Contextualized Embeddings in Named-Entity Recognition: An Empirical Study on Generalization

Bruno Taillé [1, 2], **Vincent Guigue** [2] and **Patrick Gallinari** [2]
[1] BNP Paribas [2] Sorbonne Université, CNRS, LIP6
bruno.taille@lip6.fr

Overview

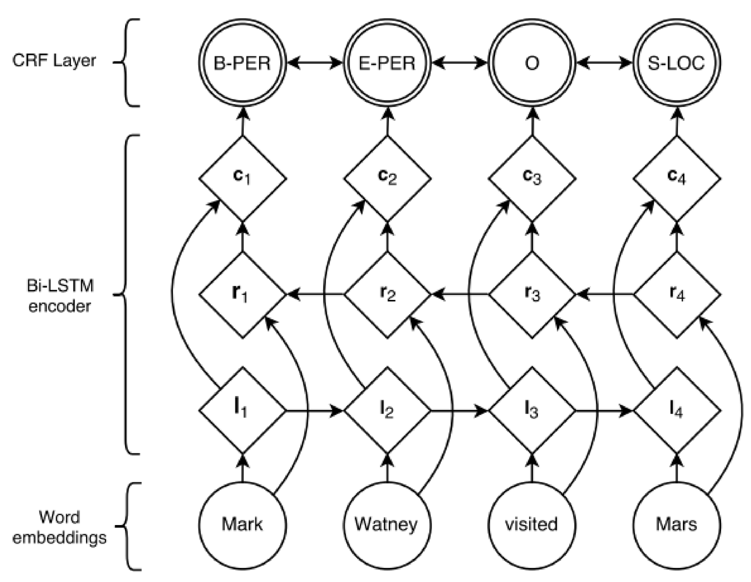
Language Model pretraining enables to compute **contextual word representations intuitively useful for generalization**, especially in Named-Entity Recognition where it is crucial to detect mentions never seen during training.

However, English NER benchmarks overestimate the importance of lexical over contextual features because of an **unrealistic lexical overlap** between train and test mentions.

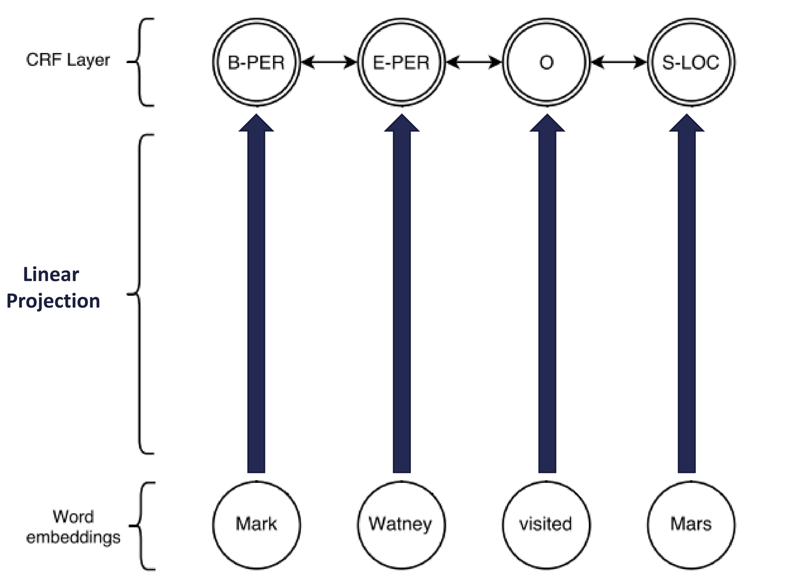
We perform an empirical analysis of the generalization capabilities of state-of-the-art contextualzed word embeddings by **separating mentions by novelty** and with **out-of-domain evaluation from CoNLL03 to OntoNotes and WNUT.**

In such setting, we show that **Language Model contextualization is particularly beneficial for unseen mentions detection, especially out-of-domain**.

