Extending Neural Question Answering with Linguistic Input Features

Fabian Hommel, Matthias Orlikowski, Philipp Cimiano, Matthias Hartung

Twitter: @morlikow, @semalytix
1. Motivation: QA for rapid information access in specialised domains
2. Task, research questions
3. Linguistic Features
4. Stanford Question Answering Dataset
5. Results (Linguistic Features vs Baseline)
6. Conclusion
I: Jane went to the hallway.
I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden
I: It started boring, but then it got interesting.
Q: What’s the sentiment?
A: positive
Q: POS tags?
A: PRP VBD JJ, CC RB PRP VBD JJ.

Kumar et al. (2016)

Green et al. (1961)

When the question has been bracketed, any un-bracketed preposition is attached to the first noun phrase in the sentence, and prepositional brackets added. For example, "Who did the Red Sox lose to on July 5?" becomes "(To [who] ) did [ the Red Sox] lose (on [July 5])?"

Month = July
Place = Boston
Day = 7
Game Serial No. = 96
(Team = Red Sox, Score = 5)
(Team = Yankees, Score = 3)

Long-term goal: A flexible, general QA system would be an effective surrogate model for bootstrapping information access in specialised domains!
Linguistic features for neural QA

- We hypothesise that the general linguistic structure of question-answer pairs is domain-agnostic to some extent in English
- Approach to QA for rapid information access in specialised domains
  - 1) **Learn general linguistic structure on large open-domain dataset (this work)**
  - 2) Adapt for specific domains
- Neural approaches led to improved performance for core NLP tasks like part-of-speech tagging (Koo et al., 2008), dependency parsing (Chen and Manning, 2014), …
- But neural models for more high-level tasks only use generic representations (word/character embeddings)
- However, e.g. Sennrich and Haddow (2016) showed that neural machine translation performance increases when adding linguistic features to word embeddings
Task and research questions

- Task is specific type of QA: reading comprehension, given a context and question, predict a span in the context as answer.

- We reimplement QANet (Yu et al. 2018) and adapt Sennrich & Haddow (2016) to include linguistic features.
  
  1. “replication study” - can we reimplement and get same performance levels?
  
  2. To what extend do linguistic features help to predict better (more precise, relevant) answers/spans?

High-level view of QANet (Yu et al. 2018)
Part of Speech tags

<table>
<thead>
<tr>
<th>Where</th>
<th>did</th>
<th>Super</th>
<th>Bowl</th>
<th>50</th>
<th>take</th>
<th>place?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV</td>
<td>VERB</td>
<td>PROPN</td>
<td>PROPN</td>
<td>NUM</td>
<td>VERB</td>
<td>NOUN</td>
</tr>
</tbody>
</table>

- Tag each token with coarse PoS tag set using spaCy library (https://spacy.io/)
- High-level, shallow linguistic information about each token
• Dependency parse using spaCy library (https://spacy.io/)
• Use dependency label to label each child token and root
• Information about position in **syntactic structure** of the sentence
super bowl 50 was an American football game to determine the champion of the national football league (NFL) for the 2015 season.

- **Shallow semantic structure** by identifying events/predicates and participants/arguments/semantic roles - *Who* did *what* to *whom*, *where*, *when* and *how*?
- e.g. PREDICATE for events, ARG0 (“agent”), ARG1 (“patient”), NOROLE for tokens without SRL
Input embeddings in neural QA

High-level view of QANet (Yu et al. 2018)
Embedding linguistic features

Input Embedding Layer

Embedding

2D Conv

ReLU

Max

Concatenation

Highway

Embeddings
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

**Ground Truth Answers**

- Santa Clara, California
- Levi's Stadium
- Levi's Stadium in the San Francisco Bay Area at Santa Clara, California

<table>
<thead>
<tr>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.5k</td>
<td>10.1k</td>
<td>10.1k</td>
<td>107.7k</td>
</tr>
</tbody>
</table>

Number of question-answer pairs in SQuAD
**Exact Match (EM)** *Percentage of predictions that match any one of the three ground truth answers exactly*

**Question:** Where did Super Bowl 50 take place?

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Ground Truth Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levi's Stadium</td>
<td>Levi's Stadium in the San Francisco Bay Area at Santa Clara, California</td>
</tr>
<tr>
<td>Santa Clara, California</td>
<td></td>
</tr>
</tbody>
</table>
**F1** Average overlap between the prediction and ground truth answer (max F1 for each question, averaged over all questions)

**Question:** Where did Super Bowl 50 take place?

**Prediction:** Levi Stadium 's

**Ground Truth Answers:**
- Levi Stadium 's
- San Francisco
- Santa Clara
- Santa Clara
- California
- California
- at
- in
- Levi Stadium 's
- Levi's
- California
- Bay Area
- in
- Santa Clara
- Levi's
- Santa's
- Levi's
Results - 1) QANet reimplementation baseline

Table 1: Impact of evaluated individual parameters and the combination of their best settings on F1 and exact match (EM) scores.
Results - 2) Linguistic features relative to baseline
Results - 2) Linguistic features relative to baseline

- PoS (100dim)
- DL (100dim)
- SRL (200dim)
- Combination

ΔF1 (0,100)  ΔEM (0,100)
Results - 2) Linguistic features relative to baseline

- PoS (100dim)
- DL (100dim)
- SRL (200dim)
- Combination

ΔF1 (0,100)  ΔEM (0,100)
Results - 2) Linguistic features relative to baseline

![Chart showing ΔF1 (0,100) and ΔEM (0,100) for PoS (100dim), DL (100dim), SRL (200dim), and Combination.](chart.png)
Results - 2) Linguistic features relative to baseline

- PoS (100dim)
- DL (100dim)
- SRL (200dim)
- Combination

\[ \Delta F_1 (0,100) \]

\[ \Delta EM (0,100) \]
Results - 2) Linguistic features relative to baseline

- PoS (100dim)
- DL (100dim)
- SRL (200dim)
- Combination

ΔF1 (0,100)  ΔEM (0,100)
Results - 2) Linguistic features relative to baseline

- Linguistic features improve over this fine-tuned baseline!
- Syntactic information helps with finding exact matches
- SRL relative low impact - too sparse & non-optimal aggregation?
- Combination is best, so DL/SRL have complementary information to PoS
- Hyperparameters best settings ( = baseline): $1.7 \Delta F1$, $1.9 \Delta EM$
Conclusion

• To what extent do neural QA models benefit from **linguistic features**?
  • Added **PoS**, **syntactic** dependencies and **semantic roles** to input representation

• Evaluation on large open-domain dataset SQuAD
  • PoS is best individual feature, but **combination best overall**
  • Higher impact on EM than on F1: proposed linguistic features seems to help with **boundary detection**, locating answer spans may depend more on word-level semantics

• Can feature engineering become cool again?

• Future work
  • Additional linguistic information (lemmatized words, NER, morphology, Sennrich and Haddow 2016)
  • Better aggregation/representation (e.g. recursive encoding layers, Socher et al. 2011)
  • Evaluate generalisation to specific domains


