SEMANTIC RELATION EXTRACTION AND CLASSIFICATION IN SCIENTIFIC PAPERS

ERIK KRISTIAN JANEZIC

IMPORTANCE

- Autonomous knowledge extraction from sceintific papers
- Generation and reall time updating of knowledge semantic networks
- Processing complex scientific queries

TASK

- SemEval 2018
- Extraction and classification of relations in abstracts of scientific papers
- Sub-tsks:
  - Classification on clean data
  - Classification on noisy data
  - Relation extraction

DATA

- 350 abstracts from different scientific fields

- Entities:

```
<text id="H01-1001"> <title>Activity detection for information access to oral
communication</title> <abstract> <entity id="H01-1001.1">Oral communication</
entity> is ubiquitous and carries important information yet it is also time
consuming to document. Given the development of <entity id="H01-1001.2">
storage media and networks</entity> one could just record and store a <entity
id="H01-1001.3">conversation</entity> for documentation. The question is,
however, how an interesting information piece would be found in a <entity id="
H01-1001.4">large database</entity> . Traditional <entity id="H01-1001.5">
information retrieval techniques</entity> use a <entity id="H01-1001.6">
```

- Relations:
  - USAGE(H01-1001.5,H01-1001.7,REVERSE)
  - USAGE(H01-1001.9,H01-1001.10)
  - PART_WHOLE(H01-1001.14,H01-1001.15,REVERSE)
  - MODEL-FEATURE(H01-1017.4,H01-1017.5)

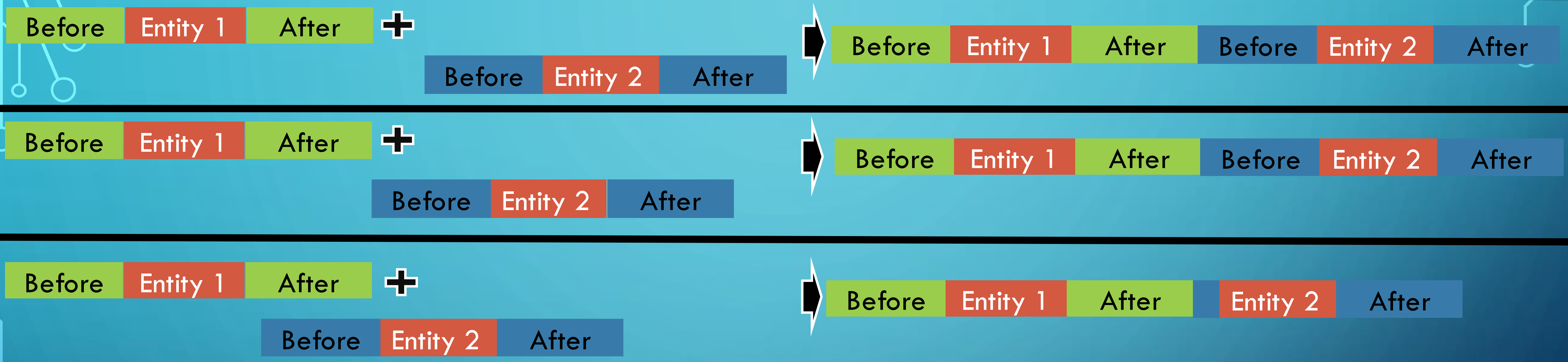
- Unbalanced relations data:

