A Natural Proof System for Natural Language

NPS4NL-5: Natural Language Inference with Natural Theorem Prover

Lasha Abzianidze Reinhard Muskens

university of groningen

Tilburg University

ESSLLI 2019 in Rīga, Latvija

Today:

- Relevant NLI datasets:
  - FraCaS
  - SICK
- Learning phase:
  - Adaptation
  - Development
- Evaluation:
  - FraCaS
  - SICK
- Demo of LangPro
- Conclusion & future work

The SICK dataset

SICK [Marelli et al., 2014b] contains Sentences Involving Compositional Knowledge:

- 10K Text-Hypothesis pairs generated semi-automatically and annotated by humans with three labels: E, C, & N.
- Contains no encyclopedic knowledge, no named entities, relatively small vocabulary, less multiword expressions and no lengthy sentences (9 words per sentence).
- Contradictions (86%) rely too much on negative words and antonyms [Lai and Hockenmaier, 2014].
- A benchmark for the SemEval-14 RTE task [Marelli et al., 2014a]: Trial (5%), Train (45%), and test (50%).
- 84% of crowd workers' labels match the majority, i.e, gold labels.

SICK construction

Original pair	
S0a: *A sea turtle is hunting for fish*	**S0b:** *The turtle followed the fish*
Normalized pair	
S1a: *A sea turtle is hunting for fish*	**S1b:** *The turtle is following the fish*
Expanded pair	
Similar meaning	
S2a: *A sea turtle is hunting for food*	**S2b:** *The turtle is following the red fish*
Logically contradictory or at least highly contrasting meaning	
S3a: *A sea turtle is not hunting for fish*	**S3b:** *The turtle isn't following the fish*
Lexically similar but different meaning	
S4a: *A fish is hunting for a turtle in the sea*	**S4b:** *The fish is following the turtle*

Normalized sentence pairs		Score	Label
S1a: *A sea turtle is hunting for fish*	**S2a:** *A sea turtle is hunting for food*	4.5	E
S3a: *A sea turtle is not hunting for fish*	**S1a:** *A sea turtle is hunting for fish*	3.4	C
S4a: *A fish is hunting for a turtle in the sea*	**S1a:** *A sea turtle is hunting for fish*	3.9	N
S2b: *The turtle is following the red fish*	**S1b:** *The turtle is following the fish*	4.6	E
S1b: *The turtle is following the fish*	**S3b:** *The turtle isn't following the fish*	4	C
S1b: *The turtle is following the fish*	**S4b:** *The fish is following the turtle*	3.8	C
S1a: *A sea turtle is hunting for fish*	**S2b:** *The turtle is following the red fish*	4	N
S1a: *A sea turtle is hunting for fish*	**S3b:** *The turtle isn't following the fish*	3.2	N
S4b: *The fish is following the turtle*	**S1a:** *A sea turtle is hunting for fish*	3.2	N
S1b: *The turtle is following the fish*	**S2a:** *A sea turtle is hunting for food*	3.9	N
S1b: *The turtle is following the fish*	**S3a:** *A sea turtle is not hunting for fish*	3.4	N
S4a: *A fish is hunting for a turtle in the sea*	**S1b:** *The turtle is following the fish*	3.5	N
S1a: *A sea turtle is hunting for fish*	**S1b:** *The turtle is following the fish*	3.8	N

SICK examples and stats

SICK-1241 GOLD: `neutral`

A woman is dancing and singing with other women

A woman is dancing and singing in the rain

SICK-341 GOLD: `contradiction`

There is no girl with a black bag on a crowded train

A girl with a black bag is on a crowded train

SICK-8381 GOLD: `entailment`

The young girl in blue is having fun on a slide

The young girl in blue is enjoying a slide

Relatedness	neutral	contradiction	entailment	Total
[1,2) range	10%	0%	0%	10% (923)
[2,3) range	13%	1%	0%	14% (1373)
[3,4) range	28%	10%	1%	29% (3872)
[4,5] range	7%	3%	27%	37% (3672)
Total	56.86% (5595)	14.47% (1424)	28.67% (2821)	9840

The FraCaS dataset

The FraCaS test suite [Cooper et al., 1996] was an early attempt to creating a semantic benchmark for NLP systems.

