Analysis of Identifying Linguistic Phenomena for Recognizing Inference in Text

Min-Yuh Day and Ya-Jung Wang

Department of Information Management, Tamkang University, Taiwan
myday@mail.tku.edu.tw, lucy19890924@gmail.com

Abstract

Recognizing Textual Entailment (RTE) is a task in which two text fragments are processed by system to determine whether the meaning of hypothesis is entailed from another text or not. Although a considerable number of studies have been made on recognizing textual entailment, little is known about the power of linguistic phenomenon for recognizing inference in text. The objective of this paper is to provide a comprehensive analysis of identifying linguistic phenomena for recognizing inference in text (RITE). In this paper, we focus on RITE-VAL System Validation subtask and propose a model by using an analysis of identifying linguistic phenomena for Recognizing Inference in Text (RITE) using the development dataset of NTCIR-11 RITE-VAL sub-task. The experimental results suggest that well identified linguistic phenomenon category could enhance the accuracy of textual entailment system.

.

Keywords: Linguistic Phenomena, Recognizing Inference in Text, Textual Entailment, Knowledge-based, Machine Learning

1. Introduction

Recognizing Textual Entailment (RTE) is a task in which a system is given two text fragments and then determines whether the meaning of hypothesis is entailed from another text [6, 17].

Since 2005, the importance of RTE in assessing semantic inference in text has been increasing. After the third PASCAL RTE Challenges in Europe, RTE became one of tasks of the Text Analysis Conference (TAC) in 2008. The RTE Challenge is a generic task that captures major semantic inference needs across many natural language processing applications, such as Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and (multi) document summarization. RTE is largely European and American project. Its counterpart in East Asia is called, Recognizing Inference in Text (RITE) [21].

RITE is a generic benchmark task that addresses major text understanding needs in variety of NLP/Information Access research areas. There are two subtasks in NTCIR-11 RITE-VAL, namely, Fact Validation (FV) and System Validation (SV). RITE-VAL task organizers provide participants with task datasets for four languages: Chinese-simplified (CS), Chinese-Traditional (CT), English (EN), and Japanese (JA). In Chinese FV Subtasks, each t2 should be tagged in one of the three labels (E, C, U). In Chinese SV-BC Subtasks, each t2 should be tagged in one of the two labels (Y, N). In Chinese SV-MC Subtasks, each t2 should be tagged in one of the four labels (F, B, C, I)[11, 21].

Linguistic phenomena were first introduced in NTCIR RITE unit-tests subtask in Japanese. Research focusing on single linguistic phenomena in recognizing textual entailment has been considered difficult because the task usually requires various types of linguistic and semantic analyses [11, 21].

Recent developments in recognizing textual entailment have heightened the need for a better understanding of linguistic phenomena-level inference. Although a considerable number of studies have been made on recognizing textual entailment, little is known about the power of linguistic phenomenon in recognizing inference in text.

The objective of this paper is to provide a comprehensive analysis of identifying linguistic phenomena for recognizing inference in text (RITE). In this paper, we focus on RITE-VAL System Validation subtask and propose a model by using an analysis of identifying linguistic phenomena for Recognizing Inference in Text (RITE) using the development dataset of NTCIR-11 RITE-VAL sub-task.

The remainder of this paper is organized as follows. Section 2 describes the research background and related works of linguistic phenomena for recognizing textual entailment. Section 3 shows the methodology and datasets for analysis. Section 4 contains the experimental result and discussion. Finally, in Section 5, we present our conclusions.

IEEE IRI 2014, August 13-15, 2014, San Francisco, California, USA
978-1-4799-5880-1/14/$31.00 ©2014 IEEE

2. Research Background and Related Works

A considerable amount of literature has been published on recognizing textual entailment. Recent developments in textual entailment recognition have heightened the need for a better understanding of linguistic phenomena-level inference [3, 4, 9-11, 13, 16, 20-22].

There are three levels of inference in the pyramid of textual entailment recognition technology, namely, (1) linguistic phenomena-level inference, (2) sentence level inference, and (3) multiple sentence level inference. Linguistic phenomena-level inference (unit test) is considered as the foundation oriented research for the textual entailment recognition. Sentence-level inference (i.e., BC, MC) would be improved from a further understanding of the linguistic phenomena-level inference [11].

