

# To Infer or not to Infer? Natural Language Inference and Computational Semantics

Stergios Chatzikyriakidis  
Centre for Linguistic Theory and Studies in Probability  
(CLASP),  
FLoV



May 9, 2019

- 1 Intro
- 2 Three Eras of NLI
- 3 Brittleness and Generalizability
- 4 Where Do we Go From Here?

- Roughly: the task of determining whether a NL hypothesis H follows from an NL premise P
  - 'inferential ability is not only a central manifestation of semantic competence but is in fact centrally constitutive of it' Cooper et al. (1996)
  - inferential ability as a means to test the semantic adequacy of NLP systems

# What Humans Infer?

- Reasoning: part of our every day routine:
  - we hear Natural Language (NL) sentences
  - we participate in dialogues
  - we read books or legal documents.
- Successfully understanding, participating or communicating with others in these situations presupposes some form of reasoning
  - about individual sentences
  - whole paragraphs of legal documents
  - small or bigger pieces of dialogue

- The variety of reasoning is difficult to be explained by a single coherent system of reasoning. *Why?*
  - because reasoning is performed in different ways in each one of them

Consider the following example:

- (1) Three representatives are needed.

(2) Three representatives are needed.

- Assume a human reasoner with expert knowledge in a legal context:
  - s/he will most probably judge that a situation where more than three representatives are provided could be compatible with the semantics of the utterance
- The same reasoner interpreting the utterance as part of a casual, everyday conversation:
  - *three* would most likely be interpreted as *exactly three*

- Reasoning can get very complicated as soon as we move to a dialogue setting:
  - A. Mont Blanc is higher than
  - B. Mt. Ararat?
  - A. Yes.
  - B. No, this is not correct. It is the other way around.
  - A. Are you...
  - B. Sure? Yes, I am.
  - A. Ok, then.

# Inference can be Incremental and Interactive

- The listener reasons based on:
  - utterances that are split between two participants
    - thus having to dynamically keep track of them.
- Furthermore, the listener must be able to compute:
  - global inferences, i.e. inferences that are based on statements/facts that are shared (agreed upon) by the dialogue participants
    - local inferences that are based on facts that are not shared by all dialogue participants.



# Can machines take it?

- The grand picture, the mother of all tasks: create computational semantics systems that will reflect this wealth of reasoning patterns
  - Start small: systems that perform reasoning on well-defined (well...) subsets of what NLI is
    - Define the task
    - Provide the system
    - Evaluate the system

- 1 Intro
- 2 Three Eras of NLI**
- 3 Brittleness and Generalizability
- 4 Where Do we Go From Here?

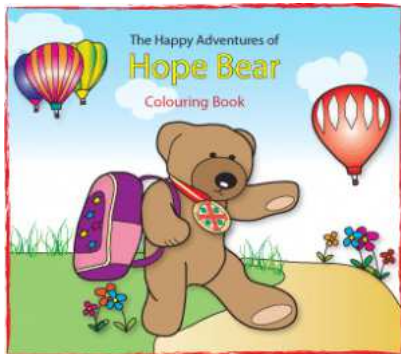
- The Symbolic Era
  - What we need is a solid formal system and good rules formalized in that system
  - We just have to find these good rules
    - It should be possible, no?

# The Three Eras of NLI

- The Symbolic Era
  - What we need is a solid formal system and good rules formalized in that system
  - We just have to find these good rules
    - It should be possible, no?
- Sure! Let us do this!

# The Three Eras of NLI

- The Symbolic Era
  - What we need is a solid formal system and good rules formalized in that system
  - We just have to find these good rules
    - It should be possible, no?
- Sure! Let us do this!



- The FraCas test suite Cooper et al. (1996)
  - a collection of mostly logical entailments. Categorization is done according to semantic category
  - Three way classification of 346 inference problems: YES (the conclusion follows), NO (the negation of the conclusion follows) and UNK (none of the two follow)

- The FraCas test suite Cooper et al. (1996)
  - a collection of mostly logical entailments. Categorization is done according to semantic category
  - Three way classification of 346 inference problems: YES (the conclusion follows), NO (the negation of the conclusion follows) and UNK (none of the two follow)

- (3) A Swede won the Nobel Prize.  
Every Swede is Scandinavian.  
Did a Scandinavian win the Nobel prize? [Yes, FraCas 049]
  
- (4) No delegate finished the report on time..  
Did any Scandinavian delegate finish the report on time?  
[No, FraCaS 070]



(5) A Scandinavian won the Nobel Prize.

