An overview of Natural Language Inference Data Collection: The way forward?

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, Staffan Larsson
Centre for Linguistic Theory and Studies in Probability (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
stergios.chatzikyriakidis@gu.se, cooper@ling.gu.se,
simon.dobnik@gu.se, sl@ling.gu.se

1 Introduction

Understanding a Natural Language (NL) sentence amounts to understanding its consequences as well as the sentences that it is a consequence of. As Cooper et al. (1996) aptly put it 'inferential ability is not only a central manifestation of semantic competence but is in fact centrally constitutive of it' . Given the latter claim, it is no surprise that the same authors argue that the best way to test the semantic adequacy of NLP systems is to look at how well they perform with regard to Natural Language Inference (NLI). It is very hard not to agree with this statement. Indeed, NLI is considered by many researchers to be the crux of computational semantics. This paper is about the datasets created for this need. In particular, we discuss the most common NLI resources arguing that all these a) fail to capture the wealth of inferential mechanisms present in NLI and b) seem to be driven by the dominant discourse in the field at the time of their creation. In light of these observations, we want to discuss the requirements that an adequate NLI platform must satisfy both in terms of the range of inference patterns found in reasoning with NL as well as the range of the data collection mechanisms that are needed in order to acquire this range of inferential patterns.

2 NLI datasets

In this section we present the merits and drawbacks of three NLI datasets which represent some of the main strategies that have been used in the data collection of inference.

2.1 The FraCaS test suite

The FraCaS test suite was built in the mid 1990's by the FraCaS project, an EU project aimed at developing a general framework for computational semantics (Cooper et al., 1996).[1] It was later recast in machine-readable form by Bill McCartney of Stanford University.[2] The data set consists of 346 problems each containing one or more statements and one yes/no-question (except for four problems, where there is no question). The total number of sentences in the data set is 1220, but since some of them are repeated in several problems, there are in total 874 unique sentences.

The FraCaS test suite has been later extended into the MultiFraCaS, which includes a translation of the test suite into German, Farsi, Mandarin and Greek.[3] There is also the Japanese FraCaS extension (JSem), which expands the original FraCaS after translation in a number of ways: it contains basic patterns of semantic inferences such as active-passive alternation and adverbial phrases not included in

[1] ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz
[2] www-nlp.stanford.edu/~wcmac/downloads/fracas.xml
[3] www.ling.gu.se/~cooper/multifracas/.

the original FraCaS and further extends the original set with constructions specific to Japanese syntax and semantics that have been discussed in the literature. Furthermore, the types of inferences are annotated.[4]

Merits: even though it contains only 346 examples, it covers a lot of inference cases and it now has the advantage of some multilinguality

Drawbacks: the examples are mostly logical inferences, the data are constructed and the dataset is very small especially with today's standards.

2.2 Recognizing Textual Entailment

The Recognizing Textual Entailment (RTE) challenges were first developed in 2004 as a means to test textual entailment, i.e. relations between a premise text and a hypothesis text. RTE uses naturally occurring data as premises and then constructs the hypothesis text based on the premise natural text. The first RTE challenge makes a binary classification of entailment, i.e. a hypothesis is either entailed or non-entailed. This is changed to a tripartite classification (hypothesis entailed, its negation entailed or no entailment) in RTE4 and RTE5 (Giampiccolo et al., 2008; Bentivogli et al., 2009).

Merits The RTE challenges remedy the unnaturalness of constructed examples by using examples from natural text and furthermore make a first step of including entailments that require presupposed information, including non-logical presuppositional inferences.

