The WHUTE System in NTCIR-11 RITE-VAL Task

Han Ren, Hongmiao Wu
School of Foreign Languages and Literature, Wuhan University
Wuhan 430072, China
{hanren, hmwu}@whu.edu.cn

Xiwen Tan, Pengyuan Wang, Donghong Ji
School of Computer, Wuhan University
Wuhan 430072, China
{xwtan, pywang, dhji}@whu.edu.cn

Jing Wan
Hubei Research Base of Language and Intelligent Information Processing, Wuhan University, Wuhan 30072, China
jingwan@whu.edu.cn

ABSTRACT
This paper describes our system of recognizing textual entailment for RITEVAL System Validation and Fact Validation subtasks at NTCIR-11. For System Validation subtask, we employ a transformation model and acquire entailment rules by extracting synonyms and inferable expressions from resources such as lexicons and knowledge bases. Also, a cascaded entailment recognition model is employed to recognize four types of entailment relations. For Fact Validation subtask, we build a pipeline approach to find texts that entails given texts. First, a retrieval model is used to search related sentences from Wikipedia documents provided, then we used the recognition model in System Validation subtask to find such sentences that entailed the given texts. Official results show that our system achieves a performance of 53.48% MacroF1 score in Chinese SVBC subtask, a 25.74% MacroF1 score in Chinese SVMC subtask, a 45.51% MacroF1 score in English FV subtask and a 38.08% MacroF1 score in Chinese FV subtask.

Keywords
Recognizing Textual Entailment, Fact Validation, System Validation, Entailment Transformation, Cascaded Entailment Classification

1. INTRODUCTION
Researches on Recognizing Textual Entailment(RTE) concern inferable relations between texts, that is, one text can be inferred from another or not. Many efforts are made in RTE community, to facilitate text understanding and inference such as studying linguistic evidences of entailment, building inference resources, and exploring inference models or algorithms as well. RTE challenges, such as TREC and NTCIR, are also organized to survey and evaluate current entailment recognition technologies.

This year, NTCIR evaluation conference holds RITEVAL textual inference challenge[3], which is the third challenge of series RITE evaluation. Different with previous challenges, RITEVAL defines a new subtask named Fact Validation(FV), that is, a system should identify whether a text is entailed by another one, which is retrieved from Wikipedia or textbook. RITEVAL also remains the traditional RTE subtask, named System Validation(SV), to evaluate performances of participating systems in judging four entailment classes: forward, bidirection, contradiction and independence. Our system participated both two subtasks and submit two runs for each subtask, which are FV-EN, FV-CS, FV-CT, SVBC-CS, SVBC-CT, SVMC-CS and SVMC-CT.

Considering that the task definition of SV subtask in RITEVAL is similar with that of multi-classification(MC) subtask in RITE-2, the system implemented in the previous challenge are smoothly updated for RITEVAL SV subtask. More specifically, we collected entailment rules and background knowledge, such as geopolitical and celebrities information for transformation model. We also employ a cascaded entailment recognition model with three classifiers to recognize four types of entailment relations in order. For FV subtask, we built a pipeline approach, that is, first we employed a retrieval model to search related sentences from Wikipedia documents provided, then we used the recognition model in SV subtask to find such sentences that entailed the given texts.

Since inference resources and background knowledge were proved to impact the performance of RTE by many researchers[1], our system extracted entailment rules and employed knowledge bases such as online dictionaries, lexicons, Penn Treebank and PropBank, that are applied in models of transformation and classification.

The rest of this paper is organized as follows. In section 2, the architecture and workflow of the system are described. Section2 also gives a more detailed explanation of each parts for each subtasks, including preprocessing, transformation and entailment recognition approaches for SV subtasks. In section 3, we describe a framework and each part of our system for FV subtasks, including key term extraction, retrieval model and entailment recognition approach. Section 4 shows the experimental results and section 5 gives some discussions about system performance and error cases. Finally, some conclusions are given in section 6.

2. SYSTEM VALIDATION

2.1 System architecture
The overall architecture of our system for SV subtask is shown in Figure 1, which contains a preprocessing model, a transformation model, a feature extraction model and three classifiers. Procedures of the system are described as follows:

1) For each text fragment and hypothesis, a preprocessing procedure is performed, including word segmentation, part-of-speech tagging, named entity recognition, syntactic dependency parsing and semantic role labeling;

2) Texts after preprocessing are aligned through transformation approach, including directional and undirectional terms;

3) In feature extraction, string, structure and linguistic feature vectors are computed according to text pairs;

4) All features are employed to judge entailment or no entailment, and then forward, bidirectional, contradiction or independence through a cascaded classifier.

2.2 Preprocessing

The preprocessing procedure includes word segmentation, Part-Of-Speech(POS) tagging, named entity recognition, syntactic parsing and shallow semantic paring.

