Natural Language Reasoning with a Natural Theorem Prover

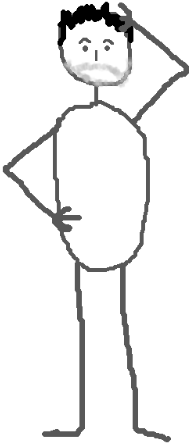
Day 1: Natural Language Inference

Lasha Abzianidze

Utrecht University

33rd ESSLLI in Gaillimh, Éire
8-12 August 2022

About me



- Masters in Language & Communication Technologies (LCT)
- PhD from Tilburg University
- Postdoc at University of Groningen & Utrecht University
- Assistant professor at Utrecht University

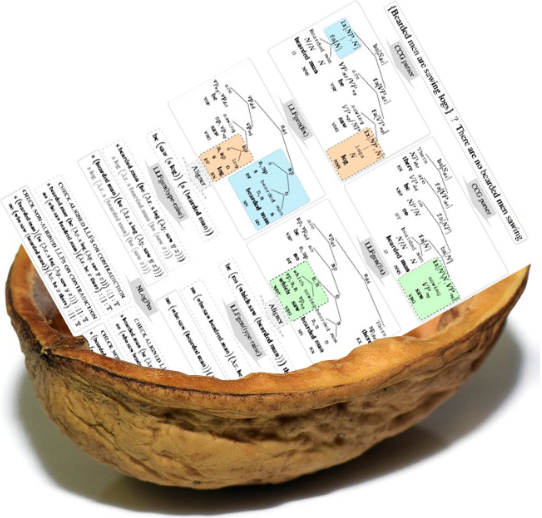
About the course

The following research questions are at the heart of the course:

- How are the meanings of natural language sentences related to each other?
- How to systematically reason with natural language sentences?
- How to get an explainable reasoning system?
- Can I use expressive but at the same time *friendly* meaning representations?

There will be many *natural* trees

Course in a nutshell



Topics per day

Mon Natural Language Inference
The task of NLI; why (natural) logic-based approaches; natural logic

Tue Semantic Tableau Method
Tableau systems for Propositional, First- & Higher-Order (& natural) logics

Wed Wide-coverage Natural Tableau
Natural tableau; getting logical forms; wide-coverage inference rules

Thu Theorem Prover for Natural Language
Automatizing Natural Tableau for NLI; FraCaS NLI dataset,
colab notebook with the prover

Fri Learning from and proving NLI problems
Learning as abduction; SICK NLI dataset; bilingual theorem proving

Course web page: naturallogic.pro/Teaching/esslli22

Course is NOT about

- Machine learning
  (but there will be some data-driven learning)
- Lexicalized formal compositional semantics:
  $[\![\text{John}]\!]@([\![\text{loves}]\!]@[\![\text{Mary}]\!]) = \texttt{love(john,mary)}$
- Only toy examples
- Proving theorems about formal logics
  (but proving natural language *theorems*)

A few prerequisites

- Syntax of first-order logic formulas
- Understanding semantics of first-order logic formulas:
  $\exists y \forall x.\texttt{love}(x, y) \rightarrow \forall x \exists y.\texttt{love}(x, y)$
- Some knowledge of (simply typed) λ-calculus
- Some familiarity with (Combinatory) Categorial Grammars

Outcome of the course

In the end of the course you will know about:

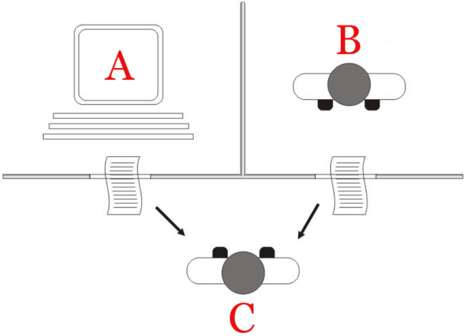
- Challenges posed by Natural Language Inference (NLI)
- Pros & cons of logic-based methods wrt NLI
- Nitty-gritty details of theorem proving with a tableau system
- Doing semantics with higher-order logic
- How to account for a semantic phenomenon in Natural Tableau
- How to use the Natural Tableau prover to solve inference problems

Natural Language Understanding

Natural language understanding is one of the main problems of AI and NLP.