Drawbacks Even though the RTE platforms have been notoriously difficult for NLI systems (especially the three way entailment tasks), most of the examples do not involve deep semantic inference but are rather complicated in terms of their syntax. Furthermore, the definition of inference assumed in a number of the examples is problematic. As Zaenen et al. (2005) have pointed out, RTE platforms suffer from cases of inference that should not be categorised as such. For these cases, a vast amount of world knowledge needs to be taken into consideration (that most importantly not every linguistic agent has). The problem is that there is no clear annotation that will distinguish different kinds of inference. Lastly, and similarly to the FraCaS, the RTE datasets are still very small (less than 1000 pairs for both the development and the test set for all challenges) compared with approaches that rely on large datasets such as Deep Learning approaches (DL).

2.3 Stanford Natural Language Inference

A recent data collection for NLI was developed in Stanford by Bowman et al. (2015) using crowdsourcing (Mechanical Turk). The subjects are given a caption of a picture and are then asked to provide: a) an alternate true caption b) an alternate possibly true caption and c) an alternate false caption. The dataset constructed out of this process contains 570k inference pairs, making SNLI two orders of magnitude bigger than the previous datasets.

Merits The size of the corpus makes it suitable for training DL approaches. The SNLI platform is an extremely useful resource, and given the current state of affairs in computational linguistics, it is the only one that is usable for approaches using DL, which are predominant at the moment. Furthermore, reasoning in SNLI is tied to specific situations.

[4]We thank an anonymous reviewer for this information as regards the Japanese FraCaS. More info can be found here: `http://researchmap.jp/community-inf/JSeM/?lang=english`. Also for a discussion on JSem please consult Kawazoe et al. (2015).

Drawbacks Situational reasoning can be also claimed to be a drawback of SNLI. What would be for example the image described by a caption "all men are human"? It appears that quantification requires evaluation of collections of image/description pairs. Furthermore, and similarly to earlier platforms, SNLI seems to capture only a fraction of the range of phenomena associated with NLI.

2.4 Some other NLI platforms

There exist a number of lesser known, theory specific or, at a first sight not NLI related datasets used for NLI. We mention briefly three here:

- the SICK dataset (Sentences Involving Compositional Knowledge) (Marelli et al., 2014) is a dataset created to test compositional distributional semantics (DS) models. Although the dataset contains examples that one would count as belonging to logical inference (negation, conjunction, disjunction, apposition, relative clauses, etc.), its focus are distributional semantic approaches and therefore it normalises several cases DS is not expected to account for. For example, named entities are turned into noun phrases standing for the class, and complex VPs and sentences involving modals and auxiliaries are simplified.[5]

- The PPDB (Paraphrase Database) relation extraction dataset (Ganitkevitch et al., 2015) is primarily a dataset on paraphrase. However, it is further annotated for entailment (unidirectional, bidirectional etc.), making it extremely useful for entailment tasks as well.

- The VQA (Visual Question Answering) corpus[6] contains open-ended questions (and answers) about images. While SNLI models inference from linguistic and perceptual data together, one could see the VQA task as inference from perceptual input alone, but guided by linguistic input (the question). The output in both cases are NL sentences[7]. We believe that inference from perceptual input should be included when constructing a state-of-the-art NLI dataset.

3 The way forward

3.1 What do we need?

We need to include more kinds of inference than were originally in the FraCaS test suite. In addition to classical logical inference we further need presuppositional inference and non-logical inference as given by implicatures and enthymemes[8]. Presuppositional inference is to some extent included implicitly in the FraCaS test suite but it is treated as logical inference implicitly.[9]

We also need to include examples of probabilistic reasoning relating to conditional probability though we need to be clear with this whether we are referring to probabilistic judgements made by individuals

[5]Given the fact that the SICK dataset involves a lot of cases one would classify as logical inference, it is not surprising to find systems based on logic being evaluated on this dataset, Abzianidze (2015); Martınez-Gómez et al. (2017).

[6]`http://www.visualqa.org/`

[7]For VQA, given an y/n-question P?, answer "yes" means inference is P and "no" means inference is $\neg P$, and given a wh-question $?x.p(x)$, the answer a means the inference is $p(a)$.