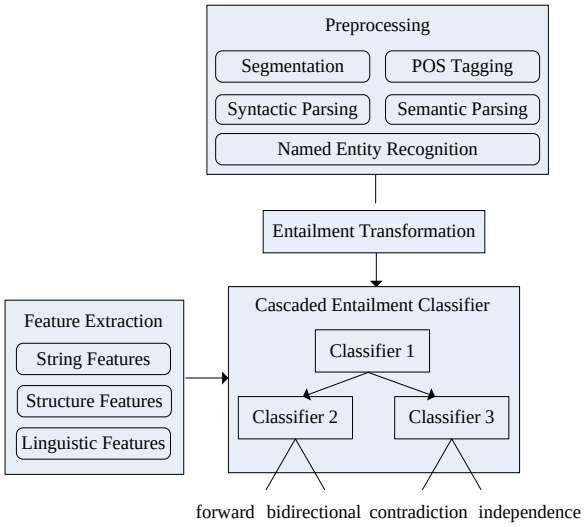

Figure 1. System architecture for SV subtask

Initially, we employ Stanford NLP tools [1], including word segmenter, POS tagger and named entity recognizer to deal with text and hypothesis in each pair. All tools are implemented by Java so that they are easily invoked by our system. In addition, we utilize a numeral normalization tool implemented in RITE-1[5], transforming temporal and Chinese numeral expressions to Arabic numerals.

The syntactic and semantic parsing model is our system for CoNLL2009[6], which labels syntactic and semantic dependency relations based on words. The reason is that syntactic and semantic dependency parsing are more flexible and precise in comparison with full parsing, hence semantic dependency relations are easier to improve the performance of our system. The annotation standard is identical with the definition in CoNLL2009, with 30 tags for the syntactic dependents and 25 tags for the semantic roles.

For training and testing of classification, three types of features are employed: string, structure and linguistic features, all of which are same with those employed in our prior system in RITE-2[7].

2.3 Entailment Transformation

Transformation is one of major strategies for entailment recognition[2, 3, 8] and frequently adopted for alignment and syntactic matching. In RTE, transformation is to search for a sequence of entailment rules, that turns a text to a hypothesis.

In our system, transformation proceeds before classification. More specifically, for each pair, text fragments in t_1 that do not exist in t_2 are picked out. Also, a counterpart of each text fragment is picked out from t_2 by searching from syntactic and semantic constituents that are same with text fragments in t_1. Then, such text fragments in t_1 are replaced. When all text fragments are replaced, the transformed pair are trained and predicted by entailment classifier.

[1] http://nlp.stanford.edu/software/

2.3.1 Undirectional Transformation

Directional transformation refers to an alternation of synonymous meaning. Compared with the prior system, we extract more synonymous expressions rather than synonymous words and named entities. For word transformation, we utilize an online resource, CIBA HANYU[2], to acquire synonyms. This resource is an online dictionary including common Chinese words and their synonyms or antonyms. The searching process is simple: we search synonyms for a word w_1 in t_1 in a pair, and then we search if any synonym is also in t_2. If such a word w_2 is in t_2 and it has the same syntactic/semantic constituent with w_1, we use w_2 to replace w_1 in t_1. The process iterates until every word in t_1 is visited. The transformation process for antonyms is similar, except that polarity values should be accumulated. For example, in pair 48 in Chinese SVMC test data, t_1 includes a word 短缺 shortage that is the antonym of 过剩 surplus in t_2. Thus the first word is transformed to the second one and the polarity is reversed for t_1.

We also introduce an extraction approach to acquire synonymous expressions from an online resource, i.e., Wikipedia. We consider two situations that synonyms often occur: one is Wikipedia redirection, the other is some expressions such as "also known as" or brackets after a term. For Wikipedia redirection, we search such terms directly to find synonyms in contents after redirection. For the second situation, heuristic rules are built to extract terms in brackets or after indicators such as "also known as". For example, the entailment judgment for pair 71 in Chinese SVBC subtask is to judge if the phrase 美国疾病控制与预防中心 Centers for Disease Control and Prevention has the same meaning with the abbreviation word CDC. We search the phrase from Wikipedia and fetch the first sentence from its introduction content: 美国疾病控制与预防中心（英文：Centers for Disease Control and Prevention，缩写：CDC）. Through rule matching we can find that CDC is the abbreviation of 美国疾病控制与预防中心. We pre-extract some synonyms from a Wikipedia dataset provided to build a entailment rule set. Before searching in Wikipedia, the system finds matching terms from this rule set.

