Annotating Logic Inference Pitfalls

Aikaterini-Lida Kalouli
aikaterini-lida.kalouli@uni-konstanz.de
University of Konstanz

Livy Real
livyreal@gmail.com
University of São Paulo

Valeria de Paiva
valeria.depaiva@nuance.com
Nuance Communications

Motivation

- Goal: compute inference relations, i.e. entailment, contradiction and neutral stances, based on the conceptual semantics of sentences
- SICK corpus (Sentences Involving Compositional Knowledge)[1] as a testing baseline for an inference system
- Investigated the data in SICK to see:
  - what lay people consider logical inferences
  - what kinds of inference are included in the corpus (for insights on implementation)

  BUT: we found many wrong annotations which led us to search the causes for the mistakes [2] and possible ways of curating the data.

SICK corpus

- 9840 pairs of English sentences used as benchmark for compositional extensions of Distributional Semantics;
- Non-abstract, every-day sentences with no complex linguistic phenomena;
- Pairs annotated for similarity degree and inference relations: entailment, contradiction, neutrality;
- 1424 pairs of contradictions ($AcBBcA$), 1300 pairs of double entailment ($AeBBeA$), 1513 pairs of single-sided entailment ($AeBBnA$) and 4992 pairs of neutrals ($AnBBnA$).

Noisy Data and How to Prevent it

1. **Problem:** Ungrammatical and non-sensical sentences, created during the corpus construction, via the normalization and expansion processes. There were mistakes during this process, despite manual checking. These mistakes are "mentally" fixed by each annotator differently.

   Example: *The black and white dog isn't running and there is no person standing behind.*: a) did the annotators assume that there should be an *it* following *behind* or did they just ignore the word *behind* altogether? and b) what normalization step caused the lack of a pronoun?

   Solution 1: Allow annotators to make personal comments on their annotations. It is easier to understand the logic of the annotation and to see what "fix" each of them chose for items that are somehow "unacceptable". A similar proposal is discussed in [3].

   Solution 2: Careful documentation of the construction process would allow us to track back which specific data comes from which specific pre-procesing step. We could then redo the step which produced the mistake and correct it at its source.

2. **Problem:** Wrong annotations with no apparent explanation. Data annotation is prone to human errors: tiredness, lack of attention or simply "don't know what to do with that item" factor.

   Example: *A= The bride and the groom are arriving after the wedding. B= The bride and the groom are leaving after the wedding.* ($AcBBnA$) Why does A contradict B but B does not contradict A? There is lack of logic here and the need to facilitate the task.

   Solution 1: A visualization tool for annotation data would facilitate the task of the annotators.

   Solution 2: Make sure the data is "cleaned" so that no personal "fixes" are needed.

   Solution 3: Use data provenance techniques, e.g. version control, that a) allow annotators to make notes about items they want to come back to, b) track down which annotation approach was used for each annotation round and c) visualize how the labels of the data change depending on the annotation approach.

3. **Problem:** No common reference to judge the pairs and presence of indefinite determiners ([4, 5]). We discussed this inference-specific issue in [6].

   Example: *A= A soccer player is kicking a ball out of the goal. B= A soccer player is kicking a ball into the goal.* ($AcBBnA$) The players are not necessarily the same.

Curating Noisy Data

- Use semi-automatic approaches that make smart use of human effort and of task-specific methods for tracing mistakes. We reduced ([7]) the manual effort of correcting SICK by defining an approach where all pairs whose sentences differ by only one word are automatically processed and checked by Princeton Wordnet ([8]);
- Develop more tools that are able to identify mistakes, like the one in [9];
- Use 'robust' statistical methods which are able to deal with this kind of noisy data, as suggested in [9].

Conclusions

- All corpora suffer from noisy annotations: the more difficult the phenomena involved, the harder and more error-prone the annotation process;
- Data curation efforts are essential to establish trustworthy baselines;
- Cleaning up data ensures that corrected mistakes can be used as guidelines for future corpora.

References

[1] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*, 2014.

[2] Aikaterini-Lida Kalouli, Livy Real, and Valeria de Paiva. Textual inference: getting logic from humans. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, 2017.

[3] Assaf Toledo, Sophia Katrenko, Stavroula Alexandropoulou, Heidi Klockmann, Asher Stern, Ido Dagan, and Yoad Winter. *Semantic Annotation for Textual Entailment Recognition*, pages 12–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[4] Annie Zaenen, Lauri Karttunen, and Richard Crouch. Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 31–36. Association for Computational Linguistics, 2005.

[5] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of ACL-08*, 2008.

[6] Aikaterini-Lida Kalouli, Livy Real, and Valeria de Paiva. Correcting contradictions. In *Proceedings of Computing Natural Language Inference (CONLI) Workshop, 19 September 2017*, 2017.

[7] Aikaterini-Lida Kalouli, Livy Real, and Valeria de Paiva. Wordnet for "easy" textual inferences. 2018, submitted.

[8] Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[9] Julie Tibshirani and Christopher D. Manning. Robust logistic regression using shift. In *Association for Computational Linguistics (ACL)*, 2014.