MACAO, 12TH OF AUGUST 2019

Prêt-à-LLOD

Universität Bielefeld

SEMALYTIX

Extending Neural Question Answering with Linguistic Input Features

Fabian Hommel, **Matthias Orlikowski**, Philipp Cimiano, Matthias Hartung

Twitter: @morlikow, @semalytix

Table of Contents

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

QA for rapid information access in specialised domains

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)

[How many games] did

[the Yankees] play (in [July])?

When the question has been bracketed, any unbracketed 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)

Green et al. (1961)

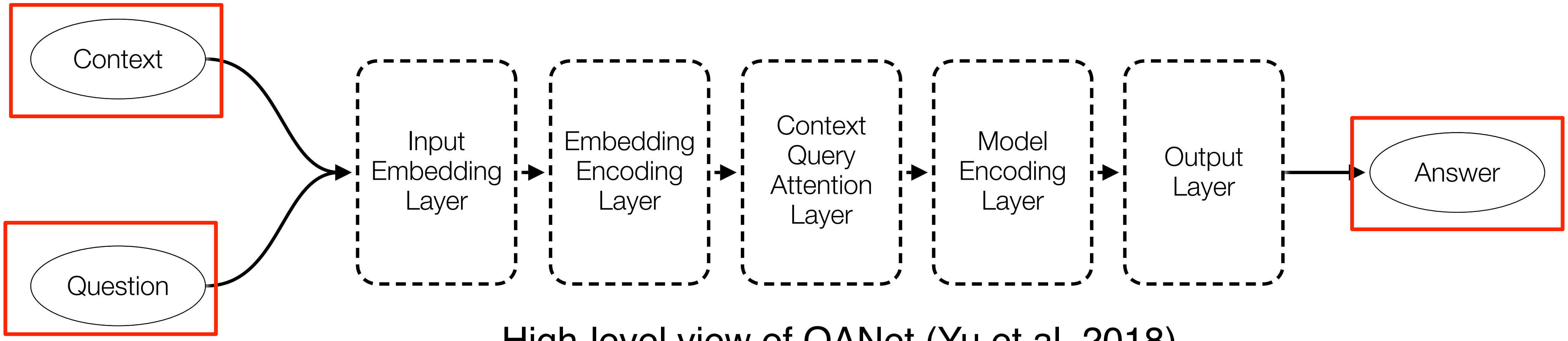
Long-term goal: A flexible, general QA system would be an effective surrogate model for bootstrapping information access in specialised domains!

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?



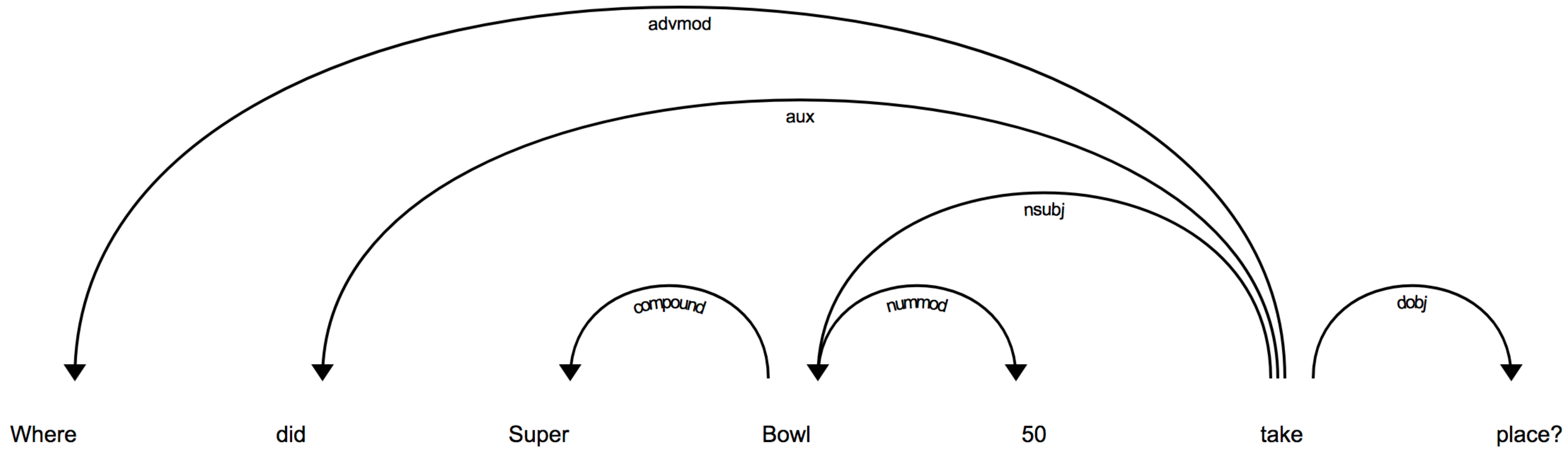
High-level view of QANet (Yu et al. 2018)

