Information Extraction From Text

MLMU Prague :: May 18

Our Vision for Text

Make Machines Comprehend Text (Practically)

Information Extraction: **Meaning** Extraction > *Fact* Extraction

Focus On Practical Problems

Natural Interface to This Technology:
Question Answering

Comprehend Text?

Read sentences to solve tasks that require some degree of understanding

Testbed: Semantic Sentence Pair Scoring

Long-term Goal: Common Text Comprehension ML Model

Because **two** sentences are much more fun than one, and their match is a semantically rich structure.

Semantic Sentence Pair Scoring

The boy is sitting near the blue ocean.
The boy is swimming in the sea.

Common Model: *convert sentences to numbers*
Specific Model: *classify the number vector*
Paraphrase: 0.41 **Topic Similarity:** 0.86

- *Academic Benchmarks*
- FAQ Search
- Answer Extraction
- Dialog Traversal
- Hypothesis Evidencing

Practical Tasks

FAQ Search - eClub FAQ

What are the working hours?

Do I need to come in person?
When do I need to come to work?
Can I work at night?

Is someone going to help me with the project?

Do I need to work on the project alone?
Will someone mentor me?
How much guidance will I get?

Vision: Google-like text box on each support page instead of wading through questions.
Related: StackOverflow-based personal assistant

Answer Extraction - live.ailao.eu

Who discovered prions?

✓ Prusiner won a Nobel Prize last year for discovering prions.

✗ Researchers, writing in the October issue of the journal Nature Medicine, describe a relatively simple way to detect prions.

1. Select Answering Sentence on Wikipedia
2. Suggest the answering portion of this sentence

Live: http://live.ailao.eu/

Dialog Traversal - Text Chat Tech. Support

Context	Response	Flag
well, can I move the drives? __EOS__ ah not like that	I guess I could just get an enclosure and copy via USB	1
well, can I move the drives? __EOS__ ah not like that	you can use "ps ax" and "kill (PID #)"	0

Context
""any apache hax around ? i just deleted all of __path__ - which package provides it ?", "reconfiguring apache do n't solve it ?"

Ranked Responses	Flag
1. "does n't seem to, no"	1
2. "you can log in but not transfer files ?"	0

1. Start a dialog using a database of 1M past dialogs
2. Select the best continuation for every reply based on past dialogs
3. If not sure, hand over to humans

Vision: Next generation website chat assistants, L1 support

Hypothesis Evaluation - Prediction Markets (Augur)

Did the Blackhawks win the Stanley Cup in 2015?

yes, rel=0% Chicago Blackhawks back in Stanley Cup final after Game 7 win over Ducks.

yes, rel=100% The Blackhawks were anointed Stanley Cup champions when the buzzer sounded after a 2-0 win in Game 6

1. Search newspapers for related articles
2. For each sentence, decide relevancy and yes/no direction
3. Aggregate evidence and declare the result

Live: **Argus** system http://argus.ailao.eu/

Vision: Automated Research, School Exam Solving

Text Comprehension Models

Text Comprehension Models

Grand Scheme:
1. Convert sentences to number vectors
2. Compare the number vectors to produce the pair score

Prerequisite:
Convert *words* to number vectors - word2vec, GloVe

Alternative: Information Retrieval offers strong baselines (tf-idf, BM25 word overlaps)

Converting Sentences to Numbers

- Average Word Vectors with (semi-)deep classifier on top
  Deep Averaging Networks

