Discriminating Between Relevant and Irrelevant Text for Fact Validation

Mirai Miura
Nara Institute of Science and Technology, Japan
miura.mirai.me1@is.naist.jp

Hiroki Ouchi
Nara Institute of Science and Technology, Japan
ouchi.hiroki.nt6@is.naist.jp

Mai Omura
Nara Institute of Science and Technology, Japan
omura.mai.oz5@is.naist.jp

Mayo Yamasaki
Nara Institute of Science and Technology, Japan
yamasaki.mayo.yc4@is.naist.jp

Akifumi Yoshimoto
Nara Institute of Science and Technology, Japan
akifumi-y@is.naist.jp

ABSTRACT

The *CL* team paticipated in the Fact Validation (FV) and System Validation (SV) subtasks in Japanese. This paper describes our systems with experimental results. In the Fact Validation subtask, a system is required to search the given documents for texts (t_1) and judge the fact validity of the given statement (t_2) based on the judgement of whether t_1 entails t_2 or not. However, if t_1 selected by the system is irrelevant to t_2, existing RTE approaches do not work well for the validity judgement. Thus, it is a key to the accurate judgement of the fact validity how to search for and select relevant t_1. Our approach first discriminates between *relevant* and *irrelevant* t_1 based on the score computed by a search engine, TSUBAKI, and then adopts different methods of judging the fact validity for each t_2. If the system regards t_1 as relevant, a simple binary classification method is adopted to judge the validity. On the other hand, if the system regard t_1 as irrelevant, a full-text search engine, Solr, is used to compute retrieval scores different from the ones computed by TSUBAKI. These retrieval scores are used as features for the binary classification. The experiments show that our approach is effective for the fact validation.

Team Name

CL

Subtasks

Fact Validation (JA)
System Validation (JA)

Keywords

information retrieval, recognizing textual entailment, document search

1. INTRODUCTION

Recognizing textual entailment (RTE) is a broad task that captures textual inference, adressed by many researchers of NLP. The task of RTE is to judge for a pair of two texts, *Text* (t_1) and *Hypothesis* (t_2), whether t_1 entails t_2 or not [2]. In the Fact Validation subtask of NTCIR-11 RITE-VAL, given a text t_2, a system identifies whether t_2 is entailed from the sentences relevant to t_2, which are retrieved from Wikipedia or textbook [3]. Some sentences with search scores, search results of a search engine TSUBAKI [4], are provided as t_1 by the task organizer. However, because those t_1 are not always relevant to t_2, it can be advisable to search for texts corresponding to t_1. In this paper, we describe the approach adopted in our system focusing on discriminating whether t_1 is relevant or not and using distinct strategies for each t_1 for the Fact Validation subtask.

2. RECOGNIZING RELEVANCE OF TEXTUAL EVIDENCE IN FACT VALIDATION

In the Fact Validation subtask, a system is required to retrieve texts t_1 including contents relevant to t_2. If t_1 has no relation to t_2, existing RTE approaches do not work well for judging the validity of t_2.

(A) t_1. 大友義鎮らは，少年使節をローマ教皇のもとに派遣した．

t_2. また，九州の大友，大村，有馬，のキリシタン 3 大名は少年使節をローマに派遣して，我が国の伝導の様子を教皇に報告した．

(B) t_1. 国会議員に認められている日本国憲法上の地位として，国会の会期中に逮捕されない．

t_2. 天皇は，日本国の象徴とされ，明治憲法の定める統治権の総攬者としての憲法上の地位を失った．

In the example (A) above, t_1 retrieved from Wikipedia is relevant to the statement of t_2. In this case, existing RTE approaches work well for judging the validity of t_2. On the other hand, because t_1 in the example (B) is irrelevant to t_2, texutual information of t_1 does not contribute to the validity judgement even if RTE approaches are applied to it. Thus, it is crucail to discriminate relevant and irrelevant t_1 before judging whether t_1 entails t_2 or not. In the Fact Validation subtasks, at most five sentences corresponding to t_1 for each t_2 are provided, which are retrieved from the given textbooks and Wikipedia by using a search engine, TSUBAKI. However, the retrieval results by TSUBAKI are not always t_1 relevant to t_2. It is troublesome to retrieve texts relevant to t_2 because almost all the texts of the given document are irrelevant to t_2. So, we propose a simple approach supplementing the search results of TSUBAKI for more accurate judgement of the fact validation.

3. SYSTEM DISCRIPTION

In this section, we describe the approach of our system focusing on a simple technique of supplementing search results of TSUBAKI for the Fact Validation subtask. The figure 1 represents our system architecture.

