Natural Language Inference

НИС «Компьютерная лингвистика»

14 апреля 2020

Работы по теме

- Основные работы: Bowman et al. 2015; Williams et al. 2018; Nie et al. 2019; Rocktäschel et al. 2016.

- Дополнительные статьи: Goldberg 2015; Dagan et al. 2006;MacCartney and Manning 2008 (входит в наш список, читаем к этому занятию!); Gururangan et al. 2018

NLI ставит задачей выявление (логических) отношений между словами (например, turtle -- linguist) и пропозициями (например, some students smoke – some students smoke cigars). В самом простом случае, отношения могут сводиться к следствию (entailment), противоречию (contradiction), отсутствию какого-либо отношения (neutral).

| Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
turtle	contradicts	linguist
Every reptile danced.	neutral	A turtle ate.
Some turtles walk.	contradicts	No turtles move.
James Byron Dean refused to move without blue jeans.	entails	James Dean didn't dance without pants.
Mitsubishi Motors Corp's new vehicle sales in the US fell 46 percent in June.	contradicts	Mitsubishi's sales rose 46 percent.
Acme Corporation reported that its CEO resigned.	entails	Acme's CEO resigned.

Формулировка задач NLI

- Можно ли вывести из посылки (premise) гипотезу (hypothesis)?
  - Опора на здравый смысл, а не на формальную логику
  - Фокус на локальных выводах, а не на цепочках логических выводов
  - Упор на разнообразие лингвистических способов выражения логических отношений
- Теоретические вопросы:
  - Zaenen et al. (2005): Local textual inference: can it bedefined or circumscribed?
  - Manning (2006): Local textual inference: it's hard to circumscribe, but you know it when you see it – and NLP needs it.
  - Crouch et al. (2006): Circumscribing is not excluding: a reply to Manning.

Решение задач NLI связано с решением целого ряда других задач:

Task	NLI framing
Paraphrase	text ≡ paraphrase
Summarization	text ⊏ summary
Information retrieval	query ⊐ document
Question answering	question ⊐ answer
	Who left? ⇒ *Someone left*
	Someone left ⊐ *Sandy left*

