Experiments of FX for NTCIR-9 RITE Japanese BC Subtask

Hiroshi Umemoto
Research & Technology Group,
Fuji Xerox Co., Ltd.
Kanagawa, Japan
hiroshi.umemoto@fujixerox.co.jp

Keigo Hattori
Research & Technology Group,
Fuji Xerox Co., Ltd.
Kanagawa, Japan
keigo.hattori@fujixerox.co.jp

ABSTRACT

We report results and analyses of our experiments for NTCIR-9 RITE Japanese BC subtask we have participated in. It is assumed that the RTE task can be analyzed at some level of accuracy with a simple string-based method using word coverage, although the task seems to require advanced natural language understanding. On the other hand, if you try to tackle the task in the manner to follow your intuition, you should consider at least syntactic features of the texts. However, it is difficult in general to obtain better results for textual inference (TI) with syntactic analysis rather than with word-level analysis. In this paper, we explain our TI method based on syntactic and semantic relations of words in the texts, and conduct experiments.

Keywords

Semantic Representation, Lexical Functional Grammar, Modality Mismatch

Team Name

FX

Subtasks/Languages

BC/Japanese

External Resources Used

EDR Japanese Word and Concept Dictionaries, IPAdic, Japanese Wikipedia, IPAL Verb

1. INTRODUCTION

Textual inference (TI) is a competence to determine whether a natural language hypothesis can be inferred from a given premise [14], and is considered as a core function to realize advanced natural language processing applications [7, 4]. In recent years, research on TI attracts a great deal of attention through AQUAINT program [3, 15] and the PASCAL Recognizing Textual Entailment (RTE) Challenges [9, 1].

It was found that the RTE task can be analyzed at some level of accuracy with a simple string-based method using word coverage, although the task seems to require advanced natural language understanding [9]. On the other hand, if you try to tackle the task in the manner to follow your intuition, you should consider at least syntactic features of the texts. However, it is difficult in general to obtain better results for information retrieval, question answering and TI with syntactic analysis rather than with word-level analysis. Some of the reasons are variety of expressions, ellipsis by anaphora [20], mismatches of entity alignment caused by metonymy [12], and even insufficient performance of syntactic analysis itself. In addition, the task requires to be recognize paraphrasing based on inference with world knowledge beyond syntactic information. In this paper, we explain our TI method based on syntactic and semantic relations of words in the texts, and conduct experiments for NTCIR-9 RITE Japanese BC subtask [19].

2. ALGORITHM

2.1 Background: Logical Entailment

First of all, we explain an algorithm of logical entailment based on the framework of formal semantics to define the meaning of a natural language sentence. Although the semantic representation (SR) of a sentence is traditionally expressed with a higher-order modal logical form that is difficult to operate, we transform it to a first-order predicate logical form with context [6]. The context in the SR could be considered as a possible world expressed by modality in the sentence. We decompose the term of a verb into terms of semantic roles with an event variable, based on neo-Davidsonian event semantics [18].

First-order predicate logic is non-deterministic in general. Then, we assign a constant into a quantified variable in the SR to obtain another SR including no quantification. The procedure of assigning a constant is known as skolemization, and the assigned constant is called a skolem constant. The SR obtained by the skolemization is proper for description logic and is deterministic [2]. We call a term representing a ternary relationship of a predicate and two arguments as a fact.

Given two sentences premise P and hypothesis H, let us determine whether P entails H according to the SRs of the both sentences.

First, we try to find possible alignments of entities in P and H. The alignments should satisfy the following conditions:

1. The part-of-speech of the word indicating an entity in P matches an entity in H.

2. If a word indicating an entity is a proper noun, the surface form of a word in H matches either a word in P or an alias in P.

3. If a word indicating an entity is not a proper noun, the

concept of a word in H is subsumed by the concept of a word in P.

Second, we decide the alignments of predicates in P and H. We regard that predicates expressing semantic roles are aligned if the semantic role of the predicate in P subsumes that in H.

We determine whether the SR of P entails the SR of H based on the alignments of entities and semantic roles. Upward monotonicity holds true if negation is out of the scope of our algorithm. If we hypothesize the ground of the sentence is always true, the logical product of the SR also holds true, and implies the combination of any terms holds true [10]. If we ignore modality in sentences, we could obtain the TI relations of sentences from first-order predicate logical forms wrapped by context corresponding to modality.

