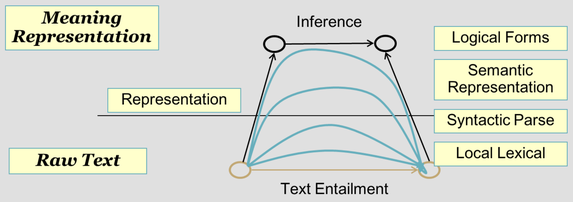
Student Response Analysis using Textual Entailment

Ashudeep Singh Devanshu Arya Amitabha Mukerjee

Indian Institute of Technology Kanpur

Introduction

Textual Entailment



- ***Textual entailment (TE)*** is a directional relation between text fragments where the relation holds whenever the truth of one text fragment follows from another text.
- TE can act as a framework for other NLP applications like QA, Summarization, IR etc.
- Textual entailment is not the same as pure logical entailment- it has a more relaxed definition: "text entails hypothesis" (t ⇒ h) if, typically, a human reading t would infer that h is most likely true.

Pascal RTE Challenges

- Held annually, since 2005. Task: To build a system that identifies entailment in text.
- In SemEval-2013 held as The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge.
- The BEETLE Dataset , 56 questions in electricity and electronics domain with 3000 student answers

```
<questionText>Why didn't bulbs A and C go out after bulb B burned
out?</questionText>
- <referenceAnswers>
    <referenceAnswer category="BEST" id="answer366"
    fileID="....">Bulbs A and C are still contained in closed paths
    with the battery</referenceAnswer>
      <referenceAnswer category="GOOD" id="answer367"
    fileID="....">Bulbs A and C are still in closed
    paths</referenceAnswer>
</referenceAnswers>
-<studentAnswers>
    <studentAnswer count="1" id="...." accuracy="correct">because
    bulb a and c were still contained within a closed path with the
    battery</studentAnswer>
    <studentAnswer count="1" id="..."
    accuracy="contradictory">they are on seperate
    circuits</studentAnswer>
</studentAnswers>
  </question>
```

The Task

A student response can be evaluated to be one out of the following:

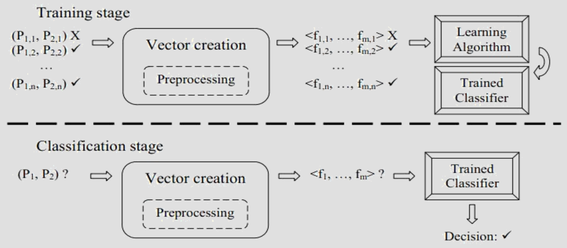
- **Correct**, if the student answer is a correct paraphrase of the reference answer
- **Partially correct/Incomplete**, if the student answer is containing some but not all information from the reference answer
- **Contradictory**, if the student answer explicitly contradicts the reference answer
- **Irrelevant**, if the student answer is not providing the necessary information
- **Non Domain**, if the student answer expresses a request for help, frustration or lack of domain

The three Tasks	
Task	**Classes**
2-way	Correct, Incorrect
3-way	Correct, Incorrect, Contradictory
5-way	Correct, Partially correct, Contradictory, Irrelevant, Non-domain

Training and Test Scenarios

- **Unseen Questions (UQ)**, all student answers to 1 to 2 randomly selected questions from each of the modules forming the training set held out as the test data set to provide a test of the system performance on new questions within the same set of domains.
- **Unseen Answers (UA)**, 4 randomly selected student answers to each of the questions in the training set is withheld to test the performance on the same questions as contained in the training set.

Our Work



Features

Baseline Features

- **Overlapping Words**
- **F1 Score**
- ***Cosine Similarity***
- **Lesk Score**

Polarity (P): Capturing the presence or absence of linguistic markers of negative polarity in both text and hypothesis, such as not, no, isn't, without, except etc.

Synonymy/ Antonymy (S/A): Using the WordNet list of synonyms/antonyms to check whether the text and hypothesis have different/same polarity synonyms/antonyms.

Number, Date and Time (N): These are designed to recognize (mis-)matches between numbers, dates and times.

Named Entity (NE): Features added to calculate overlap between named entities in student and reference answers.

Wu-Palmer WordNet Similarity (WuP): Measure of ontological similarities between each word of the reference and student answers

$$PF(C_1,C_2) = (1-\lambda).Min(N_1,N_2)-N)+\lambda(|N_1-N_2|+1)^{-1}$$

Matching Content Words (CW): Measure of the overlap amongst content words (only relating to the subject) within student and reference answer.

Results

5-Way Task

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	0.50
	K-NN	0.50	0.61	0.47	0.51

3-Way Task

Features	Classifier	UA		UQ	
		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
	K-NN (N=35)	0.63	0.65	0.49	0.51

2-Way Task

Features	Classifier	UA		UQ	
		Macro	Micro	Macro	Micro
Baseline	j48	0.78	0.80	0.73	0.73
Baseline +CW+P+N+S+A	j48	0.76	0.77	0.70	0.71
Baseline+CW+P +N+S+A+WuP	j48	0.77	0.77	0.72	0.73
	K-NN (N=35)	0.81	0.82	0.73	0.74

References

1. **Towards effective tutorial feedback for explanation questions: A dataset and baselines** NAACL 2012
2. **Learning to recognize features of valid textual entailments**. Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006.