A Search Task Dataset for German Textual Entailment

Britta D. Zeller   Sebastian Padó

Department of Computational Linguistics,
Heidelberg University, Germany

10th International Conference on Computational Semantics
March 20, 2013

EXCITEMENT

Textual Entailment (TE)

- TE is a binary relation between two texts (Text T, Hypothesis H)
- Holds if *a human reading of Text infers that Hypothesis is most likely true* [Dagan et al., 2005]. Decision problem:

  T: Mike loves Anna.

  H$_1$: Mike likes Anna.
  → Text T entails Hypothesis H$_1$

  H$_2$: Mike is Anna's husband.
  → Text T does not entail Hypothesis H$_2$
- Entailment relations are relevant in various NLP tasks:
  → "Validation": Answer Validation in QA [Peñas et al., 2008]
  → "Scoring": MT Evaluation [Padó et al., 2009]
  → "Structuring": Search Result Visualization [Berant et al., 2012]

Motivation and Goal

Main basis of research: RTE datasets

- Created by annual *Recognising Textual Entailment* workshops
- Pairs of Text and Hypothesis with positive or negative entailment
- Clean text, no grammatical errors or sloppy language
- Only available for English

These datasets are mainly used for system development

- Do their patterns apply to other languages?
- Do their patterns apply to noisier data?

Our study: Creation, analysis, and modeling of a German social media Textual Entailment dataset.

A Textual Entailment dataset from social media data

- Use case: Computer problem. Search for suitable threads in self help forums ↔ Find relevant questions

Query (H): Virus on computer

Transfer to our dataset:

- Search as test for entailment: Find first posts that entail the query
- First forum post = Text; User query = Hypothesis
- Ignore answer posts: Not helpful for query matching

A Textual Entailment dataset from social media data

- Use case: Computer problem. Search for suitable threads in self help forums ↔ Find relevant questions

Query (H): Virus on computer

Post 1 (T): Hi, I think I have a virus on my PC

Post 2: <Suggestion>

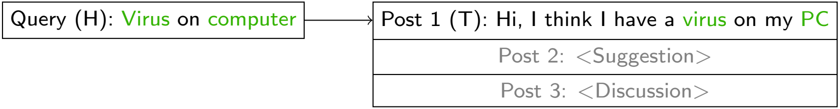
Post 3: <Discussion>

Transfer to our dataset:

- Search as test for entailment: Find first posts that entail the query
- First forum post = Text; User query = Hypothesis
- Ignore answer posts: Not helpful for query matching

A Textual Entailment dataset from social media data

- Use case: Computer problem. Search for suitable threads in self help forums ↔ Find relevant questions



Query (H): Virus on computer

Post 1 (T): Hi, I think I have a virus on my PC

Post 2: <Suggestion>

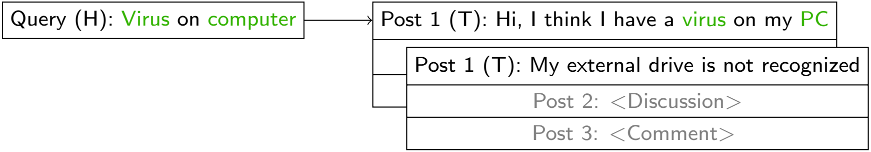
Post 3: <Discussion>

Transfer to our dataset:

- Search as test for entailment: Find first posts that entail the query
- First forum post = Text; User query = Hypothesis
- Ignore answer posts: Not helpful for query matching

A Textual Entailment dataset from social media data

- Use case: Computer problem. Search for suitable threads in self help forums ↔ Find relevant questions



Query (H): Virus on computer

Post 1 (T): Hi, I think I have a virus on my PC

Post 1 (T): My external drive is not recognized

Post 2: <Discussion>

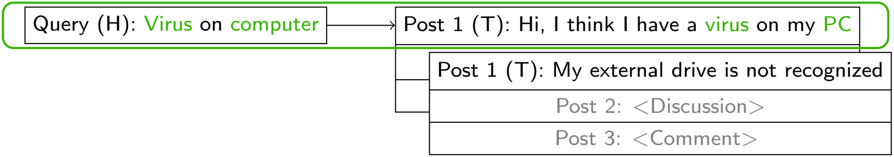
Post 3: <Comment>

Transfer to our dataset:

- Search as test for entailment: Find first posts that entail the query
- First forum post = Text; User query = Hypothesis
- Ignore answer posts: Not helpful for query matching

A Textual Entailment dataset from social media data

- Use case: Computer problem. Search for suitable threads in self help forums ↔ Find relevant questions



Query (H): Virus on computer

Post 1 (T): Hi, I think I have a virus on my PC

Post 1 (T): My external drive is not recognized

Post 2: <Discussion>

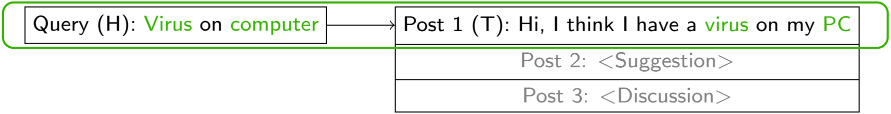
Post 3: <Comment>

Transfer to our dataset:

- Search as test for entailment: Find first posts that entail the query
- First forum post = Text; User query = Hypothesis
- Ignore answer posts: Not helpful for query matching

Link to standard tasks in Recognising Textual Entailment

- Similar to “Search” task introduced in RTE-5
  - Find entailing pairs over *several* texts
  - First post (T) textually entails query (H)



Query (H): Virus on computer

Post 1 (T): Hi, I think I have a virus on my PC

Post 2: <Suggestion>

Post 3: <Discussion>

Sample Text/Hypothesis pair from our dataset

T: Hi, mich macht das EZ Backup und Raid Zeug ganz wirr ;-) Hab Sata 1, 3 und 4 belegt. . . . Der Brenner auf Sata 4 läuft auf Slave, für ein Firmwareupdate sollte er aber auf (Secondary) Master laufen, was macht man da?
Danke, Gruß, Blondy

Hi, I'm confused about EZ Backup and Raidstuff ;-) Have Sata 1, 3 and 4 occupied. . . . The burner at Sata 4 runs as Slave, but it should be switched to (Secondary) master for a firmware update, what to do? Thanks, bye, Blondy

H: Statt auf Slave soll ein Laufwerk jetzt auf Secondary Master eingestellt werden.

A drive should get connected to Secondary Master rather than to Slave.

→ Text T textually entails Hypothesis H

Overview of our work

- Part 1: Creation of a Textual Entailment dataset
- Part 2: Dataset analysis
- Part 3: Modelling the data

Part 1: A method to create Search-task TE datasets

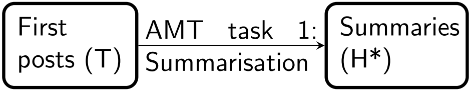
- Problem 1: We have first posts, but no queries
  - Query generation with 3 successive crowdsourcing tasks on Amazon Mechanical Turk (AMT) [Snow et al., 2008]
    Generating queries in a single task is too complex for turkers
    → Creates positive entailment pairs
- Problem 2: We need negative pairs to reflect Search task setting
  - Automatic compilation across pairs
    → Creates negative entailment pairs

⇒ Relatively small manual effort for high quality dataset

Overall procedure of entailment pair generation

First posts (T)

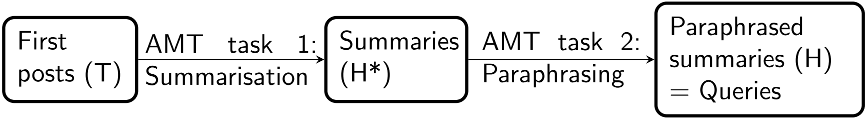
Overall procedure of entailment pair generation


First posts (T)

AMT task 1: Summarisation

Summaries (H*)

Overall procedure of entailment pair generation



First posts (T)

