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
Practical Tasks
### Vision:
Google-like text box on each support page instead of wading through questions.

### Related:
StackOverflow-based personal assistant

<table>
<thead>
<tr>
<th>What are the working hours?</th>
<th>Is someone going to help me with the project?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do I need to come in person?</td>
<td>Do I need to work on the project alone?</td>
</tr>
<tr>
<td>When do I need to come to work?</td>
<td>Will someone mentor me?</td>
</tr>
<tr>
<td>Can I work at night?</td>
<td>How much guidance will I get?</td>
</tr>
</tbody>
</table>
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

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?

<table>
<thead>
<tr>
<th>-0.101</th>
<th>What is Rohm and Haas's annual revenue?</th>
<th>Rohm and Haas, with $ billion in annual sales, makes chemicals found in such products as decorative and industrial paints, semiconductors and shampoos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.677</td>
<td>What is Rohm and Haas's annual revenue?</td>
<td>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</td>
</tr>
<tr>
<td>-7.507</td>
<td>What is Rohm and Haas's annual revenue?</td>
<td>The transaction announced today creates a global specialty chemicals company with combined annual revenues of $ billion.</td>
</tr>
</tbody>
</table>
Neural Models with Attention

\[ o_q \rightarrow \text{mean/max pooling} \rightarrow \text{Cosine} \rightarrow \text{Answer} \]

\[ o_a \rightarrow \text{mean/max pooling} \rightarrow \text{with attention} \]

\[ \tilde{h}_a(t) = h_a(t) s_{a,q}(t) \]

(Tan, 1511.04108), (Rocktäschel, 1509.06664)

---

(C) Word-by-word Attention

(A) Conditional Encoding

Premise: A wedding party taking pictures :: Someone got married

Hypothesis:
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

\[ h_x = h_L \odot h_R, \]
\[ h_+ = |h_L - h_R|, \]
\[ h_s = \sigma \left( W^{(x)} h_x + W^{(+)} h_+ + b^{(h)} \right) \]
Some Recent Results
Recent Papers


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

github.com brmson/dataset-sts
Python, Theano, Keras

---
## Ubuntu Dialogue Corpus

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

<table>
<thead>
<tr>
<th>Model</th>
<th>MC-160 one</th>
<th>all</th>
<th>MC-500 one</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>hand-crafted</td>
<td>0.842</td>
<td>0.753</td>
<td>0.721</td>
<td>0.699</td>
</tr>
<tr>
<td>Attn. Reader</td>
<td>0.481</td>
<td>0.463</td>
<td>0.444</td>
<td>0.419</td>
</tr>
<tr>
<td>Neur. Reasoner</td>
<td>0.484</td>
<td>0.476</td>
<td>0.457</td>
<td>0.456</td>
</tr>
<tr>
<td>HABCNN-TE</td>
<td>0.633  <strong>0.631</strong></td>
<td></td>
<td>0.542</td>
<td>0.529</td>
</tr>
<tr>
<td><strong>avg</strong></td>
<td>0.653</td>
<td>±0.027</td>
<td>0.556</td>
<td>±0.012</td>
</tr>
<tr>
<td><strong>DAN</strong></td>
<td>0.681</td>
<td>±0.017</td>
<td>0.577</td>
<td>±0.010</td>
</tr>
<tr>
<td><strong>RNN</strong></td>
<td>0.583</td>
<td>±0.033</td>
<td>0.533</td>
<td>±0.020</td>
</tr>
<tr>
<td><strong>Universal</strong></td>
<td><strong>0.736</strong></td>
<td>±0.033</td>
<td><strong>0.612</strong></td>
<td>±0.023</td>
</tr>
</tbody>
</table>

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

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.
Thanks for Your Attention

Jan Šedivý’s 3C Group @ FEL ČVUT

eClub Prague
http://eclubprague.com/

Ailao Labs
http://ailao.eu/