Student Response Analysis using Textual Entailment

Ashudeep Singh[1], **Devanshu Arya**[2], **Amitabha Mukerjee**[1]

[1]Computer Science and Engineering, Indian Institute of Technology Kanpur
[2]Electrical Engineering, Indian Institute of Technology Kanpur
ashudeep@iitk.ac.in, devanshu@iitk.ac.in, amit@cse.iitk.ac.in

1. Introduction

A major task in Educational NLP is to assess student responses to examination questions, homeworks and intelligent tutors. Much of the related work has been done in evaluating student essays [1][2], error detection and correction [4] and grade level text classification [3]. A subtask in student dialogue systems is Student Response Analysis (SRA) i.e. given a question and a few reference answers, the system needs to analyze student response and decide whether it is correct or else give a suitable feedback. A key requirement to accomplish this is semantic inference, for example to detect whether the student answers say the same thing as the reference answer in different words or contradict it.

2. Recognizing Textual Entailment

Recognizing Textual Entailment (RTE 2005-2012) is a series of challenges where many systems that can recognize semantic inference have been presented. This task, as a part of SemEval 2013, was aimed at bringing Educational NLP closer to semantic inference community.

3. Student Response Analysis Corpus

The corpus contains manually labeled students responses to explanation and definition questions[8]. Specifically, the data set contains a question, a reference answer and a 1-2 sentence student answer. Each student answer is labeled as one of the five judgments by a human annotator:

- **Correct**, when the student answer is a complete and correct representation of one of the reference answers.
- **Partially correct but Incomplete**, when the student answer is correct to the extent it is written, but is not complete.
- **Contradictory**, when the student answer contradicts the reference answer, i.e. both cannot be correct at the same time.
- **Irrelevant**, when the student answer is irrelevant to the reference answer although it may still be talking about the reference answer.
- **Non-Domain**, when the student answer lacks domain content, i.e. the student may be asking for 'help', 'advice' etc. like 'I don't know', 'what the book says', 'you are stupid'.

The SRA Corpus consists of 2 subsets: BEETLE and SCIENTSBANK. The BEETLE corpus contains 56 questions in basic electricity and electronics domain with 3000 student answers. The SCIENTSBANK corpus[6] contains 197 assessment questions with 10,000 student answers in 15 different science domains.

4. Main Task

The main task is to produce an assessment of student answers to explanation and definition questions asked seen in practice exercises, tests or dialogue. The main task is to assess a student's answer at 3 different levels of granularity, namely:

- **5-way task**
  - correct
  - partially correct but incomplete
  - contradictory
  - irrelevant
  - not-in-the-domain
- **3-way task**
  - correct
  - contradictory
  - incorrect : comprising of paritally-correct incomplete, irrelevant and non-domain answers.
- **2-way task**
  - correct
  - incorrect

Such recognition of partial entailment may have various utilities in the educational setting based on identifying the missing parts in the student answer, and may similarly have value in other applications such as summarization or question answering.

5. Approach

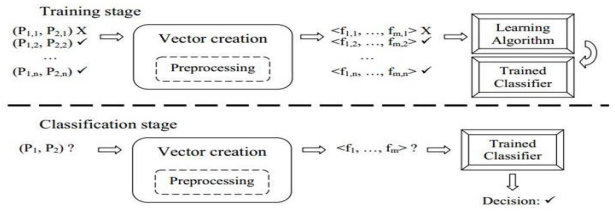

Figure 1: Figure [7] showing in brief the process of Textual Entailment Recognition via supervised machine learning

There is a correlation between textual entailment and answer correctness. In a typical answer assessment scenario, we expect a correct answer to entail the reference answer. However a student may wish to skip the details already mentioned in the question. So the problem basically is whether the answer, along with the question entails the reference answer. Let the question be q, student answer be s and the reference answer be r. Correctness means $a \wedge s \Rightarrow r$ and Contradiction means $s\wedge \Rightarrow \neg r$.

Recognizing textual entailment challenges have addressed such challenges since 2005[5]. We plan to use a RTE system along with shallow text features to train on the SRA (BEETLE and SCIENTSBANK) dataset and testing it on the test dataset provided. The evaluation metrics used will be according to the SemEval 2013, Task 7 problem statement, which are intuitively based on the $precision$ and $recall$ values for each class. The two important measures that describe the systems performance here are:

- **Macro-Average**: The average of the F1-scores of all the classes.
- **Micro-Average**: The weighted average of all the classes, where weights are the frequencies of each class in the test set.