Part of Speech tags

Where	did	Super	Bowl	50	take	place?
ADV	VERB	PROPN	PROPN	NUM	VERB	NOUN

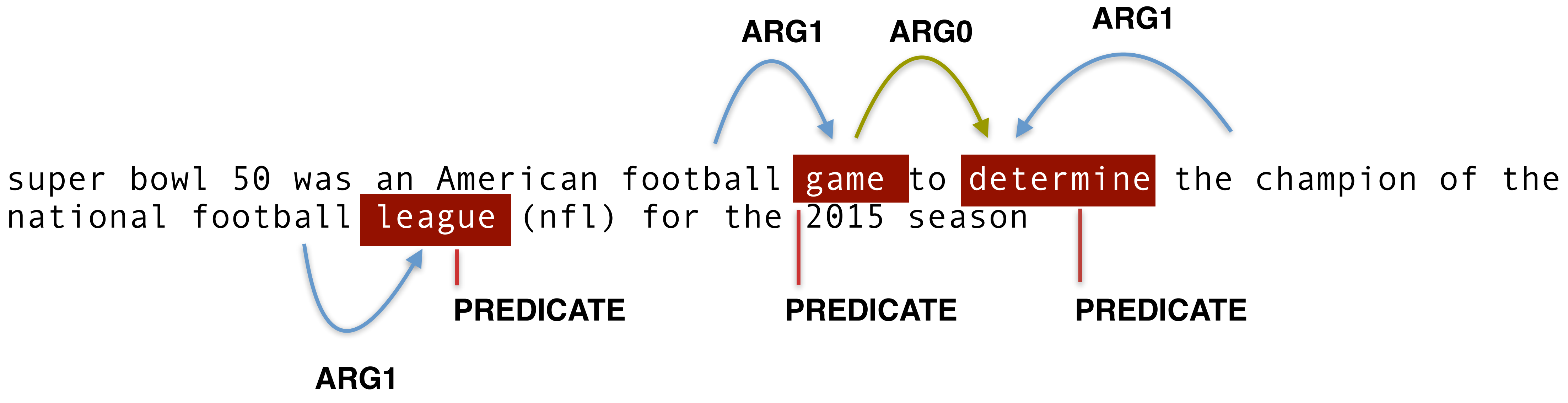
- Tag each token with coarse PoS tag set using spaCy library (https://spacy.io/)
- High-level, shallow linguistic information about each token

Dependency labels



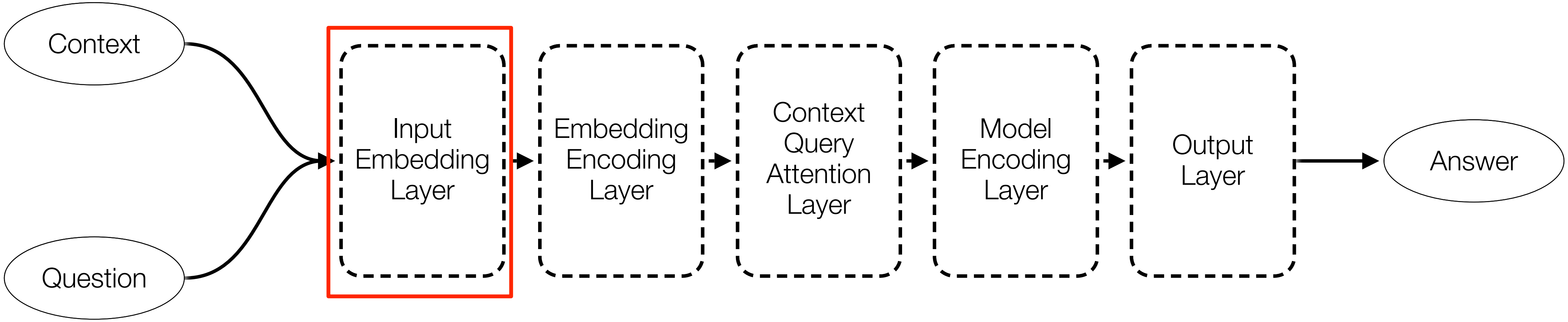
- Dependency parse using spaCy library (https://spacy.io/)
- Use dependency label to label each child token and root
- Information about position in **syntactic structure** of the sentence

Semantic role labels



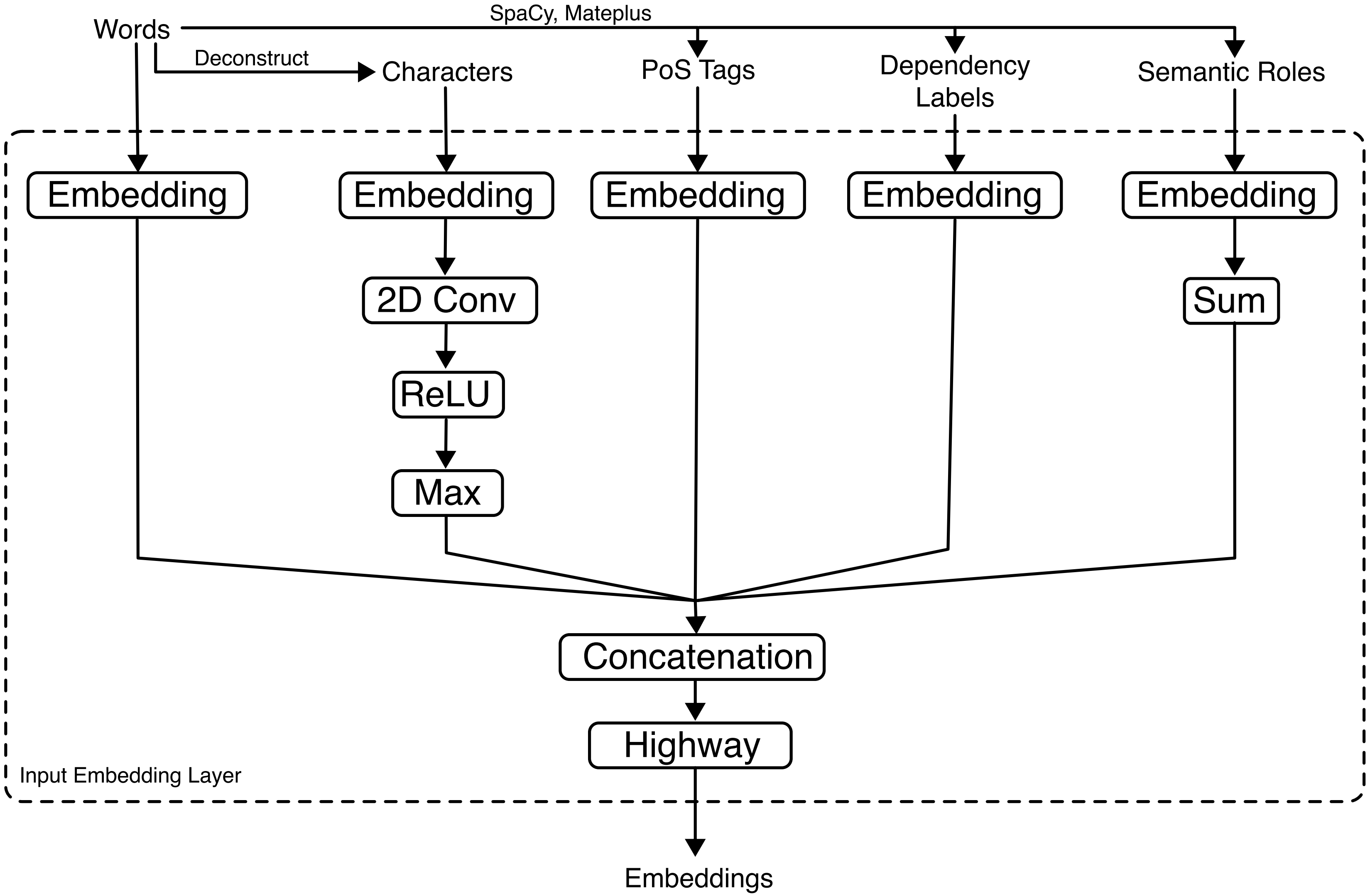
- Semantic Role Labelling (SRL) based on PropBank (Palmer et al. 2005) using *Mateplus* (https://github.com/microth/mateplus)
- **Shallow semantic structure** by identifying events/predicates and participants/arguments/semantic roles - *Who* did *what* to *whom*, *where*, *when* and *how?*
- e.g. PREDICATE 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