Every Swede is Scandinavian.

Did a Swede win the Nobel prize? [UNK, FraCaS 065]

- Other typical examples

(6) Either Smith, Jones or Anderson signed the contract.

Did John sign the contract? [UNK] (plurals, FraCaS 083)

(7) Dumbo is a large animal. Is Dumbo a small animal?

[NO] (adjectives, FraCaS 205)

(8) Smith believed that ITEL had won the contract in

1992. Did ITEL win the contract in 1992? [UNK]

(Attitudes, FraCaS 334)

- Pair a symbolic syntactic parser with logical semantics
  - Define a correspondence between abstract syntax (usually abstract syntactic trees) and semantics in some logical language

- Does extremely well in controlled domains
  - Very precise and fine-grained

- Does extremely well in controlled domains
  - Very precise and fine-grained



- Breaks down when thrown into open text or non-controlled text



- Common problem with symbolic approaches
  - This is the reason they have been largely abandoned in modern day AI or even dubbed totally useless by some contemporary AI researchers

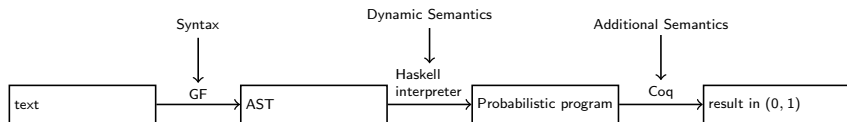
- Common problem with symbolic approaches
  - This is the reason they have been largely abandoned in modern day AI or even dubbed totally useless by some contemporary AI researchers



- Bernardy and Chatzikyriakidis (2019)
  - Building on earlier work Bernardy and Chatzikyriakidis (2017)
- In a nutshell:
  - A converter from syntax trees to types.
    - GF syntax trees to Coq types via a Haskell Program, using Monadic Dynamic Semantics
  - Type-theoretical combinators for the treatment of various linguistic phenomena (e.g. adjectives)



- DFraCoq's architecture



- Covers more of the FraCaS than any previous account (around 80%)
  - First run on the ellipsis and anaphora section
  - Overall accuracy of around 89.2%

- Comparison of DFraCoq with previous logical approaches

Section	#cases	Ours	FC	MINE	Nut	Langpro
Quantifiers	75	.96 (74)	.96	.77	.53	.93 (44)
Plurals	33	.82	.76	.67	.52	.73 (24)
Anaphora	28	.86	-	-	-	-
Ellipsis	52	.87	-	-	-	-
Adjectives	22	.95 (20)	.95	.68	.32	.73 (12)
Comparatives	31	.87	.56	.48	.45	-
Temporal	75	-	-	-	-	-
Verbs	8	.75	-	-	-	-
Attitudes	13	.92	.85	.77	.46	.92 (9)
Total	337	.89 (259)	.83 (174)	.69 (174)	.50 (174)	.85 (89)

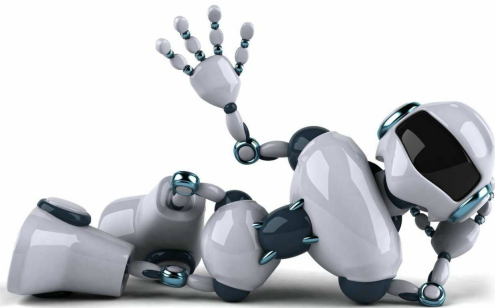
- High precision, but low recall when thrown into open text
  - Brittle systems

# The Classic Machine Learning Era

- Late 90s, all the way to approximately 2013-2014
  - Supervised Machine Learning: using hand-crafted features to train the models
    - Bag-of-Words, similarity metric based models, Maximum Entropy Classifiers, SVMs, Naive Bayes... the usual suspects
    - Enough with Logic, this will definitely work

# The Classic Machine Learning Era

- Late 90s, all the way to approximately 2013-2014
  - Supervised Machine Learning: using hand-crafted features to train the models
    - Bag-of-Words, similarity metric based models, Maximum Entropy Classifiers, SVMs, Naive Bayes... the usual suspects
    - Enough with Logic, this will definitely work



- The PASCAL Recognizing Textual Entailment Challenges (RTE)

- Platform appeared in 2005 Dagan et al. (2006)



- Platform appeared in 2005 Dagan et al. (2006)
  - New version every year until 2011 (RTE7)