- Contains 346 problems, 45% of which are multi-premised.
- Covers GQs, plurals, anaphora, ellipsis, adjectives, comparatives, temporal reference, verbs and attitudes.
- Three-way annotated by the authors of the dataset.
- Contains some ambiguous sentences and a few erroneous problems.
- Requires almost no lexical or world knowledge

Later, the FraCaS question-answer pairs where converted into an NLI format [MacCartney and Manning, 2007].

FraCaS NLI problems

FraCaS-26 GOLD: `entailment`

Most Europeans are resident in Europe
All Europeans are people
All people who are resident in Europe can travel freely within Europe
---
Most Europeans can travel freely within Europe

FraCaS-61 GOLD: `undefined`

Both female commissioners used to be in business.

Both commissioners used to be in business.

FraCaS-171 GOLD: `entailment`

John wants to know how many men work part time.
And women.

John wants to know how many women work part time.

FraCaS-87 GOLD: `entailment`

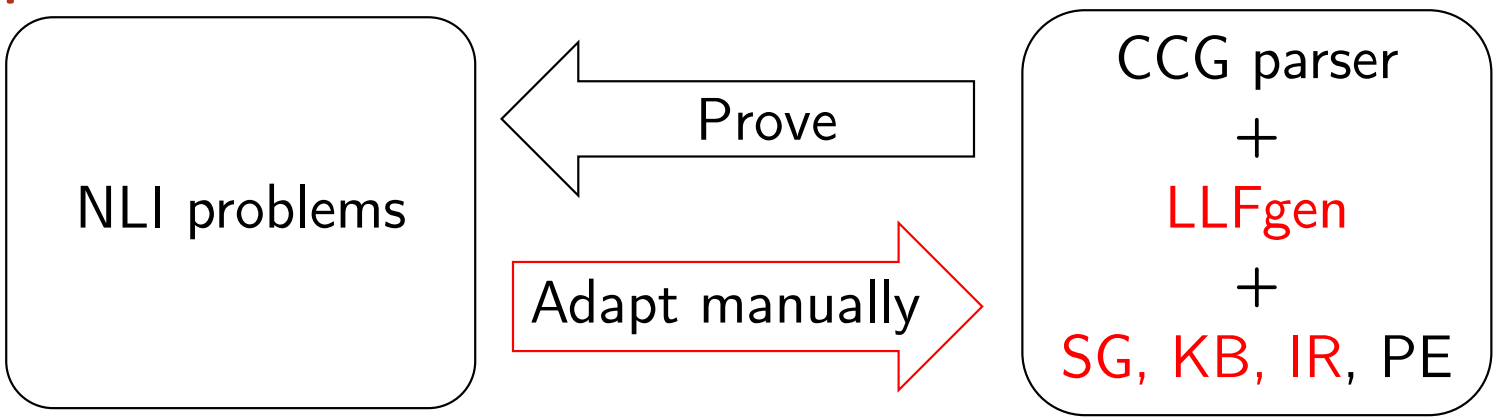
Every representative and client was at the meeting.

Every representative was at the meeting.

Learning phase

The prover LangPro is (semi-automatically) trained on the NLI datasets [Abzianidze, 2016a].

- **Adaptation**:



  Used datasets: SICK-trial and FraCaS

- **Development**:
  Finding optimal values for certain parameters of the prover based on its performance on SICK-train.

NB: Only C&C parser is used in the learning phase in order to test LangPro for an unseen parser, EasyCCG, later.

Adaptation: negative cases

We avoid fitting to the data and adopting unsound and non-general solutions.

The problems that were not solved during the adaptation:

- Sentence is not recognised as of category S or failed to be parsed
- The error is analysis is too specific to fix:

At	most	ten	commissioners	spend	time	at	home
$(S/S)/NP$	N/N	N/N	N	$(VP/PP)/NP$	N	PP/NP	N

- Lexical relation is context dependent:

SICK-4505 GOLD: `entailment`

The doctors are healing a man

The doctor is helping the patient

SICK-384 GOLD: `entailment`

A white and tan dog is running through the tall and green grass

---

A white and tan dog is running through a field

Adaptation: positive cases

The problems that were solved by upgrading one of the components of the prover:

- Treat **few** as ↓ in its 1st arg (*absolute* reading):

  FraCaS-76 GOLD: `entailment`

  Few committee members are from southern Europe

  Few female committee members are from southern Europe

- Introduce **fit** ⊑ **apply** and **food** ⊑ **meal**:

