PELTARION

Data science /

A deep dive into multilingual NLP models

February 24 2020 / 12 min read

John Moberg
AI Research Engineer

Deep learning has revolutionized NLP (natural language processing) with powerful models such as BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018) that are pre-trained on huge, unlabeled text corpora. Using transfer learning, we can now achieve good performance even when labeled data is scarce. However, most of the work to date has been focused on English, as we discuss in this blog, leaving low- and medium-resource languages such as Swedish behind. Industry and researchers are now trying to catch up, training Transformer-based models in German, Dutch, and Swedish, just to name a few.

Is this the right way to go?

An alternative approach is to train a multilingual model, that is, a single model that can handle multiple languages simultaneously. This would circumvent having to train a monolingual model for every single language, and recent results suggest that multilingual models can even achieve **better performance than monolingual models**, especially for low-resource languages. In particular, XLM-R (Conneau et al., 2019), a 100-language model introduced by Facebook AI researchers in November 2019, achieves state-of-the-art results in cross-lingual transfer *and* is competitive with English BERT on an English benchmark. In this blog post, we introduce XLM-R along with two models leading up to XLM-R: Multilingual BERT (Devlin et al., 2018) and XLM (Lample and Conneau, 2019).

PELTARION

Evaluating cross-lingual transfer with XNLI

We need a way to evaluate our model's capacity to learn multiple languages and generalize across languages. The Cross-lingual Natural Language Inference (XNLI) dataset has become a standard dataset for this purpose and is used for evaluation of all the models we discuss. It considers the case when we have plenty of English training data, but very little for other languages. More precisely, the dataset contains test data in 14 languages, but only English training data.

XNLI is based on MultiNLI and uses the English MultiNLI training set with 433 thousand examples. For evaluation, 7,500 examples are human-translated into 14 languages, yielding 105 thousand test examples in total. It is a natural language inference (NLI) task: given a premise and a hypothesis, does the premise entail or contradict the hypothesis, or is it neutral? For example, the premise and hypothesis may be *"The trophy didn't fit in the suitcase because it was too big"* and *"The suitcase was too big,"* respectively.

XNLI is commonly used in four different ways:

> **Cross-lingual transfer:** Fine-tune multilingual model on English training set and evaluate on each language's test set. This is, in some sense, the ultimate test of a model that truly understands multiple languages.
>
> **Translate-test:** Fine-tune English model on an English training set and evaluate on the machine-translated test set. This imitates the scenario where you just machine-translate your data, use an English model, and hope for the best.
>
> **Translate-train:** Machine-translate the English training set to another language and fine-tune a multilingual model on that. In this case, the test set remains in the original language.
>
> **Translate-train-all:** In this scenario, we machine-translate the training set to *all* languages and fine-tune a multilingual model on that.

Multilingual models: mBERT, XLM and XLM-R

Let's now look at some of the most prominent multilingual models today. We begin by introducing the models and follow up with results from evaluating them on XNLI.

mBERT

undersampled.

XLM

XLM (Lample and Conneau, 2019) is a Transformer-based model that, like BERT, is trained with the masked language modeling (MLM) objective. Additionally, XLM is trained with a Translation Language Modeling (TLM) objective in an attempt to force the model to learn similar representations for different languages. TLM is quite simple: input the same sentence in two different languages and mask tokens as usual. To predict a masked token, the model can then choose to use tokens from the other language.

XLM is trained with both MLM and TLM, with MLM on data from Wikipedia in the 15 XNLI languages, and TLM on several different datasets depending on the language. Note that TLM requires a dataset of parallel sentences, which may be difficult to acquire.

XLM-R

The most recent multilingual model is XLM-R (Conneau et al., 2019), where the R stands for RoBERTa (Liu et al., 2019). Based on the name, it would be natural to assume that it is XLM with RoBERTa instead of BERT, but that would be wrong.

Instead, XLM-R takes a step back from XLM, eschewing the TLM objective, and just trains RoBERTa on a huge, multilingual dataset at an enormous scale. Unlabeled text in 100 languages is extracted from CommonCrawl datasets, totaling 2.5TB of text. It is trained in a RoBERTa fashion, that is, *only* using the MLM objective. In fact, the only noteworthy difference to RoBERTa is the vocabulary size: 250 thousand tokens compared to RoBERTa's 50,000 tokens. Note that this makes the model significantly larger, 550 million parameters compared to the 355 million of RoBERTa.

Excluding the variance in scale, the central deviation between XLM and XLM-R is that XLM-R is wholly self-supervised whereas XLM requires parallel examples that can be difficult to acquire at sufficient scale. We must have faith that XLM-R can generalize across languages without supervision.

Evaluation on XNLI

PELTARION

Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																	
XLM (MLM)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																	
XLM (MLM+TLM)	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
XLM (MLM)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R	CC	1	100	**88.7**	**85.2**	**85.6**	**84.6**	**83.6**	**85.5**	**82.4**	**81.6**	**80.9**	**83.4**	**80.9**	**83.3**	**79.8**	**75.9**	**74.3**	**82.4**

The results of evaluating the models we have discussed on XNLI. Adapted from the paper by Conneau et al. (2019)

The table above, adapted from the paper by Conneau et al. (2019), contains the results of evaluating the models we have discussed on XNLI. Note that mBERT uses a smaller architecture (BERT-base), whereas XLM and XLM-R use the BERT-large architectures, so the difference in performance may be exaggerated.