The Stanford Question Answering Dataset 1.x (Rajpurkar et al. 2016)

Context (English Wikipedia excerpts, avg. length 250 tokens)

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.

Question (avg. length 10 tokens)

Where did Super Bowl 50 take place?

Ground Truth Answers

Santa Clara, California

Levi's Stadium

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

Train	Dev	Test	Total
87.5k	10.1k	10.1k	107.7k

Number of question-answer pairs in SQuAD

Evaluation Metrics - Exact Match

Exact Match (EM) *Percentage of predictions that match any one of the three ground truth answers exactly*

Where did Super Bowl 50 take place?

Levi's Stadium

Santa Clara, California

Levi's Stadium

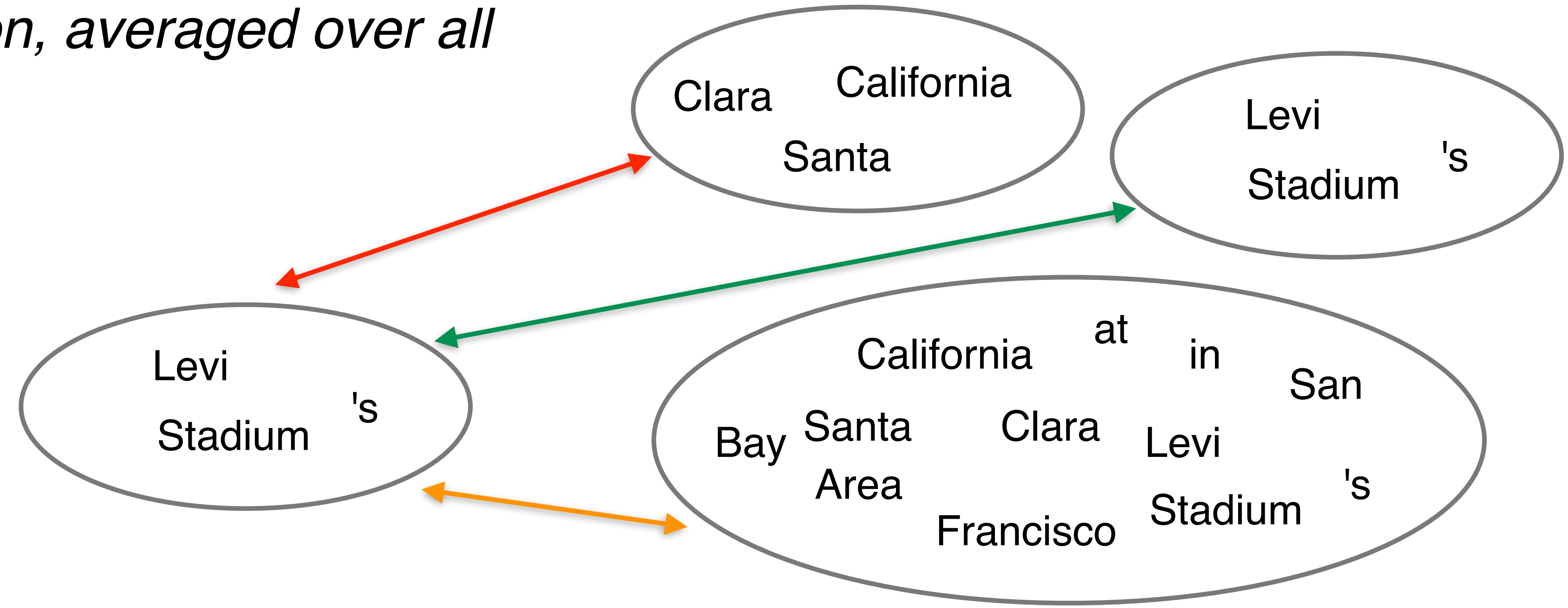
Levi's Stadium in the San Francisco Bay Area at Santa Clara, California