In the research field of textual entailment recognition, textual contradiction detection has been received a lot of attention in recent years. Condoravdi et al. [5] first argued that the detection of entailment and contradiction relations between texts is an important metric in the evaluation of text understanding systems. Ritter et al. [13] presented a case study of contradiction detection based on functional relations and proposed a model for determining whether an arbitrary phrase is function. Harabagiu et al. [9] proposed a framework which combines techniques for the processing of negation, the recognition of contrasts, and the automatic detection of antonym for recognizing contradiction in natural language texts with over 62% accuracy. Marneffe et al. [7] proposed an appropriate definition of contradiction for NLP task and provided a typology of contractions, namely, type 1 with antonym, negation, numeric, and type 2 with factive/modal, structure, lexical and world knowledge. They further defined feature sets (polarity features, number, date and time features, antonymy features, structural features, factivity features, modality features, and relational features) used to capture salient patterns of contraction.

As Cabrio et al. [3] noted, a new research interest is rising towards a deeper and better understanding of the core linguistic phenomenon in textual inference. In line with this direction, Cabrio proposed a definition for strong component-based and focused on decomposing the complexity of the textual entailment into basic phenomena and on their combination [3]. Prior studies [2-4, 15, 16] argued the incremental advances in local entailment phenomena are needed to make significant progress in the task of textual entailment.

Cabrio et al. [3] defined a component-based textual entailment architecture as a set of clearly identifiable textual entailment modules that can be singly used on specific entailment sub-problems and combined to produce a global entailment judgment. Linguistic processing and annotation (e.g., parsing, NER) can be required by a component according to the phenomenon it considers. Most textual entailment system adopted machine learning approaches with semantic and lexical syntactic features. Bar-Haim et al. [1] proposed a generic semantic inference framework that operates directly on syntactic trees that are inferred by applying entailment rules for generic linguistic phenomena.

Bentivogli et al. [2] proposed a methodology for isolating linguistic phenomena relevant to inference and created specialized data sets with monothematic Text-Hypothesis (T-H) pairs for textual entailment. Linguistic phenomena are relevant for judging the entailment relation in T-H pair of RTE data sets. Linguistic phenomena of RTE-5 T-H pairs are annotated and grouped with both macro categories and fine-grained phenomenon. In the RTE-5 T-H pairs, a total of 36 fine-grained linguistic phenomena are grouped into five macro categories, namely, lexical, lexical-syntactic, syntactic, discourse, and reasoning. A total of 8 linguistic phenomena (Identity/mismatch, format, acronymy, demonymy, synonymy, semantic opposition, hypernymy, geographical knowledge) are grouped into the lexical category. A total of 4 phenomena (Transparent head, Nominalization /verbalization, Paraphrase, Negation) are grouped into lexical-syntactic category. A total of 7 phenomena (Negation, Modifier, Argument realization, Apposition, List, Coordination, Active/passive alternation) are grouped into syntactic category. A total of 5 phenomena (Coreference, Apposition, Anaphora zero, Ellipsis, Statements) are grouped into discourse category. A total of 12 phenomena (Apposition, Modifier, Genitive, Relative clause, Elliptic expression, Meronymy, Metonymy, Membership/representative, Quantity, Temporal, Spatial, common background/general inferences) are grouped into reasoning category. The linguistic phenomena of reasoning category are the most frequent in the RTE-5 data set and required deeper inferences. Single linguistic phenomenon is directly involved in the entailment relation.

