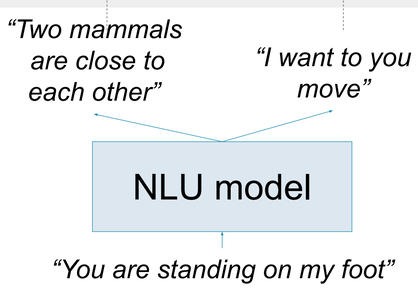
A Pragmatics-Centered Evaluation Framework for Natural Language Understanding

Damien Sileo[1;2,3], Tim Van de Cruys[2], Camille Pradel[1], Philippe Muller[3,4]
1:Synapse Développement, 2:KU Leuven, 3:IRIT, University of Toulouse, 4: Artificial and Natural Intelligence Toulouse Institute (ANITI)
damien.sileo@kuleuven.be

LREC 2022 Marseille

Semantics vs Pragmatics



Contributions

- Pragmeval, evaluation suite centered on pragmatics
- Evaluations comparing NLI and Discourse marker prediction as STILT

```
from datasets import load_dataset
dataset = load_dataset('pragmeval','pdtb')
```

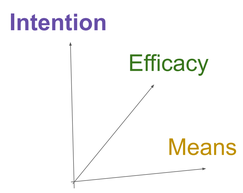
Related work

GLUE (Wang et al. 2019) semantic similarity, NLI, sentiment/cola

SuperGLUE (Wang et al. 2019) NLI, commonsense,knowledge, commitment,

DiscoEval (Chen et al. 2019) focused on discourse structure (paragraphs/nesting levels)

Datasets



Dataset	Example	Class
PDTB	*it was censorship / it was outrageous*	conjunction
STAC	*what? / i literally lost*	question-answer-pair
GUM	*Do not drink / if underage in your country*	condition
Emergent	*a meteorite landed in nicaragua. / small meteorite hits managua*	for
SwitchBoard	*well, a little different, actually*	hedge
MRDA	*yeah that's that's that's what i meant .*	acknowledge-answer
Persuasion	*Co-operation is essential for team work / lions hunt in a team*	low specificity
SarcasmV2	*don't quit your day job / [...] i was going to sell this joke. [...]*	sarcasm
Squinky	*boo ya.*	uninformative, high implicature, informal
Verifiability	*I've been a physician for 20 years.*	verifiable-experiential
EmoBank	*I wanted to be there..*	low valence, high arousal

General evaluation

model	PDTB	STAC	GUM	Emergent	SwitchB.	MRDA	Persuasion	Sarcasm	Squinky	Verif.	EmoBank
CBoW	27.4	-	20.5	59.7	3.8	0.7	70.6	61.1	75.5	74.0	64.0
BERT	48.8	48.2	40.9	79.2	38.8	22.3	74.8	77.1	87.5	86.7	76.2
BERT+MNLI	49.1	49.1	42.8	81.2	38.1	22.7	71.7	73.4	88.2	86.0	76.3
BERT+PragmEval	49.1	**57.1**	42.8	80.2	40.3	**23.1**	**76.2**	75.0	87.6	85.9	76.0
BERT+DisSent	49.4	49.0	43.9	79.8	39.2	22.0	74.7	74.9	87.5	85.9	76.2
BERT+Discovery	50.7	49.5	42.7	**81.7**	39.5	22.4	71.6	76.7	88.6	86.3	**76.6**
B+Discovery+MNLI	51.3	49.4	43.1	80.7	**40.3**	22.2	73.6	75.1	**88.9**	86.8	76.0
Human estimate	**84.0**	-	**69.3**	-	-	-	-	**80.0**	-	**87.0**	73.1

Comparison with GLUE

model	PragmEval AVG	GLUE AVG	GLUE diagnostics
BERT	61.8±.4	74.7±.2	31.7±.3
BERT+MNLI	61.7±.5	**77.0±.2**	32.5±.6
BERT+PragmEval	**63.0±.4**	75.3±.2	31.6±.3
BERT+DisSent	62.0±.4	75.1±.2	31.5±.3
B+DisSent+MNLI	62.1±.4	76.6±.1	32.4±.0
BERT+Discovery	62.4±.3	75.0±.2	31.3±.2
B+Discovery+MNLI	62.5±.4	76.6±.2	**33.3±.2**

Design choices

- Sentence of sentence pairs classification tasks
- Labels subsampling for dominant classes (MRDA, Switchboard)

Experimental setup

Jiant framework with default parameters
3 epochs, learning rate starting at 1e-4

Conclusion

➔ GLUE offers a partial view of understanding

➔ Discourse marker prediction is a useful STILT for pragmatics

➔ Combining MNLI with Discovery STILTS also help

References

(Wang et al 2019a) GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding
(Wang et al 2019b) SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems
(Chen et al 2019) Evaluation Benchmarks and Learning Criteria for Discourse-Aware Sentence Representations
(Sileo et al. 2019) Mining Discourse Markers for Unsupervised Sentence Representation Learning

Acknowledgments

Philippe Muller is partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisci plinary Institute, ANITI, as a part of France's "Investing for the Future - PIA" program.

This work is part of the CALCULUS project, which is funded by the ERC Advanced Grant H2020-ERC-2017 DG788506
https://calculus-project.eu/