We see that XLM-R performs significantly better than the other models. XLM is better than mBERT, but handles fewer languages and is based on a larger model, and so the difference may not be so large in reality. Surprisingly, cross-lingual transfer works very well: XLM-R gets 80% accuracy averaged across all languages, despite only being fine-tuned on English training data. Machine-translating the training set to all languages (totaling 6 million training examples) yields even better performance, but not by much. Considering that machine translation can be prohibitively expensive, cross-lingual transfer is very competitive.

A highlight from the XLM-R paper is their evaluation on the GLUE benchmark, a standard NLP benchmark in English, where it is shown that XLM-R is competitive with monolingual models on a monolingual benchmark, despite handling 100 languages. XLM-R achieves an average performance of 91.5, compared to 90.2, 92.0 and 92.8 of BERT, XLNet and RoBERTa, respectively. So while XLM-R doesn't beat its monolingual counterpart RoBERTa, it is remarkably close.

Our experiments with XLM-R on Swedish political data echoes this finding. XLM-R performed as good as (or slightly better than) the Swedish BERT base models provided by The Swedish Public Employment Service and The National Library of Sweden. XLM-R achieved ~80% accuracy whereas the Swedish BERT models reached ~79% accuracy. While this isn't a significant difference, it may mean that training monolingual models for small languages is unnecessary.

A look inside XLM-R

How is it that XLM-R can be fine-tuned on just one language and zero-shot transfer to other languages, having never seen labeled examples of them? Intuitively, it seems reasonable that

PELTARION

sentences. We fed the sentences through the pre-trained XLM-R and stored the representations at each layer. To study how the representations change as we move through the Transformer, we mapped them to the plane with the dimensionality reduction technique UMAP (McInnes et al., 2018). In the figure, colors correspond to different languages and symbols represent different sentences.

In the initial layers, we see clustering by language – sentences in the same language are close in embedding space. However, a few Transformer blocks later, we see that the same sentence in different languages actually maps to similar representations. This shows that XLM-R does generalize between languages, and without our supervision, has inferred that we talk about the same things even though we use different languages.

Looking at the final layer representation, we are back to clustering by language, but a bit less clearly. This is expected: XLM-R is trained to predict masked tokens, which obviously depend

PELTARION

reasonable predictions, which further strengthens our belief that XLM-R produces cross-lingual representations. To improve on this even further, our insights suggest that it may be helpful to freeze the initial layers and only fine-tune the part of the network that deals with cross-lingual representations.

Conclusions and looking forward

Multilingual models can be incredibly powerful. The latest development, XLM-R, handles 100 languages and still remains competitive with monolingual counterparts. Research is ongoing and we expect cross-lingual transfer to continue improving. Already today, and especially if this trend continues, it is hard to motivate the work and computation required to train monolingual models for medium/small-sized languages when multilingual models perform as well or better. Instead, an interesting direction could be to study the possibility of extracting smaller and more efficient monolingual models from multilingual models.

Finally, I believe that the story of multilingual models and the great success of XLM-R reinforces Sutton's thesis in The Bitter Lesson:

"Essential to these methods is that they can find good approximations, but the search for them should be by our methods, not by us. We want AI agents that can discover like we can, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done."

Instead of adding objectives and requiring difficult-to-acquire labeled data to encode our prior knowledge, we should strive for simplicity and leave it to our networks to learn the underlying structure. So far, the story of deep learning is that of end-to-end learning with massive compute and lots of data. Thankfully, the trend of transfer learning makes it so that anyone can fine-tune a good model, even without much data. XLM-R continues this successful trend.

PELTARION

John Moberg

AI Research Engineer

John is an AI Research Engineer at Peltarion, with a M.Sc. in Engineering Mathematics from Chalmers University of Technology. At Peltarion, he focuses on natural language processing, both in the context of research and in applications for clients in industry. Apart from NLP, he is especially passionate about reinforcement learning and Bayesian deep learning.

Data science topics

BERT, DEEP LEARNING, EVENTS, HEALTHCARE, RESEARCH

02/ More on Data science

PELTARION

Data science /

Building a Stack Overflow question tagging model with public BigQuery data

In this blog post, we'll show how you can shape one of these public datasets, the Stack Overflow posts dataset, to train a question tagging BERT model.

February 24 / 10 min read

Data science /

Connecting the dots in Neural Networks

February 8 / 3 min read

PELTARION

classification?

January 29 / 10 min read

Data science /

Peltarion on Microsoft's AI Show

December 3 2020 / 1 min read

Find us

Stockholm, Sweden

Contact: contact@peltarion.com

Press

Press releases

Press coverage

Press kit

Legal notice

Privacy policies

Terms