Textual Entailment
Part 4: Applications

Sebastian Pado
Institut für Computerlinguistik
Universität Heidelberg, Germany

Rui Wang
Language Technology
DFKI, Saarbrücken, Germany

Tutorial at AAAI 2013, Bellevue, WA
Thanks to Ido Dagan for permission to use slide material

Content of Part 4

- Overview: Four paradigms for using Textual Entailment in Natural Language Processing Applications
- Use Cases for two of the paradigms:
  - Use Case 1: Machine Translation Evaluation
  - Use Case 2: Entailment Graphs for Text Exploration

Overview

Applications of Textual Entailment

- Assumption (cf. Part 1): TE can cover a substantial part of the semantic processing in NLP applications
  – Mapping of semantic (sub)tasks onto textual entailment queries
- If large datasets are involved, **division of labor**:
  1. Shallow (e.g. word based) methods generate candidates
  2. Textual Entailment methods act as filter/(re)scorer
     - Integrates “deeper” algorithms / knowledge
     - Allow shallow methods to be more liberal

Applications of Textual Entailment

- Mapping of semantic (sub)tasks onto textual entailment queries
  - Part 1: What are the Text and the Hypothesis?
  - Part 2: How is the output of the TE system used?

– Main paradigms:
  - Entailment for Validation
  - Entailment for Scoring
  - Entailment for Generation
  - Entailment for Structuring

Entailment for Validation

- Example: Question Answering [Hickl et al. 2007]
  - Step 1: Identify promising answer candidates
    - Shallow methods
  - Step 2: Turn question into statement
    - Replace question word
      (who → someone, which book → a book)
  - Step 3: **Use Textual Entailment to verify that the answer candidate entails the question-as-statement**
    - Binary decision

EXCITEMENT

Example: Question Answering

Question: Who discovered Australia?
Text snippet (T): The first European to reach Australia was Willem Jansszon.
Question-as-statement (H): Someone discovered Australia.

Entailment query: The first European to reach Australia was Willem Jansszon. **⇒?** Someone discovered Australia

- Other application: Relation Extraction [Roth et al. 2009]

EXCITEMENT

Entailment for Scoring

- Example: Machine Translation Evaluation [Pado et al. 2009]
  - Step 1: Create System translation with MT system
  - Hypothesis: Good system translation is *semantically equivalent* to reference translation
  - Step 2: **Use TE to verify that the reference translation entails the system translation – and vice versa!**
    - Graded decision: Degree of semantic equivalence
      - Typically easy to obtain from Textual Entailment systems
    - Details: see **Use Case 1**

Example: MT Evaluation

MT System Translation (ST): Today I will consider this reality.
MT Reference Translation (RT) : I shall face that fact today.

Entailment query 1: ST ⇒? RT

Entailment query 2: RT ⇒? ST

- Other application: Student Answer Assessment [Nielsen et al. 2009]

Entailment for Generation

- Example: Machine Translation "Smoothing" [Mirkin et al. 2009]
  - Source language terms missing from the phrase table cannot be translated
  - Parallel corpora much smaller than monolingual corpora
- **Use entailment to generate entailed "replacements" for unknown source language terms**
  - Sentence may lose some information but is translatable
    - Prefer terms that retain maximal information
  - Requires entailment system that can generate H for given T

Example: Term Replacement in MT

unseen

T: Bulgaria, with its low-cost ski **chalets**, …
H: Bulgaria, with its low-cost ski **houses**, …

