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,awasthyp,nij}@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 processed 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
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 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).

3.2. Results
Table 1 shows the results on the CoNLL dataset. The first line shows the 0-shot performance\(^1\), 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 capacity 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, 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


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.