3.1 Discriminating Between Relevant and Irrelevant Text

We first discriminate relevant and irrelevant t_1. In the Fact Validation subtask, at most five sentences corresponding to t_1 for each t_2 with the TSUBAKI score are provided by the task organizer. We select as t_1 the sentence with the highest score among the five candidate sentences. Then, based on the assumption that the higher TSUBAKI score is, the more likely t_1 is to be relevant to t_2, we simply define the sentence as *relevant* t_1 if the TSUBAKI score of it is higher than a threshold of the TSUBAKI score we set, and discriminate between relevant and irrelevant t_1.

3.2 Using Distinct Strategies for Relevant or Irrelevant Text

We use distinct strategies for relevant or irrelevant t_1 for judging the fact validity. Because relevant t_1 is expected to include useful textual clues for the validity judgement, existing approaches of RTE can be applied to and work well for judging the fact validity. Thus, we use SVM to judge the validity as binary classification task. Features are extracted from the t_1 and t_2 after segmenting the words using Mecab[1]. In this research, we utilize simple lexical overlapped-based features as follows,

Character n-gram coverage feature
This feature is the coverage ratio of character-based n-grams in t_2 with t_2, which suggests to what degree t_2 resembles t_1 at the character level.

Morpheme n-gram coverage feature
This feature is the coverage ratio of morpheme-based n-grams in t_2 with t_2, which suggests to what degree t_2 resembles t_1 at the morpheme level.

Longest common subsequence feature
This feature is the longest common subsequence between t_1 and t_2.

In terms of irrelevant t_1, textual information of t_1 is likely to be useless for judgement of the fact validity, and existing approaches of RTE can be expected not to work well. So, we judge the validity based on another retrieval score computed by a full-text search engine, Apache Solr[2]. We use the score as the feature of the SVM classifier to judge the validity instead of three kinds of the lexical overlapped-based features mentioned earlier. The score computed by Apache Solr is distinct from the one computed by TSUBAKI. The fomula of the Apache Solr score is as follows[1],

$$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot \sum_{t\,in\,q} \{tf(t\,in\,d) \cdot idf(t)^2 \cdot doc\_len\_norm(d)\}$$

[1]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2]http://lucene.apache.org/solr/

	Training Data	Formal Run Data
relevant	142	178
irrelevant	306	336

Table 1: The number of discriminated relevant and irrelevant t_1.

	Training Data		Formal Run Data	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Proposal	66.52	63.23	58.95	55.07
relevant	67.32	60.65	61.19	55.37
irrelevant	64.79	64.68	54.49	53.58

Table 2: The results of the Fact Validation subtask of the development and formal run data. *Proposal* is the result of our proposal approach which uses different features for *relevant* and *irrelevant* t_1; *relevant* is the result of only t_1 regarded as relevant; *irrelevant* is the result of only t_1 regarded as irrelevant.

In the above formula, $coord(q,d)$ represents how many query terms q appear in the document d, $queryNorm(q)$ the normarization function of q, $tf(t\,in\,d)$ the term frequency of the term t in the document d, $idf(t)$ the inverse document frequency of the term t, $doc\_len\_norm(d)$ the normarization function of the number of words appearing in the document d. In the Fact Validation subtask, the TSUBAKI score is assigned to one sentence. On the other hand, we assign the Apache Solr score to one document. This means that the Apache Solr score takes the whole document into consideration and can capture extra-sentencial information within the document. In other words, the TSUBAKI and Apache Solr score capture different aspects of text or document. So, we assume that if irrelevant t_1 is given by TSUBAKI, the Apache Solr score take an alternative role for the validation judgement. Specifically, if the TSUBAKI score of t_1 is lower than a threshold, we regard it as irrelevant to t_2, and utilize Apache Solr to compute the alternative score for each t_2. Instead of textual information, only this score is used as the feature for SVM classifier and judge the validity based on the binary classification. In order to retrieve text and compute the score, all nouns in t_2 are extracted and used as a query. Then, we search the document set with "or" retrieval using the query.

4. EXPERIMENTAL RESULTS

In the experiment, we investigate the effectiveness of the discrimination between relevant and irrelevant t_1 and using the distinct features for each. The given training data consists of the file named "dev" and the one named "test". We use the former for setting the threshold of the TSUBAKI score and the hyperparamerter of SVM, and use the latter for the evaluation. When determining the threshold of the TSUBAKI score used for discriminating between relevant and irrelevant t_1, we use 20% of the training data and manually select the threshold. In terms of the hyperparameters of SVM, we determine them by five-fold cross-validation of the training data.