Модели NLI
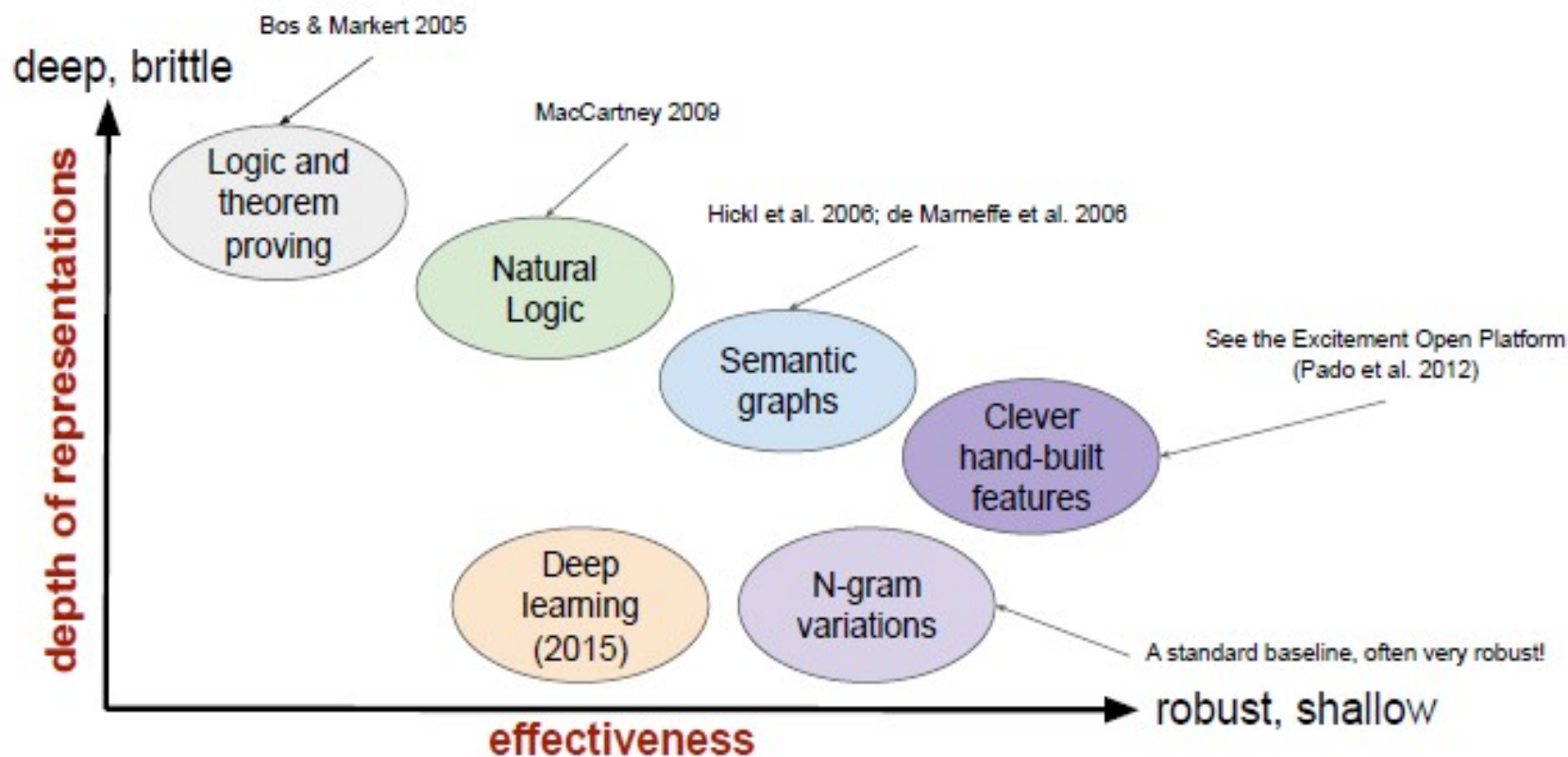
Дейтасеты

- The GLUE benchmark (diverse tasks including NLI)
  https://gluebenchmark.com
- NLI Style FEVER
  https://github.com/easonnie/combine-FEVER-NSMN/blob/master/
  other_resources/nli_fever.md
- MedNLI (derived from MIMIC III)
  https://physionet.org/physiotools/mimic-code/mednli/
- XNLI is a multilingual NLI dataset derived from MultiNLI
  https://github.com/facebookresearch/XNLI
- Diverse Natural Language Inference Collection (DNC)
  http://decomp.io/projects/diverse-natural-language-inference/
- SciTail (derived from science exam questions and Web text)
  http://data.allenai.org/scitail/
- SemEval 2013
  https://www.cs.york.ac.uk/semeval-2013/
- SemEval 2014: Sentences Involving Compositional Knowledge (SICK)
  http://alt.qcri.org/semeval2014/task1/index.php?id=data-and-tools
- The FraCaS textual inference test suite
  https://nlp.stanford.edu/~wcmac/downloads/

Обычно пространство отношений разбивается на две (следствие, не-следствие), три (следствие, не-следствие, противоречие), или четыре (тождественность, включение, обратное включение, не-следствие) категории. Какой из дизайнов кажется вам наиболее перспективным? Что вы думаете по поводу расширения наименований отношений, например, наподобие инвентаря лексических функций?

	couch / sofa	crow / bird	bird / crow	hippo / hungry	turtle / linguist
	Yes entailment		**No** non-entailment		
	Yes entailment		**Unknown** non-entailment		**No** contradiction
	P ≡ Q equivalence	**P ⊑ Q** forward	**P ⊒ Q** reverse	**P # Q** non-entailment	

SNLI

1. Bowman et al. 2015
2. Все посылки (premises) подписи из корпуса Flickr30K (Young et al. 2014).
3. Все гипотезы (hypotheses) получены с помощью краудсорсинга.
4. Некоторые предложения отражают стереотипы (Rudinger et al. 2017).
5. 550,152 train examples; 10K dev; 10K test
6. Средняя длина в токенах :
   Посылка (premise): 14.1
   Гипотеза (hypothesis): 8.3
7. Типы клауз:
   Посылка в главном предложении: 74%
   Гипотеза в главном предложении: 88.9%
8. Размер словаря: 37,026
9. 56,951 примеров были аннотированы дополнительными аннотаторами:
   58.3% примеров получили ожидаемый лейбл отношения (gold label)
   91.2% ожидаемых лейблов (gold labels) совпадают с ожидаемыми лейблами автора
   0.70 общий индекс согласия между аннотаторами (Fleiss kappa https://en.wikipedia.org/wiki/Fleiss%27_kappa)

10. https://nlp.stanford.edu/projects/snli/

Как проходил краудсорсинг

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** an **false** description of the photo.

Photo caption A little boy in an apron helps his mother cook.