USAGE	RESULT	MODEL-FEATURE	PART-WHOLE	TOPIC	COMPARE
483	72	326	234	18	95

LSTM MODEL

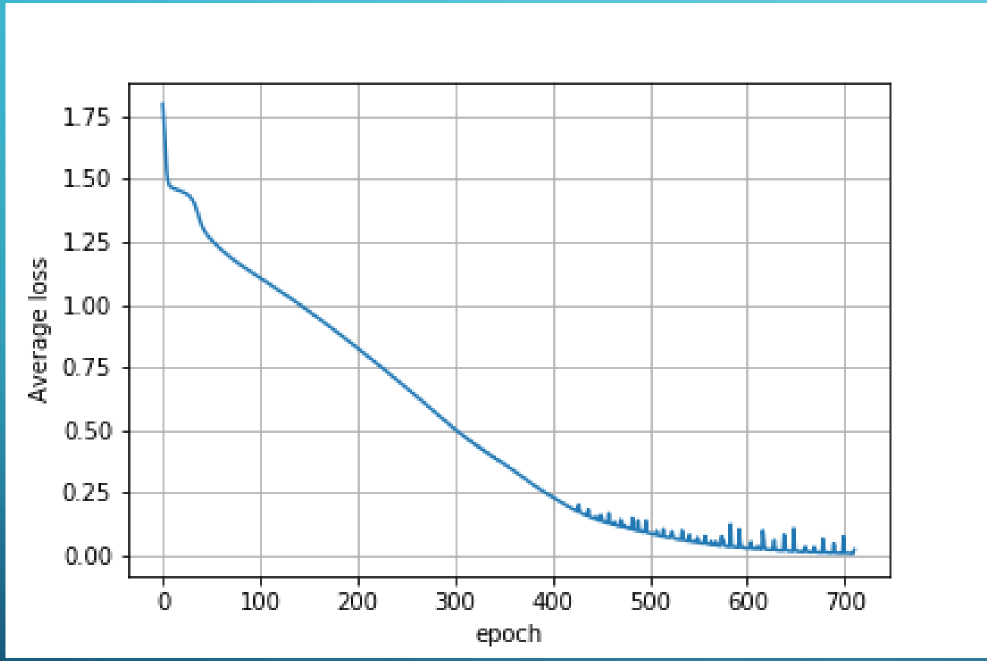
- 1 layered LSTM
- 64 dimensional hidden state
- Log-softmax transformation of the last hidden layer
- Training
  - Weighted negative log likelihood loss function
    - Weights for each relation class were calculated by: numAllRelations/numRelation
  - Stohastic gradient descent with ADAM optimization

DATA PREPARATION

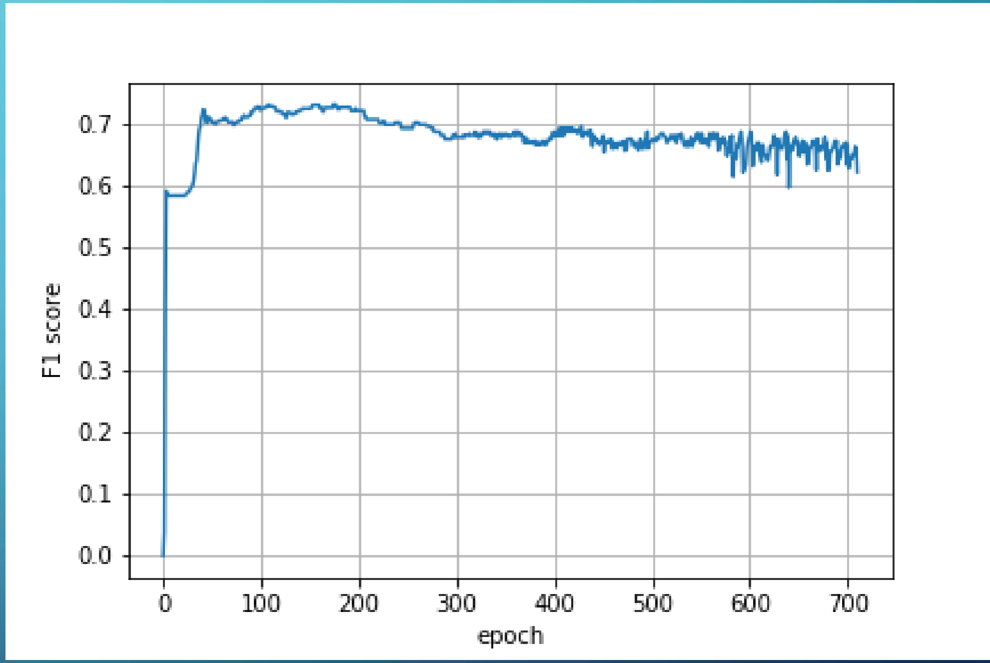


- GloVe 50 dimensional embeddings
- Mapping each word to its appropriate embedding
- If word is missing in the embedding vocabulary, ignore it

