Similarity, Attention and Memory in Semantic Sentence-Pair Scoring

Petr Baudiš ⟨pasky@ucw.cz⟩

FEE CTU Prague; Ailao

Feb 2016

Outline

Motivation

Semantic Sentence-Pair Scoring

The increase reflects lower credit losses and favorable interest rates.
The gain came as a result of fewer credit losses and lower interest rates.

The boy is sitting near the blue ocean.
The boy is wading through the blue ocean.

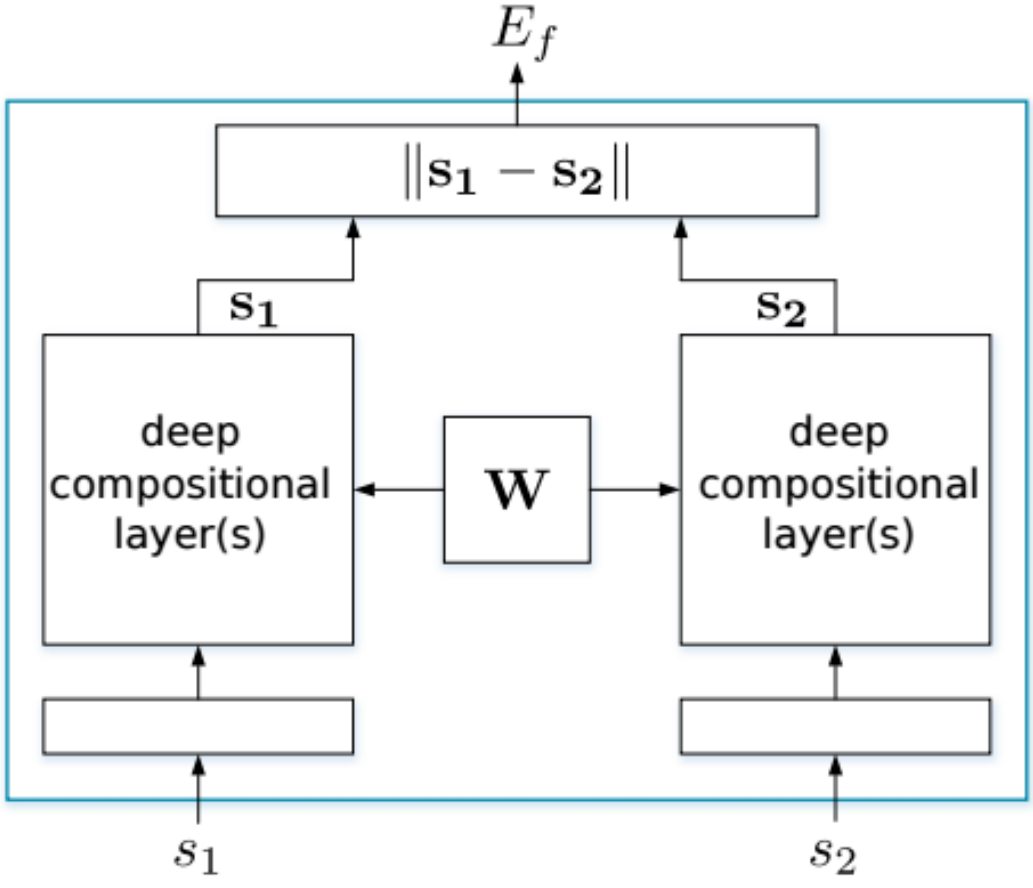
Who discovered prions?
Prusiner won a Nobel Prize last year for discovering prions.

Semantic Sentence-Pair Scoring

- Up to now: Very little cross-talk across tasks!
- Parsing+alignment methods were king before 2014
- Recently, research in CNN, RNN and attention mechanisms.
- Benchmarks for aggregate embedding methods.

Deep Learning for Sentence-Pair Scoring

(A) aggregate sentence embeddings, (B) comparison function



(Cheng and Kartsaklis, 2015)

Outline

Tasks for Sentence-Pair Scoring

- **Paraphrasing:** binary classification of independent text pairs regarding whether they say the same thing
  Ex.: Equivalent wires / headlines; duplicate questions; ...

  ✓ The increase reflects lower credit losses and favorable interest rates. — The gain came as a result of fewer credit losses and lower interest rates.

  ✗ The boy is sitting near the blue ocean.
  The boy is wading through the blue ocean.

- **Semantic Text Similarity:** similarity score of independent pairs (typically short sentences) on the scale from 0 to 5
  3: I don't like flavored beers. — I dislike many flavored drinks.

Recognizing Textual Entailment

Entailment: independent pairs of "fact" and "hypothesis" sentences as *entailment*, *contradiction* or *unknown*

ENT: A soccer game with multiple males playing.
→ Some men are playing a sport.

CONTR: A black race car starts up in front of a crowd of people.
→ A man is driving down a lonely road.

NEU: An older and younger man smiling.
→ Two men are smiling and laughing at the cats playing on the floor.

Answer Sentence Selection

Answer Sentence Selection: *bipartite ranking* — binary classification, but there many "answers" for the same "question", we sort positive answers above negative answers

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.

