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

Bruno Taillé[1,2], Vincent Guigue[1], and Patrick Gallinari[1,3]

[1] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6
[2] BNP Paribas
[3] Criteo AI Lab, Paris
{bruno.taille, vincent.guigue, patrick.gallinari}@lip6.fr

Abstract. Contextualized embeddings use unsupervised language model pretraining to compute word representations depending on their context. This is intuitively useful for generalization, especially in Named-Entity Recognition where it is crucial to detect mentions never seen during training. However, standard English benchmarks overestimate the importance of lexical over contextual features because of an unrealistic lexical overlap between train and test mentions. In this paper, we perform an empirical analysis of the generalization capabilities of state-of-the-art contextualized embeddings by separating mentions by novelty and with out-of-domain evaluation. We show that they are particularly beneficial for unseen mentions detection, especially out-of-domain. For models trained on CoNLL03, language model contextualization leads to a +1.2% maximal relative micro-F1 score increase in-domain against +13% out-of-domain on the WNUT dataset[1].

Keywords: NER · Contextualized Embeddings · Domain Adaptation

1 Introduction

Named-Entity Recognition (NER) consists in detecting textual mentions of entities and classifying them into predefined types. It is modeled as sequence labeling, the standard neural architecture of which is BiLSTM-CRF [8]. Recent improvements mainly stem from using new types of representations: learned character-level word embeddings [9] and contextualized embeddings derived from a language model (LM) [1, 6, 14].

LM pretraining enables to obtain contextual word representations and reduce the dependency of neural networks on hand-labeled data specific to tasks or domains [7, 16]. This contextualization ability can particularly benefit to NER domain adaptation which is often limited to training a network on source data and either feeding its predictions to a new classifier or finetuning it on target data [10, 17]. All the more as classical NER models have been shown to poorly generalize to unseen mentions or domains [2].

In this paper, we quantify the impact of ELMo [14], Flair [1] and BERT [6] representations on generalization to unseen mentions and new domains in NER. To better understand their effectiveness, we propose a set of experiments to distinguish the effect of unsupervised LM contextualization (C_{LM}) from task supervised contextualization (C_{NER}). We show that the former mainly benefits unseen mentions detection, all the more out-of-domain where it is even more beneficial than the latter.

[1] The code is available at https://github.com/btaille/contener

2 Lexical Overlap

Neural NER models mainly rely on lexical features in the form of word embeddings, either learned at the character-level or not. Yet, standard NER benchmarks present a large lexical overlap between mentions in the train set and dev / test sets which leads to a poor evaluation of generalization to unseen mentions as shown by Augenstein et al. [2]. They separate seen from unseen mentions and evaluate out-of-domain to focus on generalization but only study models designed before 2011 and no longer in use.

We propose to use a similar setting to analyze the impact of state-of-the-art LM pretraining methods on generalization in NER. We introduce a slightly more fine-grained novelty partition by separating unseen mentions in *partial match* and *new* categories. A mention is an *exact match* (EM) if it appears in the exact same case-sensitive form in the train set, tagged with the same type. It is a *partial match* (PM) if at least one of its non stop words appears in a mention of same type. Every other mentions are *new*.

We study lexical overlap in CoNLL03 [18] and OntoNotes [20], the two main English NER datasets, as well as WNUT17 [5] which is smaller, specific to user generated content (tweets, comments) and was designed without exact overlap. For out-of-domain evaluation, we train on CoNLL03 (news articles) and test on the larger and more diverse OntoNotes (see Table 4 for genres) and the very specific WNUT. We remap OntoNotes and WNUT entity types to match CoNLL03's [1] and denote the obtained dataset with *.