2.3.2 Directional Transformation

Undirectional transformation refers to an asymmetric meaning alternation from t_1 to t_2. We consider hypernym, hyponym and meronym relation for such transformation. We extract words with such relations from HowNet, an ontology based knowledge base in Chinese. The procedure is as follows: first we search every word in t_1 from HowNet; if it is found, the further research is proceeded that whether its hypernymous or meronymous words are also appeared in t_2; if so, the word in t_1 will be replace by the hypernymous or meronymous word in t_2. On the other hand, spatial information such as geographic information also indicate direction entailment relations. For example, t_1 of the pair 819 in Chinese SVBC test set contains the word 瑞士 Switzerland, which is a country in 欧洲 Europe appeared in t_2. Acquiring such geographic information helps to recognize spatial entailment. To acquire geographic entailment relation, we utilize a geographic knowledge base extracted before and extract rules according to geographic hypernyms and hyponyms. If t_1 contains a geographic term and t_2 contains its hypernym, the former term will be replaced by the latter one in t_1.

[2] http://hanyu.iciba.com/

2.4 Cascaded Entailment Classification

The entailment type forward is supposed to be a directional relation. Considering that most features for single classification are undirectional ones, that is, such features estimate similarity that are undirectional, we duplicate some undirectional features such as word overlap and sub tree overlap feature to make them directional, that is, considering t_1 is the text and t_2 is the hypothesis, and then t_2 the hypothesis and t_1 the text. Intuitively, if a feature gets a high score under the condition that t_1 is the text and t_2 is the hypothesis, whereas it gets a low one under the condition that t_2 is the text and t_1 is the hypothesis, it probably indicates that t_1 entails t_2 and not vice versa.

On the other hand, although feature duplication helps to recognize directional entailment relations, the combination of directional and undirectional features may result in a reducing discrimination of support vectors, because such directional features only help to judge entailment relations of forward and bidirection, while it will make noise in judging entailment relations that do not need those directional features. Alternately, a cascaded entailment recognition strategy is utilized, similar with the model in our RITE-2 system[7], that is, a text pair is first judged entailment or no entailment, and then forward, bidirectional, contradiction or independence. More specifically, for each pair (t_1, t_2), a bi-categorization classifier is employed to judge whether t_1 entails t_2. Thus the problem is equivalent with that of the 2-way judgment in SVBC subtask. In our system in NTCIR-10 RITE-2 subtask, if t_1 entails t_2, the second classifier is employed to judge whether t_2 entails t_1 or not. Different with the approach, we employ a classifier that directly judge if t_1 and t_2 has a forward or bidirection relation. Using this approach, there is no need to estimate the threshold of entailment confidence value. If t_1 does not entail t_2, the third classifier is employed to judge whether there is a contradiction or independence relation between t_1 and t_2. Finally, the output is given according to the output of the second and the third classifier.

3. FACT VALIDATION

FV subtask shows an opposite viewpoint against SV subtask, that is, given a hypothesis, FV subtask attempts to find a text that entails the hypothesis. Hence a subset of texts that is relevant to each hypothesis should be found first.

3.1 System Architecture

The overall architecture of our system for FV subtask is shown in Figure 2, which are similar with the architecture of our system for SV subtask except the retrieval model. Procedures of the system are described as follows:

1) for each hypothesis, a key term extraction model is performed to extract key terms from it;
2) a retrieval model is employed to find sentences that are relevant to such key terms;
3) after preprocessing, each hypothesis and the retrieved sentences are aligned through transformation model;
4) features of each hypothesis and the retrieved sentences are built and employed in the entailment classifier for a final decision of entailment, contradiction and unknown(for English subset, the final decision is yes or no).

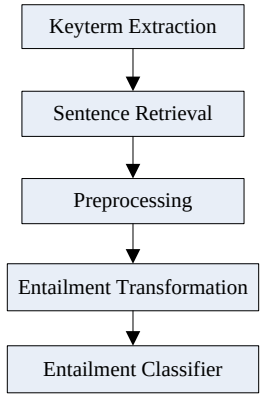

Figure 2. System architecture for FV subtask

3.2 Key Term Extraction

It is straightforward to build a query by using the whole sentence of a hypothesis. However, a text that entails a hypothesis probably has a different expression with the hypothesis, thus it is difficult to find relevant texts by using such method. Alternately, we use key terms in hypothesis to build query, since key terms in a text and its hypothesis are always same. To this end, we employ a key term extraction model, which is to acquire words and phrases that are built as queries. Since RITE organizer provides a retrieved result list for English dataset, we just implement key term extraction for Chinese dataset. After word segmentation, POS tagging and named entity recognition, words and named entities except punctuations and functional words are selected as key terms. For a better search performance, synonyms and those named entities with the same meaning are also appended as key terms.

3.3 Sentence Retrieval

The retrieval model we employ in our system is the one we used in a QA system, which showed a good performance in NTCIR-8 CCLQA task[4]. The retrieval model is based on Lucene, a free retrieval framework, and the index units are words. Same with the retrieval model in the QA system, we use BM25, which shows a better performance than VSM in answer retrieval, as the scoring approach for sentence candidate ranking.

