Learning approaches for recognizing textual entailment and finding contradiction in texts

MINH LE NGUYEN[†1] MINH PHAM[†2]
AKIRA SHIMAZU[†3]

This paper will introduce how machine learning methods and shallow semantic parsing can be applied for natural language understanding. One of the tasks in NLU is recognizing textual entailment, which is to decide whether the meaning of a text can be inferred from meaning of other one. In our work, we conduct an empirical study of the RTE task for Japanese, adopting a machine learning-based approach. We analyze the effects of using bilingual features, machine learning algorithms, and the impact of RTE resources on the performance of a RTE system. We also investigate the use of machine translation for the RTE and show that MT can be used to improve the performance of our RTE systems. We achieved promising results when attended the competitions on NTCIR-9 and NCTIR-10. The second task we would like to present in this paper is finding contradiction in texts. This task is difficult in the sense that we need to deeply understand the texts in order to find contradiction. Previous work on finding contradiction in text incorporate information derived from predicate-argument structures as features in a learning framework. In this paper, we would like to use sallow sematic parsing for these tasks using a simple rule-based framework. We discuss that our methods can be used with the learning based approaches for finding contradiction in texts.

1. Introduction

Recognizing Textual Entailment (RTE) is a fundamental task in Natural Language Understanding. It has been proposed with the aim of building a common applied semantic framework for modeling language variability [Dagan et al. 2006]. Given two text portions T (text) and H (hypothesis), the task is to determine whether the meaning of H can be inferred from the meaning of T. Textual entailment recognition is an important task for many natural language processing applications including machine translation, question answering, and text summarization. Textual entailment recognition has been paid much attention in English. There are a few systems for other language than English like Japanese and Vietnamese. In this paper, we would like to investigate how machine-learning approaches can be applied for other languages than English. We will empirically show which machine-learning model is appropriate for RTE problems in Japanese. In addition, we will indicate how bilingual features can effect to the performance of learning based RTE systems.

Another subtask of natural language understanding is to find contradiction in text (Marneffe et al., 2008). This task is necessary for many potential applications. For instance, contradiction needs to be recognized by question answering system or multi-document summarization systems (Harabagiu et al., 2006). Supervised learning model is used for the problem of recognizing contradiction (Marneffe et al., 2008). The limitation of this model is that we need to have annotated data, which require human effort and time consuming. In contrast to previous works, we focused on rule-based system for finding contradiction in texts, which would be used as an initial system for generating training data for supervised learning. The main component of our system is a contradiction detection module, which relies on the alignment of semantic role (SRL) frames extracted from the text and the hypothesis in each pair. We also consider using the binary relations extracted from the text and the hypothesis for finding the contradiction in text. Evaluation experiments on standard data sets (Giampiccolo et al., 2007; Giampiccolo et al., 2008; Bentivogli et al., 2009) show that the proposed system achieves better recall and F1 score for contradiction detection than the baseline methods, and the same recall as a state of the art supervised method for the task.

The remainder of the paper is organized as follows. Section 2 presents the proposed textual entailment recognition and Section 3 shows our method for finding contradiction in texts. Section 4 draws our conclusion and future works.

2. Textual Entailment Recognition using machine learning

2.1 The learning framework for RTE

Machine learning approaches have been applied successfully to many domains. In this paper we will show how it can be beneficial for textual entailment tasks. There are two important components for learning in RTE. The first question is that which learning method would be appropriate for this task. Typically, textual entailment recognition is referred to a binary classification method. We have investigated a number of machine learning models for RTE tasks. The machine learning models used in our work including SVM, Maximum Entropy Model, Random Forest, Boosting, and Bagging models.

Figure 1 shows the proposed machine-learning framework for RTE problem. The RTE system is divided into four main modules: bilingual enrichment, preprocessing, feature extraction, and training. At the beginning, each Japanese pair T/H is automatically translated into English using a MT engine. Then in preprocessing, both the Japanese pair and its associated translation pair are analyzed. After that, we use the features extracted from the pair and its translation for an entailment classifier to determine if the entailment relationship exists in the pair or not.