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

Answer Sentence Selection

Memory Networks (bAbI):

Mary went to the bathroom. John is in the playground.
John moved to the hallway. John picked up the football.
Mary travelled to the office. Bob went to the kitchen.
Where is Mary? A:office

Hypothesis Evidencing

Hypothesis Evidencing: Original task — like *anssel*, but scoring the question alone (typically true / false).

Argus Headline Yes/No Dataset:

Will Tiger Woods win the 2015 Masters championship? (false)

- "In a nine-hole stretch, he may make six or seven." Woods could not claim the Masters finish needed to secure a place in the WGC-Cadillac Match Play Championship in San Francisco later this month."
- Tiger Woods still the biggest star at Masters – despite being 111th in rankings.

Will the Chicago Blackhawks win the 2015 Stanley Cup? (true)

- Chicago Blackhawks back in Stanley Cup final after Game 7 win over Ducks
- The Chicago Blackhawks won their sixth Stanley Cup on Monday.

Hypothesis Evidencing

Kaggle Science Test Dataset:

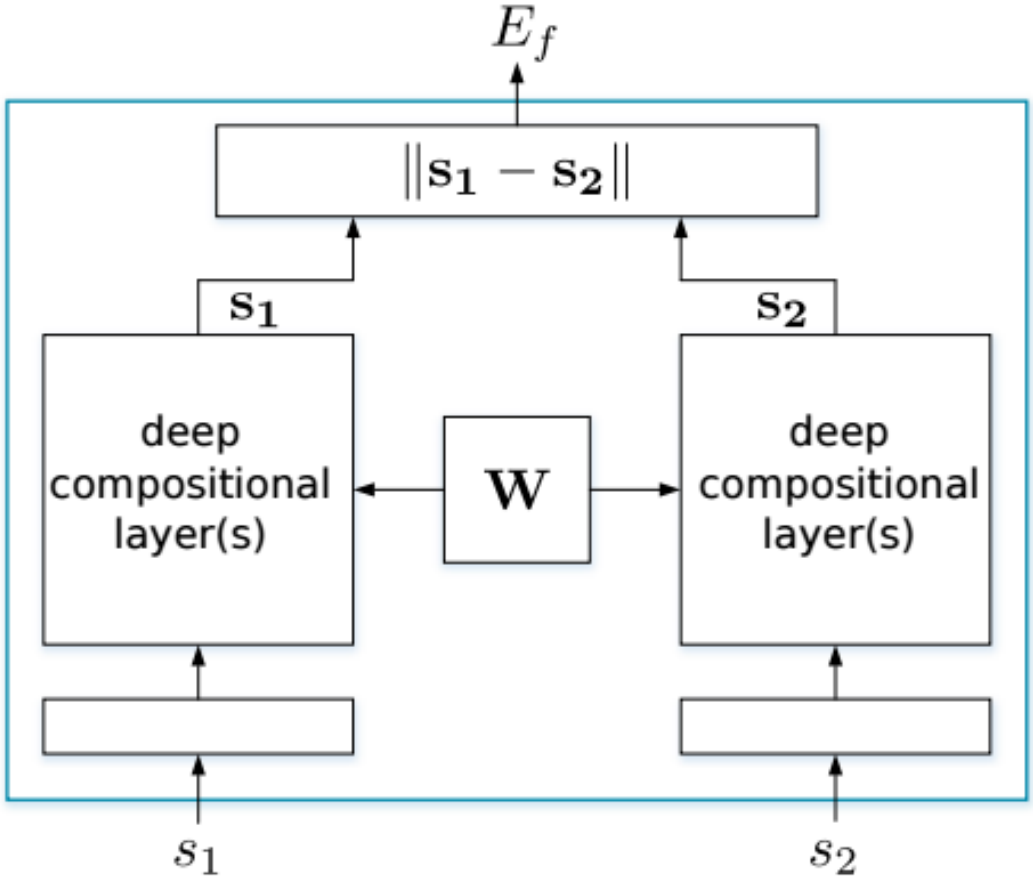
the Sun is the main source of energy for the water cycle (true)

- The Sun heats some areas more than others, which causes wind.
- The Sun's energy also drives the water cycle, which moves water over the surface of Earth.

Outline

Deep Learning for Sentence-Pair Scoring

(A) aggregate sentence embeddings, (B) comparison function

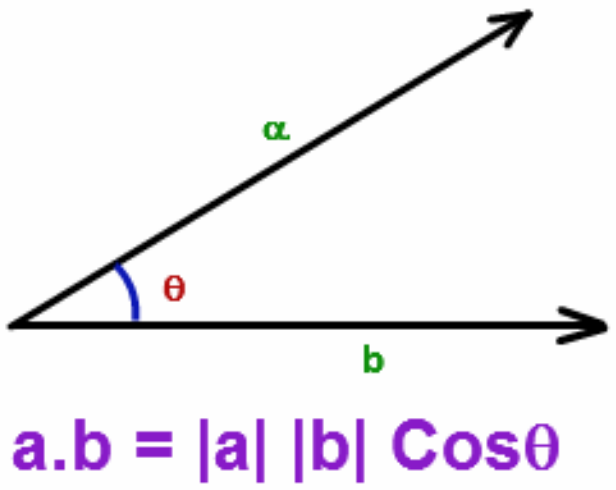


(Cheng and Kartsaklis, 2015)