[8]An enthymeme is a logic-like deductive inference with one or several premises supplied by context (Breitholtz and Villing, 2008). For example, in a discussion about which road to take, the premise "Walnut Street is shorter" may lead to the conclusion "We should take Walnut Street", assuming a hidden premise "If X is the shortest route, we should take X".

[9]For example, section 9 in the FraCaS test suite involves presuppositional rather than logical inference:

P1 Smith knew that ITEL had won the contract in 1992.

Q Did ITEL win the contract in 1992?

H ITEL won the contract in 1992. [FraCaS 334]

A more fine-grained platform for NLI would have recognised this fact and would also include a case where complement presupposition survives under negation by changing P1 into *Smith did not know that ITEL had won the contract in 1992.*

or probabilities that individual agents will make the corresponding categorical judgement. It would be an advantage to have data of both kinds.[10]

It is important that the data set be grounded in real data so that it does not just represent armchair intuitions of linguistic experts as the FraCaS test suite does. It also seems important to us that the real data should include reasoning in interactive dialogue settings and not just text (as, for example, in the textual entailment enterprise). Furthermore, there has been a significant interest in connecting language and vision recently and exploring inference in this context appears to be a promising and challenging research field. For example, Marconi (1997) distinguishes between inferential meanings of words, which enables inferences from uses of the word, and referential meaning, allowing speakers to identify the objects and situations referred to by the word.[11] A key property to explore here is how different modalities provide information for an acceptable inference. For example "A group of children enjoy their time on the beach" and "A group of adults are swimming at the beach"[12] can be considered a contradiction, provided that there exists sufficient visual information supporting this.

3.2 How to get it?

The emphasis on real data does not necessarily mean that we should abandon constructed examples as one form of input, but it does mean that constructed examples should be grounded in real data either by (i) finding corresponding examples in corpora consisting of text from different genres or the web; and (ii) by conducting experiments in which subjects are asked to evaluate or construct inferences using both textual and visual context. In the latter case we think it is useful (iii) to use crowd-sourcing and (iv) to adapt the game techniques used in GWAPs (Games with a Purpose), for example in a way they were used for construction of lexico-semantic networks (Lafourcade, 2011)[13].

3.3 How we know we got it?

In order to address the issue of verifying of NLI datasets constructed by expert semanticists in the intuitions of speakers of a language we conducted a pilot study based on the inference examples from the FraCaS suite (Cooper et al., 1996). The purpose of this evaluation is two-fold: (i) to verify whether intuitions of inference of ordinary speakers correspond to the intuitions of the authors of the dataset; and (ii) to evaluate a degree of variation of inference intuitions for individual examples which would allow us to examine how humans reason and therefore further study examples of inference that are normally considered as problematic. We selected 15 "interesting" examples (those that in our experience are either very clear in terms of the inference label and those that are problematic) covering quantification (2, 6, 12, 17, 38, 74), adjectives (199, 201, 206, 211, 212, 223), and other semantic phenomena including plurals (103), sluicing (163), and temporal reference (308). The examples were chosen so that there were 5 examples of each inference class *yes*, *no*, and *don't know+undefined*, using the labels from (MacCartney, 2007). The task was framed as a controlled crowd-sourcing data collection experiment using the Semant-o-matic tool[14], previously used to collect data in related tasks such as (Dobnik et al., 2014; Dobnik and Åstbom, 2017). In contrast to other crowd-sourcing solutions, a specialised tool allows us to target particular groups of individuals through social media and therefore combines the properties of controlled data-collection experiments with online crowd-sourcing. Examples of inference were presented

[10] It is worth repeating here that annotation for different types of inferences (not including probabilistic inference) has been done for the Japanese extension of the FraCaS, JSem, Kawazoe et al. (2015). In connection to probabilistic inference, it would be interesting to have a look at STS (Semantic Textual Similarity) datasets. These present a similarity score (ranging from 1 to 5), which can potentially be seen as a probabilistic judgement. Such scores are also present in the SICK dataset. It might be worth looking at the way these similarity scores reflect in some way the type of probabilistic inference we are discussing here.