3.4 Entailment Recognition

After getting sentence candidate list, a general RTE procedure can be made, that is, for each sentence candidate and the hypothesis, a preprocessing model is first performed for word segmentation, POS tagging, syntactic/semantic parsing and named entity recognition, then the entailment transformation model is performed to align two text pieces; after that, features are extracted for entailment classifier and a decision is made that whether the sentence candidate entails the hypothesis for English subtask. If so, the test unit is labeled as entailment, which means that the hypothesis is entailed by some sentences in the document set; otherwise, the test unit is labeled as non-entailment. For Chinese subtask, the system needs to decide that a hypothesis is entailed, contradicted or non-entailed by some sentences in the document set.

Entailment classification of the system for FV subtask is similar with that of the system for SV subtask except two changes. For Chinese FV subtask, the first tier classifier judges entailment or non-entailment relation of a text-hypothesis pair, then the second

tier classifier judges contradiction or unknown relation. For English FV subtask, a threshold is set up before entailment classification to filter those texts that have low ranking scores.

4. EXPERIMENTAL RESULTS

We participate in three subtasks, including Binary Class(SVBC), Multi Class(SVMC) and Fact Validation(FV), all of which contain traditional and simplified Chinese subtask in RITEVAL. In addition, FV subtask also defines an English subtask, in which an English training data is provided. This section reports the official RITE-3 results of these subtasks.

4.1 SVBC Subtask

For simplified Chinese SVBC subtask, we submit two runs: RITEVAL-WHUTE-CS-SVBC-01, RITEVAL-WHUTE-CS-SVBC-02. For traditional Chinese SVBC subtask, we also submit two runs: RITEVAL-WHUTE-CT-SVBC-01, RITEVAL-WHUTE-CT-SVBC-02. Since the traditional Chinese test data is identical with the simplified Chinese test data after traditional-simplified Chinese transformation, our official results of these two language versions are same in this subtask.

Our aim in this subtask is to estimate the impact of the transformation model to the entailment recognition, hence the experiments are set up as follows: the first run of each language version subtask employs the transformation model, while the second run does not. Table 1 shows the official results of these four runs, where Y denotes entailment relation, N non entailment relation, Prec. precision and Rec. recall.

Table 1. Official results of Chinese SVBC subtask

	WHUTE-CS-SVBC-01/ WHUTE-CT-SVBC-01	WHUTE-CS-SVBC-02/ WHUTE-CT-SVBC-02
MacroF1	0.5348	0.5196
Accuracy	0.5458	0.5283
Y-F1	0.6065	0.5844
Y-Prec.	0.5350	0.5223
Y-Rec.	0.7000	0.6633
N-F1	0.4631	0.4547
N-Prec.	0.5663	0.5388
N-Rec.	0.3917	0.3933

The first run achieves an outperforming performance in most cases, except for N-Rec, as shown in Table 1. More specifically, for entailment relation, run1 achieves a 1.17% performance increase of precision, a 3.67% increase of recall and a 2.21% increase of F1 metric in comparison with run2; for non entailment relation, the results of run1 show a 2.75% performance increase of precision, a 0.16% decrease of recall and a 0.84% increase of F1 metric in comparison with run2. As an overall performance, our system achieves an increasing 1.52% of MarcoF1 and 1.75% of Accuracy metric.

4.2 SVMC Subtask

For simplified Chinese SVMC subtask, we submit two runs: RITEVAL-WHUTE-CS-SVMC-01, RITEVAL-WHUTE-CS-SVMC-02. For traditional Chinese SVMC subtask, we also submit two runs: RITEVAL-WHUTE-CT-SVMC-01, RITEVAL-WHUTE-CT-SVMC-02. Since the traditional Chinese test data is identical with the simplified Chinese test data after traditional-simplified Chinese transformation, our official results of these two language versions are also same in this subtask.

Our aim in this subtask is to estimate the impact of the cascaded entailment recognition approach. Following this idea, the first run utilizes the cascaded recognition approach described in section 2.4, where three classifiers are trained for two-stage recognition, while the second run utilizes a unitary recognition approach, namely judges the entailment class directly by using a single classifier. Table 2 shows the official results, where F denotes forward entailment relation, B bidirectional relation, C contradiction relation and I independence relation.

Table 2. Official results of Chinese SVMC subtask

	WHUTE-CS-SVMC-01/ WHUTE-CT-SVMC-01	WHUTE-CS-SVMC-02/ WHUTE-CT-SVMC-02
MacroF1	0.2574	0.2430
Accuracy	0.3683	0.3508
B-F1	0.4657	0.4180
B-Prec.	0.4216	0.3902
B-Rec.	0.5200	0.4500
F-F1	0.5041	0.4946
F-Prec.	0.3472	0.3382
F-Rec.	0.9200	0.9200
C-F1	0.0599	0.0593
C-Prec.	0.2941	0.2703
C-Rec.	0.0333	0.0333
I-F1	0.000	0.000
I-Prec.	0.000	0.000
I-Rec.	0.000	0.000

Official results show that the first run achieves better performances in most cases. More specifically, for bidirection relation, the first run has a 3.14% increasing performance of precision, a 7% increase of recall and a 4.77% increase of F1 metric in comparison with the second run; for forward relation, performances of the first run have a 0.9% increase of precision, a zero raise of recall and a 0.95% increase of F1 metric compared to performances of the second run; for contradiction one, performances of the first run has a 2.38% increase of precision, a zero raise of recall and a 0.06% increase of F1 metric compared to performances of the second run; for independence one, all metrics of two runs are zero. As an overall performance, our system achieves an increasing 1.44% of MarcoF1 and 1.75% of Accuracy metric.