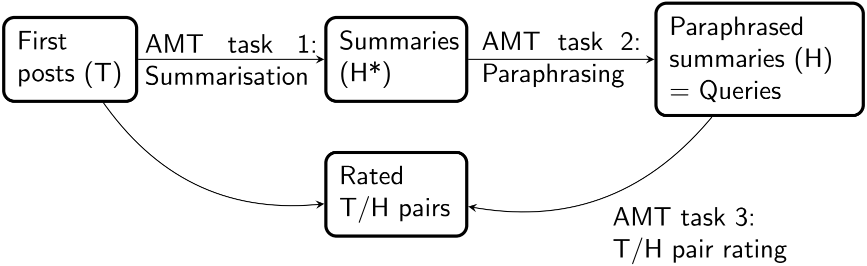
AMT task 1: Summarisation

Summaries (H*)

AMT task 2: Paraphrasing

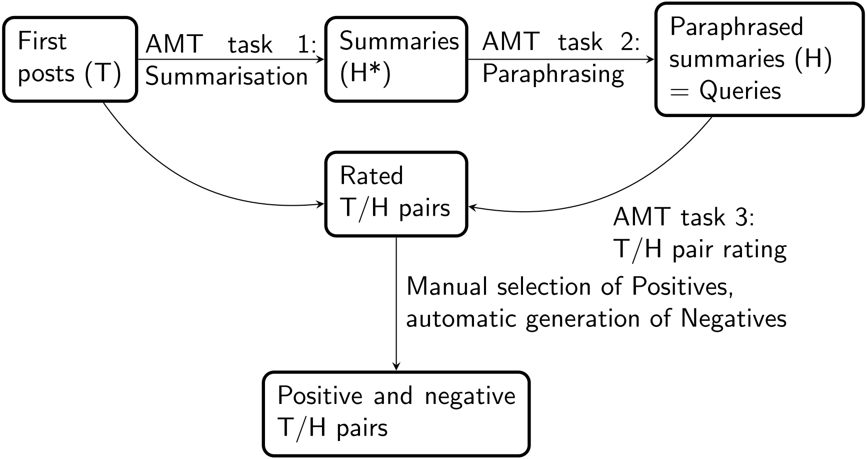
Paraphrased summaries (H) = Queries

Overall procedure of entailment pair generation



First posts (T)
AMT task 1: Summarisation
Summaries (H*)
AMT task 2: Paraphrasing
Paraphrased summaries (H) = Queries
Rated T/H pairs
AMT task 3: T/H pair rating

Overall procedure of entailment pair generation


First posts (T)
AMT task 1: Summarisation
Summaries (H*)
AMT task 2: Paraphrasing
Paraphrased summaries (H) = Queries
Rated T/H pairs
AMT task 3: T/H pair rating
Manual selection of Positives, automatic generation of Negatives
Positive and negative T/H pairs

Entailment pair generation in terms of numbers

- Starting point:
  25 first posts from computer forums
- Multiple annotations per crowdsourcing step:
  226 rated Text/Hypothesis pairs
- After manual selection and automatic generation:
  172 positive entailment pairs, 2,832 negative entailment pairs

Validation of annotated Text/Hypothesis pairs

- Quality assurance: 3 crowdsourcing ratings per entailment pair
- Rating used three categories: *perfect* (*p*), *incomplete but acceptable* (*i*), *no good* (*n*)

Ratings	p-p-p	p-p-i	p-i-i	i-i-i	...	n-i-i	n-n-i	n-n-n
Entailment	Y	Y	Y	Y	...	N	N	N
Frequency	38	45	50	20	...	21	11	7
Validation	**37**	**41**	**42**	**7**	...	**1**	**2**	**7**

- 153 pairs had no *no good* rating
    - Most pairs with at least one *p* rating were acceptable summaries
- 39 pairs had at least one *no good* rating
    - Only *n-n-n* pairs were indeed acceptable negative pairs
- We kept 127 positive and 10 negative pairs