Question **Prediction** **Ground Truth Answers**

Evaluation Metrics - F1

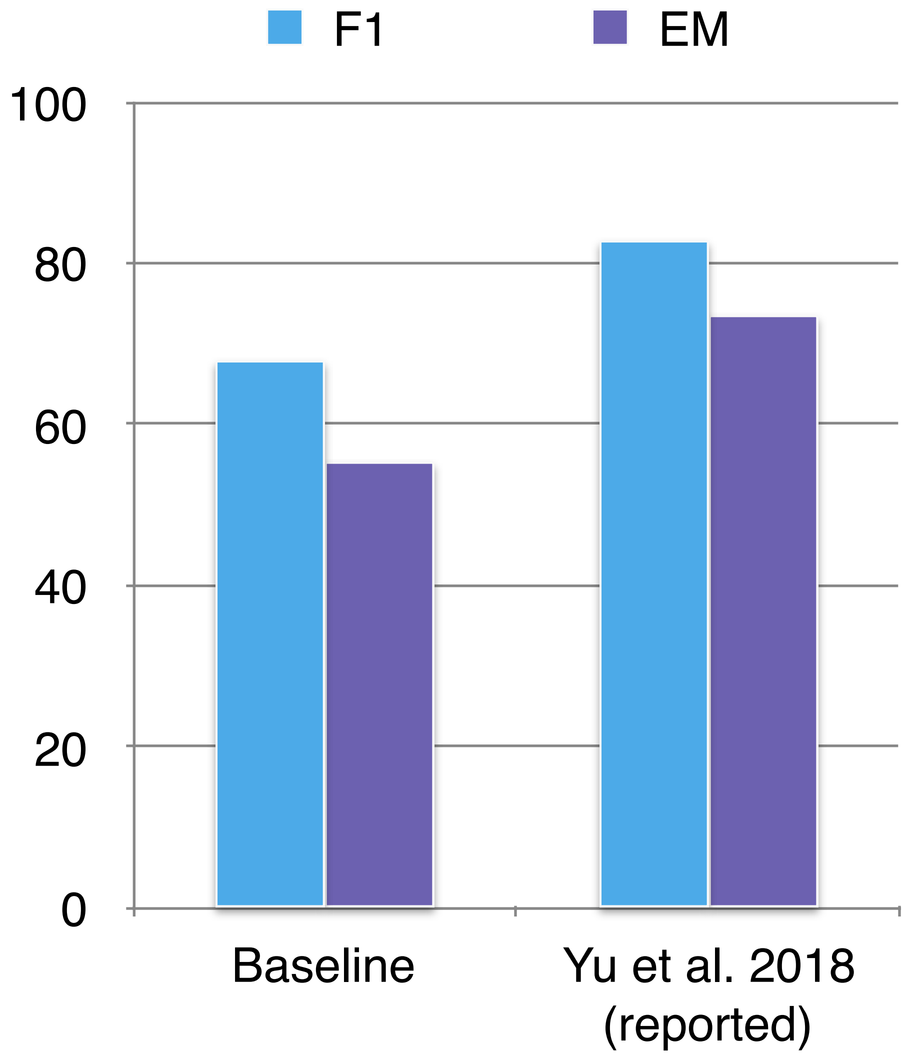
F1 *Average overlap between the prediction and ground truth answer (max F1 for each question, averaged over all questions)*

Where did Super Bowl 50 take place?