How to test whether a machine/program understands a natural language?

Use the Turing test. ⚠Too expensive! 👎 👎 👎



Juan Alberto Sánchez Margallo: https://commons.wikimedia.org/wiki/File:Test_de_Turing.jpg

Natural Language Understanding

Natural language understanding is one of the main problems of AI and NLP.

How to test whether a machine/program understands a natural language?

Given a sentence, ask what it means.

A brown cat is lying on a mat

A1: A pet, which is brown, is lying
A2: A cat is on a mat
A3: There is an animal lying on a mat

⋮

⚠Evaluation of the answers requires a system that understands natural language. 👎 👎

Natural Language Understanding

Natural language understanding is one of the main problems of AI and NLP.

How to test whether a machine/program understands a natural language?

Given a sentence, ask a yes/no/dunno question about its meaning.
A brown cat is lying on a yellow mat

Q1: Is the brown cat lying on the mat? [Yes]
Q2: Is the mat soft? [Dunno]
Q3: Is the cat jumping on the mat? [No]

⋮

⚠ This focuses on (long) questions and declarative sentences, and it is not straightforward to cover noun phrases and imperatives? 👎

A cat on the mat Feed the cat on the mat!

Natural Language Understanding

Natural language understanding is one of the main problems of AI and NLP.

How to test whether a machine/program understands a natural language?

Given two sentences $S1$ and $S2$, detect whether $S1$ entails $S2$.

$S1$: A brown cat is lying on a yellow mat
$S2$: There is an animal on a yellow-colored mat
Answer: Yes

$S1$: The cat lying on a mat
$S2$: The cat rolling on a yellow mat
Answer: No


https://www.publicdomainpictures.net

Contrasting phrases of the same category, e.g., noun phrase, declarative sentences, questions, etc.

Recognizing Textual Entailment (2005-2013)

The task of Recognizing Textual Entailment (RTE) was introduced by [Dagan et al., 2005]:

Textual entailment is defined as a *directional relationship* between pairs of text expressions, denoted by T (the entailing "Tex") and H (the entailed "Hypothesis"). We say that T entails H if humans reading T would typically infer that H is most likely true.

An RTE task:

- Given two texts, T (text) and H (hypothesis), detect textual entailment from T to H.
- RTE1 to RTE3 challenges: binary classification
- RTE4 to RTE8 challenges: 3-way classification with long texts

Initially, RTE was considered a generic task, and the data was created based on the related tasks, e.g., question answering, document summarization, and information extraction/retrieval.

RTE problems

RTE2 GOLD: `non−entailment`

Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."

A case of rabies was confirmed

RTE2 GOLD: `entailment`

About two weeks before the trial started, I was in Shapiro's office in Century City

---

Shapiro works in Century City

RTE2 GOLD: `entailment`

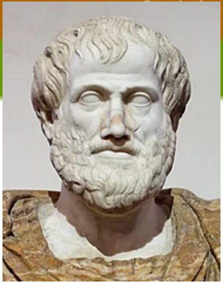
The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them

---

Alzheimer's disease is treated using drugs

Taken from [Dagan et al., 2013]

Long before RTE: Aristotle's syllogisms



Aristotle's syllogisms (4th century BC):

- 256 RTE problems
- Text consists of two sentences
- 24(!) of the problems are entailment

$\boxed{\text{OAO-3}}$ GOLD: `entailment`

Some cats have no tails
All cats are mammals

Some mammals have no tails

$\boxed{\text{IAE-3}}$ GOLD: `contradiction`

Some dogs have spots
All dogs are mammals

No mammals have spots

$\boxed{\text{IAA-1}}$ GOLD: `neutral`

Some vehicles are electric
All cars are vehicles

All cars are electric

$\boxed{\text{AAI-1}}$ GOLD: `neutral`

All canids are mammals
All chupacabras are canids

Some chupacabras are mammals

Modern NLI

The Stanford NLI (SNLI) corpus [Bowman et al., 2015]:

