Zero-Shot Learning in Named Entity Recognition with Common Sense Knowledge

Nguyen Van Hoang, Yang Yue, Dexter Neo Yuan Rong, Soeren Hougaard Mulvad

Department of Computer Science, National University of Singapore

Introduction

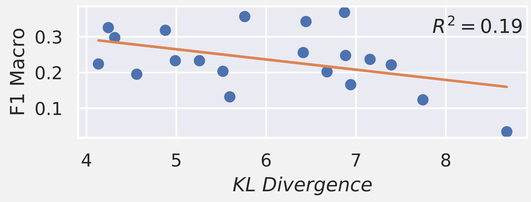
Named-entity recognition (NER) is the task of taking a block of text and annotating it with the names of the entities, such as people, organizations, time, etc. However, one of the problems of SOTA NER systems is their lack of generalization to unseen domains. Tuning NER systems to perform well in a new domain requires significant effort. We propose ZERO, a model that performs zero-shot learning in NER to **generalize to unseen domains** by incorporating pre-existing knowledge.

ZERO: Zero-shot NER

Figure 1 shows the architecture of **ZERO**. Given an input sentence, we obtain the contextual token features by passing it through LUKE (ϕ). We pass each token feature through an FCN to reduce its size to the embedding size, then compute the dot product with the embedding of each label for the output logits. The output is then passed through a softmax layer to obtain the final classification probability.

Domain Similarity vs. Performance

We run ZERO on all domain pairs to discover the **relationship** between domain similarity and performance. The KL divergence is used as a distance metric to measure similarity (the higher the less similar), and the F1 macro score is chosen as the performance metric.



The CrossNER Dataset

We use the **CrossNER dataset**. It consists of the five domains *AI*, *literature*, *music*, *politics*, and *natural science*, as well as a general news domain *Reuters* with 900-1400 labeled NER samples per domain. The domains have shared general categories such as ”person” and ”location” and their own specialized entity categories such as ”book” and ”poem” for the domain of literature.

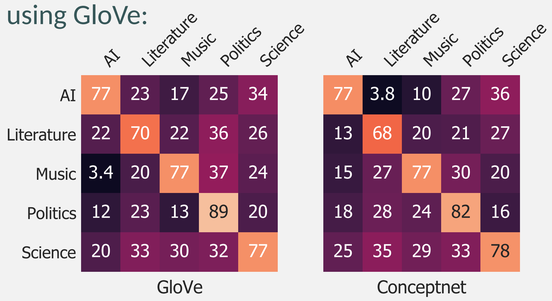
In-Domain Comparison With LUKE

We evaluate the strength of our approach by comparing it against the SOTA model **LUKE**. We find that jointly learning with additional domain features is able to achieve competitive F1 macro scores and even outperforms LUKE in the politics domain.

	AI	Literature	Music	Politics	Science
LUKE	0.7727	0.7484	0.8757	0.8709	0.7840
ZERO	0.7658	0.7177	0.8618	0.8925	0.7840

Word Embedding Comparison

Conceptnet Numberbatch is an ensemble word embedding retrofitted with the ConceptNet knowledge graph shown to be superior to **GloVe** in general. However, we observed better *macro F1 scores* of our model when using GloVe:

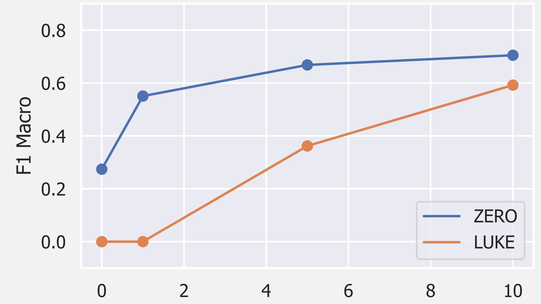


To understand why, the embeddings of all concepts were clustered into 5 clusters using *k*-Means and compared with the ground truth. GloVe performs better for our task:

	Adjusted Rand Score	V Measure
GloVe	**0.0630**	**0.1230**
Conceptnet	−0.0307	0.0971

Few-Shot Learning

We plot the performance of LUKE and ZERO against the number of examples per label used in training. Given an extremely small number of examples per label such as 1 or 5, ZERO outperforms LUKE by approximately 55% and 35% respectively.



As the number of examples per label increases the performance gap between LUKE and ZERO decreases. This shows the superiority of transferring ZERO to few-shot learning compared to a fully-supervised model like LUKE.

Conclusion

We have created ZERO, a model based on LUKE that incorporates pre-existing knowledge to perform NER. ZERO is shown to perform zero-shot learning very well while also being considerable better than the original LUKE model on few-shot learning and comparable for in-domain learning.

Model Architecture

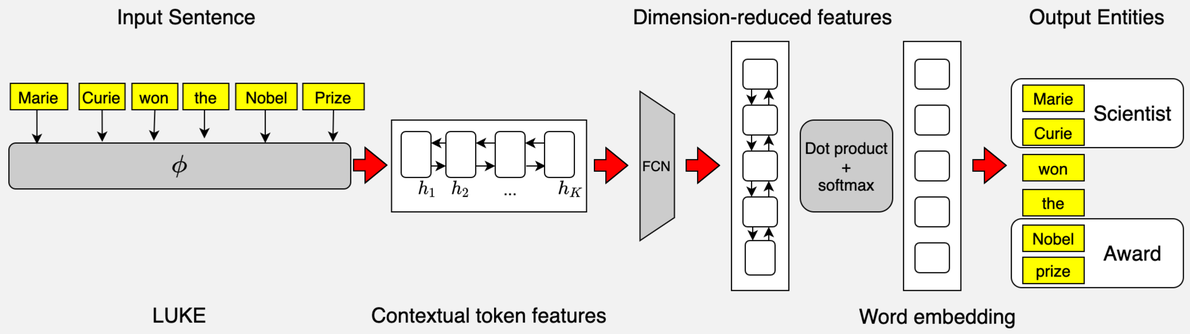


Figure 1: **ZERO** - Zero-shot learning architecture. Code accessible at `www.github.com/shmulvad/nndl2-project`.

NUS National University of Singapore