4.3 FV Subtask

For English FV subtask, we submit two runs: RITEVAL-WHUTE-EN-FV-01 and RITEVAL-WHUTE-EN-FV-02. For simplified Chinese FV subtask, we submit two runs: RITEVAL-WHUTE-CS-FV-01 and RITEVAL-WHUTE-CS-FV-02. For traditional Chinese SVMC subtask, we also submit two runs: RITEVAL-WHUTE-CT-FV-01 and RITEVAL-WHUTE-CT-FV-02. Since the traditional Chinese test data is identical with the simplified Chinese test data after traditional-simplified Chinese transformation, our official results of these two language versions are same in FV subtask.

Our aim in FV EN subtask is to estimate the impact of ranking scores in information retrieval to entailment recognition. We use two methods: the second run judge whether each retrieved text entails the hypothesis in each test pair, and the decision Y is made if at least one retrieved text entails the hypothesis; while the first

run set up a threshold of ranking score, and only those retrieved texts, each of which has a higher ranking score than the threshold, are picked out to make an entailment recognition with the hypothesis. Table 3 shows the official results of English subtask.

Table 3. Official results of English FV subtask

	WHUTE-EN-FV-01	WHUTE-EN-FV-02
MacroF1	0.4551	0.4492
Accuracy	0.5372	0.5213

Official results of FV English subtask show that the first run outperforms than the second run. More specifically, MacroF1 score of the first run has a 0.59% increasing performance than that of the second run, while accuracy score of the first run has a 1.59% increasing performance than that of the second run.

Our aim in FV Chinese subtask is to estimate the impact of different scoring model in information retrieval to entailment recognition. More specifically, the first run employs BM25 as the scoring model while the second run employs VSM with cosine similarity as the scoring method. Table 4 shows the official results of simplified and traditional Chinese FV subtask, where E-F1 denotes entailment category, C denotes contradiction category and U denotes unknown category(same with independence category in SV subtask).

Table 4. Official results of Chinese FV subtask

	WHUTE-CS-FV-01/ WHUTE-CT-FV-01-	WHUTE-CS-FV-02/ WHUTE-CT-FV-02
MacroF1	0.3808	0.3594
Accuracy	0.4192	0.3997
E-F1	0.4341	0.3920
E-Prec.	0.4734	0.4432
E-Rec.	0.4009	0.3514
C-F1	0.1873	0.1710
C-Prec.	0.3788	0.3382
C-Rec.	0.1244	0.1144
U-F1	0.5209	0.5152
U-Prec.	0.3983	0.3902
U-Rec.	0.7526	0.7579

Official results of FV Chinese subtask show that the first run achieves a better performance than the second run. More specifically, for entailment relation, the first run achieves a 3.02% increasing performance of precision, a 4.95% increase of recall and a 4.21% increase of F1 metric than the second run; for contradiction relation, the first run has a 4.06% increase of precision, a 1% increase of recall and a 1.63% increase of F1 score than the second run; for unknown relation, the first run has a 0.81% increase of precision, a 0.53% decrease of recall and a 0.57% increase of F1 score than the second run. As an overall performance, our system achieves an increasing performance 2.14% of MarcoF1 and 1.95% of Accuracy metric.

4.4 Linguistic Phenomena Based Evaluation

In gold standard data, RITEVAL organizer annotates detailed inference relations, namely linguistic phenomena. The organizer believes that annotation of linguistic phenomena can not only estimate the proposed effects to a specific sub-problem in textual entailment recognition, but also provide participants a more precise diagnostic tool to their system.

Since only linguistic phenomena contained in Chinese SV subtask are given by the official gold standard data, we compute error count and rate for each linguistic phenomenon in SVBC and SVMC subtasks, and the results are shown in table 5.

Table 5 shows that, linguistic phenomena of top three error count in the first run of Chinese SVBC subtask are inference, modifier and antonym, while linguistic phenomena of top three error rate in this run are antonym, negation and exclusion:modifier. In the second run of Chinese SVBC subtask, linguistic phenomena of top three error count are inference, modifier and synonym:lex, while linguistic phenomena of top three error rate in this run are antonym, negation and exclusion:modifier. As to Chinese SVMC subtask, linguistic phenomena of top three error count in the first run are inference, modifier and exclusion:predicate_argument, while linguistic phenomena of top three error rate in the this run are exclusion:predicate_argument, exclusion:quantity and negation. Linguistic phenomena of top three error count and rate in the second run of Chinese SVMC subtask are identical with the first run of Chinese SVMC subtask. As a counterpart, entailment judgments to coreference, case_alternation, list and clause achieve low error rates in both Chinese SVBC and SVMC subtasks.