Question **Prediction** **Ground Truth Answers**

Results - 1) QANet reimplementation baseline

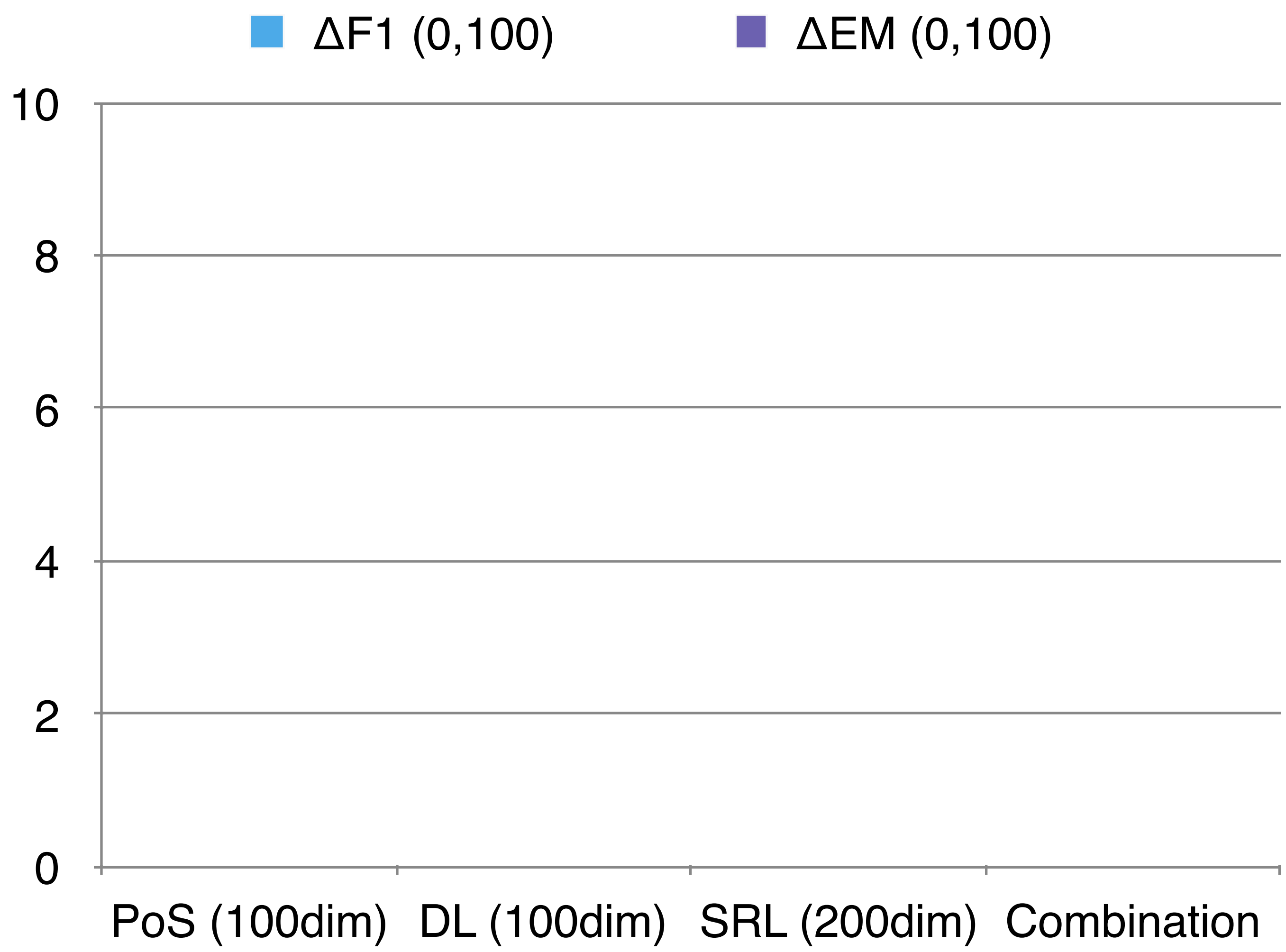


Parameter	ΔF1	ΔEM
word embeddings	**2.4**	1.8
character embeddings	1.6	1.7
# convolutional layers	1.5	**2.2**
shared wheights in encoding	1.3	1.3
# encoder blocks	0.9	0.9
# attention heads	0.6	1.2
# highway layers	0.4	1.0
model dimensionalty	0.5	0.8
pointwise feed-forward layers	0.2	0.0
combination of best settings	1.7	1.9

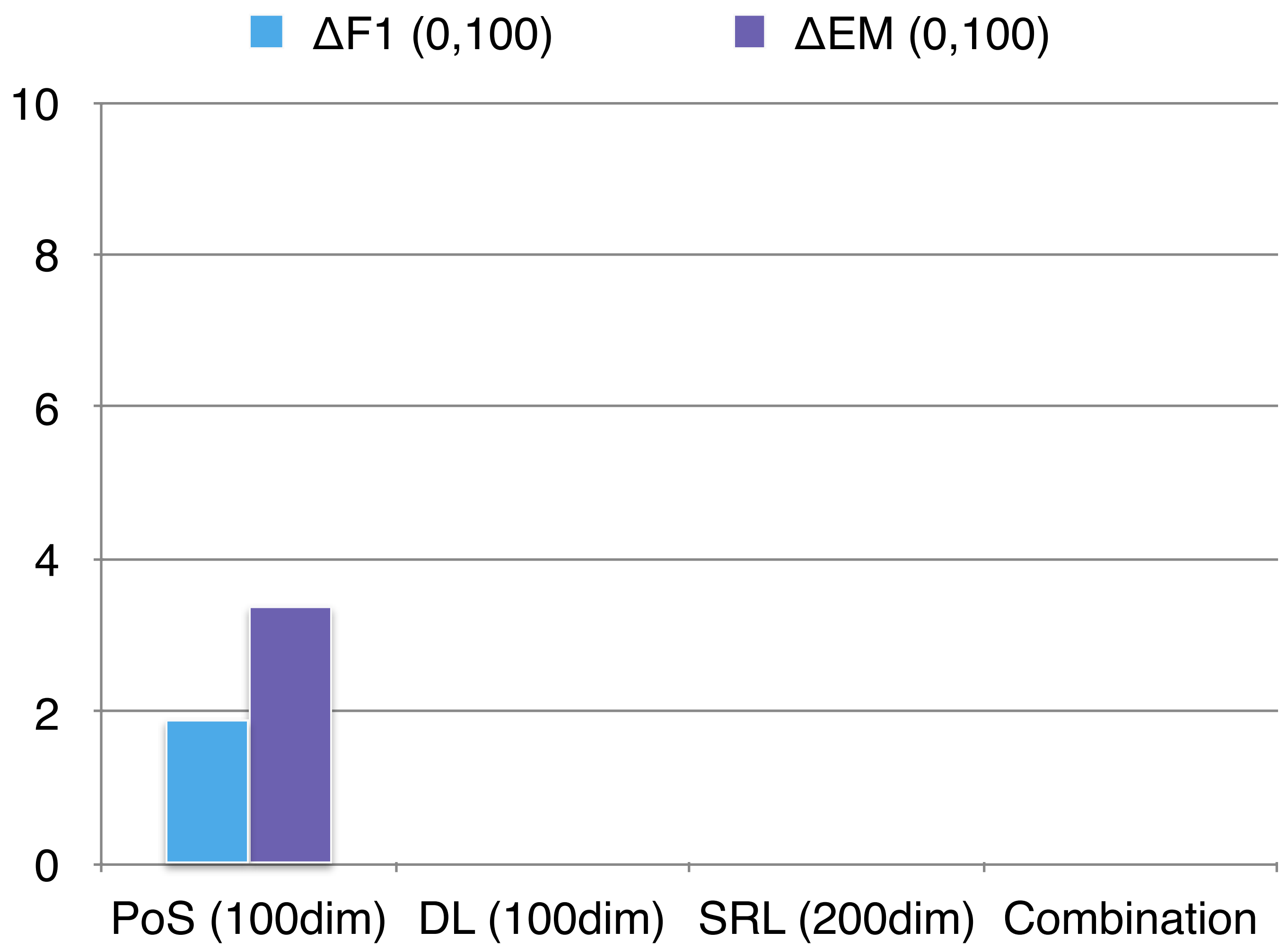
Table 1: Impact of evaluated individual parameters and the combination of their best settings on F1 and exact match (EM) scores.