Table 1. Per type lexical overlap of test mention occurrences with respective train set in-domain and with CoNLL03 train set in the out-of-domain scenario. (EM / PM = *exact / partial match*)

		CoNLL03					ON	OntoNotes*					WNUT	WNUT*			
		LOC	MISC	ORG	PER	ALL	ALL	LOC	MISC	ORG	PER	ALL	ALL	LOC	ORG	PER	ALL
Self	EM	82%	67%	54%	14%	52%	67%	87%	93%	54%	49%	69%	-	-	-	-	-
	PM	4%	11%	17%	43%	20%	24%	6%	2%	32%	36%	20%	12%	11%	5%	13%	12%
	New	14%	22%	29%	43%	28%	9%	7%	5%	14%	15%	11%	88%	89%	95%	87%	88%
CoNLL	EM	-	-	-	-	-	-	70%	78%	18%	16%	42%	-	26%	8%	1%	7%
	PM	-	-	-	-	-	-	7%	10%	45%	46%	28%	-	9%	15%	16%	14%
	New	-	-	-	-	-	-	23%	12%	38%	38%	30%	-	65%	77%	83%	78%

As reported in Table 1, the two main benchmarks for English NER mainly evaluate performance on occurrences of mentions already seen during training, although they appear in different sentences. Such lexical overlap proportions are unrealistic in real-life where the model must process orders of magnitude more documents in the inference phase than it has been trained on, to amortize the annotation cost. On the contrary, WNUT proposes a particularly challenging low-resource setting with no exact overlap.

Furthermore, the overlap depends on the entity types: Location and Miscellaneous are the most overlapping types, even out-of-domain, whereas Person and Organization present a more varied vocabulary, also more subject to evolve with time and domain.

[1] We map LOC + GPE in OntoNotes to LOC in CoNLL03 and NORP + LANGUAGE to MISC. Other mappings are self-explanatory. We drop types that have no correspondence.

3 Word Representations

Word embeddings map each word to a single vector which results in a lexical representation. We take **GloVe 840B** embeddings [13] trained on Common Crawl as the pretrained word embeddings baseline and fine-tune them as done in related work.

Character-level word embeddings are learned by a word-level neural network from character embeddings to incorporate orthographic and morphological features. We reproduce the **Char-BiLSTM** from [9]. It is trained jointly with the NER model and its outputs are concatenated to GloVe embeddings. We also experiment with the Char-CNN layer from ELMo to isolate the effect of LM contextualization and denote it **ELMo[0]**.

Contextualized word embeddings take into account the context of a word in its representation, contrary to previous representations. A LM is pretrained and used to predict the representation of a word given its context. **ELMo** [14] uses a Char-CNN to obtain a context-independent word embedding and the concatenation of a forward and backward two-layer LSTM LM for contextualization. These representations are summed with weights learned for each task as the LM is frozen after pretraining. **BERT** [6] uses WordPiece subword embeddings [21] and learns a representation modeling both left and right contexts by training a Transformer encoder [19] for Masked LM and next sentence prediction. For a fairer comparison, we use the $\text{BERT}_{\text{LARGE}}$ feature-based approach where the LM is not fine-tuned and its last four hidden layers are concatenated. **Flair** [1] uses a character-level LM for contextualization. As in ELMo, they train two opposite LSTM LMs, freeze them and concatenate the predicted states of the first and last characters of each word. Flair and ELMo are pretrained on the 1 Billion Word Benchmark [3] while BERT uses Book Corpus [22] and English Wikipedia.

4 Experiments

In order to compare the different embeddings, we feed them as input to a classifier. We first use the state-of-the-art BiLSTM-CRF [8] with hidden size 100 in each direction and present in-domain results on all datasets in Table 2.

We then report out-of-domain performance in Table 3. To better capture the intrinsic effect of LM contextualization, we introduce the Map-CRF baseline from [1] where the BiLSTM is replaced by a simple linear projection of each word embedding. We only consider domain adaptation from CoNLL03 to OntoNotes* and WNUT* assuming that labeled data is scarcer, less varied and more generic than target data in real use cases.

We use the IOBES tagging scheme for NER and no preprocessing. We fix a batch size of 64, a learning rate of 0.001 and a 0.5 dropout rate at the embedding layer and after the BiLSTM or linear projection. The maximum number of epochs is set to 100 and we use early stopping with patience 5 on validation global micro-F1. For each configuration, we use the best performing optimization method between SGD and Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We report the mean and standard deviation of five runs.

