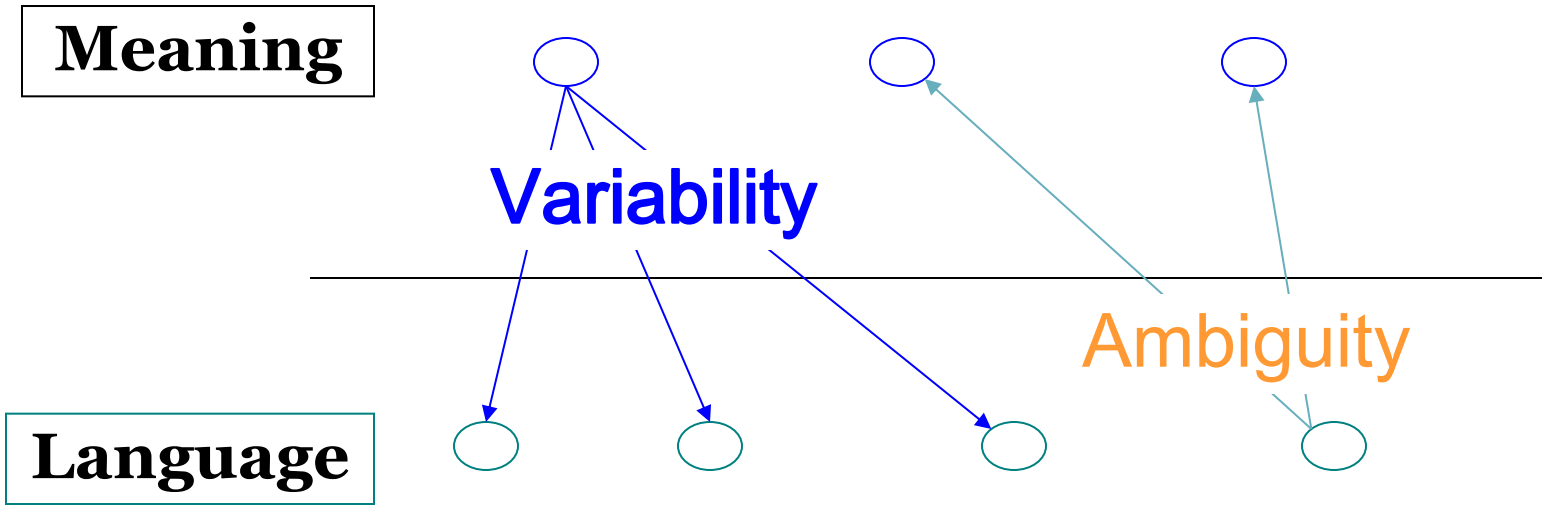
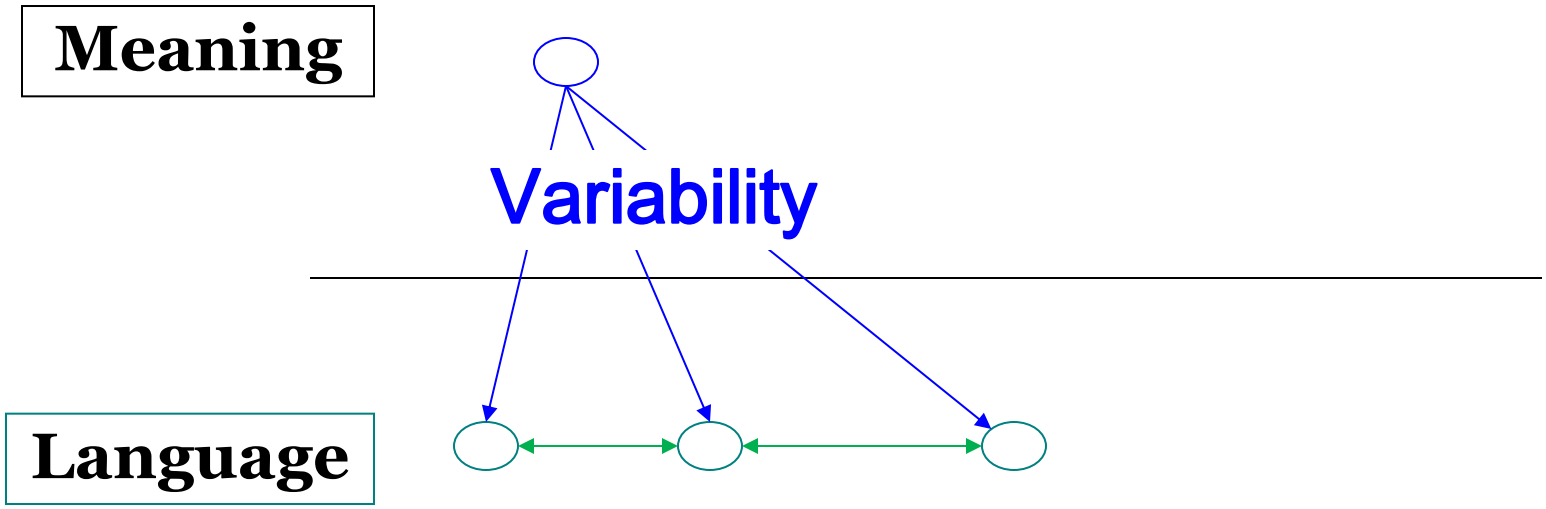
Student Response Analysis
Using Textual Entailment

