Seminar Textual Entailment

Introduction 2

Manfred Pinkal
Saarland University
Summer 2009

Light-Weight Entailment Checking

Ch. Monz, M. de Rijke (2001): Light-Weight Entailment Checking for Computational Semantics. ICoS-3, Workshop Proceedings.

- Informativity, logical definition:
  - NEW informative with respect to OLD iff
    OLD & KB → NEW does not hold
- Application to multi-document summarization

Criteria for Entailment Checking

- *Robustness and coverage of methods for generating representations*
- *Computational costs*
- *Wanted type of outcomes or output*

Criteria for Entailment Checking

- *Robustness and coverage of methods for generating representations*
  - *Deep gramatical processing, full disambiguation, FOL representations*
  - *Shallow parsing, partial disambiguation*
  - *Bag of (stemmed) words*
- *Computational costs*
- *Wanted type of outcomes or output*

Criteria for Entailment Checking

- *Robustness and coverage of methods for generating representations*
- *Computational costs*
  - *FOL entailment problem is undecidable*
- *Wanted type of outcomes or output*

Criteria for Entailment Checking

- *Robustness and coverage of methods for generating representations*
- *Computational costs*
- *Wanted type of outcomes or output*
  - *Strict binary assessment vs. approximate answers*
  - *Approximate entailment values may be both sufficient and appropriate*

Modelling entailment

$$idf_i = \log\left(\frac{N}{n_i}\right)$$

$$entscore(s_{i,d}, s_{j,d'}) = \frac{\sum_{t_k \in (s_{i,d} \cap s_{j,d'})} idf_k}{\sum_{t_k \in s_{j,d'}} idf_k}$$

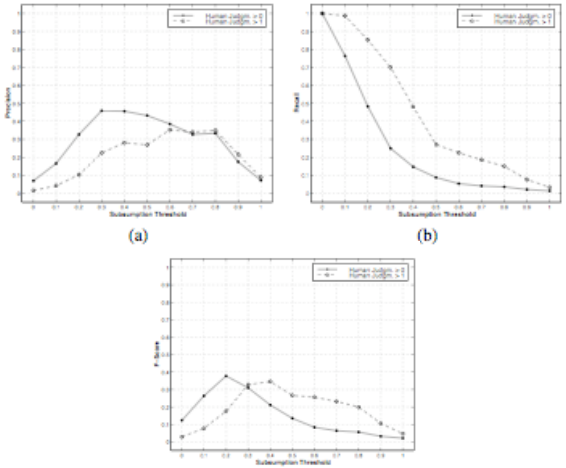
Set Entailment Threshold

„Light-weight entailment is:
reflexive
not transitive
not symetric

Experiment

- *Data-preparation:*
  - *Group newswire stories to topics*
  - *Segment into paragraphs*
  - *Tokenize, lematize*
  - *Compute topic-specific idf scores*
- *Annotation*
  - *Score 2: Full entailment*
  - *Score 1: Entailment of substantial sub-segment*
  - *Score 0: No (essential) entailed information*

Results



Discussion

Advantages

- *Wide-coverage method*
- *Empirical Evaluation*

Problems

- *Entailment vs. Topic Overlap?*
- *Adding lexical semantic knowledge:*
  - *WordNet / Predicate-argument overlap*

The RTE Challenge

A novel framework to model linguistic inference, offering:

- An intuitive, pre-theoretic concept of entailment and inference
- Approximate, wide-coverage methods for checking entailment and inference
- An evaluation method and shared tasks for recognizing textual entailment with objectively measurable results

Textual Entailment: Concept

- An intuitive, pre-theoretic concept of entailment and inference:

  „We say that *T entails H* if the meaning of *H* can be inferred from the meaning of *T*, as would typically be interpreted by people. This somewhat informal definition is based on (and assumes) *common human understanding of language* as well as *common background knowledge."* (Dagan et al. 2006)

RTE: The Shared Task

Shared task – RTE challenge:

- Central task: Determine whether a pair of a text and a hypothesis stands in (textual) entailment relation.
- Training and test material taken from
  - Information Retrieval
  - Information Extraction
  - Summarisation
  - Question Answering
- Typically, hypothesis is manually constructed. Typically, "Text" is one (possibly long) sentence.
- Development set and evaluation set with 800 TH pairs each, balanced w.r.to entailed/ not entailed.
- Development set annotated with Yes/No
- Task: Achieve maximal accuracy on evaluation set

Annotation Guidelines

- *Entailment is directional (no need of symmetric mening inclusion)*
- *H must be fully entailed by T*
- *Very probable instances annotated with YES*
- *Common knowledge is presupposed*
  - *(company has CEO, CEO is employee, employee is person)*

Annotation Process

- *Double , from RTE 2 on triple annotation*
- *Disagreement cases are filtered out (ca. 20%)*
- *More problematic cases removed in post-editing step (10-15%)*
- *Very high agreement on the remaining data set (90-95% in external evaluations)*

Evaluation Measures

- *Accuracy*
- *Average Precision on Confidence Ranking*

$$\frac{1}{R}\sum_{i=1}^{n}\frac{E(i)\times \#CorrectUpToPair(i)}{i}$$

RTE, Overview

RTE 1, 2005	*17*	*50-60%*
RTE 2, 2006	*23*	*53-75%*
RTE 3, 2007	*26*	*49-80%*

Topics for Seminar Papers

Overview:

- *Inference and Semantic Similarity*
- *Using Lexical Semantic Resources*
- *Adapting Logical Approaches*

Inference and Semantic Similarity

- Automatic Acquisition of Paraphrases: DIRT and TEASE (Lin&Pantel 2001, Szpektor et al. 2004)

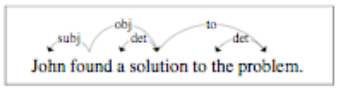


Table 2. Sample slot fillers for two paths extracted from a newspaper corpus.

"X finds a solution to Y"		*"X solves Y"*	
SLOTX	SLOTY	SLOTX	SLOTY
commission	strike	committee	problem
committee	civil war	clout	crisis
committee	crisis	government	problem
government	crisis	he	mystery
government	problem	she	problem
he	problem	petition	woe
legislator	budget deficit	researcher	mystery
sheriff	dispute	sheriff	murder

Inference and Semantic Similarity

- Context and Word-Sense specific Inference
  Pantel et al. 2007
  Mitchell&Lapata 2008
  Erk&Pado 2008
  Dinu&Wang 2009, Wang&Neumann 2007
- Inference as Directional Similarity
  Geffet&Dagan 2005
  Weeds&Weir 2003
  Bhagat et al. 2007

Using lexical-semantic resources

- Lexical knowledge for inference (Moldovan et al. 2003, Tatu et al. 2006, Clark et al. 2008)
- Frame-semantic information for textual inference (Burchardt et al. 2005, Burchardt&Frank 2006, Burchardt et al. 2008)

Adapting Logical Approaches

- Computing and exploiting polarity information (Nairn et al. 2006)
- Natural Logic (MacCartney&Manning 2007, MacCartney&Manning 2009)
- Abductive Reasoning (Hobbs et al. 1988, Hobbs et al. 1993, Raina et al. 2005)
- DRT-Based Reasoning (Bos 2001, Bos&Markert 2006)