Definitely correct   Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

> Write a sentence that follows from the given caption.

Maybe correct   Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

> Write a sentence which may be true given the caption, and may not be.

Definitely incorrect   Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."*

> Write a sentence which contradicts the caption.

Problems (optional)   *If something is wrong with the caption that makes it difficult to understand, do your best above and let us know here.*

Примеры

Как вы думаете, за счёт чего во втором и в последнем случае возникает разногласие между аннотаторами? [Не забывайте про один из принципов (слайд 4): фокус на локальных выводах, а не на цепочках логических выводов]
Какой лейбл выбрали бы вы?

Premise	Relation	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	**contradiction** c c c c c	The man is sleeping
An older and younger man smiling.	**neutral** n n e n n	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** c c c c c	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** e e e e e	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** n n e c n	A happy woman in a fairy costume holds an umbrella.

Прогресс SNLI

про F1 score: https://en.wikipedia.org/wiki/F1_score

SNLI leaderboard: Systems over time
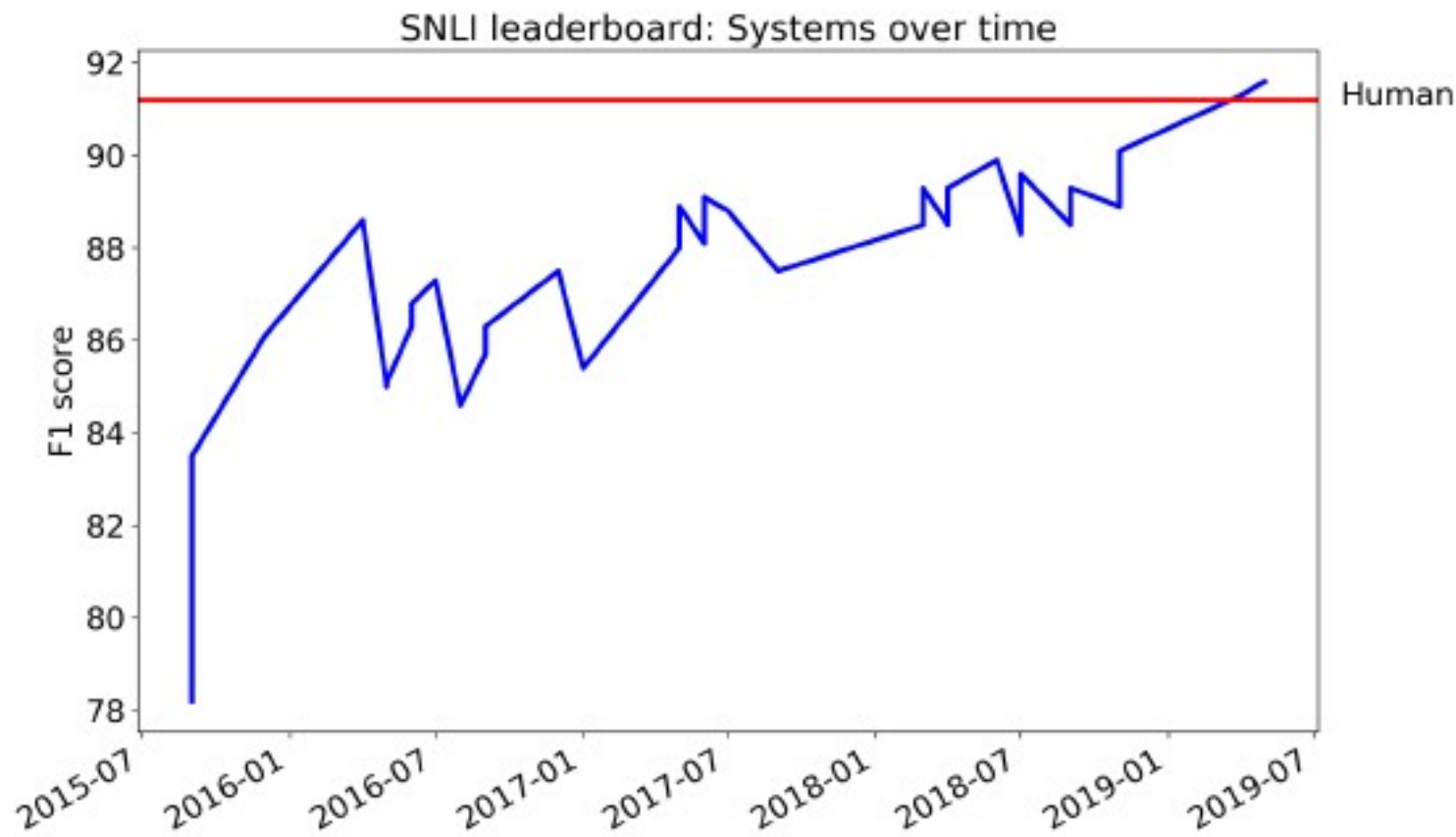
MultiNLI

