clause depending on main clause"	
L_{17}: other noun phrases	
2 sentences before	
L_{18}: case components of "predicate of main clause"	MC
L_{19}: case components of "predicate of subordinate clause depending on main clause"	
L_{20}: other noun phrases	

We investigated 379 articles, consisting of 3,695 sentences, in the corpus. There are 11,149 predicates, including verbs, adjectives, and noun+copulas. Out of these predicates, 5,530 predicates have zero pronouns, and there are 6,602 zero pronouns in total. Out of these zero pronouns, 4,986 zero pronouns have their antecedents in their articles. The remaining 1,616 zero pronouns have no antecedents in their articles, and in many cases their referents are unspecified people, that are equivalent to general pronouns in English (e.g. "They say that $\cdots$").

As to a zero pronoun which has its antecedent in the article, its antecedents are not only the directly annotated one but also indirect ones which are linked by other coreference links. In other words, these indirect antecedents are the entities which corefer to the annotated antecedent, and other zero pronouns which refer to the same referent. We handle them equally to the annotated one, because this treatment is natural to measure the distance between a zero pronoun and its antecedents. For instance, "*amai*" (sweet) in Fig.2 has a zero pronoun in its "*ga*" case, and its antecedent is "*tochiotome*" in the main clause of one sentence before. Since this antecedent "*tochiotome*" is coreferential to "*tochiotome*" in the subordinate clause of one sentence before, we regard the latter also as the antecedent. Besides, "*ookiku*" in the target sentence has a zero pronoun in its "*ga*" case, which refers "*tochiotome*", and we regard this zero pronoun as the antecedent, too.

For learning structural preference of antecedents and building a classifier, stated in the following sections, 279 articles in the corpus are used, and the rest 100 articles are reserved for the experiment.

3.2 Setting Up Location Classes

To model the distance tendency precisely, we introduce **location classes**, which are the classification of locational relations between zero pronouns and their

antecedents. Considering subordinate clauses, main clauses, embedded sentences, and so on, we established 20 location classes as described in Table 2. In Table 2, V_z means a predicate that has a zero pronoun. We call a predicate whose case component is an antecedent V_a, which is quoted in Table 2. "MC" means that V_a constitutes the main clause, and "P" means that V_z and V_a are conjunctive.

For example, let us consider "*amai*" (sweet) in Fig.2, which has a zero pronoun for "*ga*". "*utsutte-iru*" (move) is a main clause of one sentence before, and the case components are of "*utsutte-iru*", i.e. "*tyuushin*", "*saibai*", and "*tochiotome*", are located in L_{15}. "*ookiku*" (big) is a child clause of the target verb "*amai*", and the case components of "*ookiku*", i.e. "*nyohou*" and a zero pronoun referring "*tochiotome*", are in L_5.

3.3 Ordering Location Classes

We investigate how each location class is likely to have antecedents using the corpus. We calculate score of location class L as follows:

$$\frac{\#\text{ of antecedents in } L}{\#\text{ of possible antecedents in } L}$$

For a zero pronoun of "*amai*" in Fig.2, the possible antecedents in L_{15} are "*tyushin*", "*saibai*", and "*tochiotome*". In this case, since "*tochiotome*" is the antecedent, this location class has one antecedent, and three possible antecedents. These numbers are counted through the whole corpus, and the score of the location class is calculated by the above formula.

We then sort these scores and obtain location class order for each case markers of zero pronouns. Fig.3 shows the location class orders for the zero pronouns ("*ga*" and "*ni*") of "*youritsu*" (support). The first location classes for "*ga*" and "*ni*" are L_7 and L_{14}, respectively, and this exemplifies that the location class order of each case marker is different from the other orders. In addition, L_{12} is close to the target zero pronoun in the flat distance, but is not placed high in the location class order.

4 Building a Classifier

We utilize a machine learning technique to consider a lot of factors related to antecedent estimation. We employ a binary classifier to judge if a possible antecedent is eligible for an antecedent. The classifier is trained using the features shown in Table 3.