Ashudeep Singh
Devanshu Arya

Natural Language and Meaning



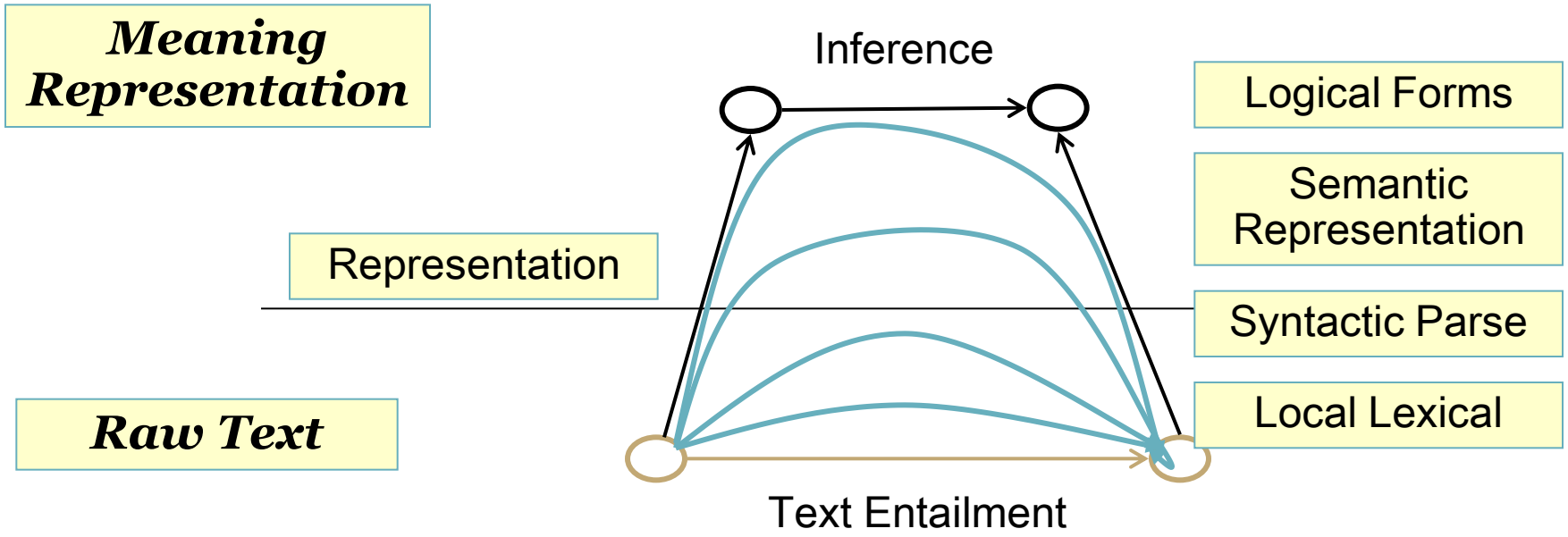
Natural Language and Meaning



Textual Entailment

- Whether one piece of text follows from another.
- Text entailment (TE) can be looked upon as mapping between variable language forms.
- TE as a framework for other NLP applications like QA, Summarization, IR etc.