Named-Entity Recognition

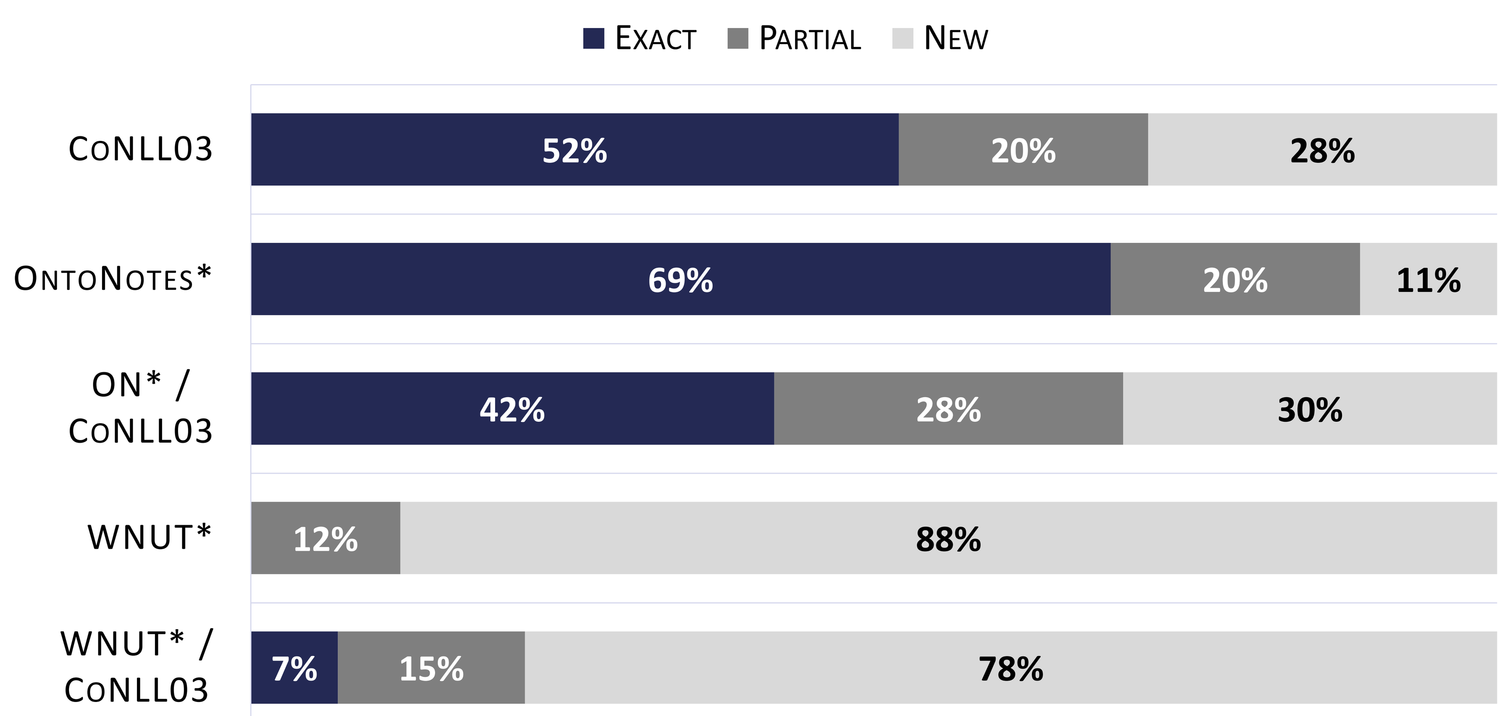


BiLSTM-CRF



Map-CRF

Lexical Overlap



Lexical overlap of test mentions with training mentions in-domain and out-of-domain when training on CoNLL03 and testing on OntoNotes or WNUT with remapped entity tags.

Explored Word Representations

Lexical : **GloVe** (Pennington 2014)

Morphological : **charBiLSTM** (Lample 2016)

ELMo[0] = charCNN from ELMo

Contextual :

ELMo (Peters 2018)
- charCNN word representation
- **Word-level** BiLSTM LM
- Fusion by weighted sum of layers

Flair (Akbik 2018)
- **Char-level** BiLSTM LM
- Concatenation of 1st and last character states

BERT (Devlin 2019)
- **Subword-level** Transformer LM
- Feature-based $BERT_{LARGE}$: LM is frozen

Results

Table 2. In-domain micro-F1 scores of the BiLSTM-CRF architecture on CoNLL03 and OntoNotes*. Results are averaged over 5 runs. Contextual embeddings are over the dashed line.