Automatic compilation of negative Text/Hypothesis pairs

- Combination of verified Hypotheses with ‘other’ Texts
- 137 distinct Hypotheses, 22 distinct Texts → 2,877 potential negative pairs
- Problem: narrow domain. Generic Hypotheses can be valid for various Texts:
  H: Computer infected with virus
- Manual check of cross-pairs with similar topics

→ 45 additional positive entailment pairs
→ 2,822 additional negative entailment pairs

⇒ Total: 172 positive pairs, 2,832 negative pairs

Part 2: Qualitative analysis of the crowdsourced data

Summary data:

- Genre-specific diction: Telegram style, ungrammatical sentences, no punctuation, no German capitalisation
- Shorter and more general than original text
- Rambling and vague posts lead to incomplete or incorrect summaries

Paraphrasing data:

- Linguistic properties:
    - Generic: Syntax changes; synonymy, hypernymy, abbreviations
    - Language-specific: Nominalisations, active/passive switches
    - Genre-specific: Same as for summaries
- Semantic errors due to ambiguous summaries
- Lack of domain knowledge by annotators

Analysis of pair rating task

Correlation of ratings and observed properties:

- *perfect*-judged pairs: Comprehensive Hypotheses, simple context
- *incomplete*-judged pairs: Short and general Hypotheses
- *no good*-judged pairs: Propagated errors
- *p-i-n*-judged pairs: Complex Texts (list of problems)

⇒ 3-step crowdsourcing setup leads to:

- High-quality Text/Hypothesis pairs
- High degree of linguistic variation
- Linguistic errors reflect noise in original data

Part 3: Modelling the dataset with TE engines

	P	R	F_1
Word overlap	38%	38%	38%
EDITS	63%	34%	**44%**

- Two language-independent models:
  - EDITS [Negri et al., 2009]: Off-the-shelf Textual Entailment system
  - Baseline: Word overlap
  - Decision task: Does the Text entail the Hypothesis?
- Word overlap as strong indicator for TE; word order also informative
- Seems to be easier than English Search task data of RTE-5
  - EDITS: 33% F_1 [Bentivogli et al., 2009]
  - Our dataset is slightly more balanced, more coherent
  - Influence of language?

Conclusions

- Resource: Freely available German social media Textual Entailment dataset
  - First test bed for TE scenarios dealing with noisy data
  - More non-English and noisy datasets needed to assess their influence
- Methodology: 3-step crowdsourcing procedure applicable to other languages and domains
- Analysis: Analysis of German pairs
  - Motivation for building language-specific knowledge resources

Bentivogli, L., Magnini, B., Dagan, I., Trang Dang, H., and Giampiccolo, D. (2009).
The fifth PASCAL recognising textual entailment challenge.
In Proceedings of TAC, Gaithersburg, MD.

Berant, J., Dagan, I., and Goldberger, J. (2012).
Learning entailment relations by global graph structure optimization.
Computational Linguistics, 38(1).

Dagan, I., Glickman, O., and Magnini, B. (2005).
The PASCAL Recognising Textual Entailment Challenge.
In Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK.

Negri, M., Kouylekov, M., Magnini, B., Mehdad, Y., and Cabrio, E. (2009).
Towards Extensible Textual Entailment Engines: the EDITS Package.
In Proceeding of IAAI, Reggio Emilia, Italy.

Padó, S., Cer, D., Galley, M., Jurafsky, D., and Manning, C. D. (2009).
Measuring machine translation quality as semantic equivalence: A metric based on entailment features.
Machine Translation, 23(2–3):181–193.

Peñas, A., Rodrigo, Á., Sama, V., and Verdejo, F. (2008).
Testing the reasoning for question answering validation.
Journal of Logic and Computation, 18:459–474.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008).
Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks.
In Proceedings of EMNLP, pages 254–263, Honolulu, HI.

Addendum: Paraphrasing examples

- Ambiguous summaries leading to incorrect paraphrases
  - SUM: Error message after Bios update, restart computer anyway?
  - PAR: I get an error message after the BIOS update, should I restart the PC?
- Lack of knowledge by annotator leading to incorrect paraphrases
  - SUM: Connection of an additional SATA device
  - PAR: I want to connect hardware made by SATA