†1 JAIST.
†2 NICT
†3 JAIST

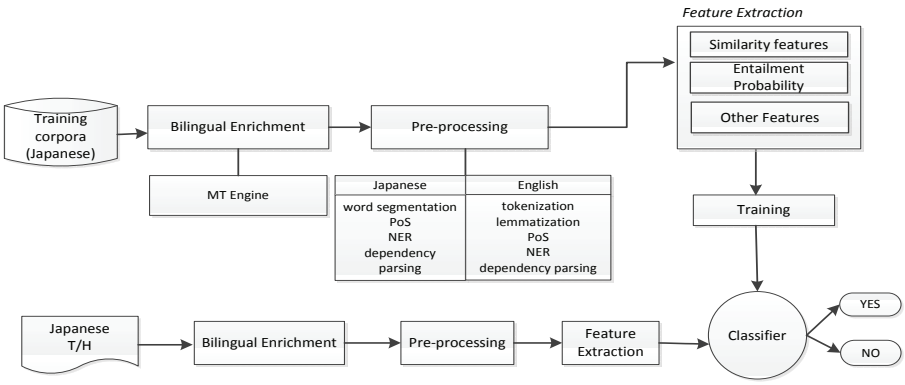

Figure 1. The proposed framework of RTE [Pham et al., 2012]

The entailment classifier is trained on the training set consisting of pairs T/H with their gold labels. In our framework, we investigate several machine learning algorithms for the RTE tasks: Support Vector Machines (SVMs) [Vapnik 1998], Maximum Entropy Model [Berger et al. 1996], and three ensemble learning algorithms: Bagging [Breiman 1996], Random Forest [Breiman 2001] and AdaBoost [Freund and Schapire 1996]. We selected Weka [Hall et al., 2009] as the tool for our machine learning models.

2.2 Feature Sets

The second component of RTE using machine learning approach is the feature space. We explored various kinds of feature sets including: Similarity Features, Entailment Probability, Dependency Relation Overlap Features, Named Entity Mismatch, and Polarity Mismatch. The detail descriptions about feature sets are shown in our previous works [Pham et al., 2012][Pham et al., 2013a]. In the scope of this paper, we briefly describe the feature set in our system as follows.

Similarity features: We used a large part of similarity features in the entailment classifier, which include: word overlap, Levenshtein distance [Malakasiotis and Androutsopoulos 2007], BLEU measures [Papineni et al. 2002], *Longest Common Subsequence String (LCS)*[Hirschberg 1977], *Jaccard Coefficient, Dice Coefficient, Manhatan Distance*, *Euclidean Distance,* Jaro-Winkler distance [Winkler 1999], and *Cosine Similarity*. These similarity measures are calculated in the pair of strings and then each similarity measure will be used as a feature for our learning model.

Entailment Probability. The entailment probability that T entails H is computed based on the probabilistic entailment model in [Glickman et al. 2005]. This feature will be used in our learning model.

Dependency Relation Overlap Features. Dependency relation overlap has been used in paraphrase identification [Wan et al. 2006]. This feature will be used in our learning framework.

Named-Entity Mismatch. In a pair T/H, if the hypothesis contains a named-entity, which does not occur in the text, the text may not entail the hypothesis. An indicator function π is used to compute the named-entity mismatch feature of T and H: $\pi(T, H) = 1$ if H contains a named-entity that does not occur in T and $\pi(T,H) = 0$, otherwise. Named-entity mismatch for both Japanese pairs and their associated English translation pairs are computed.

Polarity Mismatch. The polarity mismatch in a pair T/H may indicate that T does not entail H. We compute polarity mismatch in a pair T/H using the Polarity Weighted Word List [Pham et al., 2012]. In that list, each Japanese word is associated with a weight that indicates whether the word has positive meaning or negative meaning. We use an indicator function to capture if words in the root nodes of dependency parses of T and H have opposite polarity. The polarity mismatch is applied only on Japanese pairs.