5. DISCUSSION

In this section, we analyze performances of our system in every subtask and typical errors in our experiments. Also, some directions for further improvement are given.

5.1 System Performance

As to Chinese SVBC subtask, the usage of transformation model improves the performance of every metric in our experiments. It indicates that one factor that performance improvement of learning for entailment recognition partly is to replace synonymous expression. In fact, for text pairs with less overlapping text fragments, string similarity features such as Word Overlap and Common String Overlap probably lead to performance decline unless synonymous expressions are transformed or aligned correctly. Take the pair 62 in SVBC test data as an example, "废除" abolish in t_1 and "取消" remove in t_2 have the same meaning, but the entailment judgment is probably false if the first word is unable to transformed or aligned with the second one. Another example is the pair 188, which contains a geographic entailment relationship, that is, "亚洲" Asia in t_2 is subordinate to "全球" the whole world in t_1. If such directional entailment relation is not recognized by transformation model, the judgment will probably be false. As a statistical evidence, the linguistic phenomenon synonym:lex is in top three error count list of linguistic phenomena of the second run, whereas it does not appear at such list of the first run, mainly because the first run employs transformation model to replace synonymous words so as to keep text and hypothesis identical in form.

In the experiments of simplified MC subtask, the cascaded entailment classification improves performances of most metrics compared to the entailment recognition approach with single classifier, especially for bidirectional relation. And the reason is obvious: although generally, employing more features are conducive to improving learning performance, those features that are duplicated to judge if a hypothesis entails a text in the second run lead to more noise because those features are specific ones for recognizing bidirectional relation, whereas text pairs of other entailment classes are also featured by them. In other words, the

combination of general features for four categories and specific ones for only one category result in a reducing discrimination of support vectors. As a conclusion, the cascaded classifiers help to perform multi-categorization entailment recognition against bi-categorization one.

Table 5. Results Based on Linguistic Phenomena

Language Phenomena		CS-SVBC-01		CS-SVBC-02		CS-SVMC-01		CS-SVMC-02	
		Error Count	Error Rate	Error Count	Error Rate	Error Count	Error Rate	Error Count	Error Rate
entailment	abbreviation	10	0.40	12	0.48	13	0.52	15	0.60
	apposition	8	0.32	8	0.32	15	0.60	15	0.60
	case_alternation	7	0.25	8	0.3	11	0.41	12	0.44
	clause	14	0.24	14	0.24	19	0.32	19	0.32
	coreference	12	0.50	12	0.50	4	0.17	4	0.17
	hypernymy	12	0.44	12	0.44	19	0.70	19	0.70
	inference	99	0.54	99	0.54	109	0.59	109	0.59
	lexical_entailment	14	0.48	16	0.55	13	0.45	16	0.55
	list	13	0.35	13	0.35	10	0.27	10	0.27
	meronymy	9	0.39	9	0.39	16	0.70	16	0.70
	modifier	66	0.50	66	0.50	89	0.68	89	0.68
	paraphrase	13	0.27	16	0.33	13	0.27	16	0.32
	quantity	16	0.55	16	0.55	16	0.55	16	0.55
	relative_clause	10	0.28	10	0.28	13	0.36	13	0.36
	scrambling	14	0.40	14	0.40	20	0.57	20	0.57
	spatial	20	0.48	21	0.50	28	0.67	29	0.69
	synonymy:lex	21	0.41	31	0.61	27	0.53	36	0.71
	temporal	16	0.40	18	0.45	18	0.45	20	0.5
	transparent_head	5	0.19	5	0.19	14	0.54	14	0.54
contradiction	antonym	25	0.71	25	0.71	34	0.97	34	0.97
	exclusion:common_sense	21	0.62	21	0.62	33	0.97	33	0.97
	exclusion:modality	24	0.63	24	0.63	36	0.95	36	0.95
	exclusion:modifier	21	0.64	21	0.64	32	0.97	32	0.97
	exclusion:predicate_argument	21	0.55	21	0.55	38	1	38	1
	exclusion:quantity	10	0.34	10	0.34	29	1	29	1
	exclusion:spatial	10	0.31	10	0.31	28	0.88	28	0.88
	exclusion:temporal	15	0.44	15	0.44	33	0.97	33	0.97
	negation	19	0.68	19	0.68	28	1	28	1

It is also noticed that, in this subtask, all text pairs of independence relation are failure to be recognized, more specifically, most of them are wrong recognized as entailment relation. As a matter of fact, each text and hypothesis of such pairs have similar meaning, while many words and phrases are same between them. For example, t_1 and t_2 in pair 144 in SVMC test data are very similar except the word 二十世纪 twenty century in t_1 and the word 二十世纪开始 from the beginning of twenty century. Since 开始 begin is a temporal modifier of 二十世纪 twenty century, t_1 and t_2 have a different meaning. However, these two texts are so similar that it is difficult to decide whether such pair have entailment or independence relation due to a large proportion of similarity estimation features such as string or structure ones. In conclusion, more linguistic features should be introduced to describe detailed semantic relations in texts of each pair and the classifiers should learn from such features the relation between semantic relations and entailment judgment.