- Large corpus: 570K premise-hypothesis pairs
- Tackling entity & event co-reference by grounding in images
- Premises are image captions and hypotheses are generated by crowd workers
- Premise-hypothesis pairs are annotated by 5 crowd workers with 3 labels

The Multi-Genre NLI (MultiNLI) corpus [Williams et al., 2018]:

- Large corpus: 433K premise-hypothesis pairs
- It is modeled on the SNLI corpus
- 10 genres: Fiction, Governmnet, Slate, Telephone, Travel, 9/11,...
- Used as a sentence encoder benchmark at RepEval 2019

Examples from SNLI (arguable examples)

SNLI-3581451227.jpg#4r1c GOLD: `contradiction` ^{1c}

A little girl and boy after a wedding in a field

---

the sail boat sank in the ocean

SNLI-475816542.jpg#2r1c GOLD: `contradiction` ^{3c2e}

A black and a brown dog are running toward the camera.

A black and a brown cat are running toward the camera.

SNLI-4837051771.jpg#2r1n GOLD: `neutral` ^{2e3n}

A small ice cream stand with two people standing near it.

Two people in line to buy icecream.

SNLI-2218907190.jpg#1r1e GOLD: `entailment` ^{2c3e}

A dog begins to climb a brick staircase near plants.

A dog is going up the stairs.

Critical look at SNLI

In the test part of SNLI, 12.2% of problems get 2vs3 annotations.

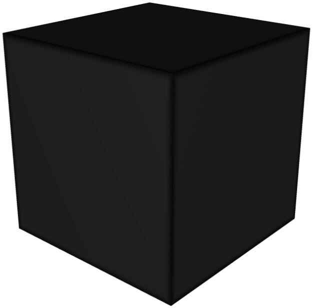
Annotation artifacts inflate systems' performance [Poliak et al., 2018, Gururangan et al., 2018]:

- Hypothesis only baselines score strikingly high wrt the majority class baseline: 69.2% vs 33.8% accuracy
- *animal*, *outdoors*, and *person* often in `entailment` hypotheses
- *tall*, *sad*, and *first* often in `neutral` hypotheses
- *cat*, *sleeping*, and *no* often in `contradiction` hypotheses

Smart black boxes

Author performance on test-SNLI (91.4%) is already suppressed by a deep neural network-based system (92.1%)

☞SNLI leaderboard



What does this mean?

Specially dedicated workshops:
BlackboxNLP

Harder challenges aka *Task-Independent Sentence Understanding*:

☞GLUE leaderboard

☞SuperGLUE leaderboard

Explainable reasoning

Explainable reasoning is a feature associated with a white box systems:

- Explain entailment by providing some sort of proof or argument
- Explain contradiction by highlighting the incompatible cases
- Explain neutral relation by providing counterexamples for entailment and contradiction

e-SNLI – NLI with natural language explanations [Camburu et al., 2018]:

- For each labelled NLI problem, crowd workers gave explanations;
- Also the word relevant for explanations were highlighted;
- An NLI system needs to predict a label and an explanation;
- How to evaluate predicted explanation automatically?
  BLEU-score is a poor metric for this purpose.

Examples from e-SNLI

SNLI-3581451227.jpg#4r1c GOLD: `contradiction`[1c]

A little girl and boy after a wedding in a field

the sail boat sank in the ocean

A girl and boy are people, not a thing, as a sail boat is. You cannot be in a field and in the ocean at the same time

SNLI-475816542.jpg#2r1c GOLD: `contradiction`[3c2e]

A black and a brown dog are running toward the camera.

A black and a brown cat are running toward the camera.

They refer to a dog, not a cat

The animal is either a cat or a dog

A dog cannot be a cat

Examples from e-SNLI (II)

SNLI-4837051771.jpg#2r1n GOLD: neutral[2e3n]

A small ice cream stand with two people standing near it.

Two people in line to buy icecream.