Tolendo et al. [18] introduced a semantic model for annotating textual entailments and explored the applicability of the proposed model to the Recognizing Textual Entailment (RTE) 1-4 corpora. They focused on valid entailments involving restrictive, intersective, and appositive modification that contribute to the recognition of the entailment. This approach concentrates on the logical aspects of textual entailment, while phenomena involving lexical semantics and world knowledge are handled by a shallow analysis. Annotations were marked in 80.65% of the entailments in the RTE 1-4 corpora of and reached cross-annotator agreement of 67.96% on average [18]. Toledo et al. [18] argued that the Recognizing Textual entailment (RTE) corpus is the resource of textual entailments with the annotated entailment as valid/invalid category. However, the RTE

categorization contains no indication of the linguistic and information processes that underlie entailment. In the lack of gold standard of inferential phenomena, entailment systems can be compared based on their performance, but not on the basis of the linguistic adequacy of their inferential process. Rooney et al. [14] conducted an investigation into the application of ensemble learning for entailment classification. Rooney et al. developed a linguistic analysis framework based on the extraction of similarity and dissimilarity features between the text and hypothesis elements of an entailment text pair.

Sammons et al. [15] proposed a model for identifying and annotating textual inference phenomena in textual entailment examples. They argued that the single global label with which RTE examples are annotated is insufficient to effectively evaluate RTE system performance. They suggested more detailed annotation and evaluation for RTE system. In the pilot RTE system analysis conducted in Sammons et al. [15], they intended to answer a research question "Does identifying the phenomena correctly help learn a better TE system?" They also argued that if a system could recognize key negation phenomena such as named entity (NE) mismatch, presence of excluding arguments correctly and consistently, it could model them as contradiction features in the final inference process to significantly improve its overall accuracy. In addition, Identifying and resolving the key entailment phenomena would boost the inference process in positive examples as well. Prior researchers showed that mismatching information between sentences is a cue of non-entailment [19]. Contradiction detection requires more precise comprehension of the consequence of sentences [8].

Watanabe et al. [20] proposed an approach by leveraging diverse lexical resources for textual entailment recognition. Nguyen et al. [12] proposed an unsupervised learning method, namely Rule-based Support-Sentence Classifier (RSSC) and Bootstrapping Support-Sentence Classifier (BSSC), to recognize agreement and contradiction semantic classes. Nguyen et al. [12] argued that word overlap method is a relatively effective indicator of sentence similarity and relatedness, however, overlap method can only be used for classification of agreement class. Two sentences with many overlap words can be totally a contradiction class. They used two criteria, namely lexical matching and negation clues, for recognize agreement and contradiction. Wu [23] proposed a light-weight Chinese textual entailment recognition system using part-of-speech information only.

Prior researches showed that the lexical, syntactic and world knowledge levels can be analyzed and exploited in order to fully identify and recognize the entailment between T and H. It is considered useful by providing results of basic linguistic analyses such as dependency parsing, predicate-argument structure analysis, and a generic entailment recognition tool [2].

Table 1 Analysis of NTCIR11 RITE-VAL FV-MC/BC development dataset (581 pairs)

FV-MC/BC	N	Y	Total
B		222 (38.2%)	222 (38.2%)
C	152 (26.2%)		152 (26.2%)
F		148 (25.5%)	148 (25.5%)
I	59 (10.2%)		59 (10.2%)
Total	211 (36.3%)	370 (63.7%)	581 (100.0%)

Although a considerable number of studies have been made on recognizing textual entailment, several attempts have been made on the linguistic phenomena-level inference, however, little is known about the power of Chinese linguistic phenomenon in recognizing inference in text.

3. Methodology

We describe the methodology used for the analysis of identifying linguistic phenomena for recognizing inference in text.

3.1. Datasets

The dataset was obtained from the organizers of NTCIR-11 RITE-VAL [11]. It's a subset of the NTCIR-11 RITE-VAL SV BC/MC data in which semantic relations are broken down into a set of single linguistic phenomena.

The organizers of NTCIR-11 RITE-VAL provide a dataset obtained from a subset of the BC data in which semantic relations are broken down into a set of single linguistic phenomena. It is considered useful by providing results of basic linguistic analyses such as dependency parsing, predicate-argument structure analysis, and a generic entailment recognition tool. A sentence pair (t1 and t2) in a part of System Validation subtask dataset has a category label related to a linguistic phenomenon.

Table 1 shows the basic analysis of NTCIR-11 RITE-VAL FV-MC/BC development dataset (581 pairs). This table indicates that it is an imbalance dataset with Y (63.7%) and N (63.7%) for BC subtask; B (38.2%), C(26.2%), F(25.5%), and I (10.2%) for MC.