In the experiments of English FV subtask, the first run with a relevant filter outperforms than the second run without it. And the reason is obvious: not all top retrieved results deserves to be considered, more specifically, only those texts that are enough relevant with a hypothesis are probably entails the hypothesis. In the experiments of Chinese FV subtask, the first run with BM25 scoring model improves performances of most metrics compared to the second run with VSM model. This experience are also proved by many document retrieval models of QA systems, such as our system in NTCIR QA task[4].

System performance on the perspective of language phenomena also shows some notable conclusions: 1)comparing linguistic phenomena of top three error count in Chinese SVBC subtask, the second run makes more mistakes on synonym:lex phenomenon recognition than the first run. In other words, the first run which employs transformation model has a better recognition performance of synonyms than the second run without transformation model. 2) all runs in SVBC and SVMC achieve good performances recognizing entailment relations with the language phenomena such as list and clause, mainly because many words and phrases in hypothesis and text are same, so that most features in the classifiers are very similar and entailment relations are easy to be judged correctly. 3) pairs having complex language phenomena such as inference and modifier are hard to be recognized. Such pairs always contain complex semantic relations, and they should be identified to achieve a better

understanding of meaning between texts in such pairs. Therefore, precise language analysis technologies, such as anaphora resolution and semantic parsing should be employed for a better performance.

5.2 Error Analysis

This subsection discusses major error types with examples shown as follows. For the convenience of case analysis, text fragments are shown instead of full texts for some examples.

Take a close view to the error cases in RITEVAL subtasks, many contradiction pairs are false judged as entailment relation. For example, the pair 1134 in Chinese SVMC test data is a contradiction one:

(1) t_1: 沥青混凝土铺面道路俗称柏油路、沥青路. Roads paving asphalt concrete are known as tar roads or asphalt roads.
t_2: 沥青混凝土铺面道路常被误认为柏油路. Roads paving asphalt concrete are often misdeemed to tar roads.

Apparently, most of the words in t_1 and t_2 are identical. Since no negative words are found, the system makes the wrong judgment of forward relation in this case. As a matter of fact, the word 俗称 known as in t_1 has an opposite meaning against the word 误认为 misdeem in t_2, whereas neither these two words are antonyms, nor these two texts have negative words. To recognize contradiction of this type, an expansion of antonym dictionaries or opinion mining technologies are necessary.

The pair 19 in Chinese SVMC test data shows another type of contradiction:

(2) t_1: 《罪与罚》是俄国文学家杜斯妥也夫斯基的长篇小说作品. Crime and Punishment is a novel of Russian writer Dostoyevski.
t_2: 《罪与罚》是俄国科学家杜斯妥也夫斯基的长篇小说作品. Crime and Punishment is a novel of Russian scientist Dostoyevski.

This pair is wrong judged as entailment, and the reason is same with the above example: two texts are very similar. Essentially, the meaning of the word 文学家 writer in t_1 is different with that of the word 科学家 scientist in t_2. Therefore, to recognize contradiction of this type, the system should find whether two different words have a synonymous or hypernymous relation; if not, these two words probably indicate a contradiction relation.

Another type of errors comes from deficient linguistic parsing. For example, the pair 927 in Chinese SVBC is an non-entailment one:

(3) t_1: 植物之所以被称为食物链的生产者，是因为它们能够透过光合作用利用无机物生产有机物并且贮存能量. Plants are viewed as the producer of food chain because they have the ability of producing organic materials using inorganic ones with photosynthesis and storing energy as well.
t_2: 生产者能够透过光合作用利用无机物生产有机物并且贮存能量. Producer have the ability of producing organic materials using inorganic ones with photosynthesis and storing energy as well.

Apparently, 它们 they in t_1 refers to 植物 plants, not 生产者 producer. The error occurs because the system does not recognize such anaphora relation which is supposed to be an evidence of entailment judgment. Another examples are:

(4) t_1: 大宪章确立了一些英国贵族享有的政治权利与自由. The Great Charter established political rights and liberties enjoyed by British nobles.
t_2: 大宪章确立了一些贵族享有的政治权利与自由. The Great Charter established political rights and liberties enjoyed by nobles.

(5) t_1:水晶宫是十九世纪的英国建筑奇观之一. The Crystal Pa-lace is one of the architectural wonders in Great Britain in nineteenth century.
t_2:水晶宫是十九世纪的英国奇观之一. The Crystal Palace is one of the wonders in Great Britain in nineteenth century.

