

Multilingual Natural Language Processing and Transformers: A Giant Step Forward

Radu Florian Taesun Moon Parul Awasthy Jian Ni
 IBM Research AI
 Yorktown Heights, NY 10598
 {raduf,tsmoon,awasthy,ni}@us.ibm.com

Recent developments in deep learning for natural language processing have opened up opportunities to develop tools and libraries for multiple languages simultaneously and also for low resource languages. Here we describe these advances as well as our experiments that show that one can build a multilingual named entity recognition system that works well on multiple languages, in addition to being able handle unseen languages.

1. Introduction

Natural language processing is a scientific discipline as well as a field of software engineering which provides a structured, statistical interface to written human language. A well-known, well-established and early application of this field is machine translation where the goal is to translate text in one language to another via a computer program without any human intervention. A more recent application with many commercial and scientific uses is sentiment analysis, which detects whether some statement or review is favorable, unfavorable or neutral toward some subject with many different shades of granularity in terms of both subject matter and sentiment.

An application that lies more in the background but is no less important is information extraction which comprises named entity recognition, relation extraction and coreference resolution. Named entity recognition detects the proper nouns in texts such as "The *UN* will hold its *General Assembly* in *New York* soon. It is expected to increase traffic significantly in *Midtown*" where the named entities (and one pronominal mention) are in italics. Relation extraction labels any relations that may exist between such named entities such as the fact that the *General Assembly* will be in a valid textual relation with *New York* that we will label here *to take place in*. Coreference resolution is the component which links together all the textual mentions in a text that refer to the same entity, in this case *General Assembly* and *It*. These three components are usually packaged together and is called an information extraction system, with wide uses in many areas that require a structured or simplified representation of large, unruly natural language corpora such that it can be pro-

cessed uniformly and quickly by downstream applications such as databases, text analytics engines and automated decision making.

All the above applications require a substantial amount of text that humans have labeled with appropriate information so that the underlying statistical models used by the NLP components have idealized output that it may adopt without having a human engineer manually encode millions or possibly billions of individual behaviors. The time and cost involved in creating this labeled data is still considerable and beyond the reach of most language communities outside a handful of the most commonly used languages such as English and Chinese. Two recent developments in deep learning for NLP give us reason to hope that the pipeline for creating tools for low resource languages is about to be greatly both simplified and improved at the same time. Namely, these are the transformer architecture (Vaswani et al., 2017) and multilingual Bidirectional Encoder Representations from Transformers (Devlin et al., 2018).

2. Prior Work

Here, we provide a brief overview prior work in NLP that relates to deep learning and multilingual NLP. Named entity recognition and its successor, mention detection, have a vast history in NLP - a full description is beyond the scope of this paper. We will touch on the deep learning research that is directly related to the results presented here.

Collobert and Weston (2008) was the first modern approach to sequence classification, including NER, that used a convolutional neural network architecture, advancing the state-of-the-art (SotA) in English CoNLL. Lample et al. (2016) introduced – what has become the standard baseline – Bidirectional LSTM (Bi-LSTM) networks to advance the SotA NER performance on the CoNLL datasets, building 4 models, one for each language.

2018 saw the introduction of strong language-model pretrained models, first with ELMo (Peters et al., 2018), then with BERT (Devlin et al., 2018). These models excel by using large amounts of unlabeled data to train neural networks that learn the structure of the

System	En	Es	De	Nl
BERT-SL E^n 0-shot	91.2	73.6	69.4	78.6
BERT-SL	91.2	87.5	82.7	90.6
BERT-ML	91.3	87.9	83.3	91.1

Table 1: Single and multi language F_1 on CoNLL’02, CoNLL’03 .

language by playing guessing games: predict the next word, predict a missing word in context, predicting the next sentence. Then, they are then used as pretrained networks to various NLP tasks, resulting in state-of-the-art results.