  SICK-4734 GOLD: `entailment`

  A man is fitting a silencer to a pistol

  A man is applying a silencer to a gun

  SICK-5110 GOLD: `entailment`

  A chef is preparing some food

  A chef is preparing a meal

Development phase

Optimal values of the following parameters are searched:

- The number of word senses to consider at the same time;
- The upper bound for the number of rule applications;
- Whether to use a term aligner:
  - Weak aligner aligns everything but terms of type `np`:

    SICK-1022 GOLD: `contradiction`

    A woman is wearing sunglasses of large size and is holding newspapers in both hands

    ---

    There is no woman wearing sunglasses of large size and holding newspapers in both hand

    SICK-727 GOLD: `contradiction`

    The man in a grey t-shirt is sitting on a rock in front of the waterfall

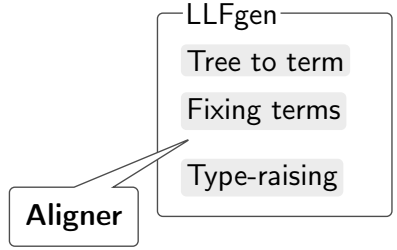
    ---

    There is no man in a grey t-shirt sitting on a rock in front of the waterfall
  - Strong aligner aligns everything but terms of type terms of type `np` with ↓arg.

    SICK-423 GOLD: `contradiction`

    Two men are not holding fishing poles

    Two men are holding fishing poles



- Efficiency criterion of tableau rules.

Efficiency criterion

Tableau rules have the following properties:

- Non-branching or branching (so called, α or β rules);
- Semantic equivalence vs proper entailment;
- Consuming (so called, γ rule) vs non-consuming;
- Producing (so called, δ rule) vs non-producing.

An example of an efficiency criterion:

$$EC = \langle \texttt{nonBr}, \texttt{semEqui}, \texttt{nonConsum}, \texttt{nonProd} \rangle$$

An efficiency vectors based on the EC efficiency criterion:

- $V_{EC}(\wedge_{\mathbb{T}}) = 1111$
- $V_{EC}(\vee_{\mathbb{T}}) = 0111$
- $V_{EC}(\exists_{\mathbb{T}}) = 1110$
- $V_{EC}(\exists_{\mathbb{F}}) = 0001$

What is the optimal efficiency criterion?