Comparing Aggregates

- **Input:** A pair of aggregate embeddings
- **Output:** Single embedding representing the whole sentence

- Cos-similarity (angle in vector space)
- Dot-product (non-normalized cos-similarity)
- Elementwise difference and product as input to an MLP *(Tai, 1503.00075)*

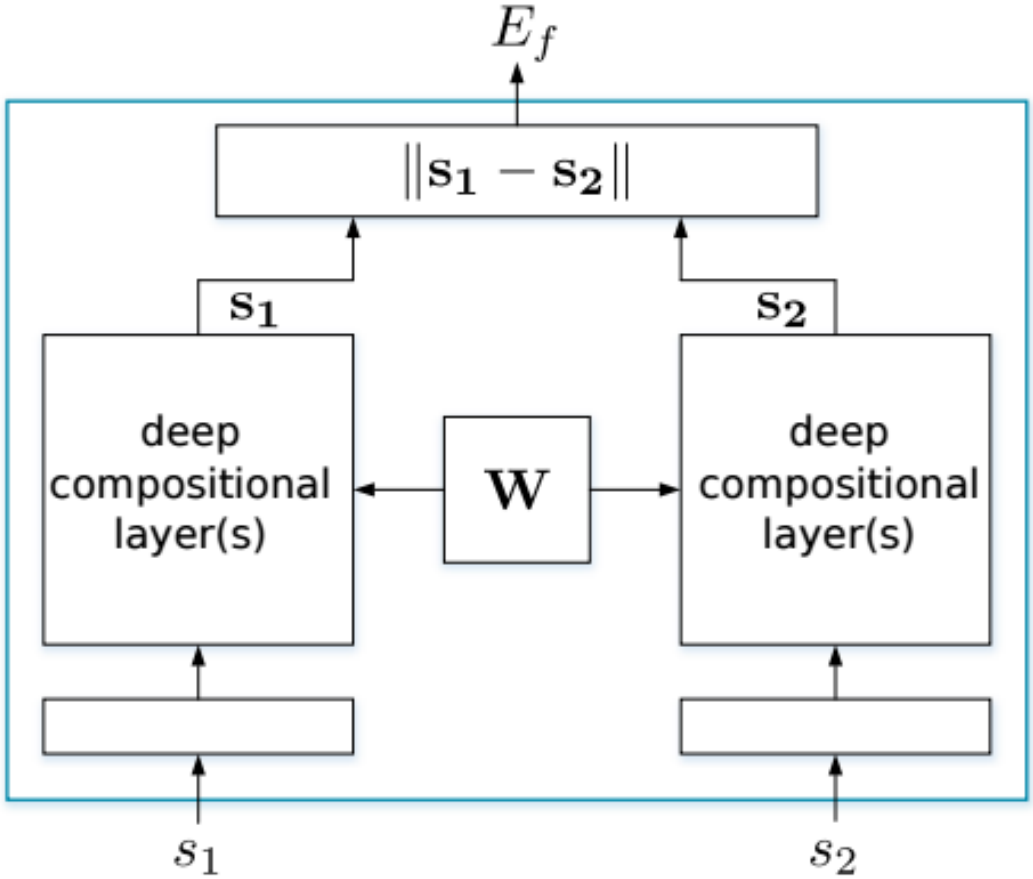


$$h_\times = h_L \odot h_R,$$
$$h_+ = |h_L - h_R|,$$
$$h_s = \sigma\left(W^{(\times)} h_\times + W^{(+)} h_+ + b^{(h)}\right)$$

Outline

Deep Learning for Sentence-Pair Scoring

(A) aggregate sentence embeddings, (B) comparison function



(Cheng and Kartsaklis, 2015)

Aggregate Embeddings

- **Input:** Sequence of word-level embeddings (word2vec, GloVe)
- **Output:** Single embedding representing the whole sentence

- Average and project
- RNN (LSTM, GRU), ReNN, CNN
- Attention-based models
- Other models (skip-thoughts, etc.)

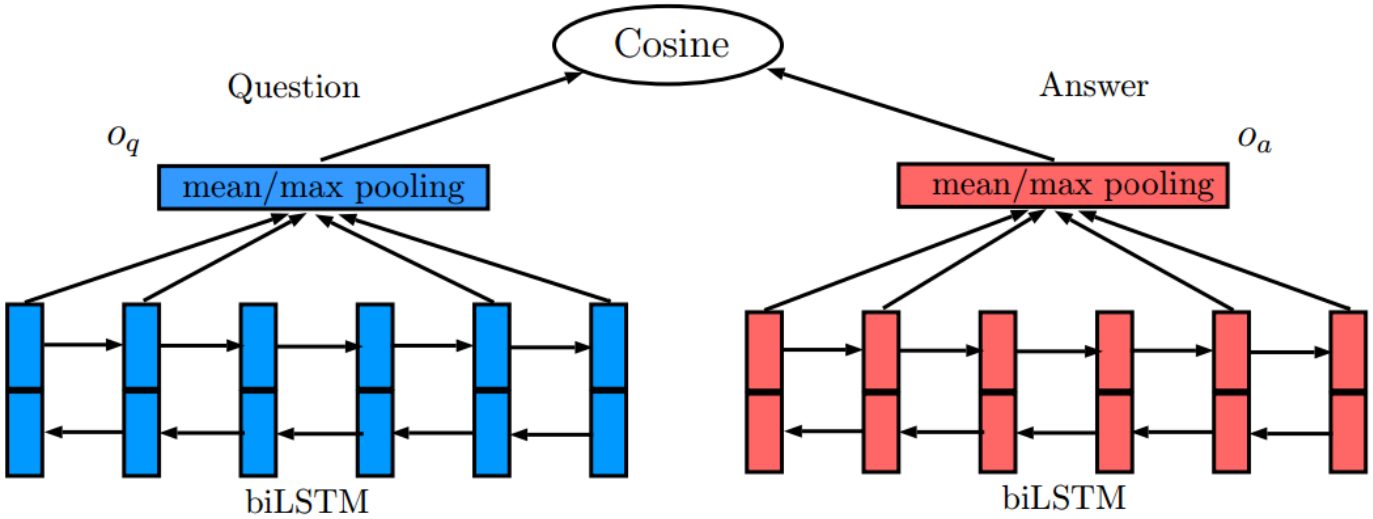
- **Considerations:** Regularization, dropout, dimensionality, objective function