Being near a stand doesn't mean you have to buy anything

Just because two people are standing near an ice cream stand, doesn't mean they are in line to buy ice cream

People who are standing near an ice cream stand are not always in line to buy ice cream

SNLI-2218907190.jpg#1r1e GOLD: entailment[2c3e]

A (dog) begins to [climb] a brick (staircase) near plants.

A (dog) is [going up] the (stairs).

Brick staircase is a paraphrase of stairs, and going up means to climb

[Climbing implies going up]

(A DOG IS CLIMBING UP THE STAIRS)

Shortcomings of NLI systems

- *Not many* NLI systems are able to reason over multiple premises [Lai et al., 2017].
- Most NLI systems do not use logic-based reasoning: poor at processing Booleans (e.g., *or*, *not*) and quantifiers (e.g., *every*, *no*).

  P1: Most boxers have been knocked out
  P2: All boxers are athletes
  P3: All athletes who has been knocked out has a broken nose
  ---
  C: Most boxers have a broken nose

- SOTA NLI systems are not explanatory (though pretty good).
- Most RTE systems can be fooled easily (i.e. not having high precision)

  SICK-1745 GOLD:: `neutral`
  T: A man is pushing the buttons of a microwave

  H: A man is being pushed toward the buttons of a microwave

Hype of NLI

But some deep learning models beat humans in NLI

Several corrections:

- ~~humans~~ → crowd workers (often underpaid)
- beat in specific domains (not the most difficult ones)
- An NLI task is a simple offshoot of a reasoning task

More representative task for reasoning is needed where a line of reasoning is evaluated: structure prediction vs classification.

Entailment trees [Dalvi et al., 2021]

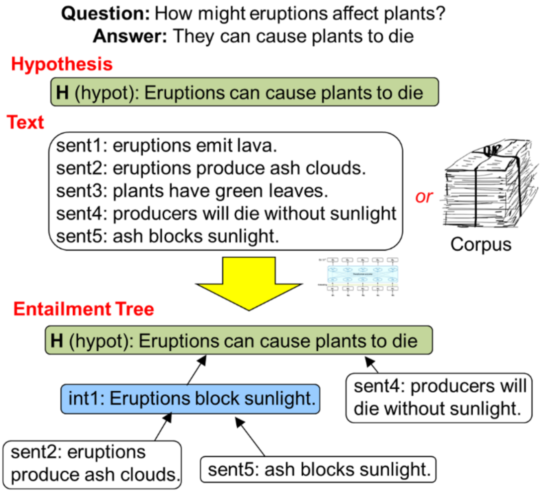
Getting closer to reasoning with generating entailment trees.

Premise sentences are leaves while the root is the hypothesis.

Possibly generating intermediate conclusions.

Evaluation scores (exact tree match):

- no-distractor premises: 35.6
- some distractor premises: 25.6
- full corpus as premises: 2.9

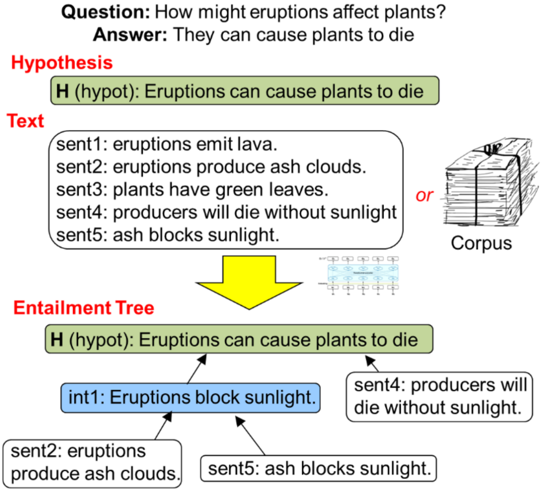


Gap between NLI and entailment trees

A reasoning task that incorporates virtues from NLI and entailment trees:

- intra-sentence reasoning;
- predicting trees or graphs.

Logic-based NLI systems can be leveraged for this purpose.



Logic-based NLI systems

Nutcracker [Bos and Markert, 2005]: CCG parsing, compositional DRS semantics, translation to FOL, FOL theorem proving & model building.