- Platform appeared in 2005 Dagan et al. (2006)
  - New version every year until 2011 (RTE7)
  - RTE concentrates on real rather than constructed data

- Platform appeared in 2005 Dagan et al. (2006)
  - New version every year until 2011 (RTE7)
  - RTE concentrates on real rather than constructed data
  - A natural piece of text is taken as a premise and then a hypothesis is constructed out of it

- Platform appeared in 2005 Dagan et al. (2006)
  - New version every year until 2011 (RTE7)
  - RTE concentrates on real rather than constructed data
  - A natural piece of text is taken as a premise and then a hypothesis is constructed out of it
    - The first RTEs had two way entailment (entailment, non-entailment), three way classification was added in the later suites

- Platform appeared in 2005 Dagan et al. (2006)
  - New version every year until 2011 (RTE7)
  - RTE concentrates on real rather than constructed data
  - A natural piece of text is taken as a premise and then a hypothesis is constructed out of it
    - The first RTEs had two way entailment (entailment, non-entailment), three way classification was added in the later suites

- P Budapest again became the focus of national political drama in the late 1980s, when Hungary led the reform movement in eastern Europe that broke the communist monopoly on political power and ushered in the possibility of multiparty politics.
- H In the late 1980s Budapest became the center of the reform movement. [follows, RTE702]

- P Budapest again became the focus of national political drama in the late 1980s, when Hungary led the reform movement in eastern Europe that broke the communist monopoly on political power and ushered in the possibility of multiparty politics.
- H In the late 1980s Budapest became the center of the reform movement. [follows, RTE702]

- P Budapest again became the focus of national political drama in the late 1980s, when Hungary led the reform movement in eastern Europe that broke the communist monopoly on political power and ushered in the possibility of multiparty politics.
- H In the late 1980s Budapest became the center of the reform movement. [follows, RTE702]
- P Like the United States, U.N. officials are also dismayed that Aristide killed a conference called by Prime Minister Robert Malval in Port-au-Prince in hopes of bringing all the feuding parties together.
- H U.N. officials take part in a conference called by Prime Minister Robert Malval. (does not follow, RTE1933)



- Performance on the RTE datasets has been in general poor
  - Around 60% accuracy for many years and throughout the challenges

# The Deep Learning Era

- The dominant (almost the only one) nowadays
  - Some sort of Artificial Neural Network model is used to deal with NLI
  - ANNs are great learners, they can approximate any function given enough data (oversimplified, for a precise definition, check Universal Approximation Theorem)
  - But enough is most of the times a lot!
    - And quality might not be great

# The Deep Learning Era

- The dominant (almost the only one) nowadays
  - Some sort of Artificial Neural Network model is used to deal with NLI
  - ANNs are great learners, they can approximate any function given enough data (oversimplified, for a precise definition, check Universal Approximation Theorem)
  - But enough is most of the times a lot!
    - And quality might not be great



- New datasets are needed
  - Both FraCaS and the RTE datasets (also SICK) are not suitable on dataset size alone

- Developed at Stanford by Bowman et al. (2015)

- Developed at Stanford by Bowman et al. (2015)
  - created using crowdsourcing (mechanical turk)

- Developed at Stanford by Bowman et al. (2015)
  - created using crowdsourcing (mechanical turk)
  - the subjects are given the caption of a picture and are asked to provide:

- Developed at Stanford by Bowman et al. (2015)
  - created using crowdsourcing (mechanical turk)
  - the subjects are given the caption of a picture and are asked to provide:
    - an alternate true caption
    - an alternate possibly true caption
    - an alternate false caption



- Developed at Stanford by Bowman et al. (2015)
  - created using crowdsourcing (mechanical turk)
  - the subjects are given the caption of a picture and are asked to provide:
    - an alternate true caption
    - an alternate possibly true caption
    - an alternate false caption

- Instructions used on Mechanical Turk

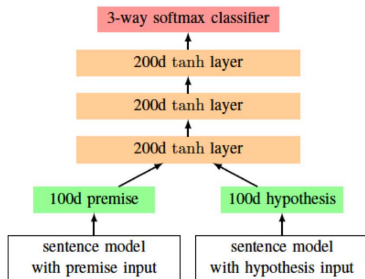
We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely a true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be a true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely a false** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the maybe correct category because it's impossible for the dogs to be both running and sitting.*