1. Williams et al. 2018

2. Train premises drawn from five genres:

   Fiction: works from 1912–2010 spanning many genres

   Government: reports, letters, speeches, etc., from government websites

   The *Slate* website

   Telephone: the Switchboard corpus

   Travel: Berlitz travel guides

3. Additional genres just for dev and test (the mismatched condition):

   The 9/11 report

   Face-to-face: The Charlotte Narrative and Conversation Collection

   Fundraising letters

   Non-fiction from Oxford University Press

   Verbatim: articles about linguistics

4. 392,702 train examples; 20K dev; 20K test

5. 19,647 examples validated by four additional annotators

   58.2% examples with unanimous gold label

   92.6% of gold labels match the author's label

6. Project page: https://www.nyu.edu/projects/bowman/multinli/

MultiNLI разметка

	Matched	Mismatched
ACTIVE/PASSIVE	15	10
ANTO	17	20
BELIEF	66	58
CONDITIONAL	23	26
COREF	30	29
LONG_SENTENCE	99	109
MODAL	144	126
NEGATION	129	104
PARAPHRASE	25	37
QUANTIFIER	125	140
QUANTITY/TIME_REASONING	15	39
TENSE_DIFFERENCE	51	18
WORD_OVERLAP	28	37
	767	753

Примеры

Annotations	Premise	Relation	Hypothesis
#MODAL, #COREF	Students of human misery can savor its underlying sadness and futility. entailment	entailment	Those who study human misery will savor the sadness and futility.
#NEGATION, #TENSE_DIFFERENCE, #CONDITIONAL	oh really it wouldn't matter if we plant them when it was starting to get warmer	contradiction	It is better to plant when it is colder.
#QUANTIFIER, #AC-TIVE/PASSIVE	They consolidated programs to increase efficiency and deploy resources more effectively	entailment	Programs to increase efficiency were consolidated.

Прогресс MultiNLI
про F1 score: https://en.wikipedia.org/wiki/F1_score
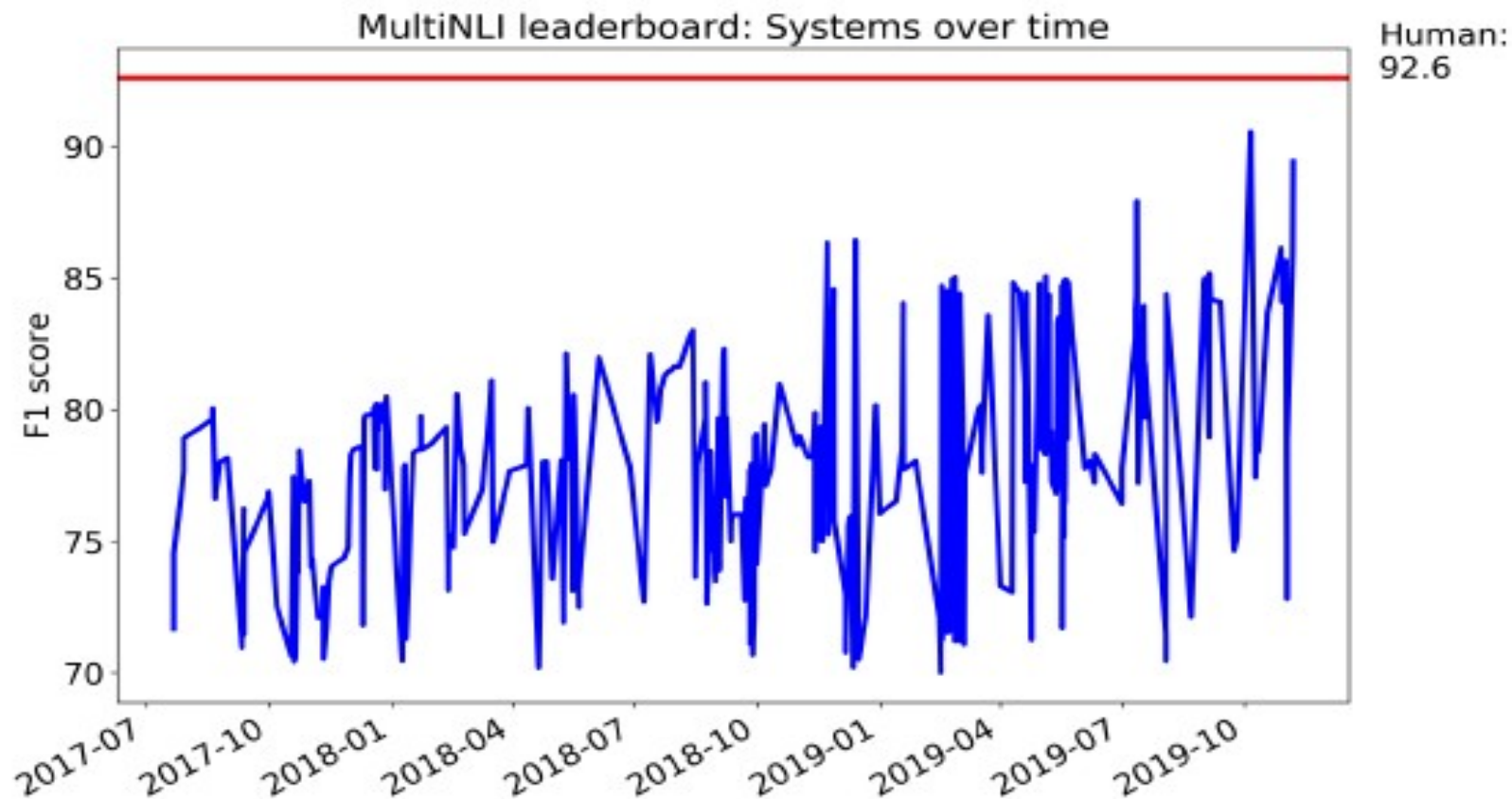
- Естественный шаг: дополнить SNLI и MultiNLI с помощью WordNet
- Но о WordNet на следующей неделе…