NatLog [MacCartney and Manning, 2007]: PCFG parsing, polarity marking, monotonicity calculus + string edits.

ccg2lambda [Mineshima et al., 2015]: CCG parsing, compositional mostly FOL semantics, Coq for theorem proving.

FraCoq [Bernardy and Chatzikyriakidis, 2017]: GF (treebank), modern/rich type theory, Coq for theorem proving.

MonaLog [Hu et al., 2019]: CCG parsing, polarity marking, monotonicity calculus.

Shortcomings of logic-based NLI systems

- Their logic is often not expressive enough to model some aspects of linguistics semantics: higher-order terms like generalized quantifiers (e.g., *few*, *most*) and subsective modifiers e.g. *competent*, *slowly*.
- Translation of linguistic semantics into formal logic is usually a complex and immense problem, e.g., NL text into first-order logic [Bos, 2008].
- After the translation, information about constituency and syntax is not available in a formal language while the information is often crucial for *shallow* reasoning, e.g., monotonicity reasoning.

  $\boxed{\text{SICK-8145}}$ GOLD: `entailment`

  T: A woman in blue has a $\boxed{\text{yellow}}$ $\boxed{\text{ball}}$ in the $\boxed{\text{mitt}}$

  H: A woman in blue has a $\boxed{\text{colored}}$ $\boxed{\text{thing}}$ in the $\boxed{\text{hand}}$

Natural logic

Natural logic is a hypothetical logic which is built in natural language and represents its integral part.

It is a theory about "the regularities governing the notion of a valid argument for reasoning in natural language" [Lakoff, 1970].

"Natural logic is a somewhat loose [...] term for [...] attempts [...] at describing basic patterns of human reasoning directly in natural language without the intermediate of some formal system" [van Benthem, 2008].

Natural logic is "the study of inference in natural language, done as close as possible to the surface forms" [Moss, 2010b].

Monotonicity reasoning

The most popular and success story of natural logic is monotonicity reasoning.

Monotonicitity reasoning is about replacing phrases in a premise in such a way that the obtained sentences are entailment of the premise.

GOLD: `entailment`

P: Every man who consumed alcohol gobbled up most snacks

H: Every young man who drank beer ate some snacks

GOLD: `entailment`

P: $3 \times [s_3(x) = x+3](2) \;\leq\; [p_3(x) = x^3]\big([m_4(x) = x \mod 4](7)\big)$

H: $2 \times [s_1(x) = x+1](1) \;\leq\; [p_4(x) = x^4]\big([m_8(x) = x \mod 8](7)\big)$

Do you see similarity between these two entailment pairs?

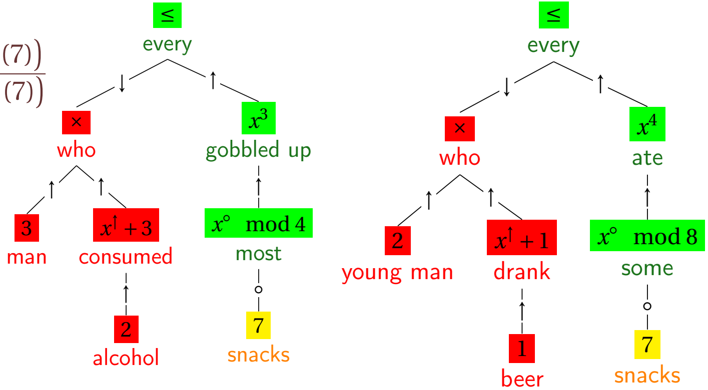
Now?