Basic Representations



- Mapping possible at different levels of the language.
  - Lexical level
  - Syntactic level
  - Semantic level
  - Logical level

The PASCAL RTE Challenge

- Held annually, since 2005
  - Task: To figure out whether text ⇒ hypothesis
  - Dataset:
    - Example of a YES result

<pair id="28" entailment="YES" task="IE" length="short">
<t>*As much as 200 mm of rain have been recorded in portions of British Columbia , on the west coast of Canada since Monday.*</t>
<h>*British Columbia is located in Canada.*</h>
</pair>

    - Example of a NO result

<pair id="20" entailment="NO" task="IE" length="short">
<t>*Blue Mountain Lumber is a subsidiary of Malaysian forestry transnational corporation, Ernslaw One.*</t>
<h>*Blue Mountain Lumber owns Ernslaw One.*</h>
</pair>

- In SemEval-2013 held as *The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge*.

Datasets

The SRA Corpus consists of 2 datasets:

- ***The* BEETLE dataset,** which is a set of transcripts of students interacting with an intelligent tutorial dialogue system for teaching conceptual knowledge in the basic electricity and electronics domain (Dzikovska et al., 2010).
  - 56 questions with 3000 student answers

- ***The Science Entailments corpus (*SciEntsBank*)*** is based on the fine-grained annotations for constructed responses to science assessment questions by Nielsen at al. (2008), which were automatically mapped to the 5-way labels as described in (Dzikovska, Nielsen and Brew, 2010).
  - 197 questions with 10,000 student answers in 15 science domains.