The TI relation is directional. In the perspective of IR, the SR of a passage P entails the SR of a hypothesis H, but the reverse relation is not always true. The assumption that upward monotonicity always holds true in the algorithm yields that a retrieved passage should indicate broader range of concept than a hypothesis. This means that no passage is fetched from a hypothesis including a semantic element that does not exist in the passages.

In the logical form of the SR, the first-class elements are the semantic role and the entity. They are clearly distinguished and not exchangeable. However, some of the events or concepts that entities represent have meanings equivalent to some sort of semantic roles. As just described, the algorithm could not determine the correct TI relations if the same content is expressed by an entity in a sentence and expressed by a semantic role in the other sentence.

2.2 Alignment and Matching Score

It is difficult in general, however, to obtain correct relationship described above from real texts, because there are a variety of obstacles described in Section 1. To apply to the RITE tasks, we evaluate a matching score of the pair texts that indicates the correspondence of ternary relationship of two entities and their role in the texts. We intend to express a degree of realization of the rigid logical entailment described above as a matching score. In other words, we hypothesize that the logarithm probability of TI relation correlates with the matching score. We assign a matching score according to the correspondence of the premise and hypothesis facts instead of conducting logical entailment operations.

The matching score of a fact is calculated with the product of the matching scores of its functor role relation and two argument words. However, if we obtain two out of three alignments between two facts in the premise side and the hypothesis side, we calculate the whole matching score of the text pair with an appropriate smoothed matching score of the hypothesis fact in order to alleviate the effects of insufficient word concept dictionaries and inadequate syntactic or semantic analyses. The correspondence of two words is assigned according to the matching level, such as exactly matched surface strings, normalized words without notational variations, synonyms, hypernyms and hyponyms defined in concept dictionaries, and overlapped substrings. Each type of role relationship has a weight, and the correspondence of two roles is assigned according to the weight of the roles if the two roles are exactly matched. Role relationship is usually analyzed with ambiguity, and we pick up roles whose matching score is the highest among the two facts.

The matching score of the pair texts is calculated with the sum of the scores of the facts. Text analyses are usually ambiguous and there could be multiple possible analyses for a single sentence. We disambiguate the analyses by picking up the one whose matching score is the highest among the possible analyses with alignments.

2.3 Penalty for Modality Mismatch

We assume statement of negation, hearsay, speculation and others as modality in this paper. In most cases, the TI relation of the text pair does not hold if their modalities do not match. Therefore, we extract modality information from SRs and determine the correspondence of the modality between the text pair. If the modalities of the pair differ, the text pair is imposed a penalty for modality mismatch. The penalty becomes higher in proportion to the matching score of the fact relating to modality. The TI relation is reversed if the penalty of the pair exceeds a limit, in spite of a high matching score evaluated with alignment analysis.

2.4 Classification for BC Subtask

The BC subtask is considered as a binary classification task whether a TI relation holds or not. We construct a binary classifier based on matching scores evaluated with relation alignment and penalties for modality mismatch. We basically determine the entailment of the text pair based on the matching score. If the matching score exceeds a threshold, we determine the entailment holds true, and if the condition is not satisfied, we determine the entailment does not hold. However, in the case we consider the mismatch of modality, if the entailment holds true based on the matching score and the penalty exceeds a threshold, then we conclude the entailment does not hold. We estimate optimal parameters with the whole development set.

In addition, if a text fails to be analyzed and there is no analysis at all, we determine the entailment based on the overlap of words between the text pair, as a fall-back treatment.

3. SYSTEM DESCRIPTION

3.1 Pipeline System

In order to implement the algorithm explained in Section 2, we employ a TI system described in [21] as a base system. The base system is realized as a pipeline of processing components shown in Figure 1. Figure 1 a. shows the pipeline of the whole system while Figure 1 b. presents the semantic rule sequence of the semantic analyzer in the system explained below.

3.2 Text Analysis

Major part of this system is constructed within Xerox Linguistic Environment (XLE)[1] that performs efficient deep parsing based on Lexical Functional Grammar (LFG) for Japanese sentences [16]. One of the analyses of LFG is f(unctional)-structure that encodes an embedded feature-value matrix, such as grammatical functions, voice, modality, tense and aspect. If the target sentence is syntactically