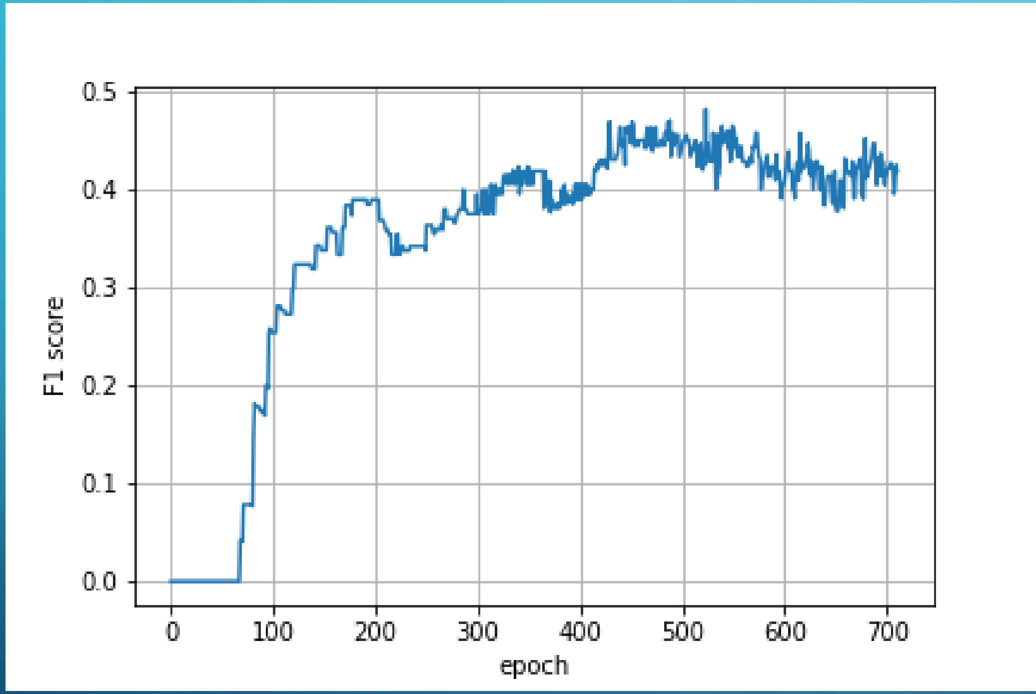
RESULTS



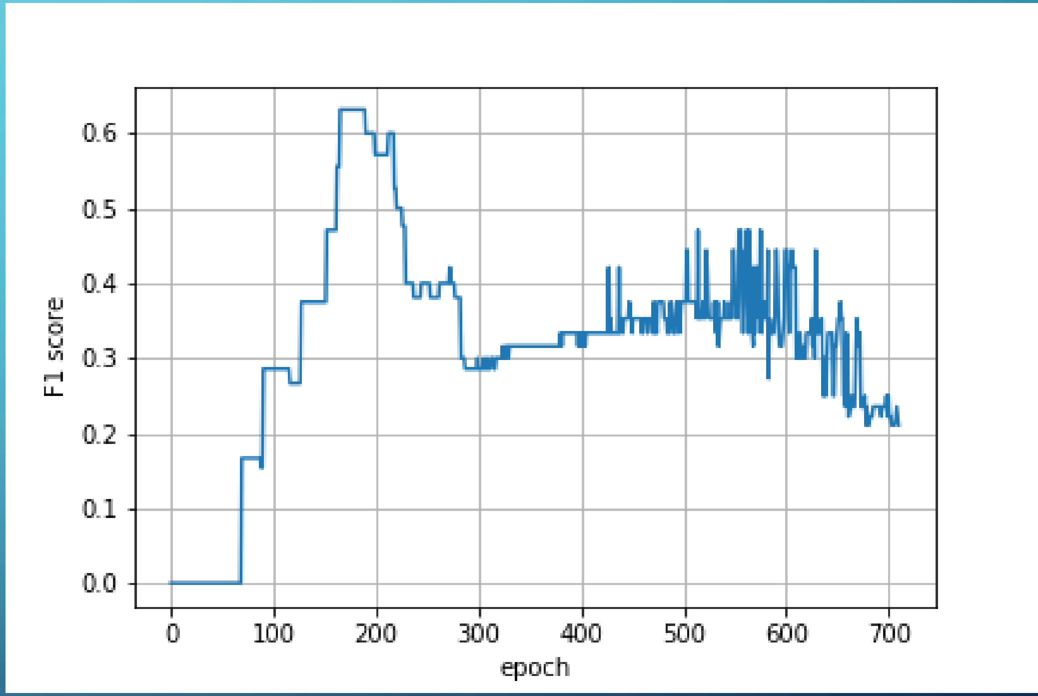
LOSS



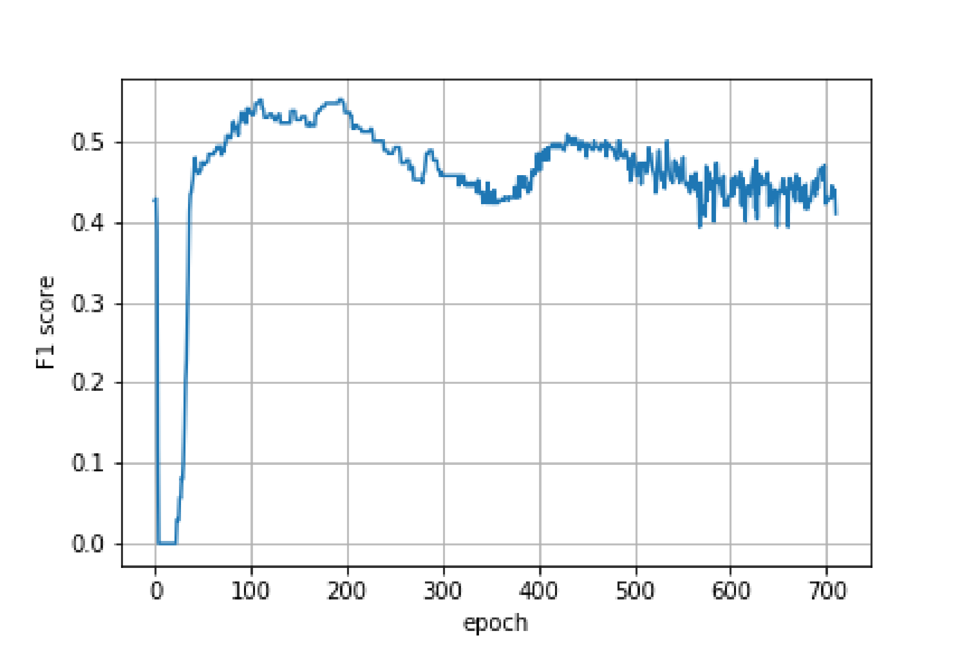
USAGE



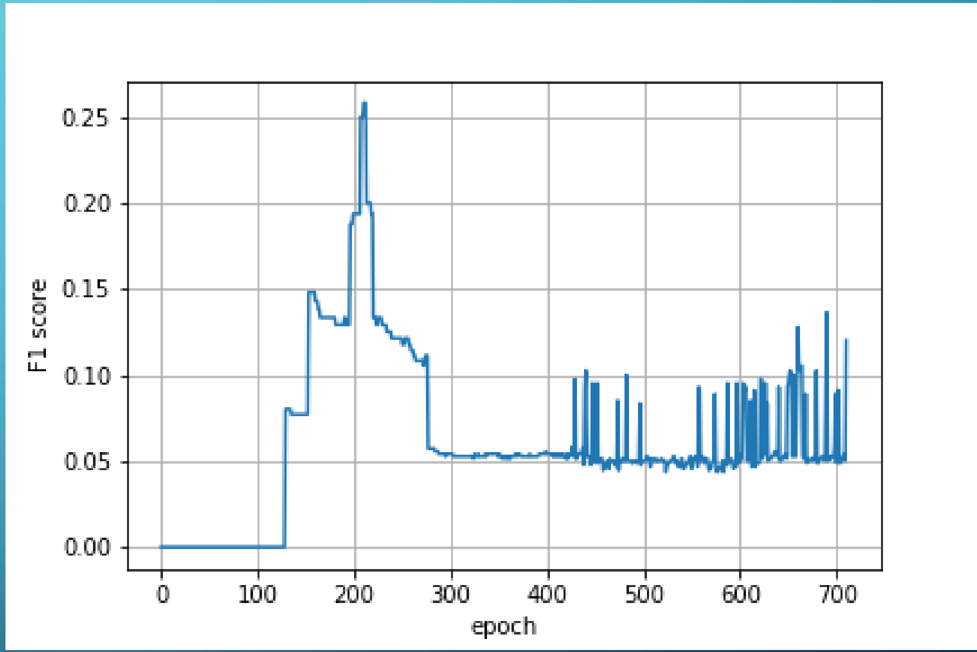
PART-WHOLE