Average and project

- Average embeddings and project the mean vector to a comparison-friendly vector space
- In *(Yu, 1412.1632)*: $s(x, y) = \sigma(x^T M \cdot y)$
- In *(Weston, 1410.3916)*: $s(x, y) = x^T U^T \cdot U y$
- Generative model: (i) the matrix projects between vector spaces of embeddings; (ii) dot-product is a measure of similarity

- Extension: Deep Averaging Networks *(Iyyer et al., 2015)*

Recurrent NNs



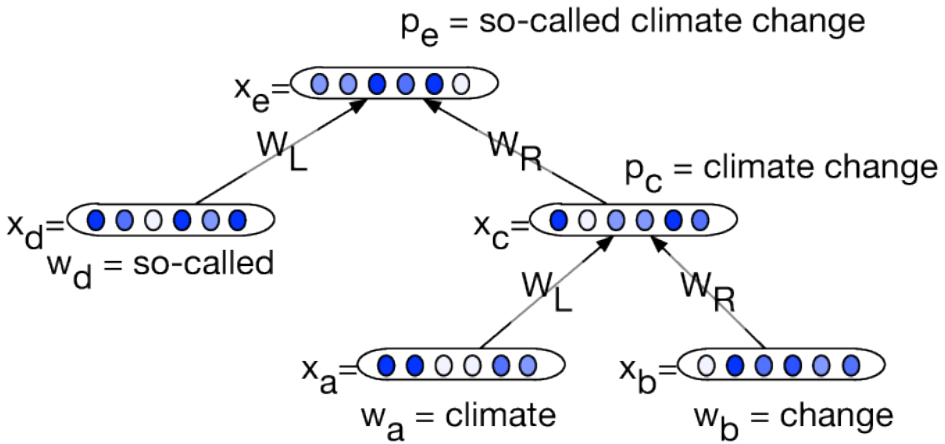
Variations: (bi)LSTM vs. GRU; pooling vs. last state

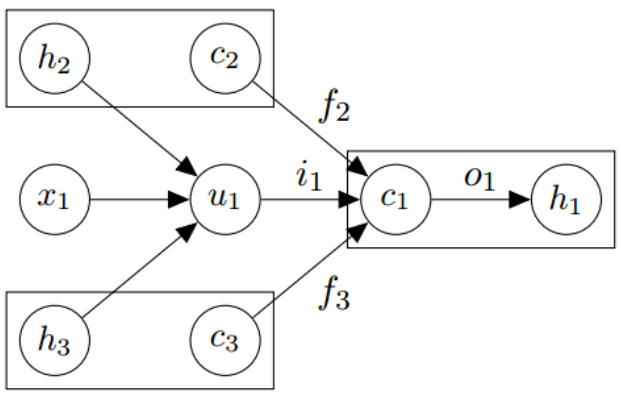
LSTM, GRU: Hidden states are called *memory units*, regulated by gates to update and forget (and retrieve)

Can also serve as just sequence preprocessing ("smearing" the context)

(Tan, 1511.04108)