		CoNLL03				OntoNotes*				WNUT*		
Embedding	Dim	EM	PM	New	All	EM	PM	New	All	PM	New	All
BERT	4096	95.7	88.8	82.2	90.5	96.9	88.6	81.1	**93.5**	77.0	53.9	**57.0**
ELMo	1024	95.9	89.2	85.8	**91.8**	97.1	88.0	79.9	**93.4**	67.7	49.5	52.1
Flair	4096	95.4	88.1	83.5	90.6	96.7	85.8	75.0	92.1	64.9	48.2	50.4
ELMo[0]	1024	95.8	87.2	83.5	90.7	96.9	85.9	75.5	92.4	72.8	45.4	49.1
GloVe + char	350	95.3	85.5	83.1	89.9	96.3	83.3	69.9	91.0	63.2	33.4	38.0
GloVe	300	95.1	85.3	81.1	89.3	96.2	82.9	63.8	90.4	59.1	28.1	32.9

Lexical overlap bias $F1_{EXACT} > F1_{PARTIAL} > F1_{NEW}$ with a wider gap out-of-domain

ELMo vs BERT vs Flair ELMo seems more stable
Flair's char-level LM is less robust to domain adaptation

ELMo[0] vs GloVe + char ELMo[0] already is an improvement over GloVe + charBiLSTM

Two contextualizations C_{NER} supervised with NER = Map-CRF to BiLSTM-CRF
C_{LM} from unsupervised LM = ELMo[0] to ELMo

Both improve generalization to unseen mentions in and out-of-domain
C_{LM} is more beneficial than C_{NER} out-of-domain, especially on genres far from source
C_{NER} and C_{LM} are complementary except in the difficult domain adaptation to WNUT*

Table 3. Micro-F1 scores of models trained on CoNLL03 and tested in-domain and out-of-domain on OntoNotes* and WNUT*. Results are averaged over 5 runs.

		CoNLL03				OntoNotes*				WNUT*			
	Emb	EM	PM	New	All	EM	PM	New	All	EM	PM	New	All
BiLSTM-CRF	BERT	95.7	88.8	82.2	90.5	95.1	82.9	73.5	**85.0**	57.4	56.3	32.4	37.6
	ELMo	95.9	89.2	85.8	**91.8**	94.3	79.2	72.4	83.4	55.8	52.7	36.5	**41.0**
	Flair	95.4	88.1	83.5	90.6	94.0	76.1	62.1	79.0	56.2	49.4	29.1	34.9
	ELMo[0]	95.8	87.2	83.5	90.7	93.6	76.8	66.1	80.5	52.3	50.8	32.6	37.6
	G + char	95.3	85.5	83.1	89.9	93.9	73.9	60.4	77.9	55.9	46.8	19.6	27.2
	GloVe	95.1	85.3	81.1	89.3	93.7	73.0	57.4	76.9	53.9	46.3	13.3	27.1
Map-CRF	BERT	93.2	85.8	73.7	86.2	93.5	77.8	67.8	80.9	57.4	53.5	33.9	38.4
	ELMo	93.7	87.2	80.1	**88.7**	93.6	79.1	69.5	**82.2**	61.1	53.0	37.5	**42.4**
	Flair	94.3	85.1	78.6	88.1	93.2	74.0	59.6	77.5	52.5	50.6	28.8	33.7
	ELMo[0]	92.2	80.5	68.6	83.4	91.6	69.6	56.8	75.0	51.9	42.6	32.4	35.8
	G + char	93.1	80.7	69.8	84.4	91.8	69.3	55.6	74.8	50.6	42.5	20.6	28.7
	GloVe	92.2	77.0	61.7	81.5	89.6	62.8	38.5	68.1	46.8	41.3	3.2	18.9

Table 4. Per-genre micro-F1 scores of the BiLSTM-CRF model trained on CoNLL03 and tested on OntoNotes* (broadcast conversation, broadcast news, news wire, magazine, telephone conversation and web text). C_{LM} mostly benefits genres furthest from the news source domain.

	bc	bn	nw	mz	tc	wb	All
BERT	87.2	88.4	84.7	82.4	84.5	79.5	**85.0**
ELMo	85.0	88.6	82.9	78.1	84.0	79.9	83.4
Flair	78.0	86.5	80.4	71.1	73.5	72.1	79.0
ELMo[0]	82.6	88.0	79.6	73.4	79.2	75.1	80.5
GloVe + char	80.4	86.3	77.0	70.7	79.7	69.2	77.9

References

Akbik, A., Blythe, D., & Vollgraf, R. Contextual string embeddings for sequence labeling. COLING 2018
Augenstein, I., Derczynski, L., & Bontcheva, K.
Generalisation in named entity recognition: A quantitative analysis. Computer Speech & Language 2017
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.
BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL 2019
Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C.
Neural architectures for named entity recognition. NAACL 2016
Moosavi, N. S., & Strube, M. Lexical features in coreference resolution: To be used with caution. ACL 2017
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L.
Deep contextualized word representations. NAACL 2018