Table 2. In-domain micro-F1 scores of the BiLSTM-CRF. We split mentions by novelty: *exact match* (EM), *partial match* (PM) and *new*. Average of 5 runs, subscript denotes standard deviation.

		CoNLL03				OntoNotes*				WNUT*		
Embedding	Dim	EM	PM	New	All	EM	PM	New	All	PM	New	All
BERT	4096	$95.7_{.1}$	$88.8_{.3}$	$82.2_{.3}$	$90.5_{.1}$	$96.9_{.2}$	$88.6_{.3}$	$81.1_{.5}$	$\mathbf{93.5}_{.2}$	$77.0_{4.6}$	$53.9_{.9}$	$\mathbf{57.0}_{1.0}$
ELMo	1024	$95.9_{.1}$	$89.2_{.5}$	$85.8_{.7}$	$\mathbf{91.8}_{.3}$	$97.1_{.2}$	$88.0_{.2}$	$79.9_{.7}$	$\mathbf{93.4}_{.2}$	$67.7_{3.2}$	$49.5_{.9}$	$52.1_{1.0}$
Flair	4096	$95.4_{.1}$	$88.1_{.6}$	$83.5_{.5}$	$90.6_{.2}$	$96.7_{.1}$	$85.8_{.5}$	$75.0_{.6}$	$92.1_{.2}$	$64.9_{.7}$	$48.2_{2.0}$	$50.4_{1.8}$
ELMo[0]	1024	$95.8_{.1}$	$87.2_{.2}$	$83.5_{.4}$	$90.7_{.1}$	$96.9_{.1}$	$85.9_{.3}$	$75.5_{.6}$	$92.4_{.1}$	$72.8_{1.3}$	$45.4_{2.8}$	$49.1_{2.3}$
GloVe + char	350	$95.3_{.3}$	$85.5_{.7}$	$83.1_{.7}$	$89.9_{.5}$	$96.3_{.1}$	$83.3_{.2}$	$69.9_{.6}$	$91.0_{.1}$	$63.2_{4.6}$	$33.4_{1.5}$	$38.0_{1.7}$
GloVe	300	$95.1_{.4}$	$85.3_{.5}$	$81.1_{.5}$	$89.3_{.4}$	$96.2_{.2}$	$82.9_{.2}$	$63.8_{.5}$	$90.4_{.2}$	$59.1_{2.9}$	$28.1_{1.5}$	$32.9_{1.2}$

4.1 General Observations

ELMo, BERT and Flair Drawing conclusions from the comparison of ELMo, BERT and Flair is difficult because there is no clear hierarchy accross datasets and they differ in dimensions, tokenization, contextualization levels and pretraining corpora. However, although BERT is particularly effective on the WNUT dataset in-domain, probably due to its subword tokenization, ELMo yields the most stable results in and out-of-domain.

Furthermore, Flair globally underperforms ELMo and BERT, particularly for unseen mentions and out-of-domain. This suggests that LM pretraining at a lexical level (word or subword) is more robust for generalization than at a character level. In fact, Flair only beats the non contextual ELMo[0] baseline with Map-CRF which indicates that character-level contextualization is less beneficial than word-level contextualization with character-level representations.

ELMo[0] vs GloVe+char Overall, ELMo[0] outperforms the GloVe+char baseline, particularly on unseen mentions, out-of-domain and on WNUT*. The main difference is the incorporation of morphological features: in ELMo[0] they are learned jointly with the LM on a huge dataset whereas the char-BiLSTM is only trained on the source NER training set. Yet, morphology is crucial to represent words never encountered during pretraining and in WNUT* around 20% of words in test mentions are out of GloVe vocabulary against 5% in CoNLL03 and 3% in OntoNotes*. This explains the poor performance of GloVe baselines on WNUT*, all the more out-of-domain, and why a model trained on CoNLL03 with ELMo outperforms one trained on WNUT* with GloVe+char. Thus, ELMo's improvement over previous state-of-the-art does not only stem from contextualization but also an effective non-contextual word representation.