Bulgarien, mit seinen günstigen Skihütten, …

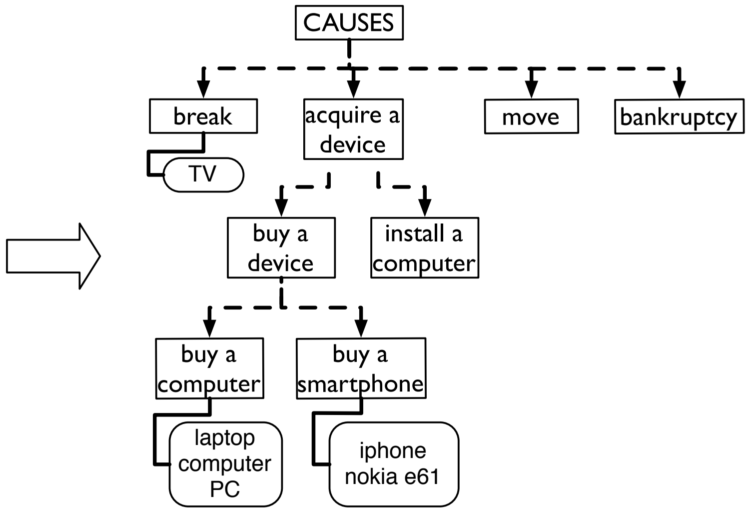
Entailment for Structuring

- Example: Information Presentation [Berant et al. 2012, **Use case 2**]
  - Starting point: Large amount of unstructured data about some concept
  - Goal: Make information easily human-accessible: Build hierarchical structure
- Step 1: Acquire atomic propositions
- Step 2: **Apply entailment queries to each pair of propositions**

- Other applications: Multi-document summarization [Harabagiu et al. 2007]

Example: Information Presentation

own a computer
we got another PC
got a laptop
bankruptcy
i couldn't download
just installed the computer
haven't used the service
TV broke
i have a nokia e 61 now
i don't use it
i'm gonna be moving
decided to buy an iphone



Use Case 1: Machine Translation Evaluation (Padó et al. 2009)

(Entailment for Scoring)

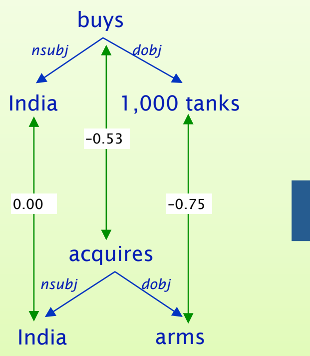
Automatic Evaluation

- Important role in Machine Translation
  - Objective *large-scale* assessment of system quality
  - Minimum Error Rate Training [Och 2002]
- Most widely used metric: BLEU
  - Pure n-gram matching
  - Problems recognizing very different translations [Callison-Burch et al. 2006, etc.]
- METEOR, TER, etc. attempt to make matching more intelligent
  - Still surface-oriented
  - Metrics should evaluate for **semantic equivalence**: TE

The Stanford Textual Entailment System

T: India buys 1,000 tanks.
H: India acquires arms.

1. Graph Alignment



2. Features

Feature	f_i	w_i
Alignment Score	-1.28	1
Alignment: good	+	0.30
Structure match	+	0.10

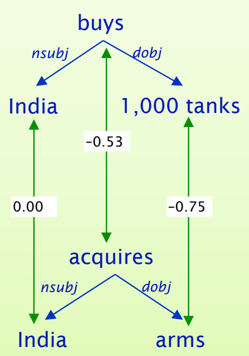
$$\text{score} = \sum_i w_i \cdot f_i = -0.88$$

3. Classification

Use for MT Evaluation

T: India buys 1,000 tanks.
H: India acquires arms.

1. Graph Alignment



2. Features

Feature	f_i	w_i
Alignment Score	-1.28	1
Alignment: good	+	0.30
Structure match	+	0.10

$$\text{score} = \sum_i w_i \cdot f_i = -0.88$$

Linear regression score =
“Degree of entailment”

Technical points

- 1. How to combine two entailment directions?
  - Option 1: Compute directions separately: Not good
  - Option 2: Combine features of both directions into one “bidirectional” regression model: Better
    - Deletion vs. addition features
- 2. How to learn feature weights?
  - Supervised learning from translation quality annotations
    - NIST OpenMT corpora: Newswire (Arabic, Chinese)
    - SMT workshop corpora: EUROPARL transcriptions (F, ES, D)

Evaluation

- Correlation with human sentence-level judgments
  - 10-fold cross validation
- Baselines:
  - BLEU
  - "TradMetrics" regression model: BLEU, TER, METEOR, NIST

Corpora	BLEU	TRADMETRICS (regression)	RTE (regression)	TRADMETRICS + RTE (regression)
NIST	60.0	65.6	63.1	**68.3**
SMT	35.9	39.6	42.3	**45.7**

RTE features and "traditional" metrics are complementary!