Recursive NNs

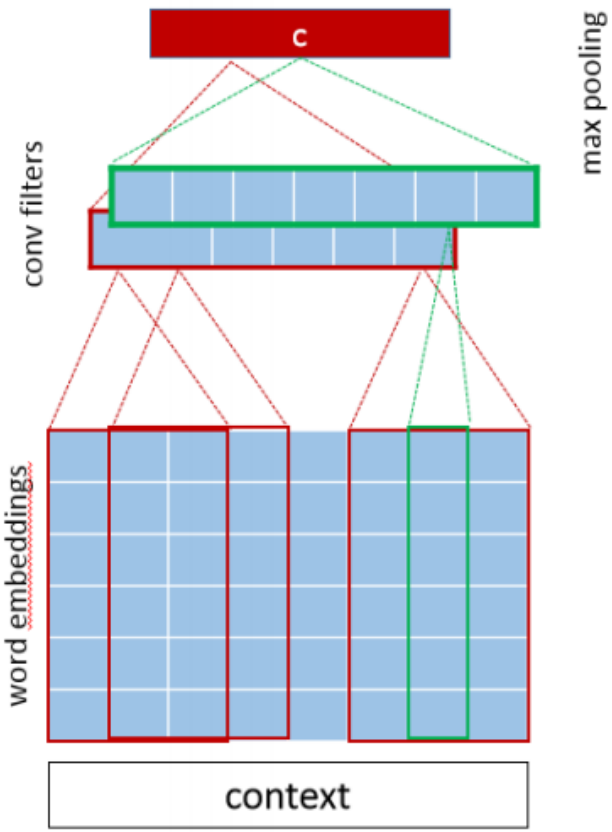


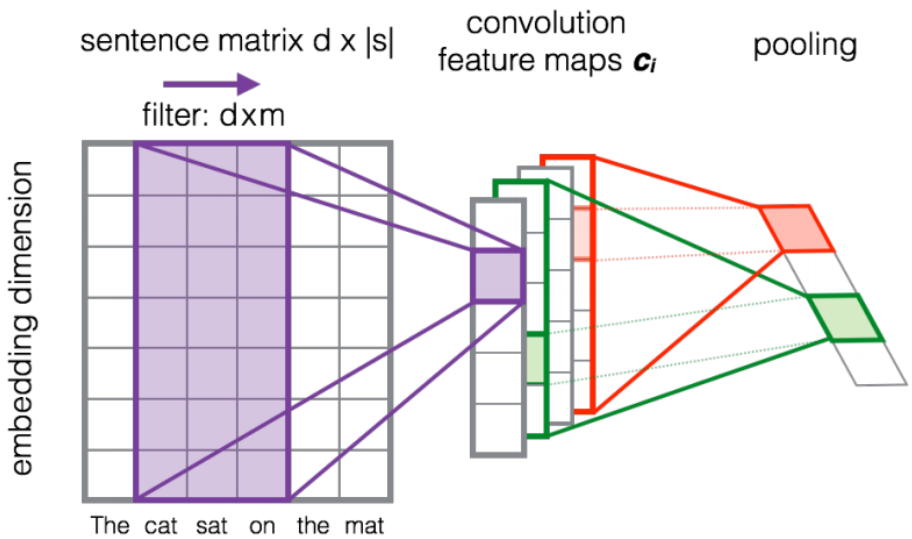


(Tai, 1503.00075)