Monotonicity reasoning in action

GOLD: `entailment`

P: $3 \times [x+3](2) \;\le\; [x^3]\big([x \mod 4](7)\big)$

H: $2 \times [x+1](1) \;\le\; [x^4]\big([x \mod 8](7)\big)$



GOLD: `entailment`

P: Every man who consumed alcohol gobbled up most snacks

H: Every young man who drank beer ate some snacks

NatLog [MacCartney, 2009]

First NLI system that introduced natural logic to the NLP community

- Semantic relation over lexical items
- composition/Chaining of the semantic relations
- Projection of the relations (for substitutions)
- Deletion and insertion of implicatives

symbol[10]	name	example	set theoretic definition[11]
$x \equiv y$	equivalence	*couch* $\equiv$ *sofa*	$x = y$
$x \sqsubset y$	forward entailment	*crow* $\sqsubset$ *bird*	$x \subset y$
$x \sqsupset y$	reverse entailment	*Asian* $\sqsupset$ *Thai*	$x \supset y$
$x \wedge y$	negation	*able* $\wedge$ *unable*	$x \cap y = \emptyset \wedge x \cup y = U$
$x \mid y$	alternation	*cat* $\mid$ *dog*	$x \cap y = \emptyset \wedge x \cup y \neq U$
$x \smile y$	cover	*animal* $\smile$ *non-ape*	$x \cap y \neq \emptyset \wedge x \cup y = U$
$x \# y$	independence	*hungry* $\#$ *hippo*	(all other cases)

NatLog [MacCartney, 2009]

First NLI system that introduced natural logic to the NLP community

- Semantic relation over lexical items
- composition/Chaining of the semantic relations
- Projection of the relations (for substitutions)
- Deletion and insertion of implicatives

$\bowtie$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$\equiv$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$\sqsubset$	$\sqsubset$	$\sqsubset$	$\equiv\sqsubset\sqsupset\mid\#$	$\mid$	$\mid$	$\sqsubset\wedge\mid\smile\#$	$\sqsubset\mid\#$
$\sqsupset$	$\sqsupset$	$\equiv\sqsubset\sqsupset\smile\#$	$\sqsupset$	$\smile$	$\sqsupset\wedge\mid\smile\#$	$\smile$	$\sqsupset\smile\#$
$\wedge$	$\wedge$	$\smile$	$\mid$	$\equiv$	$\sqsupset$	$\sqsubset$	$\#$
$\mid$	$\mid$	$\sqsubset\wedge\mid\smile\#$	$\mid$	$\sqsubset$	$\equiv\sqsubset\sqsupset\mid\#$	$\sqsubset$	$\sqsubset\mid\#$
$\smile$	$\smile$	$\smile$	$\sqsupset\wedge\mid\smile\#$	$\sqsupset$	$\sqsupset$	$\equiv\sqsubset\sqsupset\smile\#$	$\sqsupset\smile\#$
$\#$	$\#$	$\sqsubset\smile\#$	$\sqsupset\mid\#$	$\#$	$\sqsupset\mid\#$	$\sqsubset\smile\#$	$\equiv\sqsubset\sqsupset\wedge\mid\smile\#$

NatLog [MacCartney, 2009]

First NLI system that introduced natural logic to the NLP community

- Semantic relation over lexical items
- composition/Chaining of the semantic relations
- Projection of the relations (for substitutions)
- Deletion and insertion of implicatives

	projectivity for 1^{st} argument							projectivity for 2^{nd} argument						
quantifier	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
some	$\equiv$	$\sqsubset$	$\sqsupset$	$\smile^{\dagger}$	$\#$	$\smile^{\dagger}$	$\#$	$\equiv$	$\sqsubset$	$\sqsupset$	$\smile^{\dagger}$	$\#$	$\smile^{\dagger}$	$\#$
no	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid^{\dagger}$	$\#$	$\mid^{\dagger}$	$\#$	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid^{\dagger}$	$\#$	$\mid^{\dagger}$	$\#$
every	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid^{\ddagger}$	$\#$	$\mid^{\ddagger}$	$\#$	$\equiv$	$\sqsubset$	$\sqsupset$	$\mid^{\dagger}$	$\mid^{\dagger}$	$\#$	$\#$
not every	$\equiv$	$\sqsubset$	$\sqsupset$	$\smile^{\ddagger}$	$\#$	$\smile^{\ddagger}$	$\#$	$\equiv$	$\sqsupset$	$\sqsubset$	$\smile^{\dagger}$	$\smile^{\dagger}$	$\#$	$\#$
at least two	$\equiv$	$\sqsubset$	$\sqsupset$	$\#$	$\#$	$\#$	$\#$	$\equiv$	$\sqsubset$	$\sqsupset$	$\#$	$\#$	$\#$	$\#$
most	$\equiv$	$\#$	$\#$	$\#$	$\#$	$\#$	$\#$	$\equiv$	$\sqsubset$	$\sqsupset$	$\mid$	$\mid$	$\#$	$\#$