Results - 2) Linguistic features relative to baseline

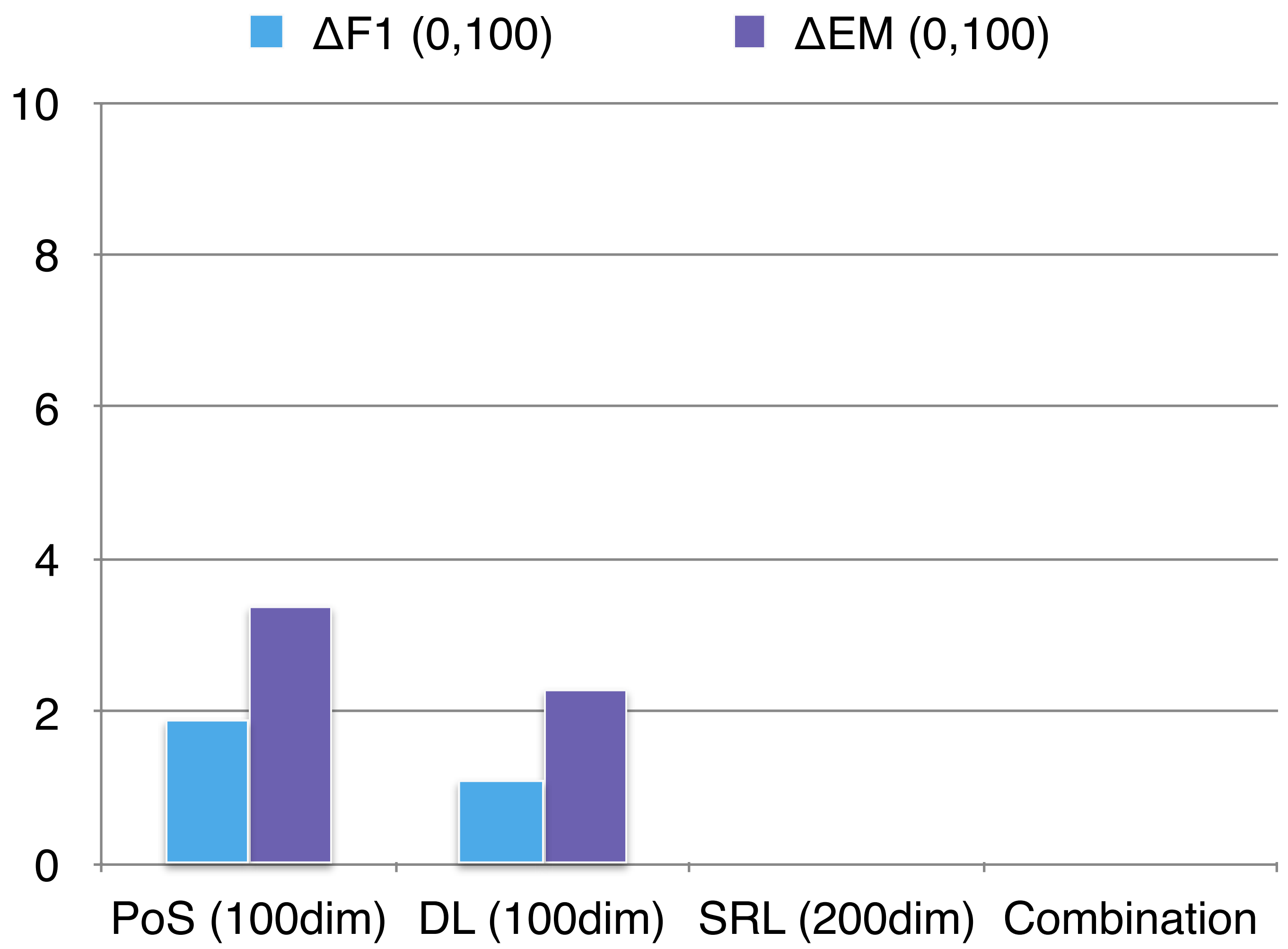
Results - 2) Linguistic features relative to baseline