- Recurrent Neural Networks

- Convolutional Neural Networks

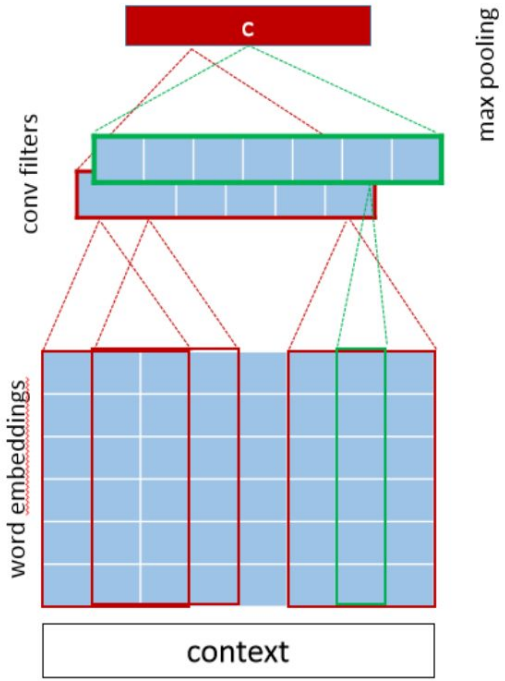
- Networks with Attention

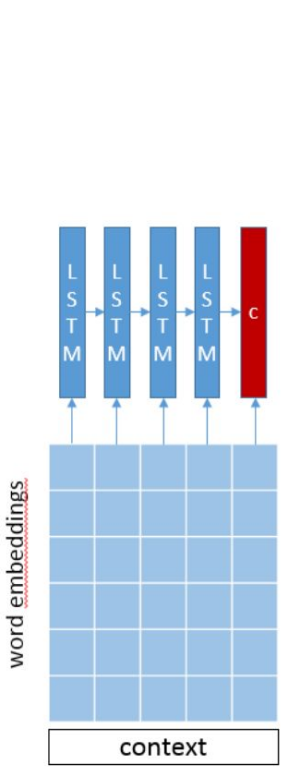
Neural Networks for Sentences

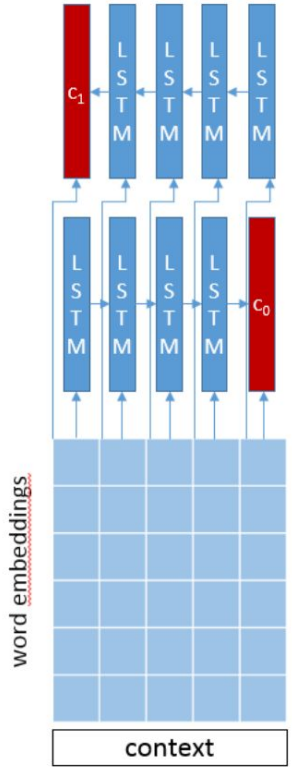
CNN

RNN with memory
(GRU, LSTM)

Bidirectional
recurrence







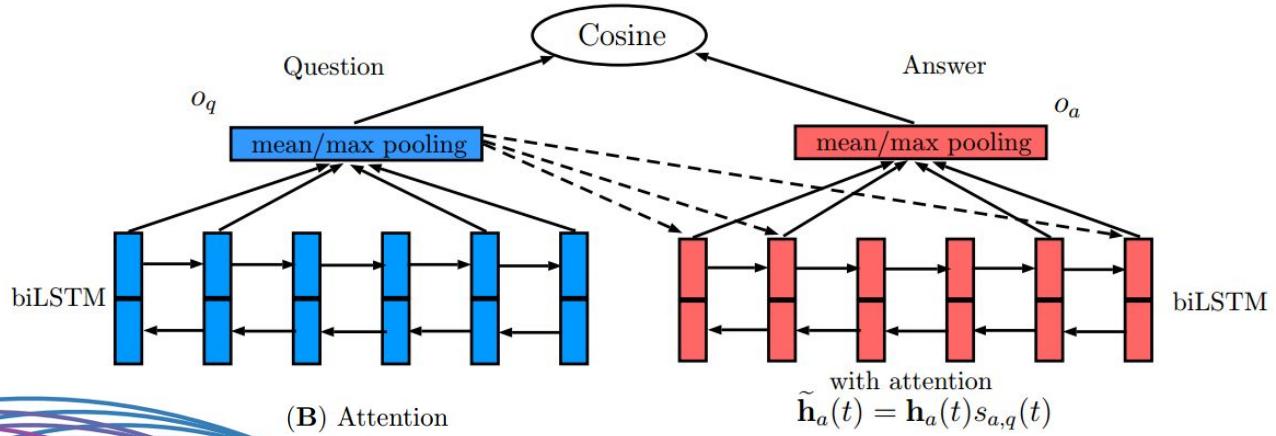
(Kadlec, 1510.03753)