Greedy search for optimal parameters

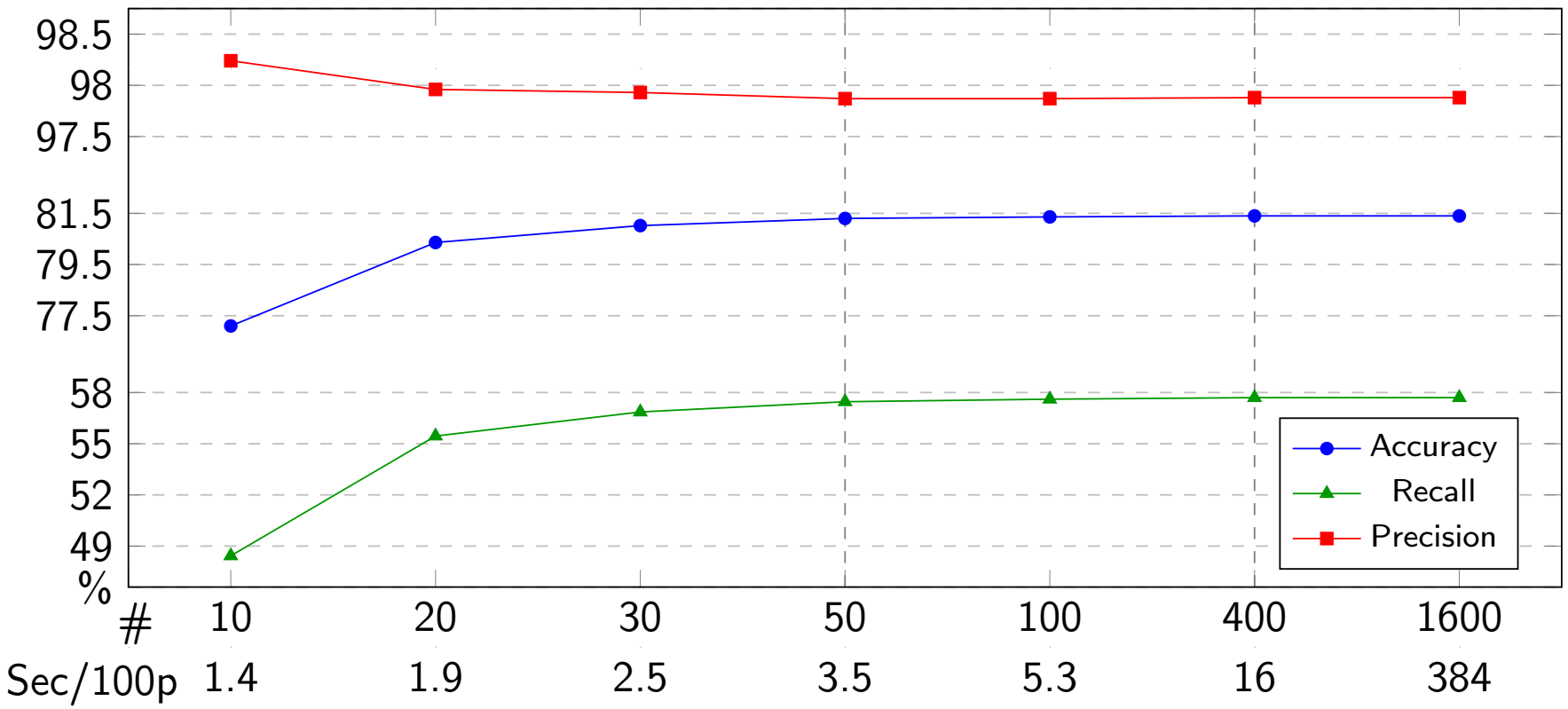
Acc%	Prec%	Rec%	Sense	Efficiency criterion	Aligner	RAL	Parser
75.09	98.5	43.6	1	[nonP,nonB,equi,nonC]	No	200	C&C
76.42	98.3	46.8	1-5	-	-	-	-
76.89	97.8	48.1	All	-	-	-	-
78.44	97.9	51.7	-	[equi,nonB,nonP,nonC]	-	-	-
79.33	97.9	53.8	-	-	Weak	-	-
81.5	97.7	59.0	-	-	Strong	-	-
81.53	97.7	59.1	-	-	Strong	400	-
81.38	98.0	58.5	-	-	Strong	400	EasyCCG
82.6	97.7	61.6	-	-	Strong	400	Both

The results are given on the SICK-train problems.

FraCaS-21 GOLD: `entailment`

The residents of member states have the right to live in Europe
All residents of member states are individuals
Every individual who has the right to live in Europe can travel freely within Europe
---
The residents of member states can travel freely within Europe

Efficient and optimal rule application numbers



The results are given on the SICK-train problems.

Solving FraCaS [Abzianidze, 2016b]

LangPro with C&C

Gold**ccLP**	yes	no	unk
yes	**51**	0	19 + 4
no	1	**14**	2
unk	1	0	**44 + 6**

P = .97, R = .71, Acc = .81

+

LangPro with EasyCCG

Gold**easyLP**	yes	no	unk
yes	**52**	0	22
no	1	**12**	4
unk	2	0	**49**

P = .96, R = .70, Acc = .80

=

LangPro

Gold**LP**	yes	no	unk
yes	**60**	0	14
no	1	**14**	2
unk	2	0	**49**

P = **.96**, R = **.81**, Acc = **.87**

FraCaS-109 GOLD: `contradiction` LP: `entailment`

Just one accountant attended the meeting

Some accountants attended the meeting

Related work (FraCaS)

[MacCartney and Manning, 2008] and [Angeli and Manning, 2014] employ a natural logic that is driven by sentence edits.

[Lewis and Steedman, 2013] employ Boxer-style [Bos et al., 2004] translation into FOL, Prover9 and distributional relation clustering.

[Mineshima et al., 2015] also uses the Boxer-style translation but some HOGQs are treated as higher-order terms. Their inference system is implemented in the proof assistant Coq.

[Tian et al., 2014] and [Dong et al., 2014] uses abstract denotations obtained from DCS trees [Liang et al., 2011]:

$$\mathbf{man} \subset \pi_{\text{subj}}\big(\mathbf{read} \cap (W_{\text{subj}} \times \mathbf{book}_{\text{obj}})\big)$$

[Bernardy and Chatzikyriakidis, 2017] uses Grammatical Framework and Coq. They use gold standard GF trees.