[1] http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html

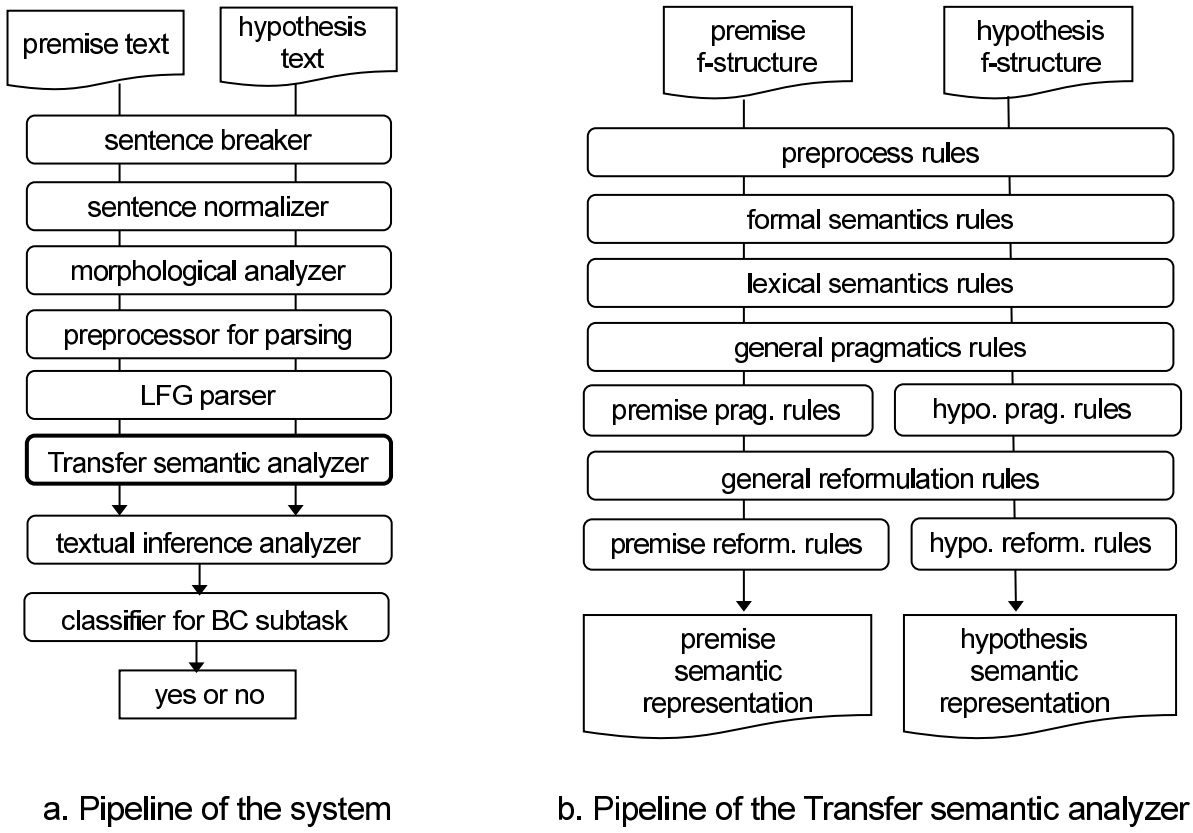

Figure 1: Pipelines of the system and the semantic analyzer

ambiguous, the multiple analyses of the sentence are held in a chart-like packed structure in XLE.

The base system performs semantic analyses with the Transfer system[2] that is built within XLE and rewrites packed f-structure efficiently without unpacking. According to semantic rewrite rules manually constructed, the Transfer system flattens f-structure, and extracts predicate-argument relations. And it produces a semantic representation (SR), converting grammatical functions into semantic roles and word skolems into corresponding concepts respectively [8]. It also handles compound words, appositive, coordination, negation, cleft sentences, functional expressions, idioms, ellipses, concept subsumption, symmetry relations, and presupposition [21].

3.3 Linguistic Resources

In morphological analysis, we use ChaSen IPA dictionary [3] with an extended lexicon including a vocabulary extracted from Wikipedia entries. The base system is equipped with a semantic lexicon that is constructed with the EDR dictionary[4] (about 260 thousand words and 410 thousand concepts), lexical information extracted from Japanese Wikipedia (about 400 thousand words from title, highlighted, infobox and redirect, and 160 thousand words from list pages) and others [21].

[2]The Transfer system is also called the packed rewriting system.
[3]http://chasen-legacy.sourceforge.jp/
[4]http://www2.nict.go.jp/r/r312/EDR/index.html

3.4 Derivation of Semantic Representation

The morphological analysis is applied to a input sentence, then the grammar analysis base on LFG is conducted. As a result, the f-structure and the c-structure are derived, which represent a predicate argument construction and a syntactic relationship respectively. The semantic analysis with the Transfer is applied to the f-structure.

