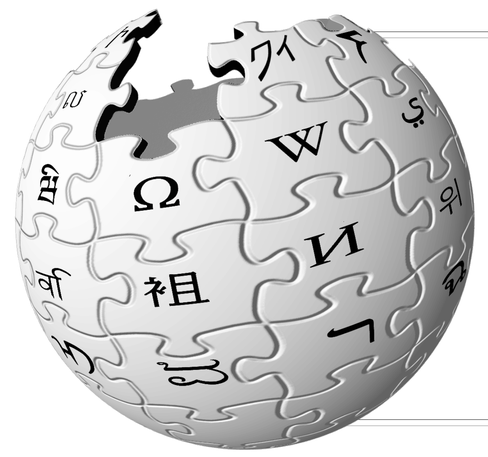
Learning Text Representations for 500K Classification Tasks on Named Entity Disambiguation

Ander Barrena, Aitor Soroa and Eneko Agirre
IXA NLP Group / University of the Basque Country, Donostia, Basque Country
ander.barrena@ehu.eus, a.soroa@ehu.eus, e.agirre@ehu.eus

We propose to break the task of NED into **500K classification tasks**, building a **word expert** model for each target mention, as opposed to a **single model** for all mentions over Wikipedia.

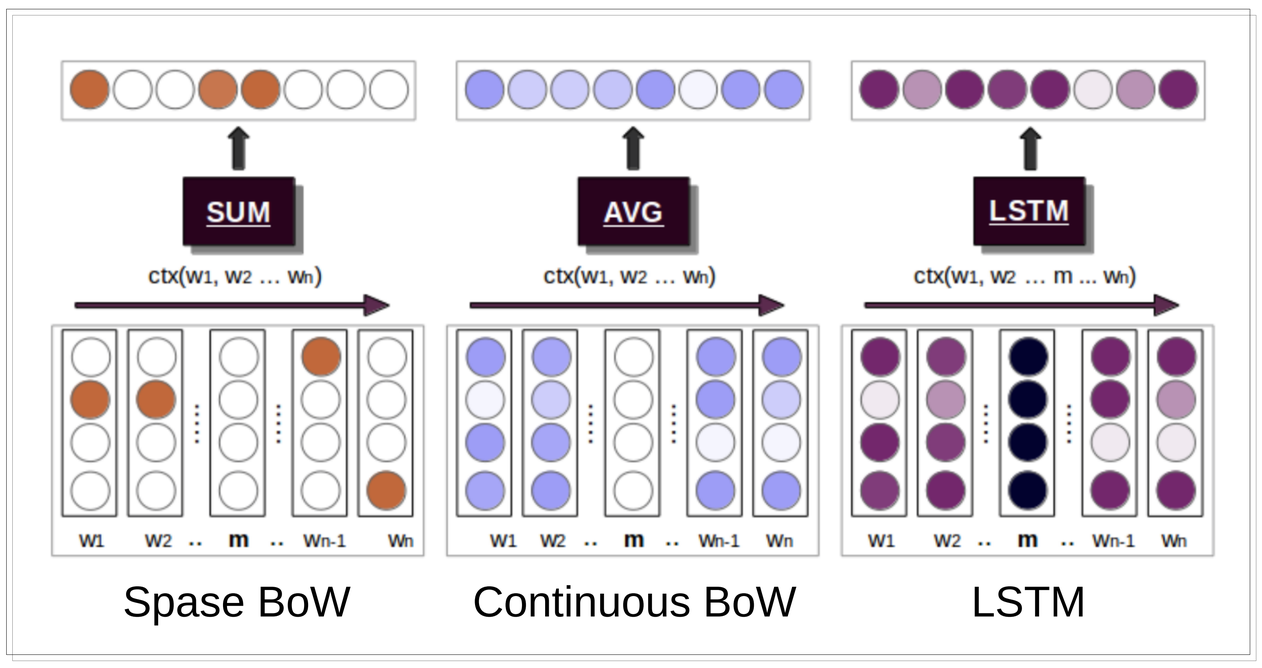
+ The Word Expert classifier needs to focus on learning a good context model for a single mention and a limited set of entities.

- Training instances follow a long tail distribution, with some mentions having a huge number of examples, but with the vast majority having very limited training data

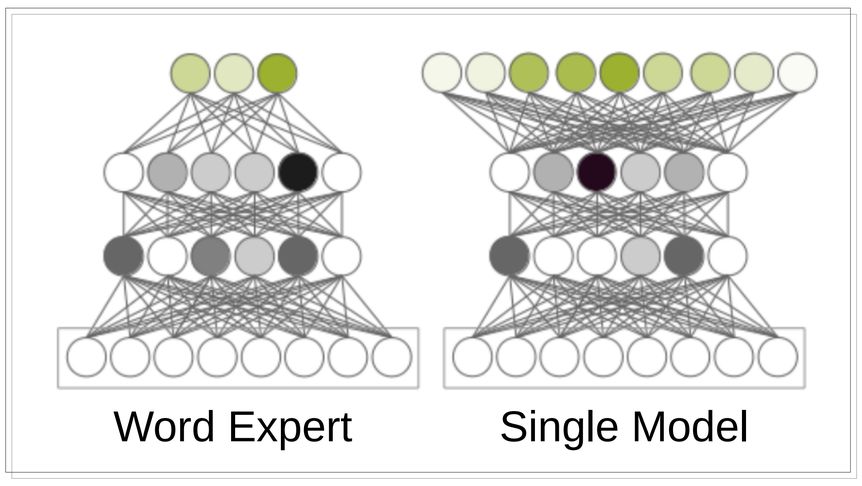


We used the **English Wikipedia** as the **only resource**. On the one hand, Wikipedia articles define the target set of entities. On the other hand, Wikipedia editors have manually added hyperlinks to articles, where the anchor text corresponds to the mention, and the url corresponds to the target entity.