3.2. Analysis of Linguistic Phenomenon

In order to understand the linguistic phenomenon categories and their related labels of BC and MC in NTCIR-11 RITE-VAL, we conduct an analysis of linguistic phenomenon category in NTCIR-11 RITE-VAL dataset. Table 2 summarizes the analysis of linguistic phenomenon (Category) in NTCIR-11 RITE-VAL development dataset (581 pairs). This analysis indicates that there are 28 linguistic phenomenon categories which are annotated in the 581 pairs of the dataset. The analysis

Table 2. Analysis of Linguistic Phenomenon (Category) in NTCIR-11 RITE-VAL Development Dataset (581 pairs).

Linguistic Phenomenon Category/Label		BC			MC				
Category ID	Category	Y	N	Total	B	C	F	I	Total
1	abbreviation	6 (1.0%)		6 (1.0%)	4 (0.7%)		2 (0.3%)		6 (1.0%)
2	antonym		20 (3.4%)	20 (3.4%)		20 (3.4%)			20 (3.4%)
3	apposition	6 (1.0%)	1 (0.2%)	7 (1.2%)	5 (0.9%)		1 (0.2%)	1 (0.2%)	7 (1.2%)
4	case_alternation	21 (3.6%)		21 (3.6%)	18 (3.1%)		3 (0.5%)		21 (3.6%)
5	clause	22 (3.8%)	3 (0.5%)	25 (4.3%)	5 (0.9%)		17 (2.9%)	3 (0.5%)	25 (4.3%)
6	coreference	9 (1.5%)	2 (0.3%)	11 (1.9%)	6 (1.0%)		3 (0.5%)	2 (0.3%)	11 (1.9%)
7	exclusion:common_sense		8 (1.4%)	8 (1.4%)		8 (1.4%)			8 (1.4%)
8	exclusion:modality		12 (2.1%)	12 (2.1%)		12 (2.1%)			12 (2.1%)
9	exclusion:modifier		14 (2.4%)	14 (2.4%)		14 (2.4%)			14 (2.4%)
10	exclusion:predicate_argument		51 (8.8%)	51 (8.8%)		51 (8.8%)			51 (8.8%)
11	exclusion:quantity		6 (1.0%)	6 (1.0%)		6 (1.0%)			6 (1.0%)
12	exclusion:spatial		14 (2.4%)	14 (2.4%)		14 (2.4%)			14 (2.4%)
13	exclusion:temporal		7 (1.2%)	7 (1.2%)		7 (1.2%)			7 (1.2%)
14	hypernymy	19 (3.3%)	11 (1.9%)	30 (5.2%)	7 (1.2%)		12 (2.1%)	11 (1.9%)	30 (5.2%)
15	inference	51 (8.8%)	24 (4.1%)	75 (12.9%)	6 (1.0%)		45 (7.7%)	24 (4.1%)	75 (12.9%)
16	lexical_entailment	11 (1.9%)	1 (0.2%)	12 (2.1%)	11 (1.9%)			1 (0.2%)	12 (2.1%)
17	list	17 (2.9%)	3 (0.5%)	20 (3.4%)	1 (0.2%)		16 (2.8%)	3 (0.5%)	20 (3.4%)
18	meronymy	2 (0.3%)	2 (0.3%)	4 (0.7%)	2 (0.3%)			2 (0.3%)	4 (0.7%)
19	modifier	34 (5.9%)	3 (0.5%)	37 (6.4%)	24 (4.1%)		10 (1.7%)	3 (0.5%)	37 (6.4%)
20	negation		20 (3.4%)	20 (3.4%)		20 (3.4%)			20 (3.4%)
21	paraphrase	47 (8.1%)		47 (8.1%)	42 (7.2%)		5 (0.9%)		47 (8.1%)
22	quantity	10 (1.7%)	1 (0.2%)	11 (1.9%)	6 (1.0%)		4 (0.7%)	1 (0.2%)	11 (1.9%)
23	relative_clause	6 (1.0%)		6 (1.0%)	5 (0.9%)		1 (0.2%)		6 (1.0%)
24	scrambling	23 (4.0%)	4 (0.7%)	27 (4.6%)	22 (3.8%)		1 (0.2%)	4 (0.7%)	27 (4.6%)
25	spatial	16 (2.8%)	2 (0.3%)	18 (3.1%)	3 (0.5%)		13 (2.2%)	2 (0.3%)	18 (3.1%)
26	synonymy:lex	47 (8.1%)	1 (0.2%)	48 (8.3%)	45 (7.7%)		2 (0.3%)	1 (0.2%)	48 (8.3%)
27	temporal	10 (1.7%)	1 (0.2%)	11 (1.9%)	1 (0.2%)		9 (1.5%)	1 (0.2%)	11 (1.9%)
28	transparent_head	13 (2.2%)		13 (2.2%)	9 (1.5%)		4 (0.7%)		13 (2.2%)
	Total	370 (63.7%)	211 (36.3%)	581 (100.0%)	222 (38.2%)	152 (26.2%)	148 (25.5%)	59 (10.2%)	581 (100.0%)