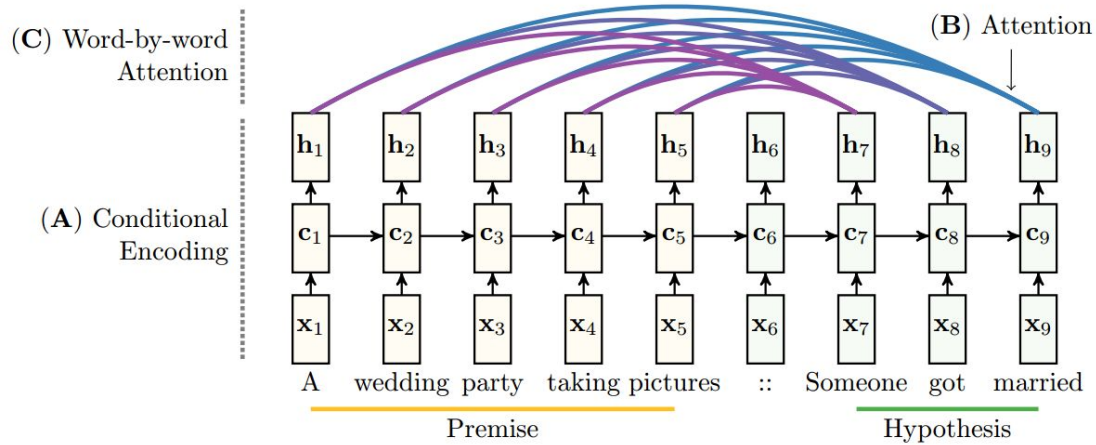
Neural Models with Attention

Idea: It is hard to fit all nuances to fixed set of numbers, what about adapting the numbers to what we need to know?

-0.101	1	What is Rohm and Haas 's annual revenue ?	Rohm and Haas , with $ billion in annual sales , makes chemicals found in such products as decorative and industrial paints , semiconductors and shampoos .
-0.677	1	What is Rohm and Haas 's annual revenue ?	The deal is the latest in a series of recent acquisitions by Rohm and Haas , a Philadelphia -based manufacturer of chemicals found in products including paints , semiconductors and shampoos , with $ billion in annual
-7.507	0	What is Rohm and Haas 's annual revenue ?	The transaction announced today creates a global specialty chemicals company with combined annual revenues of $ billion .

Neural Models with Attention





(Tan, 1511.04108),
(Rocktäschel, 1509.06664)

Universal Text Comprehension?

What about our goal to build a universal mode for text comprehension?

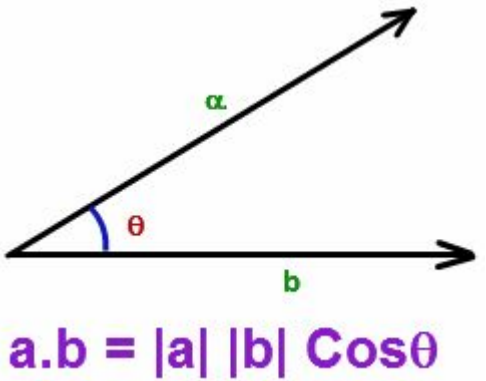
Result: RNN model trained on huge dataset (Ubuntu Dialogue) can be transferred to other tasks and almost always wins.

