MonaLog: a Lightweight System for Natural Language Inference Based on Monotonicity

Hai Hu† Qi Chen† Kyle Richardson‡ Atreyee Mukherjee† Lawrence S. Moss† Sandra Kübler†

†Indiana University, Bloomington, IN, USA ‡Allen Institute for Artificial Intelligence, Seattle, WA, USA

{huhai,qc5,atremukh,lmoss,skuebler}@indiana.edu kyler@allenai.org

Highlights

We present a light-weight inference engine based on monotonicity and natural logic that:

1. relies solely on monotonicity information, thus intuitively straight-forward;
2. can be easily hybridized with BERT (Devlin et al., 2019);
3. generates high-quality inferences for data augmentation.

Outline

Goal: determine whether a hypothesis is ***entailed*** by, or ***neutral*** to, or ***contradictory*** to a premise. For example:

id	premise	hypothesis	orig. label	corr. label
340	**A schoolgirl with a black bag is on a crowded train**	**A girl with a black bag is on a crowded train**	Entail	Entail
219	There is no girl in white dancing	A girl in white is dancing	Cntrdt	Cntrdt
294	Two girls are lying on the ground	Two girls are sitting on the ground	Neutrl	Cntrdt
743	A couple who have just got married are walking down the isle	The bride and the groom are leaving after the wedding	Entail	Neutrl
1645	A girl is on a jumping car	One girl is jumping on the car	Entail	Neutrl
1981	A truck is quickly going down a hill	A truck is quickly going up a hill	Neutrl	Cntrdt
8399	A man is playing guitar next to a drummer	A guitar is being played by a man next to a drummer	Entail	n.a.

Table 1: Examples from the SICK dataset (Marelli et al., 2014) and corrected SICK (Kalouli et al., 2017, 2018) w/ syntactic variations. n.a.: example not checked by Kalouli and her colleagues. See below for explanation.

Method: *monotonicity tagging + substitution*

- Monotonicity tagging (Hu and Moss 2018):

  Every$^{\uparrow}$ *linguistics*$^{\downarrow}$ *student*$^{\downarrow}$ *speaks*$^{\uparrow}$ *at*$^{\uparrow}$ *least*$^{\uparrow}$ *3*$^{\downarrow}$ *languages*$^{\uparrow}$
  No$^{\uparrow}$ *man*$^{\downarrow}$ *walks*$^{\downarrow}$
  Every$^{\uparrow}$ *man*$^{\downarrow}$ *and*$^{\uparrow}$ *no*$^{\uparrow}$ *woman*$^{\downarrow}$ *sleeps*$^{=}$
  If$^{\uparrow}$ *some*$^{\downarrow}$ *man*$^{\downarrow}$ *walks*$^{\downarrow}$, *then*$^{\uparrow}$ *no*$^{\uparrow}$ *woman*$^{\downarrow}$ *runs*$^{\downarrow}$
  Every$^{\uparrow}$ *man*$^{\downarrow}$ *does*$^{\downarrow}$ *n't*$^{\uparrow}$ *hit*$^{\downarrow}$ *every*$^{\downarrow}$ *dog*$^{\uparrow}$
  No$^{\uparrow}$ *man*$^{\downarrow}$ *that*$^{\downarrow}$ *likes*$^{\downarrow}$ *every*$^{\downarrow}$ *dog*$^{\uparrow}$ *sleeps*$^{\downarrow}$
  Most$^{\uparrow}$ *men*$^{=}$ *that*$^{=}$ *every*$^{=}$ *woman*$^{=}$ *hits*$^{=}$ *cried*$^{\uparrow}$

- Substitution (Hu et al. 2019): replace a token or constituent with another one and maintain the logical relation at the same time, see Figure 2.

Dataset: Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014).

- Around 10k premise-hypothesis pairs, created via rule-templates, with human annotated relations.
- Many labels are wrong (see Table 1 above). We use the corrected version from Kalouli et al. (2018).

Our code is at `https://github.com/huhailinguist/ccg2mono`. Hai Hu is partly supported by China Scholarship Council.