Seen Mentions Bias In every configuration, $F1_{exact} > F1_{partial} > F1_{new}$ with more than 10 points difference. This gap is wider out-of-domain where the context differs more from training data than in-domain. NER models thus poorly generalize to unseen mentions, and datasets with high lexical overlap only encourage this behavior. However, this generalization gap is reduced by two types of contextualization described hereafter.

Table 3. Micro-F1 scores of models trained on CoNLL03 and tested in-domain and out-of-domain on OntoNotes* and WNUT*. Average of 5 runs, subscript denotes standard deviation.

		CoNLL03				OntoNotes*				WNUT*			
	Emb	EM	PM	New	All	EM	PM	New	All	EM	PM	New	All
BiLSTM-CRF	BERT	$95.7_{.1}$	$88.8_{.3}$	$82.2_{.3}$	$90.5_{.1}$	$95.1_{.1}$	$82.9_{.5}$	$73.5_{.4}$	**$85.0_{.3}$**	$57.4_{1.0}$	$56.3_{1.2}$	$32.4_{.8}$	$37.6_{.8}$
	ELMo	$95.9_{.1}$	$89.2_{.5}$	$85.8_{.7}$	**$91.8_{.3}$**	$94.3_{.1}$	$79.2_{.2}$	$72.4_{.4}$	$83.4_{.2}$	$55.8_{1.2}$	$52.7_{1.1}$	$36.5_{1.5}$	**$41.0_{1.2}$**
	Flair	$95.4_{.1}$	$88.1_{.6}$	$83.5_{.5}$	$90.6_{.2}$	$94.0_{.3}$	$76.1_{1.1}$	$62.1_{.5}$	$79.0_{.5}$	$56.2_{2.2}$	$49.4_{3.4}$	$29.1_{3.3}$	$34.9_{2.9}$
	ELMo[0]	$95.8_{.1}$	$87.2_{.2}$	$83.5_{.4}$	$90.7_{.1}$	$93.6_{.1}$	$76.8_{.6}$	$66.1_{.3}$	$80.5_{.2}$	$52.3_{1.2}$	$50.8_{1.5}$	$32.6_{2.2}$	$37.6_{1.8}$
	G + char	$95.3_{.3}$	$85.5_{.7}$	$83.1_{.7}$	$89.9_{.5}$	$93.9_{.2}$	$73.9_{1.1}$	$60.4_{.7}$	$77.9_{.5}$	$55.9_{.8}$	$46.8_{1.8}$	$19.6_{1.6}$	$27.2_{1.3}$
	GloVe	$95.1_{.4}$	$85.3_{.5}$	$81.1_{.5}$	$89.3_{.4}$	$93.7_{.2}$	$73.0_{1.2}$	$57.4_{1.8}$	$76.9_{.9}$	$53.9_{1.2}$	$46.3_{1.5}$	$13.3_{1.4}$	$27.1_{1.0}$
Map-CRF	BERT	$93.2_{.3}$	$85.8_{.4}$	$73.7_{.8}$	$86.2_{.4}$	$93.5_{.2}$	$77.8_{.5}$	$67.8_{.9}$	$80.9_{.4}$	$57.4_{.3}$	$53.5_{2.6}$	$33.9_{.6}$	$38.4_{.4}$
	ELMo	$93.7_{.2}$	$87.2_{.6}$	$80.1_{.3}$	**$88.7_{.2}$**	$93.6_{.1}$	$79.1_{.5}$	$69.5_{.4}$	**$82.2_{.3}$**	$61.1_{.7}$	$53.0_{.9}$	$37.5_{.7}$	**$42.4_{.6}$**
	Flair	$94.3_{.1}$	$85.1_{.3}$	$78.6_{.3}$	$88.1_{.03}$	$93.2_{.1}$	$74.0_{.3}$	$59.6_{.2}$	$77.5_{.2}$	$52.5_{1.2}$	$50.6_{.4}$	$28.8_{.5}$	$33.7_{.5}$
	ELMo[0]	$92.2_{.3}$	$80.5_{1.0}$	$68.6_{.4}$	$83.4_{.4}$	$91.6_{.4}$	$69.6_{1.0}$	$56.8_{1.5}$	$75.0_{1.0}$	$51.9_{1.1}$	$42.6_{.9}$	$32.4_{.3}$	$35.8_{.4}$
	G + char	$93.1_{.3}$	$80.7_{.9}$	$69.8_{.7}$	$84.4_{.4}$	$91.8_{.3}$	$69.3_{.3}$	$55.6_{1.1}$	$74.8_{.5}$	$50.6_{.9}$	$42.5_{1.4}$	$20.6_{2.8}$	$28.7_{2.5}$
	GloVe	$92.2_{.1}$	$77.0_{.4}$	$61.7_{.3}$	$81.5_{.05}$	$89.6_{.3}$	$62.8_{.6}$	$38.5_{.4}$	$68.1_{.4}$	$46.8_{.8}$	$41.3_{.5}$	$3.2_{.2}$	$18.9_{.7}$

