Improving Relation Extraction by Pre-trained Language Representations

Christoph Alt*, Marc Hübner*, Leonhard Hennig

Speech and Language Technology Group,
German Research Center for Artificial Intelligence (DFKI)

AKBC 2019 - May 20, 2019

The Task: Relation Extraction

Examples

The measure included Aerolineas's domestic subsidiary, Austral. *(org:subsidiaries)*

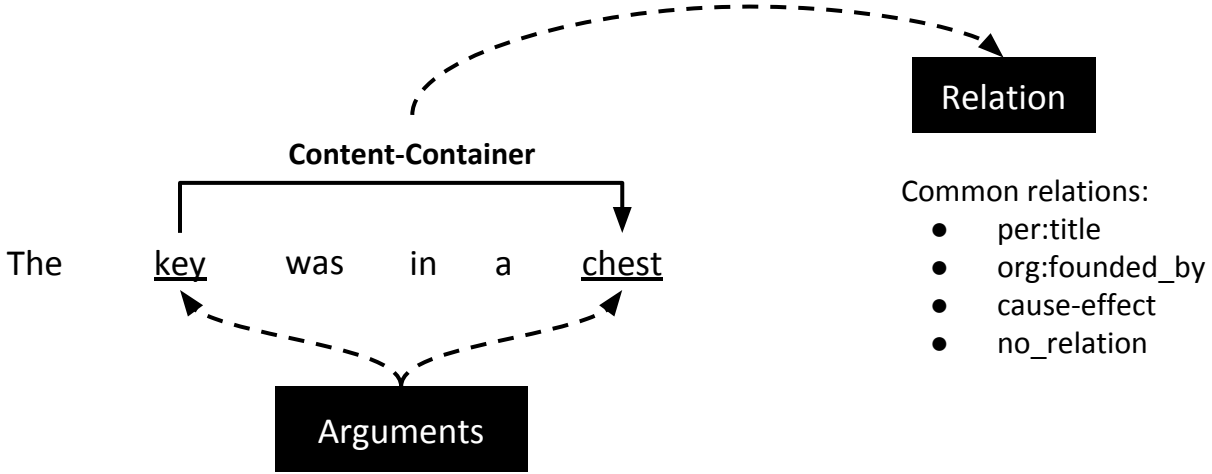
Mr. Scheider played the police chief of a resort town menaced by a shark. *(per:title)*

} TACRED

The key was in a chest. *(Content-Container)*

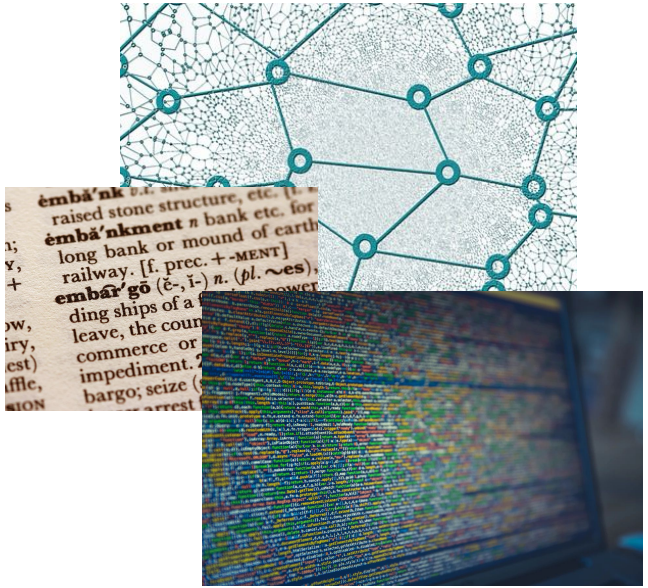
The car left the plant. *(Entity-Origin)*