- The State of the Art at this moment on the SNLI dataset is above 90% accuracy
  - Actually, the State of the Art has been surpassed tenths of times, since the first paper using NNs for SNLI (Bowman et al., 2015)

# The original approach

- Simple architecture
  - three 200d tanh layers plus a bottom layer
    - input: concatenated sentence representations
    - output: softmax classifier (3-way)
    - Achieves an accuracy close to 78%



- Tenths of systems since then, slowly advancing the SoA by proposing different NN architectures and network tweaks
- Where we are now: the BERT era (Bidirectional Encoder Representations from Transformers)
  - What it does?
    - Transfer learning: transfer knowledge obtained from one task to another (e.g. semantic similarity to NLI)
    - Pre-training: pre-train the model on two prediction tasks (one is e.g. predict the next sentence)
    - Fine-tuning for other tasks: no training from scratch is needed, just fine-tuning of parameters

- Impressive in a wide variety of tasks (basically all of the GLUE benchmarks)

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>91.1</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>81.9</b>

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The “Average” column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); BERT<sub>BASE</sub> = (L=12, H=768, A=12); BERT<sub>LARGE</sub> = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from <https://gluebenchmark.com/leaderboard> and <https://blog.openai.com/language-unsupervised/>.

- Current SoTA: BERT + Semantic Role Labelling information (Zhang et al., 2019)

- 1 Intro
- 2 Three Eras of NLI
- 3 Brittleness and Generalizability**
- 4 Where Do we Go From Here?

# Brittleness through the back door?

- At first sight, NN systems do not seem to be suffering from the brittleness problem
  - Only to some extent correct
  - Recent studies seem to propose that NN systems are also brittle, but in another sense
    - They fail to generalize outside individual datasets and are, furthermore, unable to capture certain NLI patterns, at all



# Brittleness through the back door?

- NLI systems have limited generalization ability outside the datasets that they are trained and tested on Glockner et al. (2018)
  - NLI systems break easily when, instead of being tested on the original SNLI test set, they are tested on a test set which contain sentences that differ by at most one word from sentences in the training set
    - Significant drop in accuracy, e.g. between 22 and 33 points when trained on SNLI and tested on the new dataset, is reported for three out of four state-of-the-art systems tested.
    - The system less prone to breaking is Kim et al. (2018) (5 points drop when trained on SNLI and tested on the new dataset), which utilizes external knowledge taken from WordNet Miller (1995).

# Brittleness through the back door?

- NLI systems have limited generalization ability outside the datasets that they are trained and tested on (Talman and Chatzikyriakidis, 2018)
  - Train and test six state-of-the-art NN models using train and test sets drawn from a different corpus
    - E.g. the train set is drawn from the SNLI but the test from the MultiNLI, vice versa and other similar combinations
  - The results shows an average drop of 24.9 points in accuracy for all systems, including the system by Kim et al. (2018)