We're getting something right

Ref:	U.S. Treasury Offers $14 billion of 30-Year Treasury Bonds	
Sys:	American treasury posing 14 billion from bonds with maturity 30 years	
Human: 6	RTE: 5.77	BLEU: 3.4

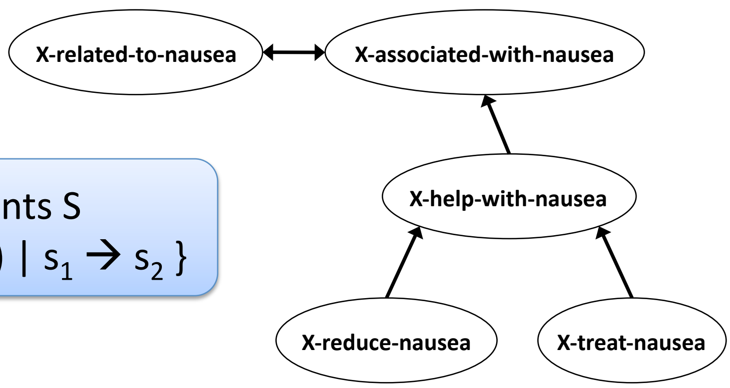
Ref:	What does BBC's Haroon Rasheed say after a visit to Lal Masjid Jamia Hafsa complex? There are no un- derground tunnels in Lal Masjid or Jamia Hafsa.	
Sys:	BBC Haroon Rasheed Lal Masjid, Jamia Hafsa after his visit to Auob Medical Complex says Lal Masjid and seminary in under a land mine	
Human: 1	RTE: 1.2	METEOR: 4.5

Use Case 2: Entailment Graphs [Berant et al. 2012]

(Entailment for Structuring)

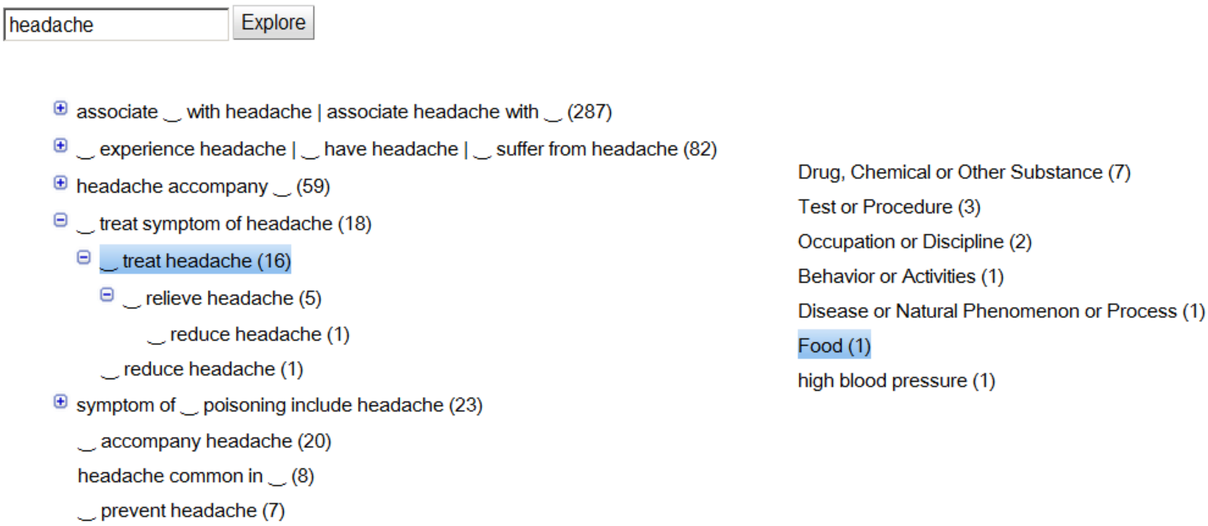
Evaluation: Information Presentation

- Guide users through facts about unfamiliar concept
  – Statements about the target concept collected "Open IE style" [Etzioni et al. 2011]
- Traditional answer: keyword-based presentation
- Proposal: Organize knowledge as **entailment graph**



