CoMeT: Integrating different levels of linguistic modeling for meaning assessment

Niels Ott, Ramon Ziai, Michael Hahn, Detmar Meurers
Collaborative Research Center 833, University of Tübingen

Overview

- Approach: Stacked classifier based on three sub-systems which vary in linguistic abstraction: CoMiC, CoSeC and three bag approaches
- Results: Competitive good performance in general, especially in unseen answers scenarios, best result in Beetle 3-way uA (73.1%)
- Our background: Meaning assessment of answers to reading comprehension questions in German second language learning

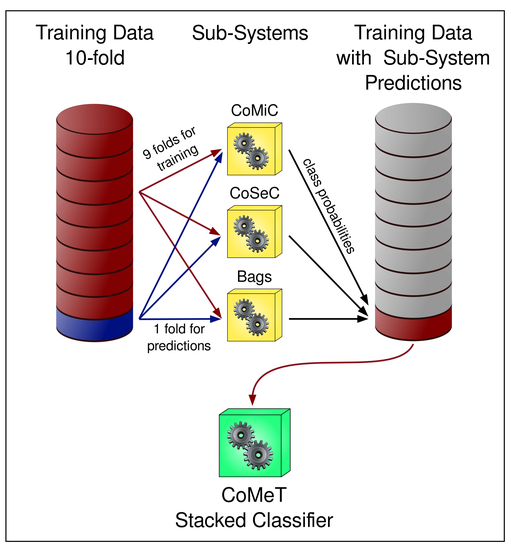
Stacked Classifier

Comparing Meaning in Tübingen (CoMeT)

- Different systems have different strengths, e.g., bag models are effective when in-domain training data is available
- Needed: a way to combine system predictions
- Majority voting is an option, but ignores information

➥ Train a classifier on individual system outputs:

  - Each sub-system produces training set predictions using 10-fold CV, as well as test set predictions using entire training set
  - CoMeT is trained on per-class probabilities of the individual system predictions and produces final system outcome
  - Classification based on Logistic Regression with ridge estimation (Weka Logistic)

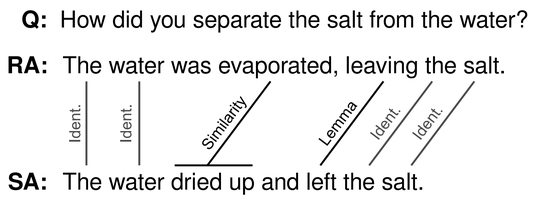


- Different system combinations for different scenarios:
  - CoMiC + CoSeC for unseen topics and unseen questions
  - CoMiC + CoSeC + bag approaches for unseen answers

Systems

Comparing Meaning in Context (CoMiC, Meurers, Ziai, Ott & Bailey 2011)

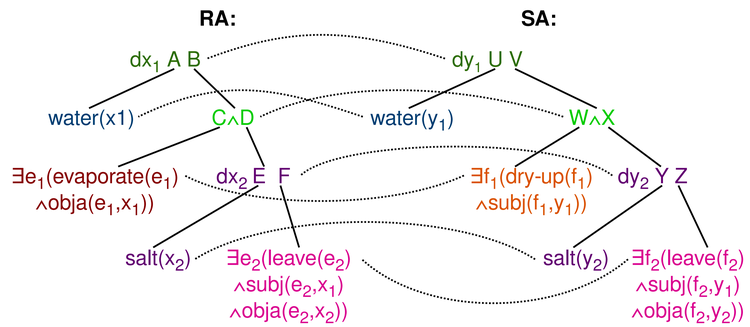
- Alignment of student and reference answers based on robust surface representations:



- Features encode number and type of alignment links (e.g., relative number of lemma alignment links).
- Classification based on Decision Trees (Weka J48)

Comparing Semantics in Context (CoSeC, Hahn & Meurers 2012)

- Alignment of student and reference answers based on semantic abstractions:



- Lexical Resource Semantics representations derived from syntactic dependencies allow alignment links between semantic units.
- Classification based on a set of numerical scores of alignments with thresholds over which the answer is considered to be correct

Bag Approaches

- Three independent approaches treat each student answer as bag of
  - words (`walks` → `walks`),
  - lemmas (`walks` → `walk`),
  - Soundex hashes (`their`, `there` → `T600`).
- All student answers are used to create each of the three bag models.
- Performance is best on known topics, when vocabulary is known.
- Classification based on SVMs with RBF kernel and optimized parameters (Weka SMO)

Evaluation

Accuracy in 3-way and 5-way tasks

System	Beetle uA	Beetle uQ	SciEntsBank uA	SciEntsBank uQ	SciEntsBank uT
3-way					
Best	73.1%	59.6%	72.0%	66.3%	63.7%
Lex. Baseline	59.5%	51.2%	55.6%	54.0%	57.7%
CoMeT	73.1%	51.8%	71.3%	54.6%	57.9%
5-way					
Best	71.5%	62.1%	64.3%	53.2%	51.2%
Lex. Baseline	51.9%	48.0%	43.7%	41.3%	41.5%
CoMeT	68.8%	48.8%	60.0%	43.7%	42.1%

Accuracy in 2-way task including sub-systems

System	Beetle uA	Beetle uQ	SciEntsBank uA	SciEntsBank uQ	SciEntsBank uT
Best	84.5%	74.1%	77.6%	74.5%	71.1%
Maj. Baseline	59.9%	58.0%	56.9%	58.9%	58.0%
Lex. Baseline	79.7%	74.0%	66.1%	67.4%	67.6%
CoMiC	76.1%	70.6%	68.0%	66.3%	68.0%
Bag of Words	83.1%	67.5%	75.9%	57.8%	59.8%
∼ of Lemmas	83.6%	67.2%	76.7%	58.3%	58.8%
∼ of Soundex	84.1%	68.4%	75.9%	57.6%	58.0%
CoSeC	62.2%	63.6%	67.2%	58.9%	62.4%
CoMeT	83.8%	70.2%	77.4%	60.3%	67.6%
CoSeC*	75.4%	70.8%	72.0%	64.9%	70.6%
CoMeT*	84.5%	71.4%	79.3%	65.4%	69.5%

Systems marked with * refer to a post-submission version where a critical bug in CoSeC was fixed, resulting in improved performance.

- CoMeT is most successful in unseen answers scenarios.
- Per-system performance confirms our hypotheses of which system is best in which setting:
  - Bag systems are best at unseen answers test sets in which vocabulary is known.
  - CoMiC and CoSeC best in unseen domains and unseen questions, since they abstract away from the surface representations.

Outlook

- Develop automatic approach to focus identification in order to pinpoint the essential parts of the student answers.
- When reading text is available, try to identify relevant parts of it and use this additional information in meaning comparison.
- Extend *n-to-m* mappings, improving the alignment performance for multi-word units such as phrasal verb constructions.