4.2 LM and NER Contextualizations

The ELMo[0] and Map-CRF baselines enable to strictly distinguish contextualization due to LM pretraining (C_{LM}: ELMo[0] to ELMo) from task supervised contextualization induced by the BiLSTM network (C_{NER}: Map to BiLSTM). In both cases, a BiLSTM incorporates syntactic information which improves generalization to unseen mentions for which context is decisive, as shown in Table 3.

Comparison However, because C_{NER} is specific to the source dataset, it is more effective in-domain whereas C_{LM} is particularly helpful out-of-domain. In the latter setting, the benefits from C_{LM} even surpass those from C_{NER}, specifically on domains further from source data such as web text in OntoNotes* (see Table 4) or WNUT*. This is again explained by the difference in quantity and quality of the corpora on which these contextualizations are learned. The much larger and more generic unlabeled corpora on which LM are pretrained lead to contextual representations more robust to domain adaptation than C_{NER} learned on a small and specific NER corpus.

Similar behaviors can be observed when comparing BERT and Flair to the GloVe baselines, although we cannot separate the effects of representation and contextualization.

Complementarity Both in-domain and out-of-domain on OntoNotes*, the two types of contextualization transfer complementary syntactic features leading to the best configuration. However, in the most difficult case of zero-shot domain adaptation from CoNLL03 to WNUT*, C_{NER} is detrimental with ELMo and BERT. This is probably due to the specificity of the target domain, excessively different from source data.

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	$87.2_{.5}$	$88.4_{.4}$	$84.7_{.2}$	$82.4_{1.2}$	$84.5_{1.1}$	$79.5_{1.0}$	$\mathbf{85.0}_{.3}$
ELMo	$85.0_{.6}$	$88.6_{.3}$	$82.9_{.3}$	$78.1_{.7}$	$84.0_{.8}$	$79.9_{.5}$	$83.4_{.2}$
Flair	$78.0_{1.1}$	$86.5_{.4}$	$80.4_{.6}$	$71.1_{.4}$	$73.5_{1.8}$	$72.1_{.8}$	$79.0_{.5}$
ELMo[0]	$82.6_{.5}$	$88.0_{.3}$	$79.6_{.5}$	$73.4_{.6}$	$79.2_{1.2}$	$75.1_{.3}$	$80.5_{.2}$
GloVe + char	$80.4_{.8}$	$86.3_{.4}$	$77.0_{1.0}$	$70.7_{.4}$	$79.7_{1.8}$	$69.2_{.8}$	$77.9_{.5}$

5 Related Work

Augenstein et al. [2] perform a quantitative study of two CRF-based models and a CNN with classical word embeddings [4] over seven NER datasets including CoNLL03 and OntoNotes. They separate performance on seen (*exact match*) and unseen mentions and show a drop in F1 on unseen mentions and out-of-domain. Although comprehensive in experiments, this analysis is limited to models dating back from 2005 to 2011. We use a similar experimental setting to draw new insights on state-of-the art architectures and word representations. We limit to the two main English NER benchmarks as well as WNUT which was specifically designed to tackle this generalization problem in the Twitter domain. These three datasets cover all the domains studied in [2].