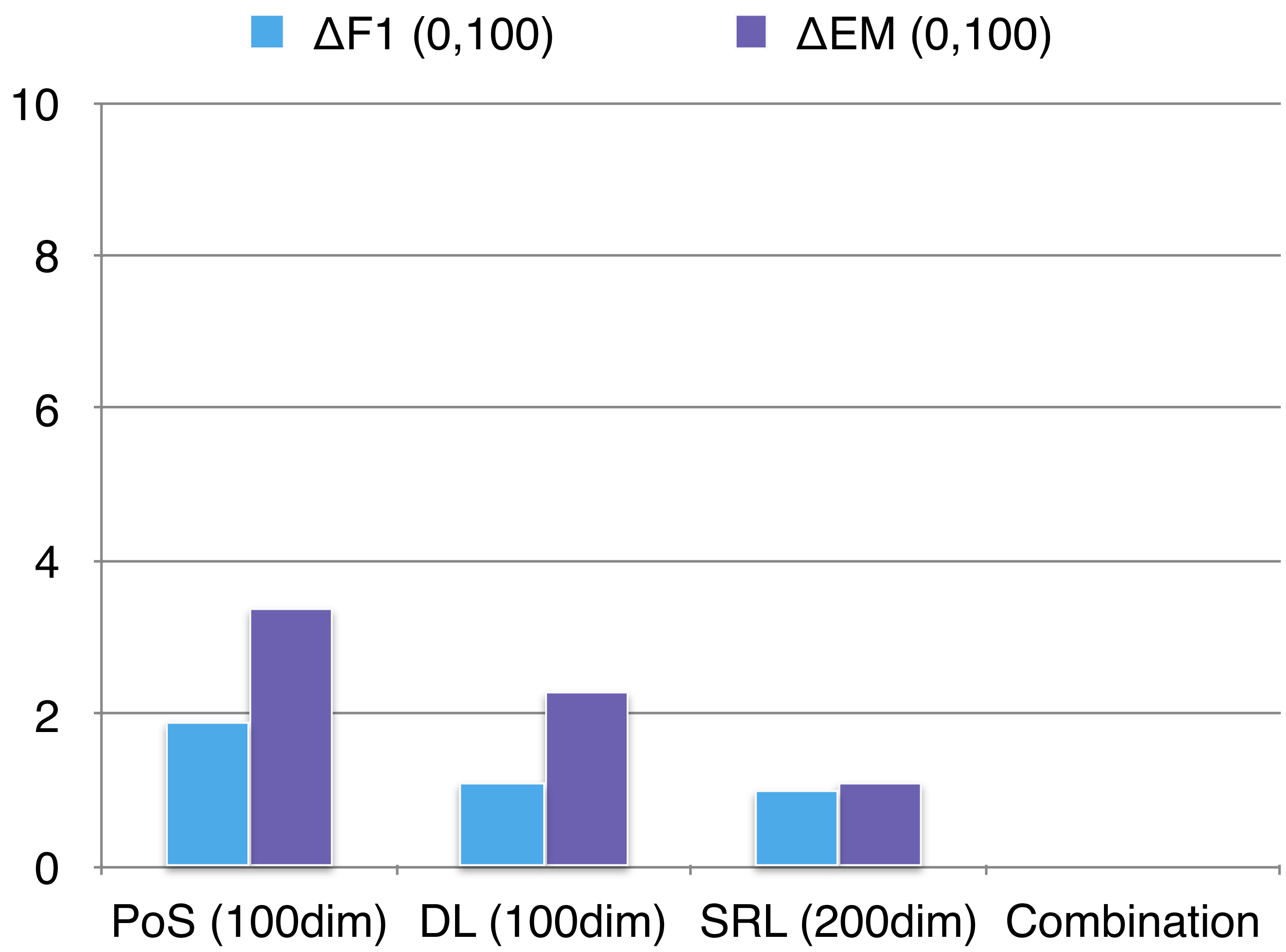
Results - 2) Linguistic features relative to baseline



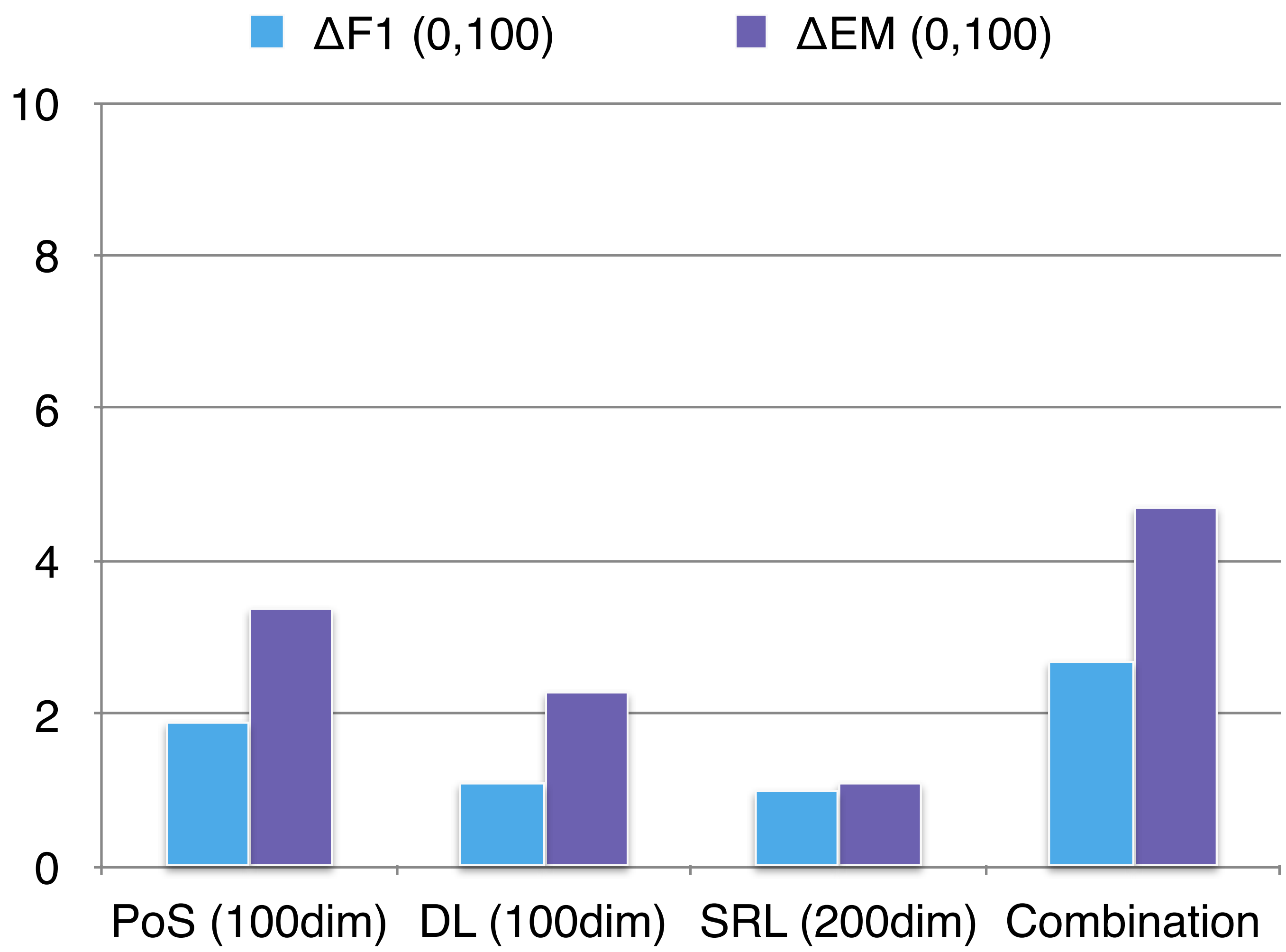
Results - 2) Linguistic features relative to baseline