Мини-дз

- Дедлайн: 21 апреля (вторник) в 20.59
- Работу можно оформить как текстовый файл, как тетрадку .ipynb...
- Работу просим отправить на daria.ryzhova@mail.ru и daschapopowa@gmail.com
- Задание:

1) предложите список отношений для русского языка и приведите хотя бы один пример на каждое отношение (например, **посылка**: *Мой слон любит коньячные торты* **отношение**: следствие **гипотеза**: *Мой слон любит торты*). Включите ли Вы дополнительную разметку на кореференцию, модальность? Расширится ли у Вас список отношений (добавятся ли пресуппозиция, тавтология, например)? – 6 баллов (то есть предполагается минимум три отношения и три соответствующих примера)

2) обоснуйте свой выбор отношений – 2 балла

3) предположите (и обоснуйте своё предположение), какие из предложенных Вами отношений будут вызывать наибольшие проблемы у людей-аннотаторов, а какие у SNLI/MultiNLI – 2 балла

Литература по теме

Johan Bos and Katja Markert. 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, Stroudsburg, PA. ACL.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, Stroudsburg, PA. Association for Computational Linguistics.

Richard Crouch, Lauri Karttunen, and Annie Zaenen. 2006. Circumscribing is not excluding: A reply to Manning. Ms., Palo Alto Research Center.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Lecture Notes in Computer Science*, volume 3944, pages 177–190. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. Ms., Bar Ilan University.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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.

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ROBERTa: A robustly optimized BERT pretraining approach. ArXiv:1907.11692.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Bill MacCartney. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Литература по теме

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Christopher D. Manning. 2006. Local textual inference: It's hard to circumscribe, but you know it when you see it – and NLP needs it. Ms., Stanford University.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D Manning. 2006. Learning to distinguish valid textual entailments. In *Proceedings of the 2nd Pascal RTE Challenge Workshop*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. UNC CHapel Hill and Facebook AI Research.

Sebastian Pado, Tae-Gil Noh, Asher Stern, and Rui Wang. 2013. Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering.*, 21(2):167–200.

Adam Poliak, Jason Naradowsky, Aparajita Haidar, Rachel Rudinger, and Benjamin Van Durme. 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.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kočiský, and Phil Plunsom. 2016. Reasoning about entailment with neural attention. ArXiv:1509.06664.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *Proceedings of ICLR 2015*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Литература по теме

Annie Zaenen, Lauri Karttunen, and Richard 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*, pages 31–36, Ann Arbor, MI. Association for Computational Linguistics.