Scoring the Pairs

Two number vectors come in, one number comes out

Idea #1: Measure vector distance (e.g. angle)

Idea #2: Use machine learning to weigh per-dimension distances



$$h_\times = h_L \odot h_R,$$
$$h_+ = |h_L - h_R|,$$
$$h_s = \sigma\left(W^{(\times)} h_\times + W^{(+)} h_+ + b^{(h)}\right)$$

Some Recent Results

Recent Papers

github.com
brmson/dataset-sts

Python, Theano, Keras

P. Baudis, J. Pichl,
T. Veselý, J. Šedivý:
Sentence Pair Scoring: Towards Unified Framework for Text Comprehension,
arXiv 1603.06127

P. Baudis, S. Stanko,
J. Šedivý:
Joint Learning of Sentence Embeddings for Relevance and Entailment,
arXiv 1605.04655

Ubuntu Dialogue Corpus

Model	MRR	1-2 R@1	1-10 R@1	1-10 R@2	1-10 R@5
* TF-IDF		0.749	0.488	0.587	0.763
* RNN		0.777	0.379	0.561	0.836
* LSTM		0.869	0.552	0.721	0.924
* MemN2N 3-hop			0.637		
avg	0.624 ±0.002	0.793 ±0.002	0.472 ±0.002	0.608 ±0.002	0.836 ±0.003
DAN	0.578 ±0.070	0.792 ±0.035	0.493 ±0.074	0.615 ±0.059	0.830 ±0.033
RNN	0.781 ±0.003	0.907 ±0.002	0.664 ±0.004	0.799 ±0.004	0.951 ±0.001
CNN	0.718 ±0.003	0.863 ±0.002	0.587 ±0.004	0.721 ±0.005	0.907 ±0.003
RNN-CNN	**0.788** ±0.001	**0.911** ±0.001	**0.672** ±0.002	**0.809** ±0.002	**0.956** ±0.001
attn1511	0.772 ±0.004	0.903 ±0.002	0.653 ±0.005	0.788 ±0.005	0.945 ±0.002

YodaQA

YodaQA: Our full-fledged QA pipeline system
Can use text (with IE models above) or knowledge bases
http://ailao.eu/yodaqa

Full-text version and *Movies* version

Answer Sentence Selection: Our models *roughly* state-of-art

For Scientists: Report confidence intervals!

Hypothesis Evaluation

Argus: 74.4% word avg, 82.3% RNN, 85.4% universal model

MCTest:

The road to Grandpa's house was long and winding. [...] Jimmy liked to collect insects on the way to his Grandpa's house, so had picked the longer path. As he went along, Jimmy found more and more insects to add to his jar. [...]. Finally, Jimmy arrived at Grandpa's house and knocked. Grandpa answered the door with a smile and welcomed Jimmy inside. They sat by the fire and talked about the insects. They watched the lightning bugs light up as night came.

1: multiple: Why did Grandpa answer the door?
A) Because he saw the insects
B) Because Jimmy was walking
*C) Because Jimmy knocked
D) Because the trip took a long time

2: one: Where do Jimmy and his Grandpa sit?
A) On insects
B) Outside
*C) By the fire
D) On the path

Model	MC-160 one	MC-160 all	MC-500 one	MC-500 all
hand-crafted	0.842	0.753	0.721	0.699
Attn. Reader	0.481	0.463	0.444	0.419
Neur. Reasoner	0.484	0.476	0.457	0.456
HABCNN-TE	0.633	**0.631**	0.542	0.529
avg	0.653 ±0.027	0.556 ±0.012	0.587 ±0.018	0.542 ±0.011
DAN	0.681 ±0.017	0.577 ±0.010	0.636 ±0.013	**0.560** ±0.007
RNN	0.583 ±0.033	0.533 ±0.020	0.539 ±0.016	0.494 ±0.012
Universal	**0.736** ±0.033	0.612 ±0.023	**0.641** ±0.017	0.538 ±0.015

Thanks for Your Attention

Jan Šedivý's 3C Group
@ FEL ČVUT

eClub Prague
http://eclubprague.com/

Ailao Labs
http://ailao.eu/