[11] While SNLI models inference from linguistic and hypothetical perceptual scenarios together, one could see the VQA task as inference from perceptual input alone, but guided by linguistic input (the question).

[12] Example from SNLI provided by a reviewer.

[13] `www.jeuxdemots.org`

[14] `http://www.dobnik.net/simon/semant-o-matic/`

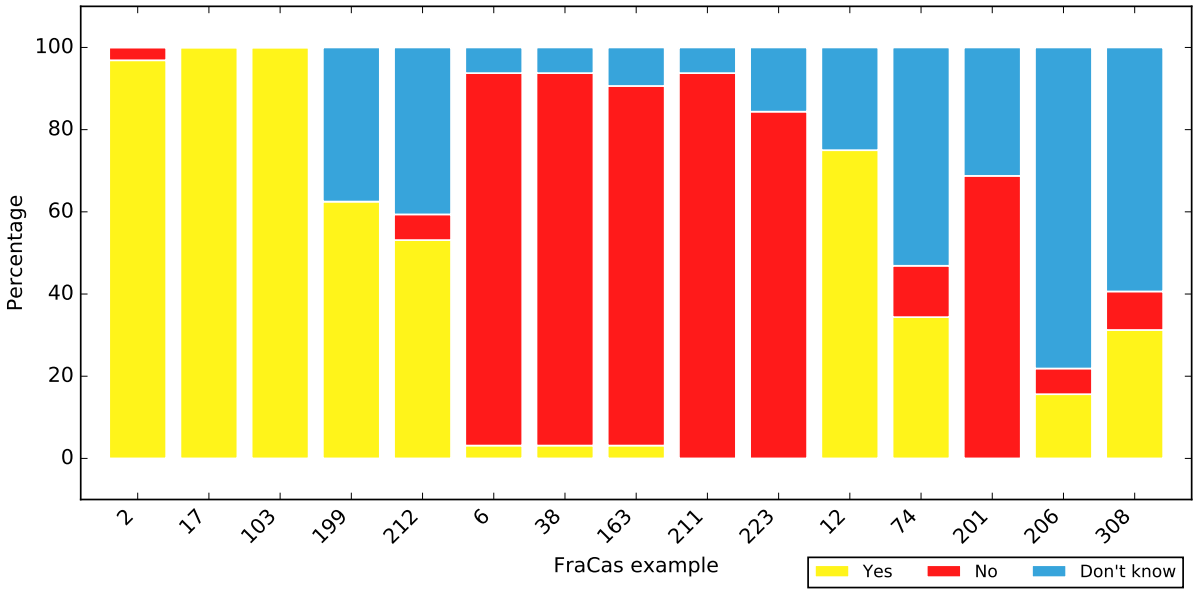

Figure 1: Judgements of 46 participants for four inference labels in FraCaS 'yes' (2, 17, 103, 199, 212), 'no' (6, 38, 163, 211, 223) and 'don't know' (74, 201, 206) + 'undefined' (12, 308).

as one or more statements (representing premises) and a question corresponding to the conclusion. Participants were instructed to answer the question choosing one of the three categories (yes, no and don't know) by only considering information contained in the sentences. Note that our answer *don't know* is different from the *don't know* category in FraCaS as it also includes cases that FraCaS labels as *undefined*. Hence, our category capture all cases where one cannot provide a *yes* and *no* answer from the information contained in the sentences but with some additional knowledge this would be possible.[15]