2.3 Data and Experimental Results

We preformed our experiment on the benchmark data for Japanese RTE subtasks: binary-class subtask (BC subtask) [Shima et al., 2011]. The BC subtask is the basic problem setting of RTE, which is to determine whether the meaning of a hypothesis H can be inferred from the meaning of a text T. The data set consists of pairs T/H along with their gold-standard labels "Y" or "N". An entailment classifier is trained on the training data and evaluates the trained classifier on the test portion. Classification accuracy and average F1-score are used to evaluate RTE methods. We evaluated the proposed systems using several machine learning approaches and feature sets. We run the system using two settings for each machine learning approaches. The first setting and the second setting use monolingual features (extracted from Japanese pairs for training and testing) and the bilingual features (extracted from both original Japanese pairs and their associated English translation pairs), respectively. To obtain English translation pairs, we used the Google API translation component. We also tried with other API (i.e. BING), however the results are not much different. Table 1 showed that our method significantly outperforms three baselines. Especially, we obtained the best accuracy when we use bagging algorithm with bilingual features (accuracy of 58.6%). Table 1 also clearly indicated that bilingual features could improve the performance of RTE systems in terms of accuracy and F-measure.

Table 1. The experimental results for other machine learning models and feature set

Method	Accuracy (%)		F-measure	
LLM Baseline	49.0		0.48	
PA-matching	49.2		0.37	
Two-stage method	51.6		0.465	
SVM-bilingual (mono)	56.4	56.2	0.564	0.56
MEM-bilingual (mono)	55.6	54.8	0.553	0.54
Bagging-bilingual (mono)	**58.6**	**57.8**	**0.586**	**0.578**
Random Forest-bi (mono)	55.8	54.2	0.558	0.542
Adaboost-bilingual (mono)	56.6	56.4	0.566	0.564

We saw that Bagging and SVM model are better than MEM and other learning methods. We selected these two models for testing on the NTCIR-9 and NTCIR-10 competitions. Table 2 shows our performance in the completion tests.

Table 2. Official Runs on BC Subtask at NTCIR9-RITE and NTICR10-RITE

Methods	NTCIR-9	NTCIR-10
SVM+bilingual features	0.580	0.762
Bagging+bilingual features	0.586	0.768

In NTICR-9, our system achieved the best result and in NTCIR 10 our system achieved the performance, which is better than the average result.

3. Finding Contradiction in text

This section will describe our study on finding contradiction in text. Unlike previous works, which focus on machine learning model for this task, we discuss how a rule-based method can be used as initial step for the learning based approach in term of generating data for supervised learning models. First, we would like to introduce an overview of our framework and then we show our experimental results on benchmark data sets.

3.1 The Approach

Figure 2 shows the architecture of the proposed system. The system takes as input a pair *(T,H)*. *T* and *H* are input to the Linguistic Analysis module, which performs text preprocessing, semantic role labeling (SRL), and relation extraction for *T* and *H*. To obtain semantic role labeling we applied the tool described in [Collobert et al., 2011]. There are two main modules in the contradiction detection component.

- The first model (SRL-based module) checks the contradiction relationship in the pair over verb frames (SRL frames).
- The second module – triple-based module utilizes binary relations extracted from *T* and *H* for classification.

The CD component is organized in a two-stage scheme. If the SRL-based module fails to check the contradiction relationship, the triple-based module will be used as a backup engine. The two-stage scheme is proposed to address the low-coverage problem of the SRL-based module. The detail of our components is described in our technical report [Pham et al., 2013b].

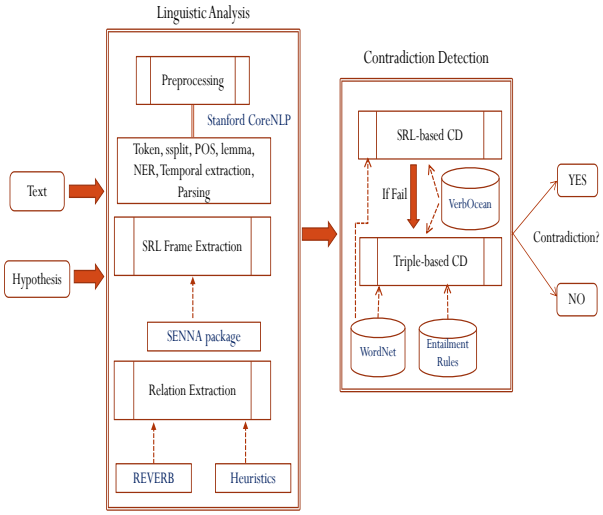

Figure 2. The checking contradiction system [Pham et al., 2013b]