MonaLog

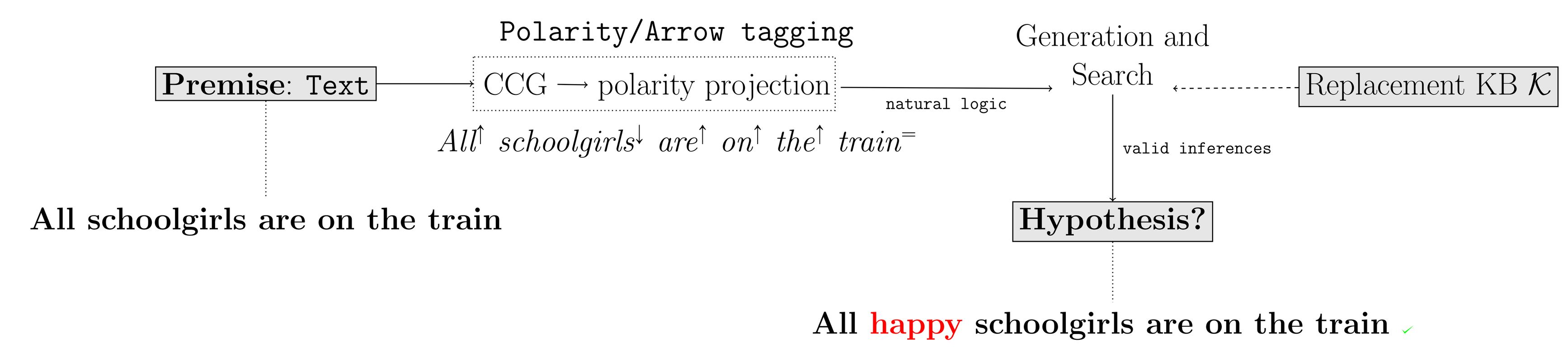

Figure 1: An illustration of our general monotonicity reasoning pipeline using an example premise and hypothesis pair: *All schoolgirls are on the train* and *All happy schoolgirls are on the train*.

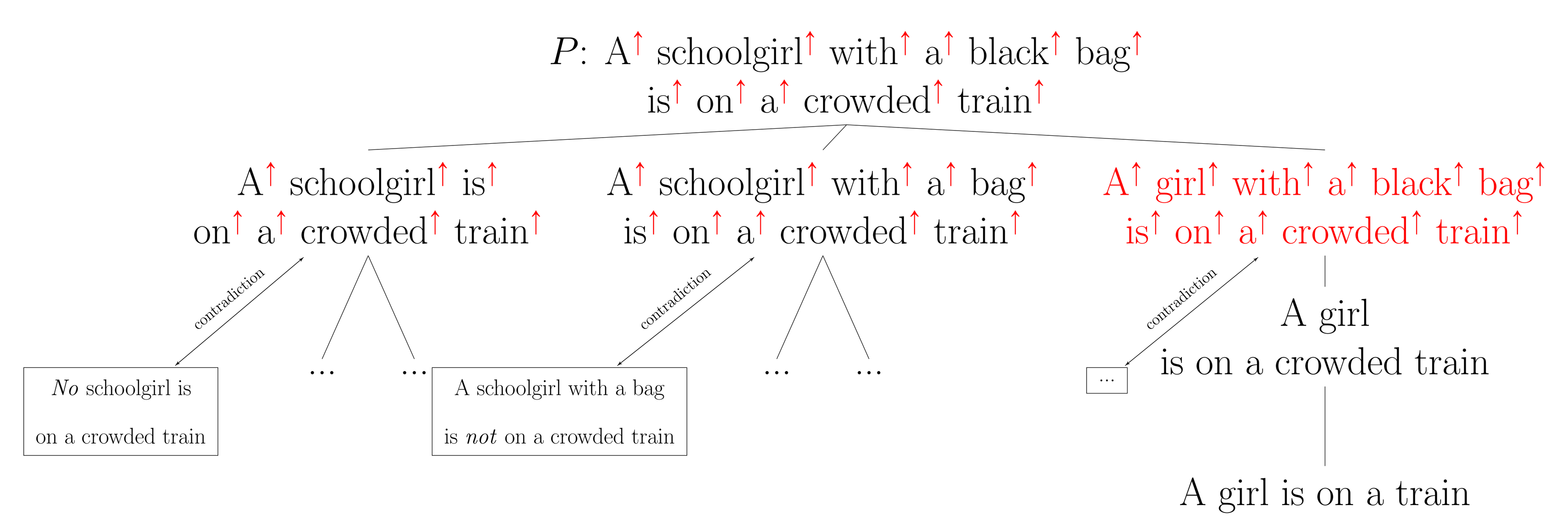

Figure 2: Example search tree for SICK 340, where P is *A schoolgirl with a black bag is on a crowded train*, with the H: *A girl with a black bag is on a crowded train*. Only one `replacement` is allowed at each step. Sentences at the nodes are generated entailments. Sentences in rectangles are the generated contradictions. In this case our system will return `entail`. The search will terminate after reaching the H in this case, but for illustrative purposes, we show entailments of depth up to 3. To exclude the influence of morphology, all sentences are represented at the lemma level in MonaLog, which is not shown here.