reveals that antonym, negation and 7 exclusion linguistic phenomena (e.g., exclusion:common_sense, exclusion:modality, exclusion:modifier, exclusion:predicate_argument, exclusion:quantity, exclusion:spatial, exclusion:temporal) are the major source of contradiction label in MC, which correspond to 26.2% of the RITE-VAL development dataset.

The ranked distribution of linguistic phenomenon category ranking in NTCIR-11 RITE-VAL development dataset is presented in Table 3. This table shows that inference is ranked top 1 linguistic phenomena category and accounts for 12.9% of the total 581 pairs.

4. Experimental Results and Discussion

In order to understand the power of linguistic phenomenon for recognizing inference in text, we conduct experiments on syntactic, semantic features and linguistic phenomenon features. A total of 28 linguistic phenomenon categories information as well as 20 features were used with SVM classifier in the experiment.

Table 4 provides the experimental results of the performance of cross validation of each feature for the NTCIR-11 RITE-VAL development dataset in Chinese Traditional SV-BC Subtask (581 pairs). We used 20 features which consist of syntactic features (e.g., dependency parser) and semantic features (e.g., WordNet, synonyms, antonyms, negation words) with SVM classifier for the experiment. The results show that the accuracy of cross validation with single feature ranges 59.55% (F20: Dependency Parser) to 64.89% (F19: AntonmCount).

The experimental results of the top three models with the combination of syntactic and semantic features as well as linguistic phenomenon category feature used with SVM classifier can be compared in Table 5 and Figure 1. The experimental results indicate that the linguistic phenomenon category feature achieves the best cross

Table 3 Analysis of Linguistic Phenomenon Category Ranking in NTCIR-11 RITE-VAL Development Dataset (581 pairs)

Rank	Category ID	Linguistic Phenomenon Category	Y	N	Total	%
1	15	inference	51	24	75	12.9%
2	10	exclusion:predicate_argument		51	51	8.8%
3	26	synonymy:lex	47	1	48	8.3%
4	21	paraphrase	47		47	8.1%
5	19	modifier	34	3	37	6.4%
6	14	hypernymy	19	11	30	5.2%
7	24	scrambling	23	4	27	4.6%
8	5	clause	22	3	25	4.3%
9	4	case_alternation	21		21	3.6%
10	2	antonym		20	20	3.4%
11	17	list	17	3	20	3.4%
12	20	negation		20	20	3.4%
13	25	spatial	16	2	18	3.1%
14	9	exclusion:modifier		14	14	2.4%
15	12	exclusion:spatial		14	14	2.4%
16	28	transparent_head	13		13	2.2%
17	8	exclusion:modality		12	12	2.1%
18	16	lexical_entailment	11	1	12	2.1%
19	6	coreference	9	2	11	1.9%
20	22	quantity	10	1	11	1.9%
21	27	temporal	10	1	11	1.9%
22	7	exclusion:common_sense		8	8	1.4%
23	3	apposition	6	1	7	1.2%
24	13	exclusion:temporal		7	7	1.2%
25	1	abbreviation	6		6	1.0%
26	11	exclusion:quantity		6	6	1.0%
27	23	relative_clause	6		6	1.0%
28	18	meronymy	2	2	4	0.7%
		Total	370	211	581	100.0%

validation accuracy (81.41%) and outperform the top three models with the best combination of syntactic and semantic features (74.25%). The experimental results suggest that the single feature of linguistic phenomenon category enhance 7.16% of the accuracy of textual entailment system compared to traditional combination of syntactic and semantic features.