The f-structure is a nested matrix construction, and is converted into a flat one. The SR of the input sentence is represented as a conjunctive proposition of facts. The scopes of quantification and modality are expressed explicitly as a context entity, and a fact is a first-order predicate logic term, in which the functor is a role and the two arguments are skolems, wrapped by a context. Incidentally, quantification has no ambiguity through the skolemization.

The Transfer system applies rewriting rules one by one to the set of facts of the input sentence. It may apply multiple different rewriting rules to a single fact, and then preserves each of the rewritten results in a particular choice space. The SR of the sentence is expressed as a one logical term, which is a conjunctive proposition of facts in different choices. The Transfer system does not treat each of the choice spaces individually, but a compact chart structure, although the choices are expanded exponentially with multiple rewriting

rules.

The SR includes explicit facts expressing appositive, coordination, idioms, symmetry relations, presupposition, and others, which are not expressed literally in the sentence. Modality is expressed with a fact, one of whose arguments is a fact. In practice, the modality fact has an argument of the predicate of a fact due to the limitation of first-order predicate logic.

3.5 Inference Module

The Transfer system also performs determining TI relations according to the algorithm in Section 2. The semantic rewrite rules for the hypothesis are partly different from rules for the premise, and the SR of the hypothesis is not equal to the SR of the premise that is identical to the hypothesis. The inference module detects syntactic alternations related to case, part-of-speech and voice based on SRs. It assigns matching scores and penalties to the combinations of facts, determines corresponding facts, and calculates the total score of all the facts in the sentence. It also applies rewriting rules regarding modality, and marks facts for modality expressions. It assigns a penalty to the corresponding fact marked as modality. In the framework of the Transfer system, a score is also expressed as a fact.

If there are multiple analyses for a single sentence, the score of each analysis is calculated individually. The inference module picks up a choice space where the total matching score is the highest, and removes the facts of the other choices. If there are multiple analyses, both of which have the highest matching score, we pick up an arbitrary choice among them.

It determines the entailment relationship according to the comparison of the calculated matching score and its threshold selected though experiments with the development set. In the case that it cannot calculate the score due to the lack of an analysis, it calculate the correspondence of word overlap with the result of the morphological analysis. The threshold of word overlap is also selected through experiments with the development set.

4. EXPERIMENTS

4.1 Settings

We conducted experiments against the development set, and selected the threshold of the matching score. We decided to use the thresholds 0.3 and 0.6, which correspond to high recall and high precision respectively, in order to obtain accurate results. Similarly, we selected 0.4 as the threshold of the penalty and as the threshold of the word overlap through the results of experiments against the development set.

We conducted three runs explained below. In the first run (run-01), we applied the optimal score parameter 0.3 and did not use penalty at all. In the second run (run-02), the score parameter was the same as run-01, but use penalty with its parameter 0.4. In the third run (run-03), we applied the next optimal score parameter 0.6 and the same penalty parameter as run-02.

4.2 Results

The evaluation results of the test set are shown in Table 1. The accuracy of the classification based on word overlap against 20 fall-back cases that have no analyses is 0.60. As a reference, the results of the development set represents a similar tendency of that of the test set.

5. ANALYSES OF THE RESULTS

5.1 Distribution of Matching Score

The distribution of the matching score is shown in Table 2. The difference between the matching scores of true cases and false ones are not significant. The average and median of matching scores of true and false cases suggest that the maximum accuracy obtained by classifying with the threshold of the matching score is about 50%. The result of run-01 in Figure 2 also supports this suggestion. The matching score seems to contribute less than expected unfortunately. Compared with run-02 and run-03, though their matching thresholds are different, their accuracies do not differ a lot. Run-01 and run-03 resulted in almost the same accuracy, although run-03 was higher precision than run-01.

The matching score almost depends on the number of facts, partly because it is not normalized according to the length of the sentence explicitly. Although the correspondences of words and roles are determined according to concepts and role types, the weight of each fact is not considered. The proposed method based on parts-of-speech of words and semantic roles does not distinguish vital facts to determine the entailment relationship, partly because the precision of the analysis seems to be not sufficient. Specifically, it is difficult to recognize and to weight essential dependency relations.