3.2 Experimental Results

In experiments, we evaluate the proposed method on the test sets of the three-way subtask at RTE-3, RTE-4, and RTE-5 competitions (Giampiccolo et al., 2007; Giampiccolo et al., 2008; Bentivogli et al., 2009). The development sets provided at each competition are used to turn threshold values in two CD modules of the system. We use Precision, Recall, and F1 score of the contradiction label as evaluation measures. The first baseline method is the method presented in (Marneffe et al., 2008), which employed supervised machine learning techniques for the CD task. The second baseline is the BLUE system of Boeing's team (Clark and Harrison, 2009) at RTE-4 and RTE-5 competitions. We also compare the results achieved by our system with average results of submitted systems for RTE-3, RTE-4 and RTE-5 challenges. In order to assess the effectiveness of the two-stage systems scheme, we separately run each CD module on the three data sets and compare the results with those of the combined system.

Table 3. Experimental results on the benchmark data

Method	RTE-3			RTE-4			RTE-5		
	F1	P	R	F1	P	R	F1	P	R
Marneffe	21.04	19.44	22.95	-			-		
Blue system	-			16.13	41.67	10.0	11.54	42.86	6.67
Average result	14.28	10.72	11.69	13.63	25.26	13.47	14.79	26.40	13.70
SRL-based	14.28	13.41	15.27	19.55	22.41	17.33	19.23	22.72	16.67
Triple-based	13.59	22.58	9.72	14.49	26.3	10.0	16.67	19.48	16.67
Our system	**16.27**	**14.0**	**19.44**	**22.82**	**23.0**	**22.67**	**24.4**	**21.14**	**28.89**

The first module will compare the SRL-based contradiction score of each pair with a threshold. If the score is greater than or equal to the threshold, it determines that the contradiction relation exists in the pair. Similarly, the second module recognizes the contradiction relationship by using triple-base contradiction scores, which are calculated on the pair.

Table 3 provides experimental results achieved on test sets of RTE-3, RTE-4, and RTE-5 challenges by our system and baseline methods. Table 3 shows that the proposed system consistently obtained better recall values and F1 scores than those of baseline methods except the supervised machine learning-based method in (Marneffe et al., 2008). The results suggest that it would be helpful if we could combine the supervised learning approach with other methods. Table 3 indicated that the SRL-based module achieved better recall and F1 score than those of the triple-based module. It could be explained that the information contained in shallow semantic representations is richer than that of extraction triples, so the SRL-based module covers more contradiction phenomena than the triple-based module. It indicated that our system could recognize more contradiction phenomena than the baseline methods. Furthermore, the combined system consistently obtained better recall and F1 score than each individual module. Experimental results also show that the second backup module increases the coverage of contradiction phenomena for our system. Our system achieved the best precision in RTE-3, RTE-4, and RTE-5 in comparison with other strong baseline systems. The results strongly suggested that our method could perform in any text documents to find contradiction because we do not need training data. We can exploit our system to obtain labeled data for supervised learning frameworks.

4. Conclusions

This paper presents a machine learning approach for textual entailment recognition and semantic processing for finding contradiction in text. We have shown that the use of bilingual information with machine learning models is essential for improving the accuracy of RTE. Ensemble learning model (bagging) achieved the best performance in comparison with other machine learning models. We also compared the machine learning approach for finding contradiction in texts with our rule-based model using semantic processing. Experimental results on the benchmark data showed that our approach can be used as an initial step for generating training data for the learning based approach in the task of finding contradiction in texts.

Reference

[Androutsopolos 2010] I. Androutsopoulos and P. Malakasitotis, A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research 38*, 135–187, 2010.

