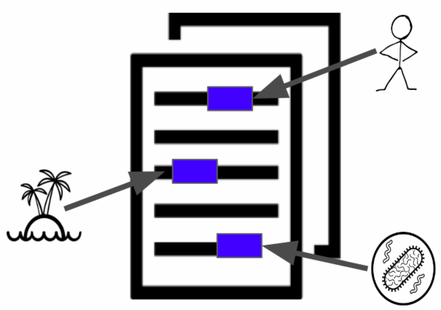
An Analysis of Deep Contextual Word Embeddings and Neural Architectures for Toponym Mention Detection in Scientific Publications (UNH at SemEval-2019 Task 12: Toponym Resolution in Scientific Papers)

Matthew Magnusson, Laura Dietz
{mfm2, dietz}@cs.unh.edu
University of New Hampshire
Department of Computer Science

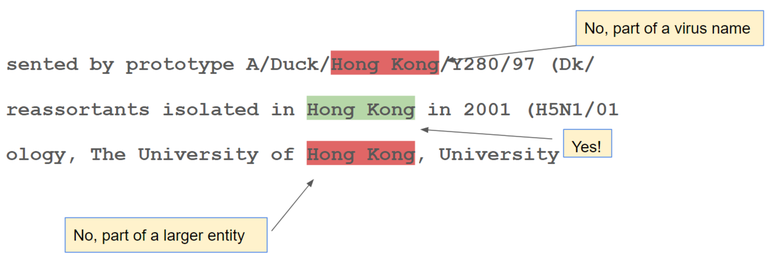
Motivation

- We started with SemEval Task 12 and expanded our work
- Large amount of knowledge stored in scientific publications
- Knowledge is often "locked up" in PDF format
- Challenging for automated information extraction and processing



Toponyms

- Toponyms are geographic place names in text
- Often found in scientific publications
- e.g. identifying virus outbreak location in journals

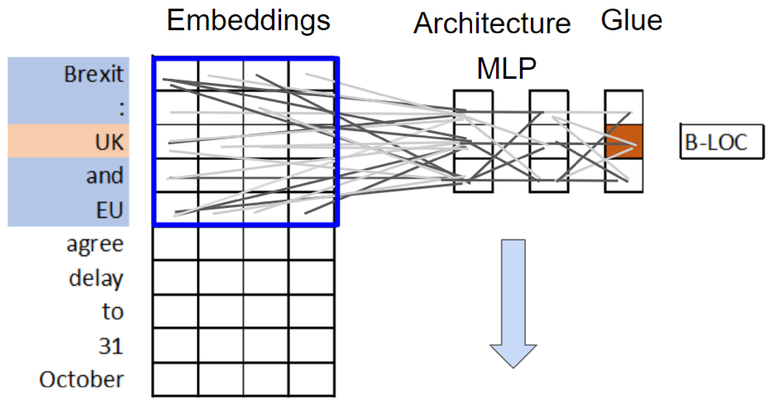


- **Task**: Given the text of a scientific publication (extracted from a PDF), label the character locations of the toponyms.

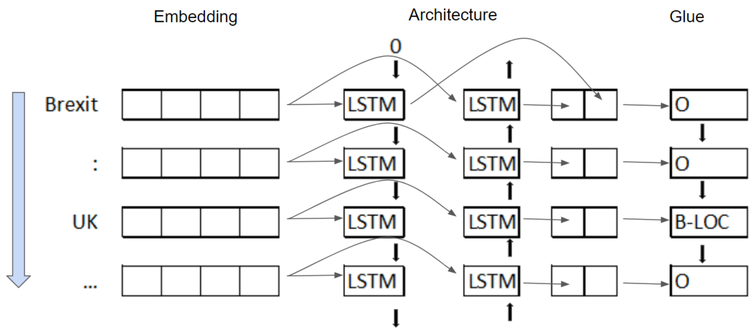
This material is based upon work supported by the National Science Foundation under Grant No. 1846017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Models

Sliding Window MLP (ELMo)



Bi-LSTM with CRF (ELMo) Huang et al. (2015)



Data Set

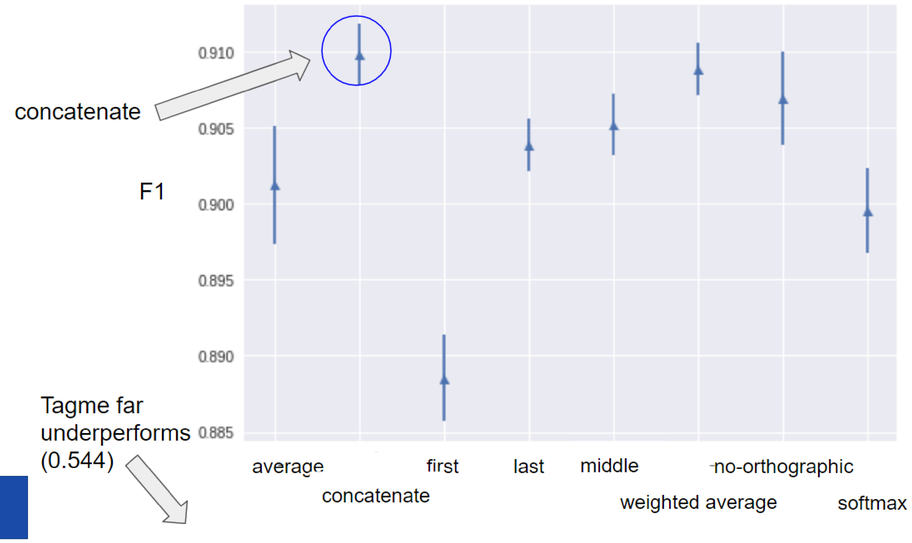
- Full texts of open access journal articles from PubMed Central (PMC) by Weissenbacher (2019)

	Documents	Tokens	Toponyms
Train	72	396,668	3,637
Valid	32	179,443	2,141
Test	45	253,159	4,616
Total	149	829,720	10,394

Results

Embedding	Model	Glue	F1
ELMo	bi-LSTM	CRF	**0.91**
ELMO +word2vec	bi-LSTM	CRF	0.91
BERT	bi-LSTM	CRF	0.9
word2vec	bi-LSTM	CRF	0.88
word2vec	MLP	Softmax	0.86
ELMo	CNN-ELMO	Softmax	0.84
	TagMe - Baseline		0.54

Standard Error on ELMo bi-LSTM model



Conclusions

- Non-contextual, domain-specific word embeddings underperform deep contextual embeddings trained on a general large-scale corpus for state-of-art bi-LSTM models.
- The neural model with the best performance is bi-LSTM with CRF using concatenated ELMo contextual embeddings.