The classifier is trained using Support Vector Machines (SVM). Training data are created from "Relevance-tagged Corpus". We treat the closest correct antecedent as a positive example and possible antecedents between a zero pronoun and the positive antecedent as negative examples. Zero pronouns that have their closest correct antecedents in more than two sentences before are not used for training. In the case of "*amai*" (sweet) in Fig.2, two "*tochiotome*" in one sentence before and the zero pronoun referring "*tochiotome*" in "*ookiku*" are positive examples, and the other nouns are negative.

Table 3. Feature set for the classifier.

features related to both zero pronoun and possible antecedent
· similarity between possible antecedent and examples of case slot (0 - 1)
· location class of possible antecedent (L_1, $\cdots$, L_{20})
· possible antecedent is located before zero pronoun (yes, no)
· possible antecedent depends over zero pronoun (yes, no)
features of possible antecedent
· case marker of possible antecedent (*ga*, *wo*, *ni*, $\cdots$)
· predicate of possible antecedent is in a noun-modifying clause (yes, no)
· possible antecedent is in the first sentence of the article (yes, no)
· possible antecedent depends on main clause (yes, no)
· possible antecedent is marked by a topic marker (yes, no)
· possible antecedent belongs to <agent> (yes, no)
· strength of predicate clause of possible antecedent (0, 1, $\cdots$, 6)
features of zero pronoun
· case marker of zero pronoun (*ga*, *wo*, *ni*)
· predicate of zero pronoun is in a noun-modifying clause (yes, no)
· voice of predicate of zero pronoun (active, passive, causative)
· type of predicate of zero pronoun (verb, adjective, noun+copula)
· head of zero pronoun is a verbal noun (yes, no)
· examples of the corresponding case slot belong to <agent> (yes, no)

5 Estimation of Antecedents of Zero Pronouns

We estimate antecedents of zero pronouns based on examples in the case frames and the classifier. We examine possible antecedents according to the location class orders, 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.

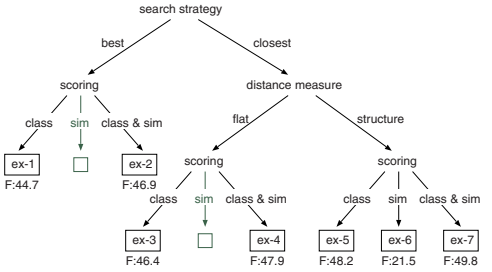
For example, "*youritsu*" (support) in Fig. 3 has zero pronouns in "*ga*" and "*ni*". The ordered possible antecedents for "*ga*" are L_7:"*Minsyutou*", L_{14}: "*Jimintou*"(ϕ *ga*), L_{14}:"*Ishihara chiji*"(ϕ *wo*), $\cdots$. The first candidate "*Minsyutou* (similarity: 0.73)", which is labeled as positive by the classifier, and whose similarity to the case frame examples exceeds the threshold (0.60), is determined as the antecedent.

6 Experiments

We conducted experiments of our zero pronoun resolution system using "Relevance-tagged corpus". To make article length uniform, we used 10 sentences at the beginning of each article. We used the SVM package, TinySVM[1], 2nd-order polynomial kernel. For testing, the system was given 100 articles that have correct dependency structure.

To illustrate the effectiveness of our approach, our experiments are performed under 7 configurations (Fig.4) using three parameters: search strategy, distance measure, and scoring. In Fig.4, ex-7 corresponds to our approach, ex-1 is similar to the approach suggested by [6], and ex-3 is similar to [3]. Experimental results are also shown in Fig.4. The accuracies (F-measure) are calculated by evaluating both detection and antecedent estimation of zero pronouns together. Our

[1] http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/



search strategy	
closest	select the closest candidate whose score exceeds a threshold
best	select the best-scored candidate in the current sentence and preceding two sentences
distance measure	
structure	use the location class orders
flat	measure the distance by words from a zero pronoun
scoring	
class	use raw score returned by the classifier
sim	use only similarity to examples of a case frame
class&sim	use both of them as we proposed

Fig. 4. Configurations of our experiments.

approach (ex-7) achieved 53.9% (496/921) in precision and 46.3% (496/1072) in recall, and outperformed the other methods significantly.