Vaswani et al. (2017) is the most important new development in neural network architectures for NLP which relies solely on attention mechanisms while dispensing entirely with recurrence and convolution. A particular instantiation of this architecture is BERT (Devlin et al., 2018) which trained a transformer-based architecture on large amounts of unlabeled text, with a cloze and next sentence prediction objectives, then feeding the sentence/paragraph embeddings to a linear feedforward layer, again surpassing the SotA in many tasks.

Akbik et al. (2018) extends the ELMo framework by computing Bi-LSTM sequences at character level for the entire sentence, then combines the token aligned pieces to feed into a bidirectional LSTM layer, together with the word embeddings, and obtaining SotA results on CoNLL and OntoNotes.

Multilingual Work

The resource problem or the fact that a considerable amount of time and money has to be spent in creating human labeled corpora to be used as training data for each given domain, language and NLP component has been plaguing the field since the earliest statistical models were defined and developed. As such, there has been much interesting cross-lingual induction of NLP tools, i.e. harnessing existing work in machine translation or cross-lingual dictionaries to induce NLP tools in a language without such tools from a language that does have such tools. An important early work is Yarowsky et al. (2001).

Following the development of pretrained word embeddings, interest has shifted to using these word embeddings in a multilingual setting (Ruder et al., 2017). Sil et al. (2015) trained a joint mention detection model on English and Spanish, resulting in better performance on the Spanish data. Akbik et al. (2018) did experiments by training Flair on all CoNLL’02 and ’03 languages and providing one model on their github page of their system, Flair Github (2019). Xie et al. (2018) aligned monolingual embeddings from English to Spanish, German, and Dutch, and then translated the English CoNLL dataset into these languages, and built a self-attentive Bi-LSTM-CRF model using the translated languages, creating 0-shot NER systems. Pires

System	En	Ar	Zh
BERT E^n 0-shot	87.9	10.7	65.2
BERT SL	87.9	68.7	72.9
BERT ML	88.3	69.9	74.1

Table 2: Comparison of Monolingual and Multilingual RE performance (F_1 score).

System	En	Pt	De	Es	It	Ja	Fr	Ar
SIRE	87	82	76	85	77	82	74	61
BERT E^n	93	77	73	80	72	62	66	36
BERT ML	93	87	84	89	82	84	81	74

Table 3: Performance on the KLUE dataset, 8 languages.

et al. (2019) used multilingual BERT and techniques similar to our zero-shot baseline to obtain SotA numbers for zero-shot on all four CoNLL languages.

Conneau et al. (2018) developed a task specifically for the multilingual setting where NLP practitioners could test knowledge transfer across languages in an unsupervised manner on a problem known as natural language inference. This is a problem where an NLP system must decide when given two sentences whether the second sentence entails the first, contradicts the first or is neither. Usually, a system would train only on labeled English sentence pairs and then be evaluated on sentence pairs from 14 difference languages that are not English such as French, Chinese, Urdu, etc. The knowledge transfer is implemented by harnessing machine translation systems that either translate non-English languages into English during decoding or translate the English training data into the target language so that a new language specific system can be trained.

3. Multilingual Named Entity Recognition

3.1. Data and Framework

We experiment on the Dutch, English, German and Spanish CoNLL data sets (Tjong Kim Sang, 2002; Sang and Meulder, 2003), the OntoNotes dataset (Weischedel et al., 2011), and the KLUE dataset, a multilingual NER dataset used in Watson NLP, and use the multilingual BERT embeddings provided by (Devlin et al., 2018). One main advantage of this method is that they create a universal vocabulary that spans the most frequent 104 languages in Wikipedia, effectively allowing us to feed many languages as input to the system.

To evaluate our hypotheses, we run two types of experiments: one in which we only train the system on the English dataset from each corpus (which is typically the largest) and then test on the other languages, and the second in which we train on all the datasets. We compare the results with the systems obtained by training on each individual language separately, as it is the common practice nowadays.