6. Train and Test Scenarios

The dataset was used to create two scenarios for traing and testin purposes, namely Unseen Answers(UA) and Unseen Questios(UQ).

- In **Unseen Answers** case, all student answered to 1 to 2 randomly selected questions from each of the modules forming the training set whuch was held out as the test data set. The objective for creating such dataset was to test the system performance on new questions within the same set of domains.
- In **Unseen Questions** case, 4 randomly selected student answers to each of the questions in the training set was withheld. This would serve as test data for system performance on the same questions as contained in the training set.

7. Features

To build a system that evaluates student answers, we use the labeled training set to train our system on. For this, we require numerous text-similarity and overlap features that can well describe how close the students answer is to the reference answer. In the further sub-sections we describe the features that defined our feature-set on which the system was trained.

7.1. Baseline Features

It includes all of the features provided by the task organizers as baseline. There are four types of lexically-driven text similarity measures, and each is computed by comparing the learner response to both the expected answer(s) and the question, resulting in eight features in total – four in comparison the question, four with the maximum overlapping reference answer. The four types of measures include:

- **Overlapping Words** : It is simply the number of overlapping words between the student and reference answers.
- **Cosine Similarity** : Representing each answer as a bag-of-words vector, cosine similarity is the cosine of the angle between the vectors.
- **F1 Score** : Defining precision and recall on the number of hits i.e. co-occurences of words, F1-score is the harmonic mean of these precision, recall values.
- **Lesk Score** : Simplified lesk score is used to compare the overlap in meanings of the words in answers. [9]

Following are the features we added to the system to make it perform better:

7.2. Matching Content words

We created a vocabulary of content words i.e. the words dealing with the subject talked about in the question-answering assignment. As the data-set was from electronics and electrical domain question-answers, we used a hand-made dictionary of such words.

7.3. Polarity

Our training feature set includes a polarity feature that captures the presence (or absence) of linguistic markers of negative polarity in both text and hypothesis, such as not, no, few, without, except etc.. If there is simultaneous

existence or non-existence of a negative polarity word in both the students answers and the reference answers or the question, the polarity feature contained a value of 1 otherwise a negative value was assigned.

7.4. Antonymy

This feature checks whether an aligned pair of words in respective answers and questions appear to be antonymous by consulting WordNet ontology[12]. If there is an occurrence of such pair of antonym words, then it checks the polarity of the word preceding these words. For example, if there are words in the text and hypothesis as good and bad than it assigns a boolean positive to this feature. However, if bad is preceeded by not, than this feature returns a boolean negative.

7.5. Synonymy

Similar to antonymy features, this feature checks for the presence of synonym words in the texts and then checks the polarity of the preceding word. The assignment of boolean digits is done in the same way as in Antonym Features.

7.6. Number, Date and Time Features

These features recognize (mis-)matches between numbers, dates, and times, between the texts. If the numerical data in both the text lies in the same range than it is assigned a positive boolean otherwise a negative value is given.

7.7. Wu-Palmer Similarity

For each word in the student's answer we add the highest Wu-Palmer similarity with a word from the reference answer. Wu-Palmer similarity is a measure that uses WordNet ontology to say how similar two words are.[11]

$\frac{2.depth(lcs)}{depth(s_1)+depth(s_2)}$, where lcs is the least common subsequence, s_1 and s_2 are the strings.

8. Results

We ran our system on the test-sets according to the two scenarios – **UA** (Unseen Answers) and **UQ** (Unseen Questions). We incrementally added features to our system and observed the Micro and Macro-average of the F1-scores. Following are the results of our system compared with the Baseline and top 3 entries of each task:
The features have been represented as abbreviations:

- **CW** – Matching Content Words
- **P** – Polarity feature
- **S** – Synonymy feature
- **A** – Anonymy feature
- **N** – Number, date and Time features
- **WuP** – Wu-Palmer similarity

The classifiers used are:

- **j48** – Weka implementation of Decision Trees
- **k-NN** – k-Nearest Neighbors using k=35.