Comparison on FraCaS

Sec (Sing/All)	Single-premised (Acc %)							Overall (Acc %)				
	BL	NL07,08	LS P/G	NLI	T14a,b	M15	**LP**	BL	LS P/G	T14a,b	M15	**LP**
1 GQs (44/74)	45	84 **98**	70 89	95	80 93	82	93	50	62 85	80 **95**	78	**95**
2 Plur (24/33)	58	~~42~~ **75**	-	~~38~~	-	67	**75**	61	-	-	67	**73**
5 Adj (15/22)	40	60 80	-	**87**	-	**87**	**87**	41	-	-	68	**77**
9 Att (9/13)	67	~~56~~ 89	-	~~22~~	-	78	**100**	62	-	-	77	**92**
1,2,5,9 (92/142)	50	- **88**	-	-	-	78	**88**	52	-	-	74	**87**

NL07 [MacCartney and Manning, 2007], **NL08** [MacCartney and Manning, 2008], **NLI** [Angeli and Manning, 2014], **LS** [Lewis and Steedman, 2013], **M15** [Mineshima et al., 2015], **T14a** [Tian et al., 2014] and **T14b** [Dong et al., 2014]

Advantages of our approach over the related ones include:

- Reasoning (with the semantic tableau) over multiple-premises;
- Logical forms close to surface forms;
- Underlying expressive high-order logic.

Curing SICK [Abzianidze, 2015]

Gold SICK-test \ LangPro	Ent	Cont	Neut
Entailment	**805**	0	609
Contradiction	2	**482**	236
Neutral	26	7	**2760**

P=**97.4%**, R=**60.3%**, Acc=**82.14%**

Mainly the usage of WordNet and noisy gold labels are blamed for false proofs.

ID	G/LP	Premise	Conclusion
1405	N/E	A **prawn** is being cut by a woman	A woman is cutting **shrimps**
4443	N/E	A man is singing to a **girl**	A man is singing to a **woman**
2870	N/C	Two people are riding a **motorcycle**	Nobody is riding a **bike**
8913	N/C	A couple is not looking at a map	A couple is looking at a map
363	C/C	P: A soccer ball is not rolling into a goal net C: A soccer ball is rolling into a goal net	

False neutrals

Reason for false neutrals are knowledge sparsity (ca 50%), a lack of rules (ca 25%), wrong labels and parsing mistakes.

ID G/LP	Premise	Conclusion
4974 E/N	Someone is holding a hedgehog	Someone is holding a small animal
6258 E/N	P: A policeman is sitting on a **motorcycle** C: The cop is sitting on a police **bike**	
4553 E/N	P: A man is emptying a container made of plastic C: A man is emptying a plastic container	
4720 E/N	A monkey is practicing martial arts	A chimp is practicing martial arts
6447 C/N	P: [A small boy [in a yellow shirt]] is laughing on the beach C: There is no small boy [in a yellow shirt [laughing on the beach]]	

Comparison on SICK

SemEval-14 systems	Prec%	Rec%	Acc%	(+LP)	NWS%
Baseline (majority)	-	-	56.69		39.7
Illinois-LH	81.56	**81.87**	**84.57**	(+0.65)	72.8
ECNU	84.37	74.37	83.64	(+1.77)	72.7
UNAL-NLP	81.99	76.80	83.05	(+1.48)	71.2
SemantiKLUE	85.40	69.63	82.32	(+2.84)	71.5
The Meaning Factory	93.63	60.64	81.59	(+2.78)	73.0
UTexas (Prob-FOL)	**97.87**	38.71	73.23	(+9.44)	62.5
LangPro	**97.35**	60.31	82.14		**74.8**

RTE systems	Acc%
Prob-FOL	76.52
Prob-FOL*+Rules	**85.10**
Nutcracker+PPDB	79.60
ABCNN-3	**86.20**
LSTM RNN+SNLI	80.80

Gold\System	E	C	N
Entailment	2	−2	0
Contradiction	−2	2	0
Neutral	−1	−1	1

"Hard" problems

The problems from SICK-test that were proved correctly by both ccLangPro and easyLangPro but failed by all the top five systems at the SemEval-14 task.