<u>Input</u>: Set of statements S
<u>Goal</u>: Find E = $\{ (s_1,s_2) \mid s_1 \rightarrow s_2 \}$

BIU Healthcare Explorer [Adler et al. 2012]

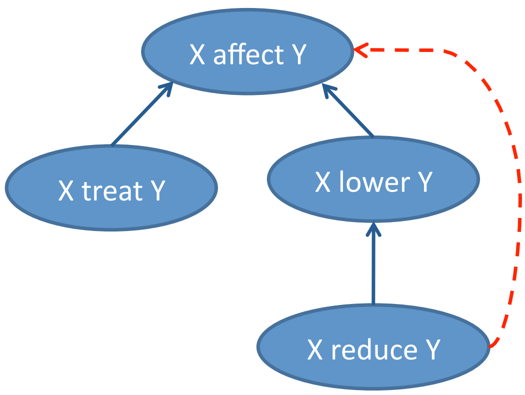


http://irsrv2.cs.biu.ac.il:8080/exploration/

Building Graphs

- Naïve graph construction: Decide entailment for each pair of statements
- Problem: “Local” decisions are not guaranteed to conform to properties of the entailment relation: **transitivity**

X affect Y ⇒ X treat Y	✔
X treat Y ⇒ X affect Y	✘
. . .	
X lower Y ⇒ X affect Y	✔
X reduce Y ⇒ X lower Y	✔
X reduce Y ⇒ X affect Y	✘



Learning Entailment Graphs

- Input: Corpus C
- Output: Entailment graph G = (P,E)
  1. Extract statements S from C
  2. Use a local entailment classifier to estimate $P_{ij} = P(s_i \rightarrow s_j)$ for each pair (s_i, s_j)
     - Techniques from Part 2
  3. **Find the most probable transitive graph**
     - **Part 1: Define objective function for graph**
     - **Part 2: Identify best graph**

Graph Objective Function

$$\hat{G} = \arg\max \sum_{i \neq j} w_{ij} \cdot x_{ij}$$

$x_{ij} = \begin{cases} 1 & i \rightarrow j \\ 0 & \text{else} \end{cases}$

$$w_{ij} = \log \frac{p_{ij} \cdot \theta}{(1 - p_{ij}) \cdot (1 - \theta)}$$

θ: "density" prior

- Still assumes independence between edges

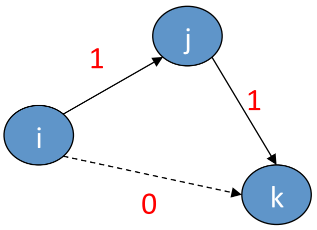
Integer Linear Program

$$\hat{G} = \arg\max \sum_{i \neq j} w_{ij} \cdot \boxed{x_{ij}}$$

$$\forall i, j, k : x_{ij} + x_{jk} - x_{ik} \leq 1$$

$$x_{ij} \in \{0, 1\}$$

1+1-0 = 2 > 1



- NP hard
  - Optimization: Decompose sparse graph
    - Details: [Berant et al. 2012]