Features	Classifier	UA		UQ	
		Macro	Micro	Macro	Micro
Baseline	j48	0.42	0.48	0.41	0.46
Baseline+CW	j48	0.46	0.53	0.39	0.45
	Logistic	0.45	0.54	0.43	0.51
Baseline+CW+P+N	j48	0.49	0.57	0.50	0.51
Baseline+CW+P+N+S+A+WuP	j48	0.47	0.57	0.49	0.5
	K-NN	0.50	0.61	0.47	0.51
ETS$_2$	j48	0.619	0.705	0.552	0.614
CoMeT$_1$	-	0.569	0.675	0.300	0.445
EHUALM$_1$	-	0.526	0.566	0.300	0.416

Table 1: Comparison of our system's results with the top-3 entries in the SemEval 2013 5-way task

Features	Classifier	UA		UA	
		Macro	Micro	Macro	Micro
Baseline	j48	0.55	0.58	0.48	0.50
Baseline+CW+P+N+S+A+WuP	j48	0.57	0.60	0.46	0.48
	KNN	0.63	0.65	0.49	0.51
ETS$_2$[1] **[10]**	j48	0.71	0.723	0.585	0.597
CoMeT$_1$	-	0.715	0.728	0.466	0.488
ETS$_1$	j48	0.592	0.619	0.521	0.512

Table 2: Comparison of our system's results with the top-3 entries in the SemEval 2013 3-way task

Features	Classifier	UA		UQ	
		Macro	Micro	Macro	Micro
Baseline	j48	0.78	0.80	0.73	0.73
Baseline+CW+P+N	j48	0.76	0.77	0.7	0.71
Baseline+CW+P+N+S+A+WuP	j48	0.77	0.77	0.72	0.73
	K-NN	0.81	0.82	0.73	0.74
ETS$_2$	j48	0.833	-	0.702	-
CoMeT$_1$	-	0.833	-	0.695	-
CU$_1$	-	0.778	-	0.689	-

Table 3: Comparison of our system's results with the top-3 entries in the SemEval 2013 5-way task

9. Conclusion

The recognition of textual entailment in text has been hot for a few years now and its application in Student Response Analysis looks promising. We presented our feature set that we used to define our systems training parameters. The results reflect the system's ability to correctly evaluate student answers in majority cases. But its quite sure that there is a long way to go before we can develop robust systems that do the task in everyday classroom scenario.

10. References

[1] Yigal Attali and Jill Burstein. 2006. *Automated essay scoring with e-rater v.2*. The Journal of Technology, Learning, and Assessment, 43, February.

[2] Mark D. Shermis and Jill Burstein, editors. 2013. *Handbook on Automated Essay Evaluation: Current Applications and New Directions.* Routledge

[3] Sarah Petersen and Mari Ostendorf. 2009. *A machine learning approach to reading level assessment.* Computer, Speech and Language, 231:8906.

[4] Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel R. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[5] Dagan, I., Glickman, O., & Magnini, B. 2006. *The pascal recognising textual entailment challenge.* In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment $pp.177 - 190$. Springer Berlin Heidelberg.

[6] Rodney D. Nielsen, Wayne Ward, James H. Martin, and Martha Palmer. 2008b. *Annotating studentsunderstanding of science concepts.* In Proceedings of the Sixth International Language Resources and Evaluation Conference, $LREC08$, Marrakech, Morocco.

[7] Androutsopoulos, Ion, and Prodromos Malakasiotis. "A survey of paraphrasing and textual entailment methods." arXiv preprint arXiv:0912.3747 (2009).

[8] Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. *Towards effective tutorial feedback for explanation questions: A dataset and baselines.* In Proc. of 2012 Conference of NAACL: Human Language Technologies, pages 20010.

[9] Banerjee, S., & Pedersen, T. (2002). *An adapted Lesk algorithm for word sense disambiguation using WordNet. In Computational linguistics and intelligent text processing (pp. 136-145).* Springer Berlin Heidelberg.

[10] Heilman, M., & Madnani, N. ETS: Domain Adaptation and Stacking for Short Answer Scoring. SemEval 2013.

[11] Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004 (pp. 38-41). Association for Computational Linguistics.

[12] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.