# How our NLI systems are doing: brittleness through the backdoor

Train	Dev	Test	Test Accuracy	Delta	Model
<b>SNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>86.1</b>		600D BiLSTM-max
<b>SNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>86.6</b>		600D HBMP Talman et al. (2018)
<b>SNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>88.0</b>		600D ESIM Chen et al. (2017)
<b>SNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>88.6</b>		300D KIM Kim et al. (2018)
SNLI	SNLI	MultiNLI-m	55.7*	-30.4	600D BiLSTM-max
SNLI	SNLI	MultiNLI-m	56.3*	-30.3	600D HBMP
SNLI	SNLI	MultiNLI-m	59.2*	-28.8	600D ESIM
SNLI	SNLI	MultiNLI-m	61.7*	-26.9	300D KIM
SNLI	SNLI	SICK	54.5	-31.6	600D BiLSTM-max
SNLI	SNLI	SICK	53.1	-33.5	600D HBMP
SNLI	SNLI	SICK	54.3	-33.7	600D ESIM
SNLI	SNLI	SICK	55.8	-32.8	300D KIM
<b>MultiNLI</b>	<b>MultiNLI-m</b>	<b>MultiNLI-m</b>	<b>73.1*</b>		<b>600D BiLSTM-max</b>
<b>MultiNLI</b>	<b>MultiNLI-m</b>	<b>MultiNLI-m</b>	<b>73.2*</b>		<b>600D HBMP</b>
<b>MultiNLI</b>	<b>MultiNLI-m</b>	<b>MultiNLI-m</b>	<b>76.8*</b>		<b>600D ESIM</b>
<b>MultiNLI</b>	<b>MultiNLI-m</b>	<b>MultiNLI-m</b>	<b>77.3*</b>		<b>300D KIM</b>
MultiNLI	MultiNLI-m	SNLI	63.8	-9.3	600D BiLSTM-max
MultiNLI	MultiNLI-m	SNLI	65.3	-7.9	600D HBMP
MultiNLI	MultiNLI-m	SNLI	66.4	-10.4	600D ESIM
MultiNLI	MultiNLI-m	SNLI	68.5	-8.8	300D KIM
MultiNLI	MultiNLI-m	SICK	54.1	-19.0	600D BiLSTM-max
MultiNLI	MultiNLI-m	SICK	54.1	-19.1	600D HBMP
MultiNLI	MultiNLI-m	SICK	47.9	-28.9	600D ESIM
MultiNLI	MultiNLI-m	SICK	50.9	-26.4	300D KIM
<b>SNLI+MultiNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>86.1</b>		<b>600D BiLSTM-max</b>
<b>SNLI+MultiNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>86.1</b>		<b>600D HBMP</b>
<b>SNLI+MultiNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>87.5</b>		<b>600D ESIM</b>
<b>SNLI+MultiNLI</b>	<b>SNLI</b>	<b>SNLI</b>	<b>86.2</b>		<b>300D KIM</b>
SNLI+MultiNLI	SNLI	SICK	54.5	-31.6	600D BiLSTM-max
SNLI+MultiNLI	SNLI	SICK	55.0	-31.1	600D HBMP
SNLI+MultiNLI	SNLI	SICK	54.5	-33.0	600D ESIM
SNLI+MultiNLI	SNLI	SICK	54.6	-31.6	300D KIM

- Test accuracies (%). For results highlighted in bold the training data include examples from the same corpus as the test data. For the other cases, the training and test data involve separate corpora.

# The issue of generalizability

- BERT seems to be doing much better than the other models in our examples
  - Did not exist when the other negative results papers appeared
- Still when moved from more similar to less similar datasets (this is to be expected, of course, though the drop is huge)

- 1 Intro
- 2 Three Eras of NLI
- 3 Brittleness and Generalizability
- 4 Where Do we Go From Here?**

# Where do we go from here?

- Keep all research directions open
  - There seems to be still use for research directions that were deemed useless in the field
- Attempt integration
  - Hybrid approaches
    - Symbolic + DL: On what level, what are the gains if at all
  - Get a clear idea of what approach works well with what, but most importantly, what approach DOES NOT work well with what

# Where do we go from here? Better datasets


- Reflect in a more accurate way the range of inference patterns found in human reasoning with NL
  - Most importantly: we need datasets that can test NLP systems on reasoning with dialogue
- Probabilistic datasets?
  - NLI datasets that give probability scores for an inference instead of a three way classification
  - Needed in order to test probabilistic semantics systems like the ones currently developed at CLASP (e.g. Cooper et al. (2015); Bernardy et al. (2018))

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. A type-theoretical system for the fracas test suite: Grammatical framework meets coq. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*, 2017.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. A wide-coverage symbolic natural language inference system. In *Submitted*, 2019.

Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, and Shalom Lappin. A compositional bayesian semantics for natural language. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 1–10, 2018.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In 



*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. Association for Computational Linguistics, 2017.

R. Cooper, D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman. Using the framework. *Technical Report LRE 62-051r*, 1996. <http://www.cogsci.ed.ac.uk/fracas/>.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. Probabilistic type theory and natural language semantics. *LiLT (Linguistic Issues in Language Technology)*, 10, 2015.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer, 2006.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*, 2018.

Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak.

Semantic sentence matching with densely-connected recurrent and co-attentive information. *arXiv preprint arXiv:1805.11360*, 2018.

George A Miller. Wordnet: a lexical database for english.

*Communications of the ACM*, 38(11):39–41, 1995.

Aarne Talman and Stergios Chatzikyriakidis. Testing the generalization power of neural network models across nli benchmarks. *arXiv preprint arXiv:1810.09774*, 2018.

Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. Natural language inference with hierarchical bilstm max pooling architecture. *arXiv preprint arXiv:1808.08762*, 2018.

Zhuosheng Zhang, Yuwei Wu, Zuchao Li, Shexia He, Hai Zhao, Xi Zhou, and Xiang Zhou. I know what you want: Semantic learning for text comprehension. *arXiv preprint arXiv:1809.02794*, 2019.