Moosavi and Strube raise a similar lexical overlap issue in Coreference Resolution on the CoNLL2012 dataset. They first show that for out-of-domain evaluation the performance gap between Deep Learning models and a rule-based system fades away [11]. They then add linguistic features (such as gender, NER, POS...) to improve out-of-domain generalization [12]. Nevertheless, such features are obtained using models in turn based on lexical features and at least for NER the same lexical overlap issue arises.

Finally, Pires et al. [15] concurrently evaluate the cross-lingual generalization capability of Multilingual BERT for NER and POS tagging. Our work on monolingual generalization to unseen mentions and domains naturally complements this study.

6 Conclusion

NER benchmarks are biased towards seen mentions, at the opposite of real-life applications. Hence the necessity to disentangle performance on seen and unseen mentions and test out-of-domain. In such setting, we show that contextualization from LM pretraining is particularly beneficial for generalization to unseen mentions, all the more out-of-domain where it surpasses supervised contextualization.

Despite this improvement, unseen mentions detection remains challenging and further work could explore attention or regularization mechanisms to better incorporate context and improve generalization. Furthermore, we can investigate how to best incorporate target data to improve this LM pretraining zero-shot domain adaptation baseline.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018), http://www.aclweb.org/anthology/C18-1139
2. Augenstein, I., Derczynski, L., Bontcheva, K.: Generalisation in named entity recognition: A quantitative analysis. Computer Speech & Language **44**, 61–83 (7 2017), https://linkinghub.elsevier.com/retrieve/pii/S088523081630002X
3. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T.: One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. arXiv preprint arXiv:1312.3005 (2013), https://arxiv.org/pdf/1312.3005.pdf
4. Collobert, R., Weston, J.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**, 2493–2537 (2011), http://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf
5. Derczynski, L., Nichols, E., Van Erp, M., Limsopatham, N.: Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In: 3rd Workshop on Noisy User-generated Text. pp. 140–147 (2017), https://www.aclweb.org/anthology/W17-4418
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019), https://www.aclweb.org/anthology/N19-1423
7. Howard, J., Ruder, S.: Universal Language Model Fine-tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 328–339 (2018), http://aclweb.org/anthology/P18-1031
8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991 (2015), https://arxiv.org/pdf/1508.01991.pdf
9. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. In: Proceedings of NAACL-HLT 2016. pp. 260–270 (2016), https://www.aclweb.org/anthology/N16-1030
10. Lee, J.Y., Dernoncourt, F., Szolovits, P.: Transfer Learning for Named-Entity Recognition with Neural Networks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). pp. 4470–4473 (2018), http://www.lrec-conf.org/proceedings/lrec2018/pdf/878.pdf
11. Moosavi, N.S., Strube, M.: Lexical Features in Coreference Resolution: To be Used With Caution. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 14–19 (2017), https://www.aclweb.org/anthology/P17-2003
12. Moosavi, N.S., Strube, M.: Using Linguistic Features to Improve the Generalization Capability of Neural Coreference Resolvers. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 193–203 (2018), http://aclweb.org/anthology/D18-1018
13. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. In: EMNLP 2014 (2014), https://nlp.stanford.edu/pubs/glove.pdf
14. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237 (2 2018), http://www.aclweb.org/anthology/N18-1202
15. Pires, T., Schlinger, E., Garrette, D.: How multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4996–5001 (2019), https://www.aclweb.org/anthology/P19-1493

16. Radford, A., Salimans, T., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training p. 12 (2018), https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
17. Rodriguez, J.D., Caldwell, A., Liu, A.: Transfer Learning for Entity Recognition of Novel Classes. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1974–1985 (2018), http://aclweb.org/anthology/C18-1168
18. Sang, E.F.T.K., De Meulder, F.: Introduction to the CoNLL-2003 shared task. In: Proceedings of the seventh Conference on Natural Language Learning at NAACL-HLT 2003. vol. 4, pp. 142–147 (2003), https://www.aclweb.org/anthology/W03-0419
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017), https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
20. Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M.: OntoNotes Release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia, PA (2013), https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf
21. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, ., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144 (2016), https://arxiv.org/pdf/1609.08144.pdf
22. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 19–27 (2015), https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Zhu_Aligning_Books_and_ICCV_2015_paper.pdf