Experiment 1: inference engine

system	P	R	acc.
On **uncorrected** SICK			
majority baseline	–	–	56.36
hypothesis-only baseline (Poliak et al., 2018)	–	–	56.87
MonaLog (this work)			
MonaLog + all transformations	83.75	70.66	77.19
Hybrid: MonaLog + BERT	83.09	85.46	85.38
ML/DL-based systems			
BERT (base, uncased)	86.81	85.37	86.74
Yin and Schütze (2017)	–	–	**87.1**
Beltagy et al. (2016)	–	–	85.1
Logic-based systems			
Bjerva et al. (2014)	93.6	60.6	81.6
Abzianidze (2015)	97.95	58.11	81.35
Martínez-Gómez et al. (2017)	97.04	63.64	83.13
Yanaka et al. (2018)	84.2	77.3	84.3
On **corrected** SICK			
MonaLog + existential trans.	89.43	71.53	79.11
MonaLog + pass2act	89.42	72.18	80.25
MonaLog + all transformations	89.91	74.23	81.66
Hybrid: MonaLog + BERT	85.65	87.33	**85.95**
BERT (base, uncased)	84.62	84.27	85.00

Table 2: Performance on original/corrected SICK test set. P / R for MonaLog averaged across three labels. Results involving BERT are averaged across six runs.

Discussion:

- Important to perform syntactic transformations first;
- Data correction plays an important role;
- Well above the majority baseline and not too far below other logic-based models;
- Hybrid with BERT:
  - Trust MonaLog if it predicts E or C, otherwise use predictions from BERT;
  - Improves accuracy on corrected SICK.

Experiment 1: inference engine

id	premise	hypothesis	SICK	corr. SICK	Mona
359	There is no dog chasing another or holding a stick in its mouth	Two dogs are running and carrying an object in their mouths	N	n.a.	C
1402	A man is crying	A man is screaming	N	n.a.	E
1760	A flute is being played by a girl	There is no woman playing a flute	N	n.a.	C
2897	The man is lifting weights	The man is lowering barbells	N	n.a.	C
2922	A herd of caribous is not crossing a road	A herd of deer is crossing a street	N	n.a.	C
3403	A man is folding a tortilla	A man is unfolding a tortilla	N	n.a.	C
4333	A woman is picking a can	A woman is taking a can	E	N	E
5138	A man is doing a card trick	A man is doing a magic trick	N	n.a.	E
5793	A man is cutting a fish	A woman is slicing a fish	N	n.a.	C

Table 3: Examples of incorrect answers by MonaLog; n.a. = the problem has not been checked in corr. SICK. C: contradiction; E: entailment; N: neutral.

- Some are mistakes in the original SICK, e.g., 359, 1760, 2897, etc.
- Some are hard to determine, e.g., 1402, 5793.

→ highlight the precision of MonaLog + need for high-quality annotation.

Experiment 2: data generation

MonaLog can generate inferences (= entailments + contradictions) from a given input sentence (see Figure 2). We train BERT on SICK.train plus the generated data in multiple settings and test on SICK.

label	premise	hypothesis	comm.
E	A woman be not cooking something	A person be not cooking something	correct
E	A man be talk to a woman who be seat beside he and be drive a car	A man be talk	correct
E	A south African plane be not fly in a blue sky	A south African plane be not fly in a very blue sky in a blue sky	unnat.
C	No panda be climb	Some panda be climb	correct
C	A man on stage be sing into a microphone	A man be not sing into a microphone	correct
C	No man rapidly be chop some mushroom with a knife	Some man rapidly be chop some mushroom with a knife with a knife	unnat.
E	Few$^{\uparrow}$ people$^{\downarrow}$ be$^{\downarrow}$ eat$^{\downarrow}$ at$^{\downarrow}$ red$^{\downarrow}$ table$^{\downarrow}$ in$^{\downarrow}$ a$^{\downarrow}$ restaurant$^{\downarrow}$ without$^{\downarrow}$ light$^{\uparrow}$	Few$^{\uparrow}$ large$^{\downarrow}$ people$^{\downarrow}$ be$^{\downarrow}$ eat$^{\downarrow}$ at$^{\downarrow}$ red$^{\downarrow}$ table$^{\downarrow}$ in$^{\downarrow}$ a$^{\downarrow}$ Asian$^{\downarrow}$ restaurant$^{\downarrow}$ without$^{\downarrow}$ light$^{\uparrow}$	correct

Table 4: Sentence pairs generated by MonaLog, lemmatized.

training data	# E	# N	# C	acc.
SICK.train: baseline	1.2k	2.5k	0.7k	85.00
1/4 gen. + SICK.train	8k	2.5k	4k	85.30
1/2 gen. + SICK.train	15k	2.5k	7k	85.81
all gen. + SICK.train	30k	2.5k	14k	86.51
E, C prob. threshold = 0.95	30k	2.5k	14k	86.71
Hybrid baseline	1.2k	2.5k	0.7k	85.95
Hybrid + all gen.	30k	2.5k	14k	87.16
Hybrid + all gen. + threshold	30k	2.5k	14k	**87.49**

Table 5: Results of BERT trained on MonaLog-generated entailments and contradictions plus SICK.train.

We see a 2 percent boost in accuracy, despite the highly skewed generated data, suggesting that we have generated high-quality data that are useful for a machine learner.

Future work

- Machine-learning methods for monotonicity tagging, and the handling of syntactic transformations;
- Explore ways of generating neutral statements, non-lemmatized sentences; sentence filtering methods;
- A fully corrected SICK dataset.