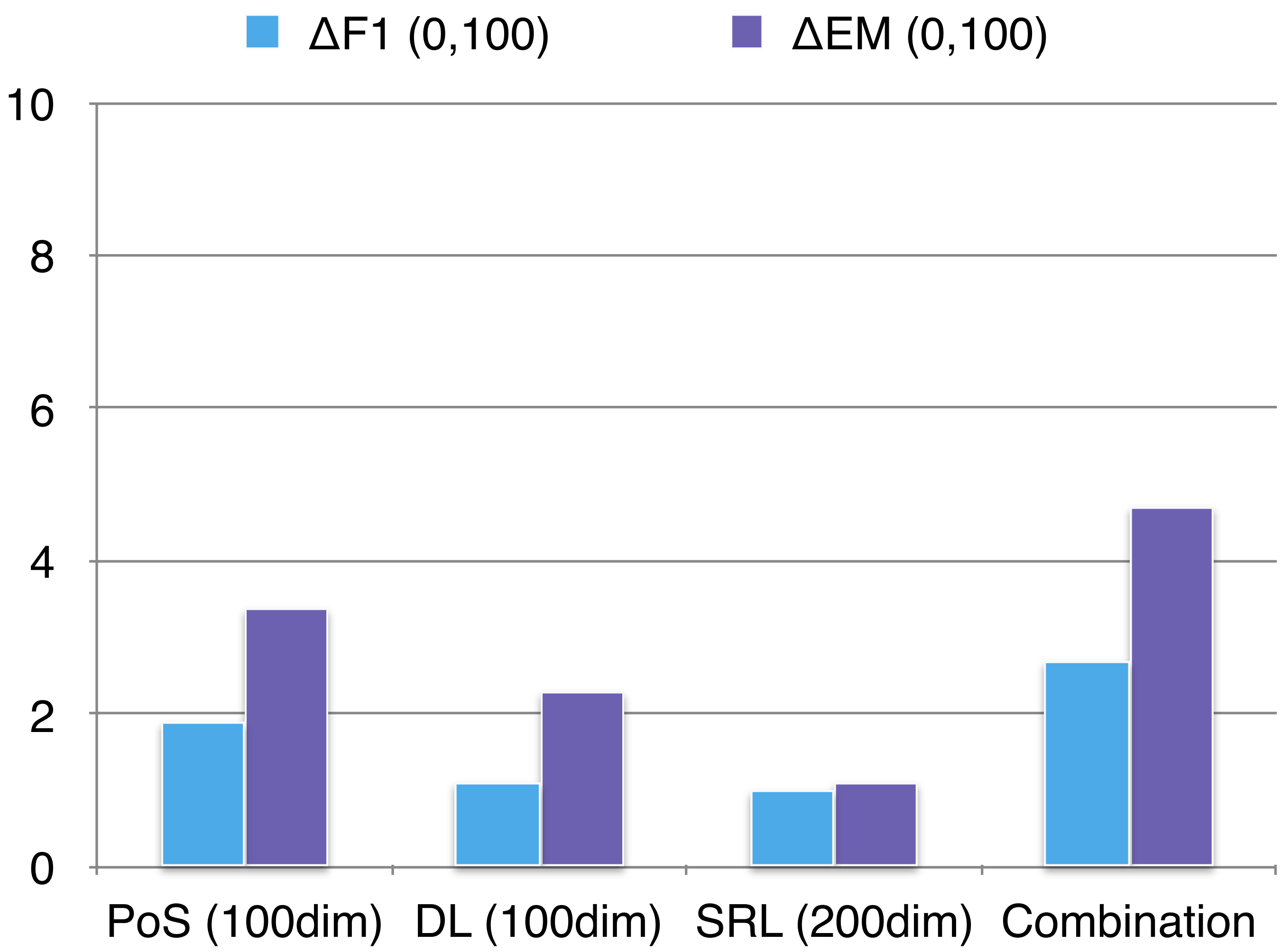
Results - 2) Linguistic features relative to baseline



Results - 2) Linguistic features relative to baseline



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* ΔF1, *1.9* ΔEM

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

References

GREEN JR, Bert F; et al. (1961). "Baseball: an automatic question-answerer" (PDF). Western Joint IRE-AIEE-ACM Computer Conference: 219–224.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of The 33rd International Conference on Machine Learn- ing*, volume 48 of *Proceedings of Machine Learning Research*, pages 1378–1387, New York, New York, USA. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Nat- ural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 83–91.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehen- sion. *CoRR*, abs/1804.09541.

Fabian Hommel, Analytics Engineer

fabian.hommel@semalytix.de

Matthias Orlikowski, NLP Product/
Solutions Engineer

matthias.orlikowski@semalytix.de

@morlikow

Dr. Matthias Hartung, Co-Founder & Chief Research & Development Officer

hartung@semalytix.de

Prof. Dr. Philipp Cimiano, Co-Founder & Chief Technology Officer, Head of Semantic Computing Group at Bielefeld University

cimiano@semalytix.de

www.semalytix.com

www.pret-a-llod.eu

www.sc.cit-ec.uni-bielefeld.de