5. Conclusion

In this paper, we report the comprehensive analysis of identifying linguistic phenomena for recognizing inference in text (RITE). We have proposed a model by using an analysis of identifying linguistic phenomena for Recognizing Inference in Text (RITE) using the development dataset of NTCIR-11 RITE-VAL System Validation sub-task. The experimental results suggest that well identified linguistic phenomenon category could enhance the accuracy of textual entailment system.

The contributions of this paper are three fold:

(1) We proposed a model with the analysis of identifying the Chinese linguistic phenomena for

Table 4 Experimental results of the performance of cross validation of each feature for the NTCIR-11 RITE-VAL development dataset in Chinese Traditional SV-BC Subtask (581 pairs)

Feature ID	Feature Name	Closed Test (SV-BC)	Cross Validation (SV-BC)
F01	CharLengthT1	**66.44%**	61.27%
F02	CharLengthT2	65.58%	60.93%
F03	CharLengthDifference	64.54%	59.72%
F04	CharLengthRatio	64.03%	63.68%
F05	LCSSequence	64.72%	60.07%
F06	WordLengthT1	64.37%	63.51%
F07	WordLengthT2	64.89%	60.76%
F08	WordLengthDifference	65.40%	62.48%
F09	WordLengthRatio	64.03%	63.51%
F10	CharBasedED	64.03%	60.93%
F11	WordBasedEDC	64.20%	63.51%
F12	NounCount	63.68%	63.17%
F13	VerbCount	63.68%	63.68%
F14	WordSemainticSimilarity	64.03%	60.93%
F15	WordNetSimilarity	65.06%	63.17%
F16	WordNetSimilarityRatio	65.23%	63.17%
F17	WordNetSimilarityMin	65.23%	63.17%
F18	NegationCountCard	64.03%	63.34%
F19	AntonymCount	65.06%	**64.89%**
F20	Dependency Parser	64.72%	59.55%

Table 5 Experimental results of three models and linguistic phenomenon (581 pairs)

Models	Cross Validation
Model1	74.25%
Model2	66.27%
Model3	66.09%
Linguistic phenomenon (Category)	**81.41%**

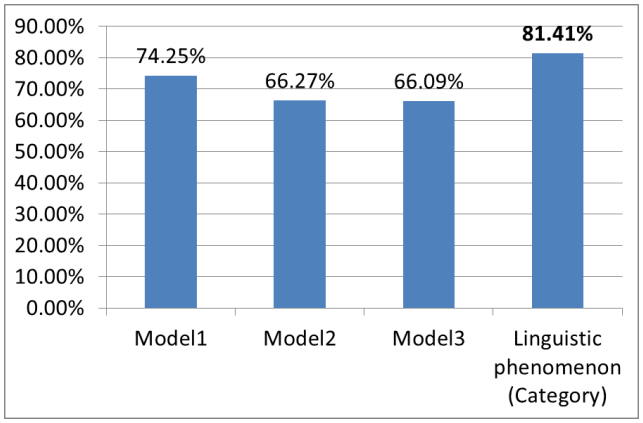

Figure 1 Analysis of linguistic phenomenon and top three models cross validation of NTCIR-11 RITE-VAL development dataset (Chinese Traditional SV-BC Subtask)(581 pairs)

recognizing interference in text for NTCIR-11 RITE-VAL system validation sub-task.

(2) We confirmed the power of linguistic phenomenon for recognizing inference in text in line with prior research in RTE.

(3) We thoroughly evaluate our proposed model in the context of the system validation subtasks of the NTCIR-11 RITE-VAL. The results demonstrate the efficacy of the proposed model for the NTCIR-11 RITE-VAL.

6. Acknowledgement

This research was supported in part by the National Science Council of Taiwan under Grants NSC101-3113-P-032-001 and TKU research grant.