Experimental Evaluation

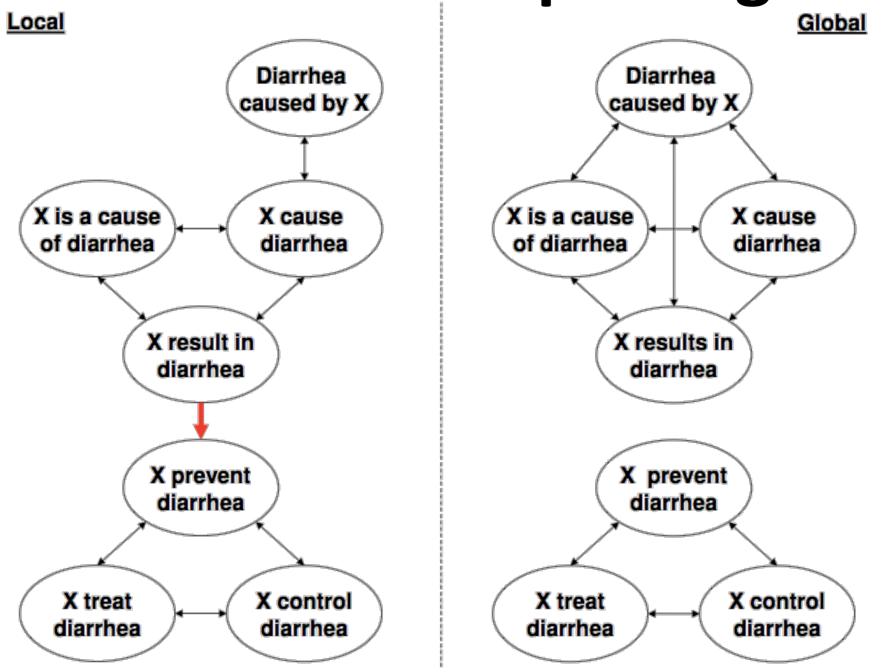
- 50 million word tokens **healthcare** corpus
- Gold standard entailment graphs for 23 medical concepts
  - Smoking, seizure, headache, lungs, diarrhea, chemotherapy, HPV, Salmonella, Asthma, etc.
- Evaluation: F_1 on learned edges vs. gold standard
- Baselines:
  - WordNet as source of entailments between predicates
  - “Local” model without enforcing transitivity

Results

	Recall	Precision	F_1
WordNet	10.8	44.1	13.2
Local	**53.5**	38.0	39.8
Global (ILP)	46.0	**50.1**	**43.8**

- Global algorithm avoids false positives
  - High precision

Illustration – Graph Fragment



Take-home Message

- Many applications can be mapped (partially) onto Textual Entailment
  - Four paradigms: verify, score, generate, structure
  - Large datasets: Division of labor between shallow methods (generators) and Textual Entailment (filter)
- Two Use Cases:
  - MT Evaluation: TE to measure semantic equivalence
  - Entailment Graphs: Global learning for information presentation

Reference List

- Berant, J., Dagan, I., and Goldberger, J. (2012). Learning entailment relations by global graph structure optimization. Computational Linguistics, 38(1):73–111.
- Callison-Burch, C., Osborne, M., and Koehn, P. P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In Proceedings of EACL, pages 249–256.
- Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam, M. (2011). Open information extraction: the second generation. Proceedings of IJCAI, pages 3–10.
- Harabagiu, S., Hickl, A., and Lacatusu, F. (2007). Satisfying information needs with multi-document summaries. Information Processing and Management, 43(6):1619–1642.

Reference List

- Hickl, A. and Bensley, J. (2007). A Discourse Commitment-Based Framework for Recognizing Textual Entailment. Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 171–176.
- Mirkin, S., Specia, L., Cancedda, N., Dagan, I., Dymetman, M., and Szpektor, I. (2009). Source-Language Entailment Modeling for Translating Unknown Terms. Proceedings of ACL, pages 791–799.
- Nielsen, R. D., Ward, W., and Martin, J. H. (2009). Recognizing Entailment in Intelligent Tutoring Systems. Natural Language Engineering, 15(4):479–501.
- Och, F. J. (2003). Minimum error rate training in statistical machine translation. In Proceedings of ACL, pages 160–167.

Reference List

- Padó, S., Cer, D., Galley, M., Manning, C. D., and Jurafsky, D. (2009). Measuring Machine Translation Quality as Semantic Equivalence: A Metric based on Entailment Features. Machine Translation, 23(2–3):181–193.
- Roth, D., Sammons, M., and Vydiswaran, V. V. (2009). A Framework for Entailed Relation Recognition. Proceedings of ACL, pages 57-60.