We also evaluated each accuracy of detection and antecedent estimation of zero pronouns under ex-7 setting. For the detection, we attained 87.1% in precision, 74.8% in recall, and 80.5% in F-measure. For the antecedent estimation, the accuracy was 61.8%. We compare our accuracies with [5], whose experiments are similar to ours. They achieved 48.9% in precision, 88.2% in recall, and 62.9% in F-measure for zero pronoun detection, and 54.0% accuracy for antecedent estimation on 30 newspaper articles. It is difficult to directly compare their results with ours due to the difference of the size of the test articles, but our method gave improvements over theirs in F-measure of zero pronoun detection by 17.6% and in accuracy of antecedent estimation by 7.8%. In particular, the significant improvement of zero pronoun detection indicates that the automatically constructed case frames are more effective than hand-crafted case frames which are used by [5].

Some major errors are shown in the following.

Errors caused by analysis limitation

The current system analyzes only predicates, and this means that it only handles some parts of all the relations in an article.

syusyou-wa *Syakaitou-no* *ritou* *mondai-ni-tsuite*, *tairyou ritou-niwa*
prime minister-TM Socialist Party-of secession problem-about much secession acc.

itara-nai-tono *mitoushi-wo* *nobeta*.
(not) cause that prospect acc. state

(The prime minister stated his prospect about the secession problem of the Socialist Party that ϕ would not cause secession of many members.)

In this example, "*itara-nai*" has a zero pronoun of "*ga*". Its antecedent is identified as "*syusyou*" by the system, but the correct antecedent is "*Syakaitou*" (the Socialist Party). This is because "*Syakaitou*" is not included in any case components of the predicates and is not ranked high in the location class order. To cope with this problem, it is necessary to deal with not only predicates but also verbal nouns. When analyzing them in this example, "*ritou*", which is in high rank for "*itara-nai*", has "*Syakaitou*" in its "*ga*" case, and "*itara-nai*" can be analyzed correctly as a result. Like this example, it is necessary to clarify and utilize a lot of relations in sentences, and this will lead to an improvement in the accuracy.

Detection errors of zero pronouns

Our system tends to recognize false zero pronouns.

shireikan-wa, ···	Russian-*gun-no*	*sensya*	50*dai-wo*	*hakai-shita-to*	*happyou-shita*.
general-TM	Russian army of	tank	50 acc.	destroy that	announce

(The general announced that ϕ had destroyed 50 tanks of Russian army.)

In this example, since the selected case frame of "*happyou*" (announce) has vacant "*ni*" slot, the system erroneously identifies it as a zero pronoun. This problem is not attributed to case frame errors, but context dependence. That is to say, in this context we do not think of whom the general announced the destruction to. To handle this problem, we need to incorporate context dependent features into the classifier.

7 Conclusion

We have described a Japanese zero pronoun resolution system. This system detects zero pronouns and restricts their possible antecedents based on automatically constructed case frames. To prefer close possible antecedents from a zero pronoun, we also introduced structural preference of antecedents. The experimental results showed that our approach significantly outperformed the previous work and the baseline methods.

References

1. McCarthy, J.F., Lehnert, W.G.: Using decision trees for coreference resolution. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. (1995) 1050–1055
2. Murata, M., Isahara, H., Nagao, M.: Pronoun resolution in Japanese sentences using surface expressions and examples. In: Proceedings of the ACL'99 Workshop on Coreference and Its Applications. (1999) 39–46
3. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics **27** (2001) 521–544
4. Kawahara, D., Kurohashi, S.: Fertilization of case frame dictionary for robust Japanese case analysis. In: Proceedings of the 19th International Conference on Computational Linguistics. (2002) 425–431
5. Seki, K., Fujii, A., Ishikawa, T.: A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In: Proceedings of the 19th International Conference on Computational Linguistics. (2002) 911–917
6. Aone, C., Bennett, S.W.: Evaluating automated and manual acquisition of anaphora resolution strategies. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. (1995) 122–129
7. Kurohashi, S., Nagao, M.: 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. (1994)
8. Kawahara, D., Kurohashi, S., Hasida, K.: Construction of a Japanese relevance-tagged corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation. (2002) 2008–2013