NatLog [MacCartney, 2009]

First NLI system that introduced natural logic to the NLP community

- Semantic relation over lexical items
- composition/Chaining of the semantic relations
- Projection of the relations (for substitutions)
- Deletion and insertion of implicatives

signature	$\beta(\text{DEL}(\cdot))$	$\beta(\text{INS}(\cdot))$	example
$+/-$	$\equiv$	$\equiv$	*he managed to escape* $\equiv$ *he escaped*
$+/\circ$	$\sqsubset$	$\sqsupset$	*he was forced to sell* $\sqsubset$ *he sold*
$\circ/-$	$\sqsupset$	$\sqsubset$	*he was permitted to live* $\sqsupset$ *he lived*
$-/+$	$\wedge$	$\wedge$	*he failed to pay* $\wedge$ *he paid*
$-/\circ$	$\mid$	$\mid$	*he refused to fight* $\mid$ *he fought*
$\circ/+$	$\smile$	$\smile$	*he hesitated to ask* $\smile$ *he asked*
$\circ/\circ$	$\#$	$\#$	*he believed he had won* $\#$ *he had won*

NatLog in action

P: John refused to move without blue jeans

H: John didn't dance without pants

Detect polarities of the words in the premise

	Sentence & Atomic edit	Lexical	Projected	Overall
(S0)	John refused to / move without / blue jeans			
	(E1) del(refused to)	$\mid$	$\mid$	$\mid$
(S1)	John moved without / blue jeans			
	(E2) ins(didn't)	$\wedge$	$\wedge$	$\sqsubset$
(S2)	John didn't / moved without / blue jeans			
	(E3) sub(move, dance)	$\sqsupset$	$\sqsubset$	$\sqsubset$
(S3)	John didn't / dance without / blue jeans			
	(E4) del(blue)	$\sqsubset$	$\sqsubset$	$\sqsubset$
(S4)	John didn't / dance without / jeans			
	(E5) sub(jeans, pants)	$\sqsubset$	$\sqsubset$	$\sqsubset$
(S5)	John didn't dance without pants			

Shortcomings of NatLog

- Cannot account for paraphrases:

  John bought a car from Bill
  Bill sold a car to John

  A student wrote an essay
  An essay was written by a student

- *Weaker* than first-order logic:

  Not all bird fly
  Some birds does not fly

- Not all ordered string edits lead to a correct relation

- The word-alignment and -substitution nature of reasoning falls short of processing multiple premises

Related work

Other works on monotonicity reasoning and natural logic:

- First study of monotonicity reasoning as a formal calculus [Van Benthem, 1986, van Benthem, 1987, Sánchez-Valencia, 1991]
- Moving from syllogistic logics towards natural logic [Moss, 2010a]
- A tableau proof system for a fragment of natural logic [Muskens, 2010]
- Formal system for extended monotonicity reasoning [MacCartney and Manning, 2008, Icard, 2012, Icard and Moss, 2014]

Implementations:

- Monotonicity-based inference system for a fragment of English, operating on categorical grammar derivation trees [Fyodorov et al., 2003, Zamansky et al., 2006]
- Implementation of syllogistic logic with monotonicity [Eijck, 2005]
- Two implementations of extended syllogistic logics [Hemann et al., 2015]
- Natural language inference using polarity-marked parse trees [Hu et al., 2019]

Conclusion

The NLI task is popular task (with many datasets) for evaluating reasoning capacity of NLP systems.