We recruited 46 participants, mostly employees (academic and non-academic) at University of Gothenburg some of whom are native speakers of English but the majority of participants had Swedish as their first language with excellent command of English. The results of their judgements are shown in Figure 1. Overall there is a strong agreement with the FraCas score on *yes* and *no* classes. Sometimes examples of the *yes* and *no* classes are labelled as *don't know*. Most variation exists in the *don't know* class, where some participants are likely to answer *yes*, possibly because they are bringing in additional background knowledge which allows them to draw a conclusion. The results also point out to examples within the FraCaS suite where the choice of the label may not be straightforward. For example, both 199 and 201 involve interpretation of adjectival scope ("former successful university student"). Results also indicate that the label *don't know* may also be assigned in cases where there are several premises and relations between premises which are likely contribute to high cognitive load and participants simply give up making an inference (example 212). We are extending the evaluation of the FraCas in our ongoing and future work by including (i) more examples; (ii) more participants; (ii) participants with different backgrounds (students who attended a computational semantics class vs non-experts); and (iii) different languages (currently Slovenian and Greek). Figure 1 may be given probabilistic interpretation as internalised beliefs of an agent about the most likely conclusion to draw after observing drawing of conclusions of other agents (Cooper et al., 2015). An alternative probabilistic model of inference is to query individual participants directly to indicate their degree of belief (rather than using 3 categories used in FraCaS) that an inference holds using a gradient scale with a help of an unlabelled slider (as in Dobnik and Åstbom, 2017) which we intend to explore in our future work.

[15]The task can be accessed at `https://linux.dobnik.net/simon/experiments/semant-o-matic/interface.en.flov.php?conversation=presentation_test`.

References

Abzianidze, L. (2015). A tableau prover for natural logic and language. In *Proceedings of EMNLP15*.

Bentivogli, L., P. Clark, I. Dagan, and D. Giampiccolo (2009). The fifth PASCAL recognizing textual entailment challenge. In *TAC*.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning (2015). A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pp. 632–642.

Breitholtz, E. and J. Villing (2008). Can aristotelian enthymemes decrease the cognitive load of a dialogue system user. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2008,'LonDial')*, pp. 94–100.

Cooper, R., D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman (1996). Using the framework. Technical report LRE 62-051r, The FraCaS consortium. `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

Cooper, R., S. Dobnik, S. Lappin, and S. Larsson (2015, November). Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology — LiLT 10*(4), 1–43.

Dobnik, S. and A. Åstbom (2017, August 15–17). (Perceptual) grounding as interaction. In V. Petukhova and Y. Tian (Eds.), *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany, pp. 17–26.

Dobnik, S., J. D. Kelleher, and C. Koniaris (2014, 1–3 September). Priming and alignment of frame of reference in situated conversation. In V. Rieser and P. Muller (Eds.), *Proceedings of DialWatt – Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, Edinburgh, pp. 43–52.

Ganitkevitch, E., P. Pavlick, J. Rastogi, B. Van Durme, and C. Callison-Burch (2015, July 26–31). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, Beijing, China, pp. 425–430. Association for Computational Linguistics.

Giampiccolo, D., H. T. Dang, B. Magnini, I. Dagan, E. Cabrio, and W. B. Dolan (2008). The fourth PASCAL recognizing textual entailment challenge. In *TAC*.

Kawazoe, A., R. Tanaka, K. Mineshima, and D. Bekki (2015). An inference problem set for evaluating semantic theories and semantic processing systems for Japanese. In *JSAI International Symposium on Artificial Intelligence*, pp. 58–65. Springer.

Lafourcade, M. (2011, December 8). *Lexique et analyse sémantique de textes - structures, acquisitions, calculs, et jeux de mots*. tel-00649851, Université Montpellier II - Sciences et Techniques du Languedoc.

MacCartney, B. (2007). The FraCaS textual inference problem set. Online resource, `http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`, Stanford University.

Marconi, D. (1997). *Lexical competence*. Cambridge, Mass.: MIT Press.

Marelli, M., S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pp. 216–223.

Martınez-Gómez, P., K. Mineshima, Y. Miyao, and D. Bekki (2017). On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*.

Zaenen, A., L. Karttunen, and R. Crouch (2005). Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pp. 31–36. Association for Computational Linguistics.