5.2 Effect of the Penalty of Modality Mismatch

Next, we look into the effect of the penalty of modality mismatch. Compared run-02 to run-01, penalty for modality mismatch seems to contribute better accuracy. Some cases of false positive in run-01 are affected by the penalty, and the number of true negative in run-02 is more than that in run-01. It can be expected that selecting a proper threshold of the penalty does not worsen the recall a lot but improves the precision a little.

Looking into modality expressions precisely, we found cases where the analysis was not adequate but the entailment judgement happened to be correct. Although we conducted a relatively simple treatment for modality, it could be expected that integrating more detailed analyses on modality improves the accuracy. We also believe that integrating patterns expressing contradiction in addition to the difference of modality is effective to achieve better accuracy on the pairs whose syntactic and semantic constructions are similar.

5.3 Error Analyses

The proposed method is basically intended to recognize the entailment appropriately against the pairs whose syntactic and semantic constructions are similar. However, the correspondences of words are highly weighted, and we found cases where the judgements of the entailment were not correct although the syntactic constructions are similar. It is very hard to compromise the balance of the rigorous logical entailment and the proposed loose matching method against cases where the syntactic constructions are similar. We also found cases where synonyms and paraphrasal expressions were not treated properly as well as anaphora.

Many cases where the inference and interpretation based on world knowledge were not treated at all, although they

Table 1: Evaluation results of the test set

run	accuracy	precision	recall	f-score	true positive	false positive	true negative	false negative
01	0.510	0.506	0.796	0.619	199	194	51	56
02	0.524	0.517	0.724	0.603	181	169	69	81
03	0.520	0.517	0.624	0.565	156	146	94	104

Table 2: Distribution of Matching Scores and Penalties

run	cases	# of cases	ave. # of facts	ave. score	mid. score	ave. penalty
total	true	250	6.25	1.99	1.30	0.11
	false	250	6.33	2.39	1.28	0.16
run-01	true positive	199	6.86	2.49	1.93	0.14
	false positive	194	6.99	3.04	2.00	0.21
	false negative	51	3.96	0.12	0.20	0.01
	true negative	56	3.98	0.11	0.18	0.02
run-02	true positive	181	6.68	2.39	1.90	0.02
	false positive	169	6.86	2.86	1.80	0.02
	false negative	69	5.17	0.99	0.20	0.33
	true negative	81	5.22	1.42	0.20	0.45
run-03	true positive	156	6.89	2.72	2.10	0.02
	false positive	146	7.12	3.26	2.20	0.02
	false negative	94	5.24	0.84	0.20	0.25
	true negative	104	5.24	1.20	0.20	0.36

are substantially difficult cases, of course. It is rare to grasp interpretations, which are vague and implicitly expressed contents, with only words and syntactic or semantic relations literally expressed in a sentence.

We suspect there are cases where human judgements could be largely different in individuals. It might be required to consider the degree and the classification of entailment relationship to improve the quality of the test collections.

6. RELATED WORK

[13] developed a dialog based QA system (Dialog Helpsystem) that analyzes natural texts with KNP, a Japanese dependency parser, and performs flexible matching based on probabilistic model. Inheriting the architecture of Dialog Helpsystem, [12] developed a QA system with large knowledge base. [20] present a study on a question answering system using Kura, a Japanese paraphrase engine, and report its precision falls behind that of a baseline system based on bag-of-words model, and argues that anaphora resolution is important and the contribution of paraphrases is less than expected. [17] propose a TI system with a newer version of KNP, a Japanese syntactic and case structure analyzer, and disclose their test set[5]. Compared with these systems, our system processes ambiguous analyses efficiently and eliminates elements from queries in the perspective of TI. While various TI systems for English texts are proposed through AQUAINT program and the RTE Challenges, [11] show that TI information is used to either filter or rank answers returned by a QA system, accuracy can be increased by as much as 20% overall. [3] present a TI system, and our proposed system shares the use of the Transfer system with it. [5] propose a TI system with LFG and FrameNet.

7. CONCLUSION

We explained our TI method based on syntactic and semantic relations of words in the texts, and showed results and analyses of our experiments for NTCIR-9 RITE Japanese BC subtask. The base system is realized as a pipeline of processing units. We evaluate a matching score of the pair texts that indicates the correspondence of ternary relationship of two entities and their role in the texts, and hypothesize that the probability of TI relation correlates with the matching score. We extract modality information from SRs and determine the correspondence of the modality in the text pair, and the text pair is imposed a penalty for modality mismatch. The TI relation is reversed if the pair is imposed a penalty, in spite of a high matching score evaluated with alignment analysis.