ID	G	Text	Hypothesis
247	C	T: The woman is not wearing glasses or a headdress H: A woman is wearing an Egyptian headdress	
406	E	T: A group of scouts are hiking through the grass H: People are walking	
2895	C	The man isn't lifting weights	The man is lifting barbells
3527	E	T: A person is jotting something with a pencil H: A person is writing	
3570	C	The piece of paper is not being cut	Paper is being cut with scissors
3608	N	T: A monkey is riding a bike H: A bike is being ridden over a monkey	
3806	E	A man in a hat is playing a harp	A man is playing an instrument
4479	E	The boy is playing the piano	The boy is playing a musical instrument

Introducing a new tableau rule

Let us add a new rule to Natural Tableau and LangPro:

We want introduce a rule in order to account for the entailment:

GOLD: `entailment`

Most women are working

Most women are rich

There is a woman who is working and is rich

This rule will help:

MOST2

$[\vec{M}_1] : \textbf{most}\ N_{\mathrm{n}}\ W_{\mathrm{vp}} : [] : \mathbb{T}$

$[\vec{M}_2] : \textbf{most}\ N_{\mathrm{n}}\ R_{\mathrm{vp}} : [] : \mathbb{T}$

$N_{\mathrm{n}} : [c_e] : \mathbb{T}$

$[\vec{M}_1] : W_{\mathrm{vp}} : [c_e] : \mathbb{T}$

$[\vec{M}_2] : R_{\mathrm{vp}} : [c_e] : \mathbb{T}$

c_e is fresh and $W \neq R$

Conclusion

Natural Tableau is a wide-coverage but still logic-based reasoning system inspired by Natural Logic.

It represents a proof-theoretic approach to NLI.

Natural tableau was successfully scaled up for the NLI task:

CCG parser + LLFgen + theorem prover

Pros and cons of Natural Tableau:

- ✔ Employs higher-order logic to model linguistic semantics;
- ✔ Allows deep logical and shallow (e.g. monotonicity) reasoning;
- ✔ Getting logical form is similar to syntactic parsing;
- ✘ Heavily hinges on CCG parsing;
- ✔ Proofs are highly reliable ($\leq 3\%$ false proofs);
- ✘ Suffers from multi-sense words;
- ✘ No fully automated learning from data yet;
- ✔ Its decision procedure is transparent and explanatory;

Future work

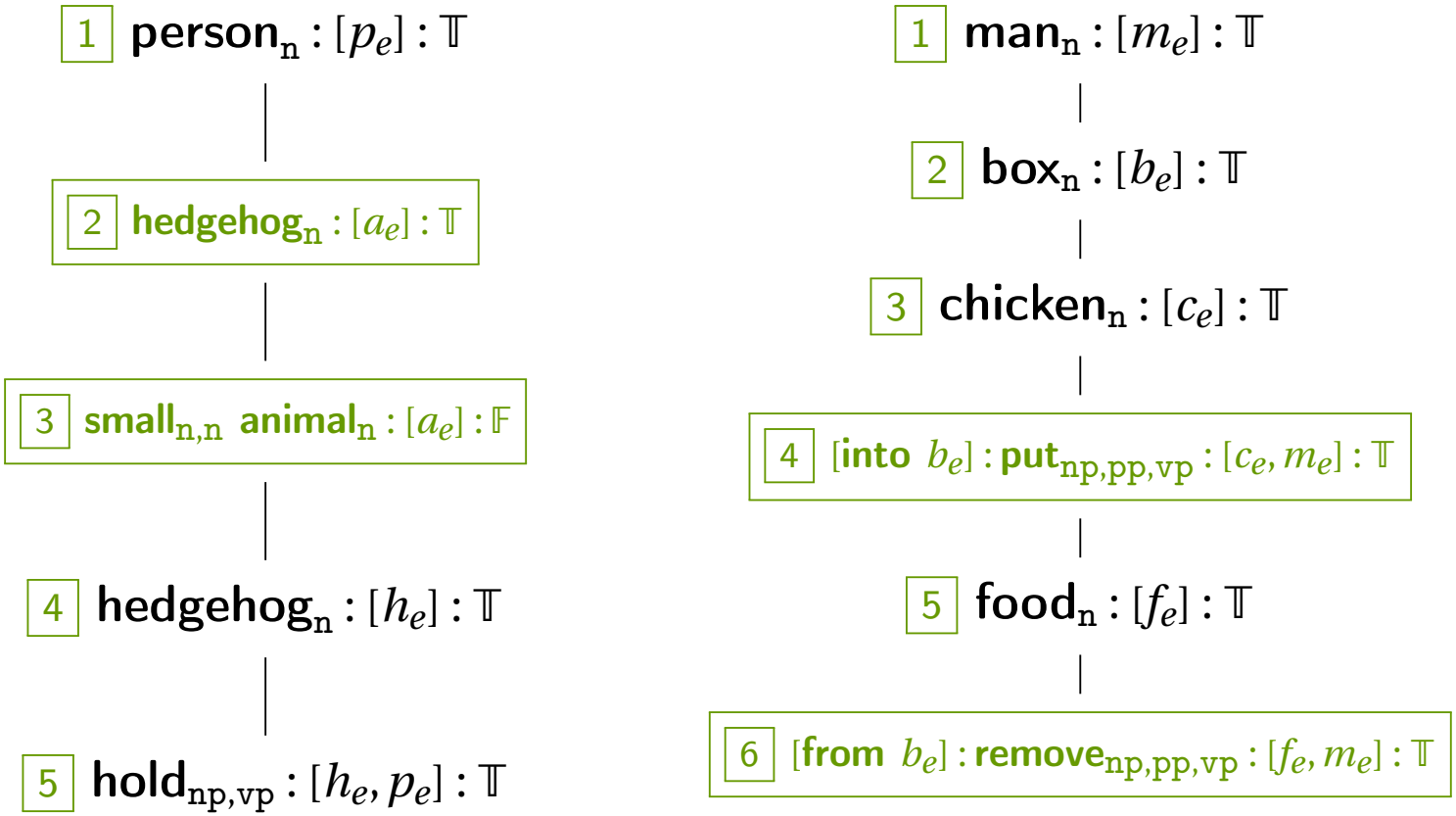
There are really many directions for future work:

- Explore different types of RTE data, e.g., the newswire or human generated data [Bowman et al., 2015];
- Incorporate more knowledge in KB, e.g., paraphrase database [Ganitkevitch et al., 2013].
- Model different phenomena: comparatives, anaphora, cardinals, etc.
- Pairing with distributional semantics: $R(w_1, w_2, r)$ and weighted closure branches;
- Acquisition of lexical knowledge: abductive reasoning;
- Generate LLFs from Universal Dependency trees

  \+ the Universal Semantic Tagging [?]

  $\rightarrow^{?}$ Multilingual Natural Tableau

Inference to the best explanation



Thank you

Thank you for coming here in the early mornings and listening me repeating tableau, tableau, tableau, . . . , tableau!

References I


Abzianidze, L. (2015). A tableau prover for natural logic and language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502, Lisbon, Portugal. Association for Computational Linguistics.


Abzianidze, L. (2016a). *A natural proof system for natural language*. PhD thesis, Tilburg University.


Abzianidze, L. (2016b). Natural solution to fracas entailment problems. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 64–74, Berlin, Germany. Association for Computational Linguistics.


Angeli, G. and Manning, C. D. (2014). Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.


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., Clark, S., Steedman, M., Curran, J. R., and Hockenmaier, J. (2004). Wide-coverage semantic representations from a ccg parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, pages 1240–1246, Geneva, Switzerland.


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.

References II


Cooper, R., Crouch, D., Eijck, J. V., Fox, C., Genabith, J. V., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. (1996). *FraCaS: A Framework for Computational Semantics*. Deliverable D16.


Dong, Y., Tian, R., and Miyao, Y. (2014). Encoding generalized quantifiers in dependency-based compositional semantics. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation*, pages 585–594, Phuket,Thailand. Department of Linguistics, Chulalongkorn University.


Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.


Lai, A. and Hockenmaier, J. (2014). Illinois-lh: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.


Lewis, M. and Steedman, M. (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics (TACL)*, 1:179–192.


Liang, P., Jordan, M. I., and Klein, D. (2011). Learning dependency-based compositional semantics. In *Association for Computational Linguistics (ACL)*, pages 590–599.


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.

References III

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.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014a). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of SemEval 2014 (International Workshop on Semantic Evaluation)*, pages 1–8, East Stroudsburg PA. ACL.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014b). A sick cure for the evaluation of compositional distributional semantic models. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Tian, R., Miyao, Y., and Matsuzaki, T. (2014). Logical inference on dependency-based compositional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 79–89, Baltimore, Maryland. Association for Computational Linguistics.