but is too simple for evaluating reasoning

Logic-based systems can be used to create more challenging reasoning datasets

Monotonicity reasoning, the signature of natural logic

- Polarity marking
- Word replacements according to semantic regularities
- NatLog: inspired by natural logic but not sufficiently *logical* or *robust*

References I

Bernardy, J.-P. and Chatzikyriakidis, S. (2017). A type-theoretical system for the FraCaS test suite: Grammatical framework meets coq. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Bos, J. (2008). Wide-coverage semantic analysis with boxer. In Bos, J. and Delmonte, R., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

Bos, J. and Markert, K. (2005). Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. (2018). e-snli: Natural language inference with natural language explanations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Dagan, I., Glickman, O., and Magnini, B. (2005). The pascal recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

References II

Dalvi, B., Jansen, P., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., and Clark, P. (2021). Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eijck, J. V. (2005). Syllogistics = monotonicity + symmetry + existential import.

Fyodorov, Y., Winter, Y., and Francez, N. (2003). Order-based inference in natural logic. *Logic Journal of the IGPL*, 11(4):385–416.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hemann, J., Swords, C., and Moss, L. (2015). Two advances in the implementations of extended syllogistic logics. In Balduccini1, M., Mileo, A., Ovchinnikova, E., Russo, A., and Schüuller, P., editors, *Joint Proceedings of the 2nd Workshop on Natural Language Processing and Automated Reasoning, and the 2nd International Workshop on Learning and Nonmonotonic Reasoning at LPNMR 2015*, pages 1–14.

Hu, H., Chen, Q., and Moss, L. (2019). Natural language inference with monotonicity. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 8–15, Gothenburg, Sweden. Association for Computational Linguistics.

Icard, T. F. (2012). Inclusion and exclusion in natural language. *Studia Logica*, 100(4):705–725.

Icard, T. F. and Moss, L. S. (2014). Recent progress on monotonicity. *Linguistic Issues in Language Technology*, 9.

References III

Lai, A., Bisk, Y., and Hockenmaier, J. (2017). Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Lakoff, G. (1970). Linguistics and natural logic. In Davidson, D. and Harman, G., editors, *Semantics of Natural Language*, volume 40 of *Synthese Library*, pages 545–665. Springer Netherlands.

MacCartney, B. (2009). *Natural language inference*. Phd thesis, Stanford University.

MacCartney, B. and Manning, C. D. (2007). Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 193–200, Stroudsburg, PA, USA. Association for Computational Linguistics.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In Scott, D. and Uszkoreit, H., editors, *COLING*, pages 521–528.

Mineshima, K., Martínez-Gómez, P., Miyao, Y., and Bekki, D. (2015). Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061, Lisbon, Portugal. Association for Computational Linguistics.

Moss, L. S. (2010a). Logics for natural language inference. Expanded version of lecture notes from a course at ESSLLI 2010.

Moss, L. S. (2010b). Natural logic and semantics. In Aloni, M., Bastiaanse, H., de Jager, T., and Schulz, K., editors, *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, pages 84–93. Springer Berlin Heidelberg, Berlin, Heidelberg.

References IV

Muskens, R. (2010). An analytic tableau system for natural logic. In Aloni, M., Bastiaanse, H., de Jager, T., and Schulz, K., editors, *Logic, Language and Meaning*, volume 6042 of *Lecture Notes in Computer Science*, pages 104–113. Springer Berlin Heidelberg.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Sánchez-Valencia, V. (1991). Categorial grammar and natural reasoning. ILTI Publication Series for Logic, Semantics, and Philosophy of Language LP-91-08, University of Amsterdam.

Van Benthem, J. (1986). *Essays in Logical Semantics*, volume 29 of *Studies in Linguistics and Philosophy*. Springer Netherlands.

van Benthem, J. (1987). Meaning: Interpretation and inference. *Synthese*, 73(3):451–470.

van Benthem, J. (2008). A brief history of natural logic. In *Technical Report PP-2008-05*. Institute for Logic, Language & Computation.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Zamansky, A., Francez, N., and Winter, Y. (2006). A 'natural logic' inference system using the lambek calculus. *Journal of Logic, Language and Information*, 15(3):273–295.