7. References

[1] R. Bar-Haim, I. Dagan, I. Greental, and E. Shnarch, "Semantic inference at the lexical-syntactic level," in *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 2007, p. 871.

[2] L. Bentivogli, E. Cabrio, I. Dagan, D. Giampiccolo, M. L. Leggio, and B. Magnini, "Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference," in *LREC*, 2010.

[3] E. Cabrio and B. Magnini, "Towards component-based textual entailment," in *Proceedings of the Ninth International Conference on Computational Semantics*, 2011, pp. 320-324.

[4] E. Cabrio and B. Magnini, "Decomposing Semantic Inferences," *Linguistic Issues in Language Technology,* vol. 9, 2013.

[5] C. Condoravdi, D. Crouch, V. De Paiva, R. Stolle, and D. G. Bobrow, "Entailment, intensionality and text understanding," in *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*, 2003, pp. 38-45.

[6] I. Dagan, B. Dolan, B. Magnini, and D. Roth, "Recognizing textual entailment: Rational, evaluation and approaches," *Natural Language Engineering,* vol. 16, pp. 105-+, Jan 2010.

[7] M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning, "Finding Contradictions in Text," in *ACL*, 2008, pp. 1039-1047.

[8] M.-C. de Marneffe, A. R. Rafferty, and C. D. Manning, "Identifying conflicting information in texts," *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation,* 2011.

[9] S. Harabagiu, A. Hickl, and F. Lacatusu, "Negation, contrast and contradiction in text processing," in *AAAI*, 2006, pp. 755-762.

[10] B. Magnini and E. Cabrio, "Combining specialized entailment engines," *Proceedings of LTC'09,* 2009.

[11] S. Matsuyoshi, Y. Miyao, T. Shibata, C.-J. Lin, C.-W. Shih, Y. Watanabe, and T. Mitamura, "NTCIR-11 RITE-VAL," 2014.

[12] H.-M. Nguyen and K. Shirai, "Recognition of Agreement and Contradiction between Sentences in Support-Sentence Retrieval," 2013.

[13] A. Ritter, D. Downey, S. Soderland, and O. Etzioni, "It's a contradiction---no, it's not: a case study using functional relations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 11-20.

[14] N. Rooney, H. Wang, and P. S. Taylor, "An investigation into the application of ensemble learning for entailment classification," *Information Processing & Management,* vol. 50, pp. 87-103, Jan 2014.

[15] M. Sammons, V. Vydiswaran, and D. Roth, "Ask not what textual entailment can do for you," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1199-1208.

[16] M. Sammons, V. Vydiswaran, and D. Roth, "Recognizing textual entailment," *Multilingual Natural Language Applications: From Theory to Practice. Prentice Hall, Jun,* 2011.

[17] 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 the 9th NII Test Collection for Information Retrieval Workshop (NTCIR'11)*, 2011, pp. 291-301.

[18] A. Toledo, S. Alexandropoupou, S. Chesney, S. Katrenko, H. Klockmann, P. Kokke, B. Kruit, and Y. Winter, "Towards a Semantic Model for Textual Entailment Annotation," *Linguistic Issues in Language Technology,* vol. 9, 2014.

[19] L. Vanderwende and W. B. Dolan, "What syntax can contribute in the entailment task," *Machine Learning Challenges,* vol. 3944, pp. 205-216, 2006.

[20] Y. Watanabe, J. Mizuno, E. Nichols, K. Narisawa, K. Nabeshima, N. Okazaki, and K. Inui, "Leveraging Diverse Lexical Resources for Textual Entailment Recognition," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 11, p. 18, 2012.

[21] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C.-W. Lee, C.-J. Lin, S. Shi, T. Mitamura, and N. Kando, "Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10," in *Proceedings of the 10th NTCIR Conference*, 2013, pp. 385-404.

[22] H. Weisman, J. Berant, I. Szpektor, and I. Dagan, "Learning verb inference rules from linguistically-motivated evidence," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 194-204.

[23] Y. C. Wu, "Integrating statistical and lexical information for recognizing textual entailments in text," *Knowledge-Based Systems,* vol. 40, pp. 27-35, Mar 2013.