TreeLSTM: Semantic Text Similarity specific innovation

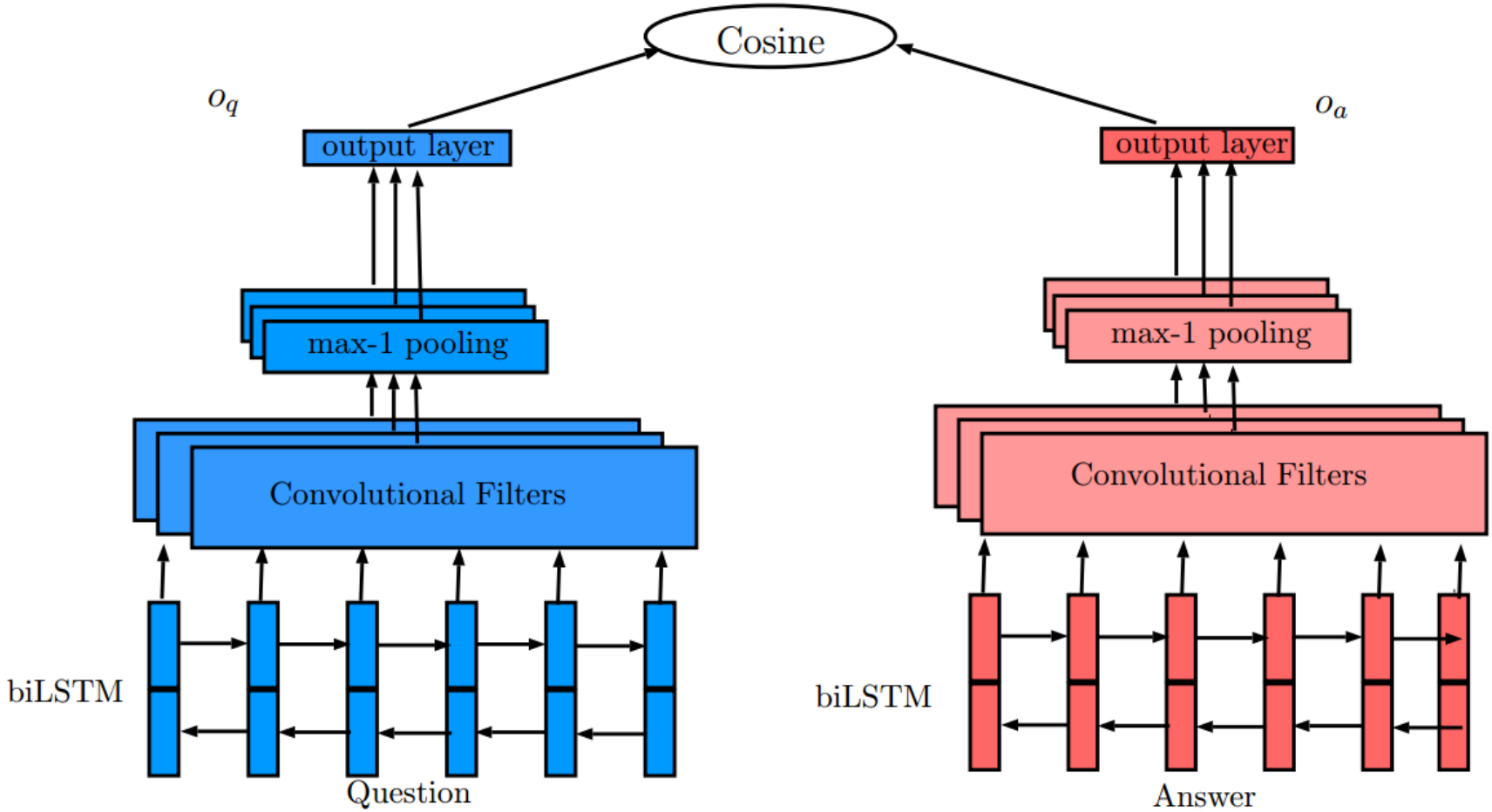
Convolutional NNs





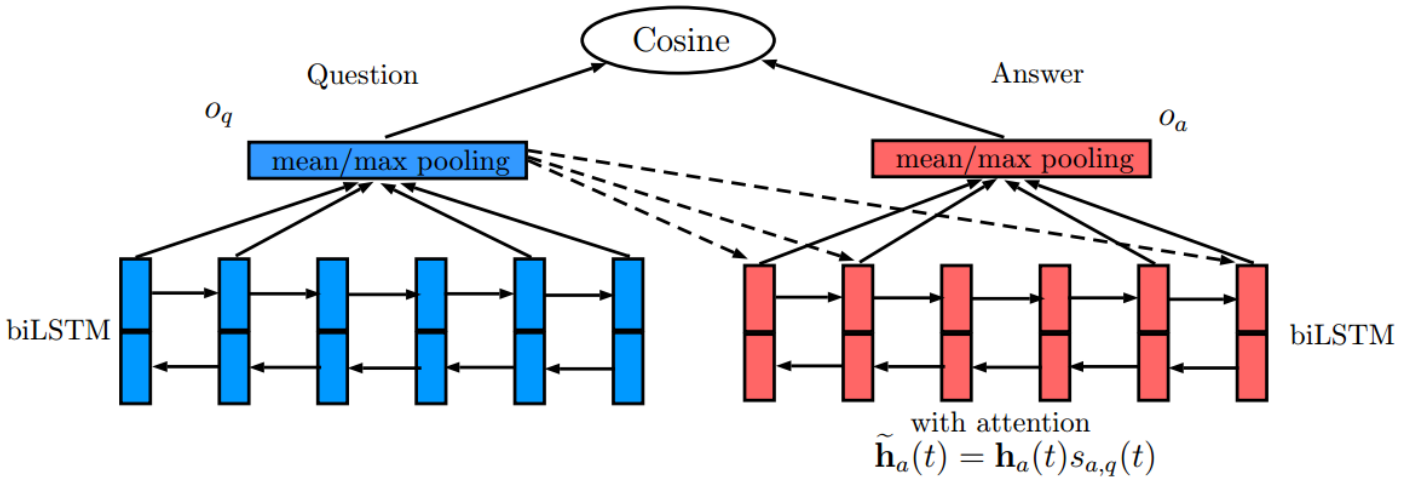
(Kadlec, 1510.03753), (Severyn and Moschitti, 2015)

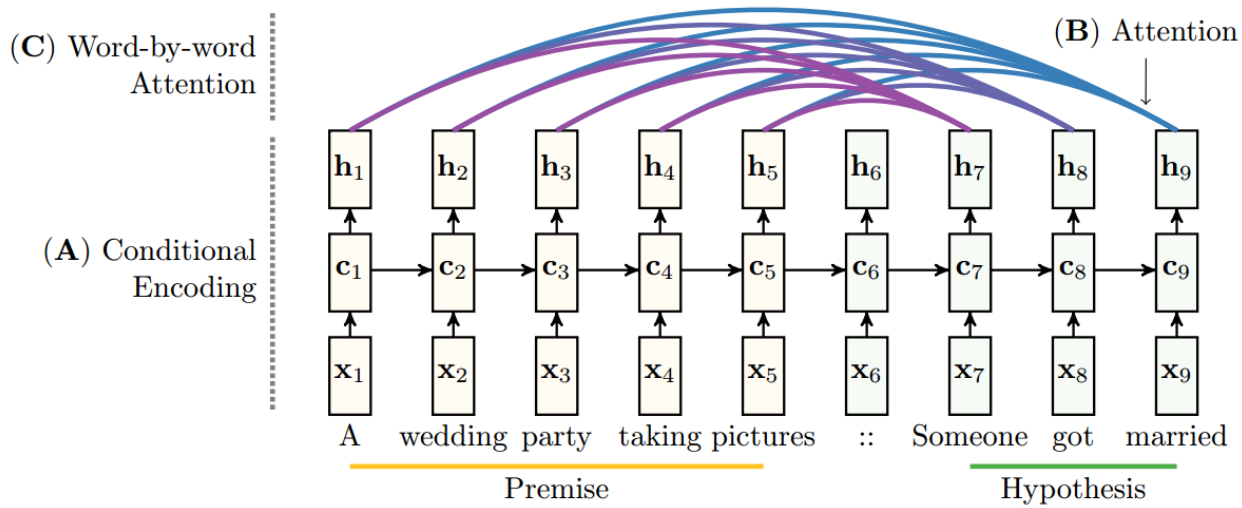
Convolutional NNs



(Tan, 1511.04108)