The proposed method is basically intended to provide a proper judgement against a text pair whose syntactic or semantic constructions are similar. Although it realizes a relatively simple treatment for modality, it is expected to improve the precision a little without worsening the recall a lot with integrating more detailed analyses and a proper threshold. We also believe that integrating patterns expressing contradiction in addition to the difference of modality is effective to achieve better accuracy.

However, the correspondences of words are highly weighted, and we found cases where the judgements of the entailment were not correct although the syntactic constructions are similar. It is very hard to compromise the balance of the rigorous logical entailment and the proposed loose matching method against cases where the syntactic constructions are similar. We also found cases where synonyms and paraphrasal expressions were not treated properly as well as anaphora. It is rare to grasp interpretations, which are vague and implicitly expressed contents, with only words and syntactic or semantic relations literally expressed in a sentence.

[5]http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/rte.html

8. REFERENCES

[1] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, 2009.

[2] D. Bobrow, C. Condoravdi, R. Crouch, R. Kaplan, L. Karttunen, T. H. King, V. de Paiva, and A. Zaenen. A basic logic for textual inference. In *Proceedings of the AAAI Workshop on Inference for Textual Question Answering*, 2005.

[3] D. G. Bobrow, B. Cheslow, C. Condoravdi, L. Karttunen, T. H. King, R. Nairn, V. de Paiva, C. Price, and A. Zaenen. PARC's Bridge and question answering system. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF07) Workshop*, pages 46–66, 2007.

[4] J. Bos. Let's not argue about semantics. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 2835–2840, 2008.

[5] A. Burchardt and A. Frank. Approaching textual entailment with lfg and framenet frames. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.

[6] C. Condoravdi, D. Crouch, J. Everett, V. Paiva, R. Stolle, D. Bobrow, and M. van den Berg. Preventing existence. In *FOIS '01: Proceedings of the International Conference on Formal Ontology in Information Systems*, pages 162–173, 2001.

[7] R. Cooper, D. Crouch, J. V. Eijck, C. Fox, J. V. Genabith, J. Jaspars, H. Kamp, M. Pinkal, D. Milward, M. Poesio, and S. Pulman. Using the framework. Technical report, The FRACAS Consortium, 1996.

[8] D. Crouch and T. H. King. Semantics via f-structure rewriting. In *Proceedings of LFG06 Conference*, pages 145–165. CSLI On-line Publications, 2006.

[9] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. *Lecture Notes in Computer Science*, 3944:177 – 190, 2006.

[10] L. T. F. Gamut. *Logic, Language, and Meaning*. The University of Chicago Press, 1991.

[11] S. Harabagiu and A. Hickl. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905 – 912, 2006.

[12] Y. Kiyota, S. Kurohashi, and F. Kido. Dialog navigator: A question answering system based on large text knowledge base. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 460–466, 2002.

[13] S. Kurohashi and W. Higasa. Dialogue helpsystem based on flexible matching of user query with natural language knowledge base. In *Proceedings of the 1st ACL SIGdial Workshop on Discourse and Dialogue*, pages 141–149, 2000.

[14] B. MacCartney and C. D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, 2007.

[15] B. MacCartney and C. D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 521–528, 2008.

[16] H. Masuichi, T. Ohkuma, H. Yoshimura, and Y. Harada. Japanese parser on the basis of the Lexical-Functional Grammar formalism and its evaluation. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 298–309, 2003.

[17] M. Odani, T. Shibata, and S. Kurohashi. Canonicalization of paraphrase expressions onto predicate-argument structure and its application to recognizing textual entailment. In *Proceedings of the 15th Annual Meeting of the Association for Natural Language Processing (NLP2009)*, pages 260–263, 2009. (in Japanese).

[18] T. Parsons. *Events in the Semantics of English*. MIT Press, 1990.

[19] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. Overview of ntcir-9 rite: Recognizing inference in text. In *Proceedings of NTCIR-9*, 2011. (to appear).

[20] T. Takahashi, K. Nawata, K. Inui, and Y. Matsumoto. Effects of structural matching and paraphrasing in question answering. *IEICE Transactions on Information and Systems*, E86-D(9):1677–1685, 2003.

[21] H. Umemoto, D. Sugihara, T. Ohkuma, and H. Masuichi. Detecting Japanese textual entailment using LFG analysis and lexical resources. In *Special Interest Group Technical Reports of IPSJ 2008-NL-188*, pages 57–64, 2008. (in Japanese).