The Task

- Given student's textual answer to a system's question – asses the answer relative to a reference answer

```
<question qtype="Q_EXPLAIN_SPECIFIC" ……>
  <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 entailment perspective: Student answer should paraphrase or entail the reference

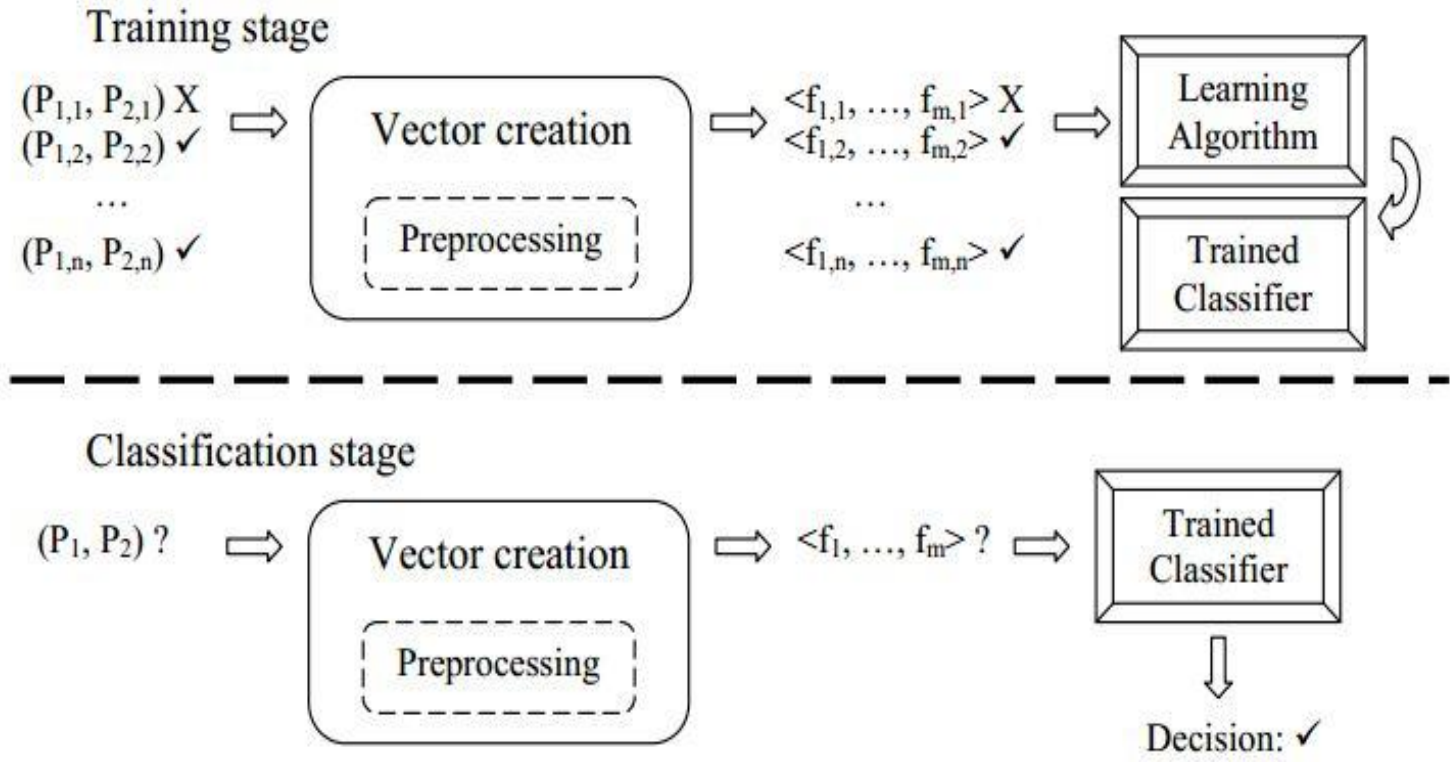
The Task

- Classifying the student responses as a 5-way task
  - **Correct,** if the student answer is a complete and correct paraphrase of the reference answer;
  - **Partially_correct_incomplete,** if the student answer is a partially correct answer 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 "irrelevant", talking about domain content but not providing the necessary information;
  - **Non_domain,** if the student answer expresses a request for help, frustration or lack of domain knowledge - e.g., "*I don't know*", "*as the book says*", "*you are stupid*".

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 will be held out as the test data set.
  - It will 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.
  - serves as test data for system performance on the ***same questions as contained in the training set***.

Textual Entailment Recognition via supervised machine learning



Baseline Features

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

Compared the student answers to reference answers and questions, resulting in total features (the four values indicated above for the comparison with the question and the highest of each value from the comparisons with each possible expected answer).

"*Towards effective tutorial feedback for explanation questions: A dataset and baselines*" NAACL 2012

Results - Baseline

	Precision	Recall	F-score
correct	0.61	0.71	0.66
partially_correct _incomplete	0.26	0.25	0.26
contradictory	0.38	0.28	0.32
irrelevant	0.13	0.11	0.12
non_domain	0.6	0.9	0.72
macroaverage	**0.4**	**0.45**	**0.41**
microaverage	**0.46**	**0.48**	**0.46**

Unseen Answers (UA)

	Precision	Recall	F–Score
correct	0.608	0.709	0.655
partially_correct _incomplete	0.261	0.25	0.255
contradictory	0.382	0.279	0.322
irrelevant	0.133	0.105	0.118
non_domain	0.6	0.9	0.72
macroaverage	**0.397**	**0.449**	**0.414**
microaverage	**0.457**	**0.48**	**0.463**

Unseen Questions (UQ)

Features Used:

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

Classifier: Weka J48 Decision tree

Additional Features

- **Polarity features**: capturing the presence (or absence) of linguistic markers of negative polarity in both text and hypothesis, such as not, no, few, without, except etc.
- **Antonymy features**: Using the WordNet list of antonyms to check whether the text and hypothesis have different/same polarity antonyms.
- **Number, date and time features**: These are designed to recognize (mis-)matches between numbers, dates and times.

Learning to recognize features of valid textual entailments.
Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006.

Results

precision	recall	fmeasure	
correct	0.633	0.8125	0.711
partially_correct_ incomplete	0.461	0.3125	0.372
contradictory	0.446	0.4054	0.425
irrelevant	0.125	0.0588	0.08
non_domain	0.643	0.7826	0.706
macroaverage	**0.461**	**0.4744**	**0.459**
microaverage	**0.522**	**0.5513**	**0.528**

Unseen Answers (UA)

precision	recall	fmeasure	
correct	0.589	0.701	0.64
partially_correct _incomplete	0.259	0.244	0.251
contradictory	0.385	0.275	0.321
irrelevant	0	0	0
non_domain	0.643	0.9	0.75
macroaverage	**0.375**	**0.424**	**0.392**
microaverage	**0.448**	**0.471**	**0.454**

Unseen Questions (UQ)

Features Used:
Baseline +

- Antonymy features
- Polarity
- Numeric, date, time matching

Classifier: Weka J48 Decision tree

Thank You!!!

Questions