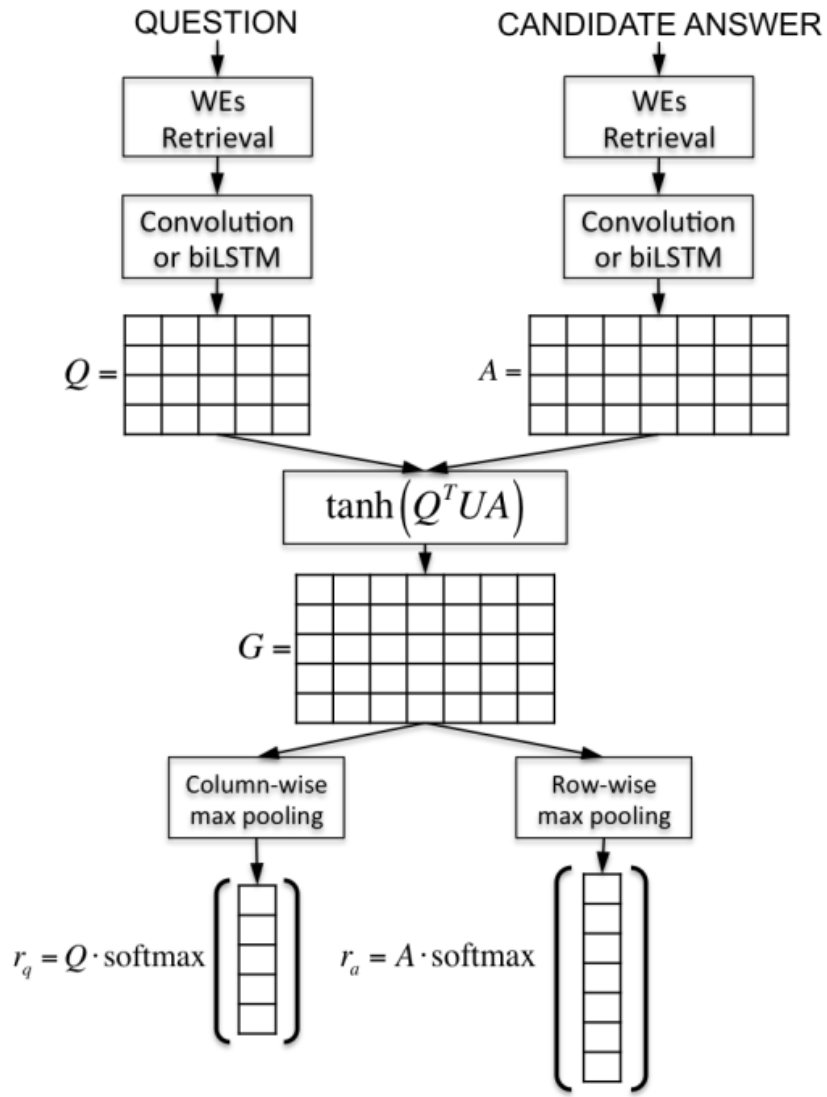
Attention-based Models





(Tan, 1511.04108), (Rocktäschel, 1509.06664)

Attentive Pooling



(dos Santos, 1602.03609)