[Breiman 96] L.Breiman Bagging predictors. *Machine Learning 24*, 123–140, 1996

[Breinman 2001] L. Breinman. Random forests. *Machine Learning 45,* 1, 5–32, 2001

[Bentivoli 09] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini. 2009. The fifth pascal recognizing textual entailment challenge. *In Proceedings of TAC Workshop*.

[Berant 2011] J. Berant, I. Dagan, and J. Goldberger. Global learning of typed entailment rules. In *Proceedings of ACL-HLT 2011*

[Burchardt 09]A. Burchardt, M. Pennacchiotti, S. Thater, and M. Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering*, 15(Special Issue 04):527–550.

[Chklovski 04] T. Chklovski and P. Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, July.

[Clark and Harrison 2009] P. Clark and P. Harrison. 2009. Recognizing textual entailment with logical inference. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

[Collobert et al., 2011] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. pp* 2493–2537, November.

[Chang and Lin 11] C.C. Chang and J.C. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[Dagan 07] I. Dagan, D. Roth, and F. Massimo. 2007. A tutorial on textual entailment..

[Freund 96] Y. Freund and R.E. Shapire. 1996. Experiments with a new boosting algorithm. In *Proceedings of 13th International Conference on Machine Learning*. 148–156.

[Fellbaum 98] C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

[Glikeman 05] O. Glikcman, I. Dagan, and M. Koppel KOPPEL. Web based probabilistic textual entailment. *In Proceedings of the 1st RTE Workshop*. Southampton, UK, 2005.

[Giampiccolo 2007] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. 2007. The third pascal recogniz- ing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9.

[Giampiccolo 2008] D. Giampiccolo, H.T. Dang, B. Magnini, I. Dagan, E. Cabrio, and B. Dolan. 2008. The fourth Pascal recognizing textual entailment challenge. In *Proceedings of TAC 2008 Workshop*.

[Harabagiu 06] S. Harabagiu and A. Hicki, 2006. Methods for using textual entailment in open-domain question answering. *In Proceedings of ACL '06*. 905–912.

[Hall et al., 2009] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten (2009); The WEKA Data Mining Software: An Update; *SIGKDD Explorations, Volume 11, Issue 1.*

[Hirschberg 77] S.D. Hirschberg. 1977. Algorithms for the longest common subsequence problem. *J. ACM 24*, 664–675.

[Harbagiu et al., 2006] S. Harabagiu, A. Hickl, and F. Lacatusu. 2006. Negation, contrast, and contradiction in text processing. In *Proceedings AAAI 2006*.

[Marneffe et al., 2008] M.D. Marneffe, A.N. Rafferty, and C.D. Manning. 2008. Finding contradictions in text. *In Proceedings of ACL 2008*.

[Pham et al., 2012] M. Pham, M.L. Nguyen, A. Shimazu. An Empirical Study of Recognizing Textual Entailment in Japanese Text, ACM *Transaction on Asian Language Information Processing (ACM-TALIP), 11.4. 2012*

[Pham et al., 2013a] M. Pham, M.L. Nguyen, A. Shimazu. JAIST Participation at NTCIR-10 RITE-2, Proceedings of the 10th NTCIR Conference, June 18-21, 2013, Tokyo, Japan

[Pham et al., 2013b] M. Pham, M.L. Nguyen, A. Shimazu, "Using Shallow Semantic Parsing and Relation Extraction for Finding Contradiction in Text", *Research Technical Report-JAIST, IS-RR-2013-002, pp. 1-10, 2013*

[Shima et al., 2011] H. Shima, H.Kanayama, C.W. Lee, C.J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. 2011. Overview of ntcir9 rite: Recognizing inference in text. *In Proceedings of NITCIR-9 2011*.

[Vapnik 1998] V. Vapnik. 1998. *Statistical learning theory*. John Wiley.

[Wan et al., 2006] S. Wan, M. Dras, R. Dale, and C. Paris Using dependency-based features to take the "paraphrase" out of paraphrase. In *Proceedings of ALTW* 2006.