Learning text representations for 500K mentions



NED models

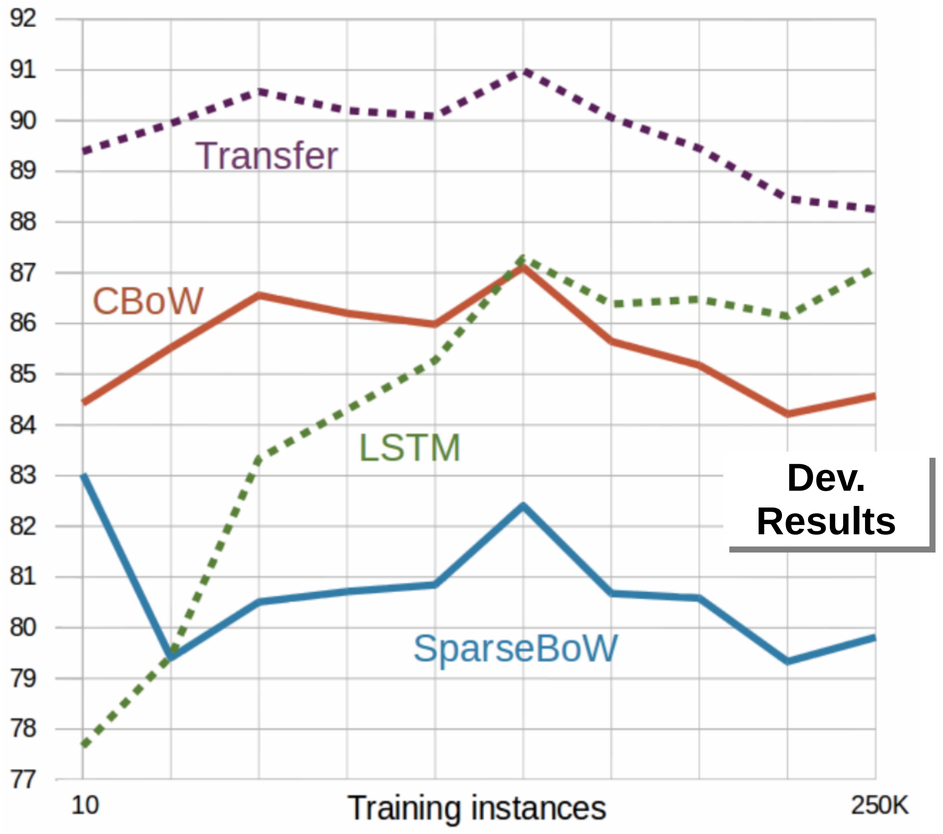


Data augmentation and Transfer Learning

Data augmentation: In order to have a larger number of training instances, we augment the training set for target mention with the contexts of other mentions that occur as anchors of one of the candidates. We combined both **original** and **augmented** classifiers, yielding better results in development.

$$P(e|\mathbf{c}) = P(e|\mathbf{c})_{orig} P(e|\mathbf{c})_{aug}$$

Transfer Learning: We reuse the context layer of the LSTM which was learned alongside the single model instead of learning a separate LSTM layer for each word expert. When training the word experts, we keep the LSTM layer fixed.



Method	AIDA (testb)
Local models	
(Lazic et al., 2015) sup.	79.7
Sparse BoW	86.72±0.23
Continuous BoW	89.39±0.44
LSTM	88.44±0.26
Transfer LSTM	**91.19**±0.07
(Lazic et al., 2015)† semi-sup.	86.4 †
(Yamada et al., 2016)*	87.2*
(Ganea and Hofmann, 2017)*	88.8*
Local & Global models	
(Chisholm and Hachey, 2015)*	88.7*
(Globerson et al., 2016)*	91.0*
(Yamada et al., 2016)*	91.5*
(Ganea and Hofmann, 2017)*	92.2*

* for systems trained in domain data.
† for systems using semi-supervised methods.

Current State of the art.

AIDA (testa) dev.	
Sparse BoW	83.28
CBoW	86.19
LSTM	84.35
Transfer LSTM	**86.87**
Single Model	45.95

Method	tac10	tac11	tac12
Local models			
(Lazic et al., 2015) sup.	—	74.5	68.7
Sparse BoW	85.82	80.25	63.12
Continuous BoW	86.96	81.55	67.49
LSTM	86.73	81.44	67.32
Transfer LSTM	**87.32**	**84.41**	**72.58**
(Lazic et al., 2015)† semi-sup.	—	79.3†	74.2†
(Chang et al., 2016)*	84.5*	—	—
(Yamada et al., 2016)*	84.6*	—	—
Local & Global models			
(Cucerzan, 2012)*	—	—	72.0*
(Chisholm and Hachey, 2015)*	80.7*	—	—
(Globerson et al., 2016)*	87.2*	84.3*	82.4*
(Yamada et al., 2016)*	85.2*		

* for systems trained in domain data.
† for systems using semi-supervised methods.

Code to reproduce results:
https://github.com/anderbarrena/500kNED