Outline

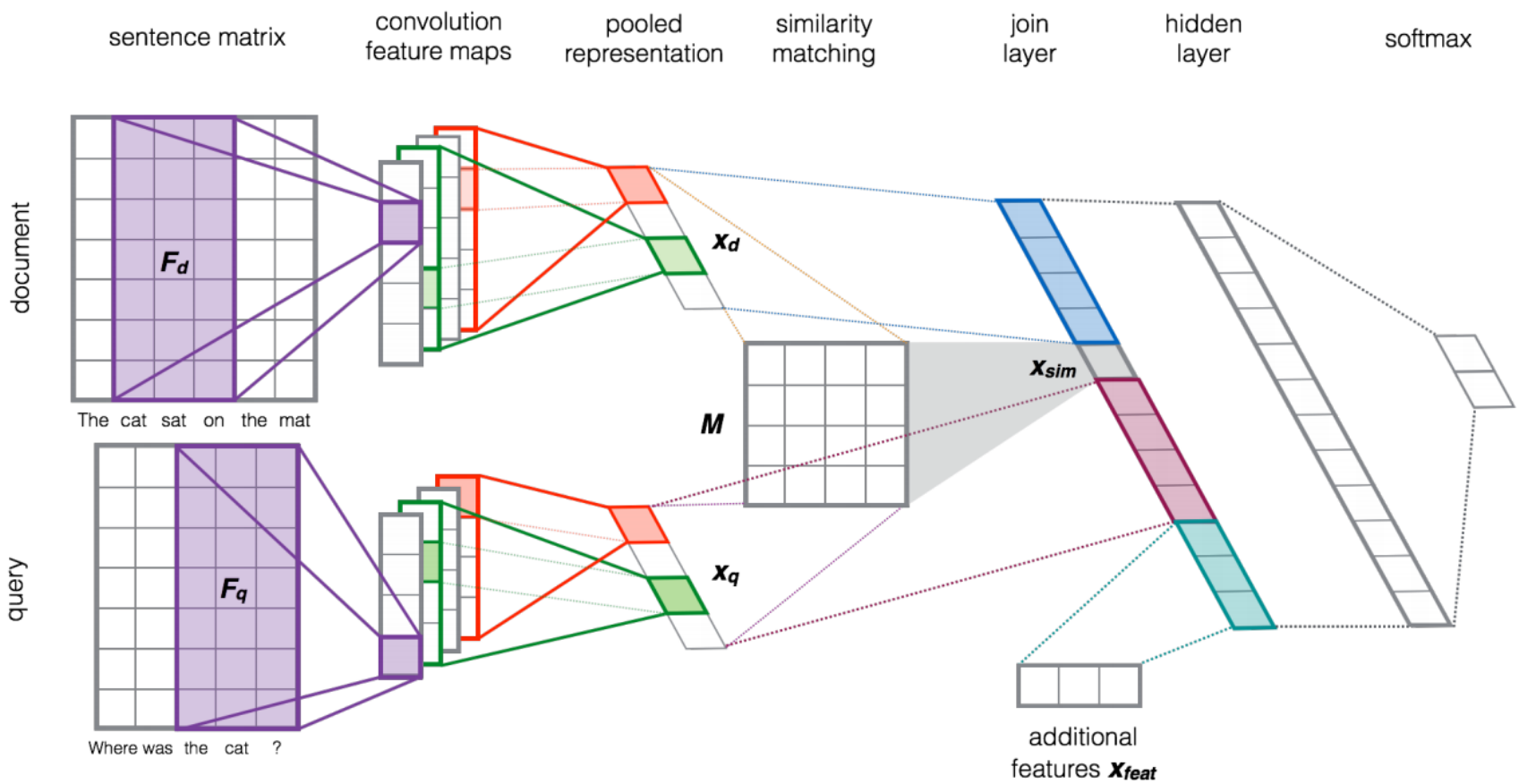
Answer Sentence Selection

Yih (2013) - LCLR	Yih et al. (2013)	0.709	0.770
Yu (2014) - TRAIN-ALL bigram+count	Yu et al. (2014)	0.711	0.785
W&N (2015) - Three-Layer BLSTM+BM25	Wang and Nyberg (2015)	0.713	0.791
Feng (2015) - Architecture-II	Tan et al. (2015)	0.711	0.800
S&M (2015)	Severyn and Moschitti (2015)	0.746	0.808
W&I (2015)	Wang and Ittycheriah (2015)	0.746	0.820
Tan (2015) - QA-LSTM/CNN+attention	Tan et al. (2015)	0.728	0.832
dos Santos (2016) - Attentive Pooling CNN	dos Santos et al. (2016)	0.753	0.851

MAP, MRR; aclwiki "Question Answering (State of the art)"

Answer Sentence Selection

Practice is sometimes not so rosy:



(Severyn and Moschitti, 2015)

Ubuntu Dialogue

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

	Baselines from [1]			Our Architectures			
	TF-IDF	RNN	LSTM	CNN	LSTM	Bi-LSTM	Ensemble
1 in 2 R@1	65.9%	76.8%	87.8%	84.8%	90.1%	89.5%	**91.5%**
1 in 10 R@1	41.0%	40.3%	60.4%	54.9%	63.8%	63.0%	**68.3%**
1 in 10 R@2	54.5%	54.7%	74.5%	68.4%	78.4%	78.0%	**81.8%**
1 in 10 R@5	70.8%	81.9%	92.6%	89.6%	94.9%	94.4%	**95.7%**

"Our best classifier is an ensemble of 11 LSTMs, 7 Bi-LSTMs and 10 CNNs trained with different meta-parameters."

(Kadlec, 1510.03753)

Outline

KeraSTS

Our Contribution: `https://github.com/brmson/dataset-sts`

- Software framework for many datasets
- Deep learning models using the KeraSTS toolkit
- Full Module/Task separation
- TODO: Attention screenshot?

Datasets

New datasets:

- Answer Selection "YodaQA" (much harder than pre-existing)
- Argus Headline Yes/No Questions (Silvestr Stanko)
- Kaggle Science Exams (work-in-progress)

Research

- Better datasets (supersede *anssel-wang*)
- Result variance and error bars
- Enrichment of input sequence with word matches
- Modular approach

Outline

Model Outlook

- Pretraining input modules on huge datasets.
- Extra supervision (like answer position in sentence).
- Learn to save per-sentence facts in short-term memory? (MemN2N; fully-differentiable shift-reduce parsing.)

Conclusion

- New universal paradigm for more advanced NLP models.
- Multi-task training and auxiliary supervision will be fun.
- KeraSTS makes marrying models and tasks super-easy!
- What other tasks / applications do occur to you?
  Do you have a dataset for us?

Thank you for your attention!

pasky@ucw.cz

Collaborators: Silvestr Stanko, Jiří Nádvorník
Supervised by: Jan Šedivý