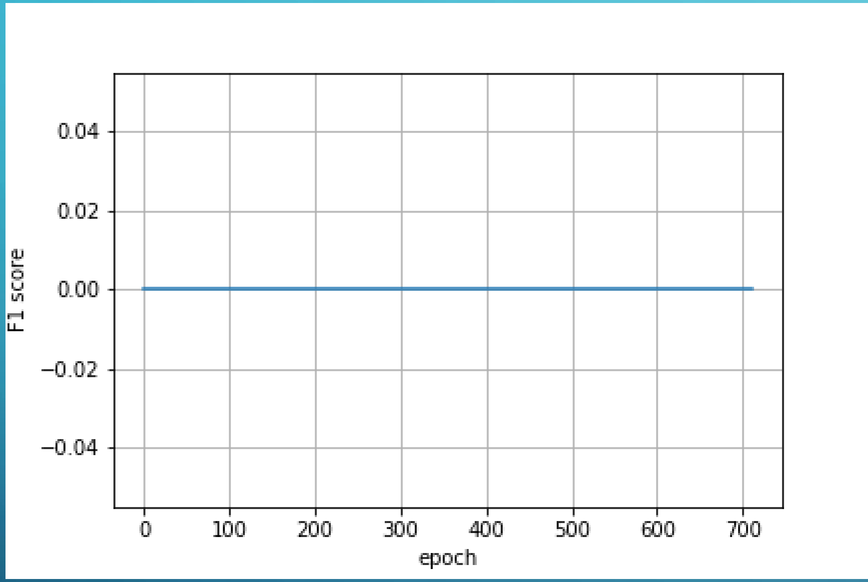
RESULT



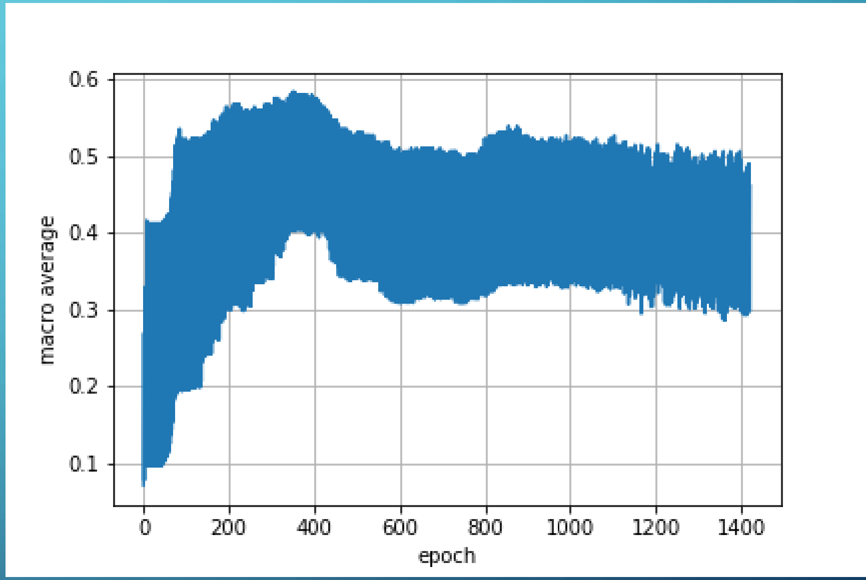
MODEL-FEATURE



COMPARE



TOPIC



MACRO AVERAGE

FEATURE PLANS

- Include POS tags and WordNet descriptions in word rpresentations
- Expermient with combination of CNNs and LSTMS
- Research sparse word representations and hierarchical temporal memory in more detile and see how it could be used in combination with other models