3.2. Results

Table 1 shows the results on the CoNLL dataset. The first line shows the 0-shot performance¹, which is the system trained only on English. The system that was trained on data from all languages outperforms each system trained only on its own language by an average of 0.4 F_1 , which is the standard measure for NER - the hyperbolic mean of precision and recall. The 0-shot system is behind language-specific systems by 13 F_1 , which is not too bad, given that the system was not exposed to the languages at all.

Table 2 shows the results obtained on the OntoNotes corpus. The interesting part here is that the languages do not share the script at all. Surprisingly, the English-trained system performs very well on Chinese, only being 7.5 F_1 behind the language-specific system, basically delivering 90% of the performance. The multilingual system is again better than the single-language systems by 0.9 F_1 , including 0.4 in English, which shows that even the dataset with the largest data size can be improved using this approach.

Finally, Table 3 shows the results of running the BERT multilingual across 8 languages: English, Brazilian Portuguese, German, Spanish, Italian, Japanese, French, and Arabic. We compare here against a feature-based system developed at IBM Research - SIRE (Statistical Information and Relation Extraction) (Florian et al., 2004), which is not deep-learning based, and is representative of the best non deep learning statistical systems.

The multilingual BERT outperforms SIRE by a large margin - 10.7 F_1 on average. The English-trained BERT system is behind SIRE by 8.2 F_1 absolute (89.5% relative) and 14.4 F_1 (82.8% relative) behind the multitrained system, even though it did not have access to any of the foreign language labeled data.

3.3. Observations and Comments

These results show that if one does not have the resources to create labeled training data in a large variety of languages, they can build the data in English, and then use the trained BERT system to also have the capability of processing other languages. If one has the resources, then they can train a truly multilingual system that will perform very well across languages.

We also note that this technique is applicable to most NLP problems, not only named entity recognition - we have applied it successfully, for instance, to sentiment classification and relation extraction as well.

4. Conclusion

Multilingual pretraining in the form of multilingual BERT opens up exciting opportunities and hints at a new modus operandi for low resource languages and multilingual NLP in general. As we have shown here,

¹0-shot is used in literature to mean the system did not have any training data in that category.

one can obtain very good performance with a system that was trained only on English, and even better performance if the system is trained on multiple languages.

On three datasets, the multilingual BERT system outperformed the language-based BERT systems, and was much better than a feature-based statistical approach (SIRE). As a proxy for a single-language system, the English-trained BERT system performed at about 80-90% of the full multilingual BERT system, showing that, in cases where the resources are not there to build multiple language datasets, this is an effective approach to build a system that can tackle multiple languages at once.

We are looking forward to more research into better representation of languages that will lead to even better performance across all NLP tasks.

References

- A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/C18-1139>.
- R. Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of International Conference on Machine Learning*, 2008.
- A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053, 2018. URL <http://arxiv.org/abs/1809.05053>.
- J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.
- R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, and S. Roukos. A statistical model for multilingual entity detection and tracking. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/N04-1001>.
- Z. R. Github. very simple framework for state-of-the-art natural language processing (nlp). <https://github.com/zaladoresearch/flair>, 2019.

- G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016. URL <http://arxiv.org/abs/1603.01360>.
- M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL <http://arxiv.org/abs/1802.05365>.
- T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P19-1493>.
- S. Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*, 2017.
- E. F. T. K. Sang and F. D. Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0306050, 2003. URL <http://arxiv.org/abs/cs.CL/0306050>.
- A. Sil, G. Dinu, and R. Florian. Proceedings of the 2015 text analysis conference, TAC 2015. NIST, 2015. URL <https://tac.nist.gov/publications/2015/papers.html>.
- E. F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118853.1118877. URL <https://doi.org/10.3115/1118853.1118877>.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, and N. Xue. *OntoNotes: A Large Training Corpus for Enhanced Processing*. 01 2011.
- J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical*
- Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D18-1034>.
- D. Yarowsky, G. Ngai, and R. Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research, HLT '01*, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. doi: 10.3115/1072133.1072187. URL <https://doi.org/10.3115/1072133.1072187>.