We firstly select the given sentence with the highest TSUBAKI score as t_1 for each t_2. Then, we discriminate the two

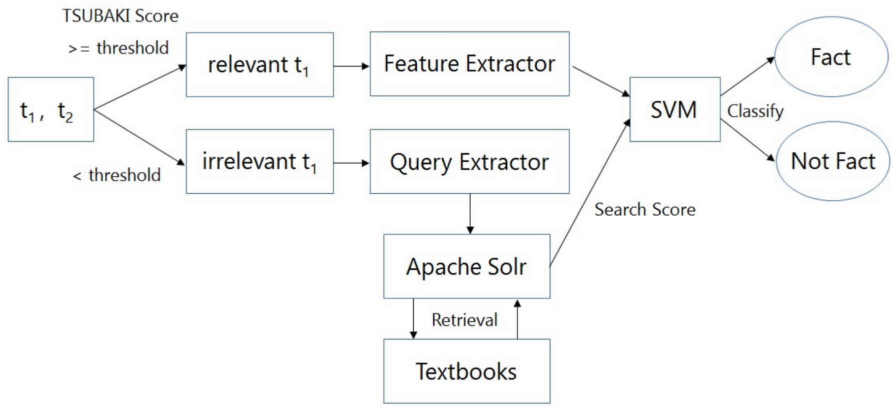

Figure 1: System architecture

	Training Data		Formal Run Data	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Proposal	66.52	63.23	58.95	55.07
All relevant	63.39	46.05	61.99	46.82
All irrelevant	62.50	61.13	54.97	54.35

Table 3: The comparison of the results. *Proposal* is the result of our proposal approach; *All relevant* is the result of all t_1 regarded as relevant; *All irrelevant* is the result of all t_1 regarded as irrelevant.

	Accuracy	Macro-F1
system1	76.50	64.17
system2	71.86	62.27
system3	71.86	65.27

Table 4: The results of the System Validation subtask

kinds of t_1 based on the threshold of the TSUBAKI score, and extract the distinct features for each. In order to judge the fact validity, we use *libsvm*[3] as the implementation of SVM classifier. In the Fact Validation and System Validation subtasks, we submit the three systems which adopt the same approach but have the different hyperparameter of SVM.

The table 1 shows the number of relevant and irrelevant t_1 discriminated based on the threshold of the TSUBAKI score. The number of irrelevant t_1 is twice as many as relevant one. The table 2 represents the result of the Fact Validatin subtask using the training data and formal run data set. Proposal in the table 2 is our proposal system discriminating relevant and irrelevant t_1 and using different features for each. The macro-F1 value is 55.07%. *relevant* in the table 2 is the result of only *relevant* t_1 in the proposal approach, and *irrelevant* in the table 2 the result of only irrelevant ones. Each macro-F1 value is 55.37%, 53.58%. The table 3 shows the comparison of the results. *Proposal* in the table 3 is our proposal method, *All relevant* is the case that all t_1 for all t_2 are regarded as *relevant* and the textual features used for SVM. *All irrelevant* is the case that all t_1 for all t_2 are regarded as *irrelevant* and the Apache Solr score used as the feature for SVM. The best Macro-F1 among the three is our proposal approach *Proposal*, 55.07, which means that it is effective to disctriminate relevant and irrelevant t_1 and use distinct features. Considering the result of *All relevant*, 46.82, it is advisable not to select t_1 based on only the TSUBAKI score and use lexical overlapped-based features for a classifier, because such t_1 are likely not to include useful information for judging the fact validity of t_2. On the other hand, the macro-F1 of *All irrelevant* is not so low as that of *All relevant* but lower than that of *Proposal*. This means that the Apache Solr score works well as the feature of SVM but do not reach the effectiveness of our proposal approach.

The table 4 shows the results of the System Validation subtask. In the System Validation subtask, we use the same approach and features as the ones in the Fact Validation subtask. The only difference among system1-3 in table 4 is the hyperparameters of SVM, which means that the same approach is adopted. The hyperparameter of each system are set as 1.0, 1.1, 1.2.

5. CONCLUSIONS

We introduced our approach to the Fact Validation in the NTCIR-11 RITE-VAL shared task. Our approach first discriminates relevant and irrelevant t_1, and then uses distinct features for SVM classifier, one is simple lexical overlapped-based features and the other is the score computed by using Apache Solr. Although the discrimination between relevant and irrelevant t_1 was made simply based on the TSUBAKI score in this research, the result was better than those of non-discrimination approaches. This shows that using textual information of relevant t_1 is useful for the validation judgement. Besides, in the case of irrelevant t_1, it is advisable to use alternative information. As an immediate future work, by adopting more sophisticated methods for the discrimination and feature engineering, the result can be more improved.

[3]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

6. REFERENCES

[1] S. Abe. Excel で学ぶ！ Lucene のスコア計算. Technical report, 株式会社 ロンウィット, 2011.

[2] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, 2006.

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

[4] K. Shinzato, T. Shibata, D. Kawahara, and S. Kurohashi. Tsubaki: An open search engine infrastructure for developing information access methodology. *Journal of information processing*, 20(1):216–227, 2012.