} SemEval 2010 Task 8



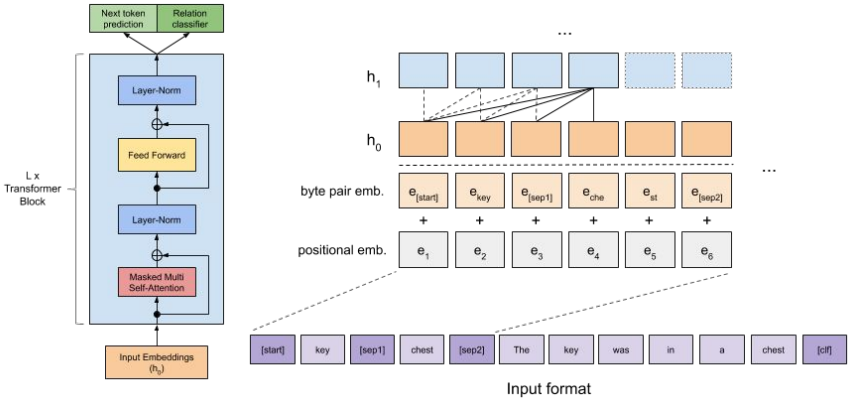
Current RE Methods ...

- **… rely on explicit features** (POS, DEP, NER, ...)
  - Requires an additional preprocessing step
  - Error propagation of automated labeling
- **… require large amounts of labeled data**
  - Typically limited training examples available
  - Specific to languages and domains
- **… rely on task-specific architectures**
  - Dataset-specific components
    (e.g. tree-pruning or piecewise splitting)
  - Requires extensive hyperparameter tuning



Our Approach

- **Relation Extraction via LM**
  - Use pre-trained OpenAI GPT
  - Specific input format for RE
  - Fine-tune a LM to the RE task
  - Reduce overfitting and catastrophic forgetting via aux. LM objective
- **Language modeling**
  - Provides implicit features via pre-training
  - Extracts syntactic and semantic knowledge from unlabeled data
- **General purpose architecture**
  - Task adaption only requires change to input format
  - *Future work:* Extend with supporting facts in natural language



Key Insights

- Improved supervised RE performance compared to state-of-the-art approaches (on TACRED and SemEval 2010 Task 8)
- Increased sample-efficiency, achieving baseline performance with only 20% of the data
- Language models perform better on generic entities (SemEval) than on named entities (TACRED)
- Entity Masking improved generalization for named entities

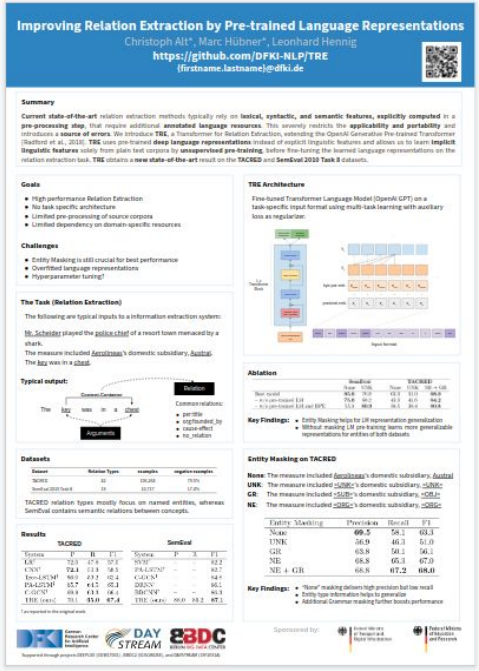
TACRED

System	P	R	F1
LR†	72.0	47.8	57.5
CNN†	**72.1**	50.3	59.2
Tree-LSTM†	66.0	59.2	62.4
PA-LSTM†	65.7	64.5	65.1
C-GCN†	69.9	63.3	66.4
TRE (ours)	70.1	**65.0**	**67.4**

SemEval 2010 Task 8

System	P	R	F1
SVM†	–	–	82.2
PA-LSTM†	–	–	82.7
C-GCN†	–	–	84.8
DRNN†	–	–	86.1
BRCNN†	–	–	86.3
TRE (ours)	88.0	86.2	**87.1**

Come to our Poster!!!



Github: https://github.com/DFKI-NLP/TRE

Contact: {christoph.alt, marc.huebner, leonhard.hennig}@dfki.de