Example (4) comes from the pair 165 and (5) comes from the pair 257 in Chinese SVBC test data. t_1 entails t_2 in (5), although the word 建筑 architectural are removed, whereas t_1 does not entail t_2 in (5), when the word 英国 British are removed. In fact, the word 英国 British has a constraint relation with the word 权利 rights and the word 自由 liberties, while 建筑 architectural shows a modification relation to the word 奇观 wonders. It indicates that precise semantic relations are helpful for identifying entailment relation. As a direction, precise language processing technologies, such as anaphora resolution and semantic parsing, are supposed to be applied.

Background knowledge also impacts system performance. Take the pair 67 in Chinese SVBC test data as an example:

(6) t_1: 瑞典通过现代宪法, 瑞典国王在议会的一切权力被废除, 但内阁每月仍会向国王在王宫内正式汇报. The Swedish government had passed the modern constitution abrogating all privileges of Swedish King in the parliament, but the cabinet would still report to Swedish King formally in the palace every month.
t_2: 瑞典成为君主立宪国家. Sweden became a constitutional monarchy state.

In this case, most words in the two texts are different, thus the system makes a wrong judgment that t_1 does not entail t_2. In fact, background knowledge is required in this case, that is, the definition of 君主立宪 constitution monarchy should be provided, and the system should decide if t_1 entails such definition. To this end, efforts should be made in two aspects: 1)employing various resources and language technologies such as dictionaries, knowledge bases, retrieval and summarization approaches to find background knowledge required; 2)modeling background knowledge based inference in order to find semantic relations between texts in pairs.

6. CONCLUSION

In this paper, we describe our system for RITEVAL subtask at NTCIR-11. For System Validation subtask, we employ a transformation model and acquire entailment rules by extracting synonyms and inferable expressions from resources such as lexicons and knowledge bases. We also employ a cascaded entailment recognition model including three classifiers to recognize four types of entailment relations. For Fact Validation subtask, we build a pipeline approach to find texts that entails given texts. We employed a retrieval model to search related sentences from Wikipedia documents provided, then we used the

recognition model in SV subtask to find such sentences that entailed the given texts. Official results show improving performances of such models by comparing run1(having some models) and run2(removing some models) in each subtask.

Error cases in our experiments also indicate some improving directions: 1)contradiction is not only represented by antonyms or negative words, but also implied in opposite expressions, hence an expansion of antonym dictionaries or recognizing entailment relations such as hypernym, synonym and meronymy are necessary; 2)precise language processing technologies, such as anaphora resolution and semantic parsing, are supposed to be applied to find semantic relations that are helpful for identifying textual entailment; 3)background knowledge are required to be extracted and modeled in order to find semantic relations that can not be found in given texts.

7. ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China(Grant Nos. 61402341, 61373108, 61173062) and China Postdoctoral Science Foundation funded project (2014M552073).

8. REFERENCES

[1] Androutsopoulos, I. and Malakasiotis, P. 2010. A Survey of Paraphrasing and Textual Entailment Methods. Journal of Artificial Intelligence Research 38: 135-187.

[2] Bar-Haim, R., Berant, J., Dagan, I., Greental, I., Mirkin, s., Shnarch, E. and Szpektor, I. Efficient Semantic Deduction and Approximate Matching over Compact Parse Forests. *In proceedings of the Fourth PASCAL Challenges Workshop on Recognizing Textual Entailment*. Gaithersburg, Maryland, USA, 2008.

[3] Matsuyoshi, S., Miyao, Y., Shibata, T., Lin, C.-J., Shih, C.-W., Watanabe, Y. and Mitamura, T. Overview of the NTCIR-11 Recognizing Inference in TExt and Validation(RITE-VAL) Task. *In proceedings of the 11th NTCIR Workshop*. Tokyo, Japan, 2014.

[4] Ren, H., Ji, D. and Wan, J. WHU Question Answering System at NTCIR-8 ACLIA Task. *In proceedings of the 8th NTCIR conference*. Tokyo, Japan, 2010.

[5] Ren, H., Ji, D. and Wan, J. The WHUTE System in NTCIR-9 RITE Task. *In Proceedings of the 8th NTCIR workshop meeting*. Tokyo, Japan, 2011.

[6] Ren, H., Ji, D., Wan, J. and Zhang, M. Parsing Syntactic and Semantic Dependencies for Multiple Languages with A Pipeline Approach. *In Proceedings of the 13th Conference on Computational Natural Language Learning*. Boulder, Colorado, USA, 2009.

[7] Ren, H., Wu, H., Lv, C., Ji, D. and Wan, J. The WHUTE System in NTCIR-11 RITE Task. *In proceedings of the 10th NTCIR Conference*. Tokyo, Japan, 2013.

[8] Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W., Lin, C.-J., Shi, S., Mitamura, T., Kando, N., Shima, H. and Takeda, K. Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop. *In Proceedings of the 10th NTCIR Workshop*. Tokyo, Japan, 2013.