🗃The Current Best of Universal Word Embeddings and Sentence Embeddings

medium.com/huggingface/universal-word-sentence-embeddings-ce48ddc8fc3a

Thomas Wolf May 13, 2018

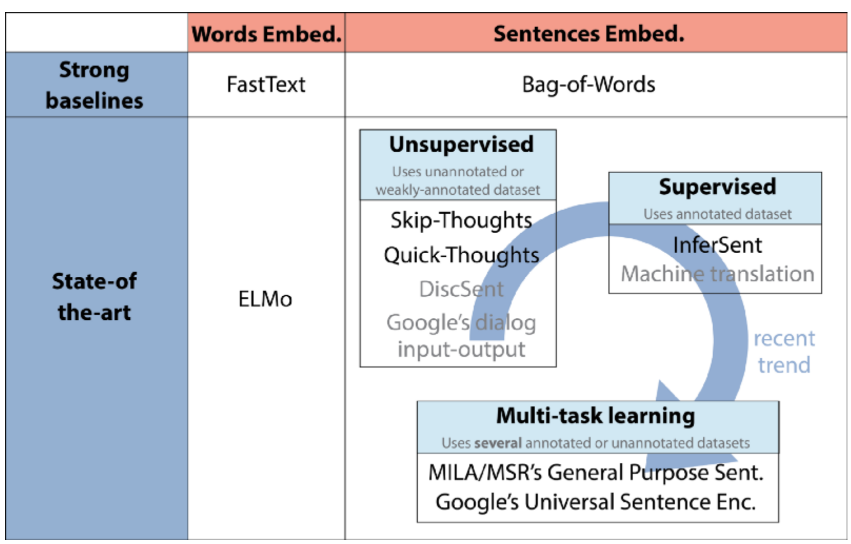
Word and sentence embeddings have become an essential part of any Deep-Learning-based natural language processing systems.

They encode words and sentences 🛏 in fixed-length dense vectors 📐 to drastically improve the processing of textual data.

A huge trend is the quest for **Universal Embeddings:** embeddings that are pre-trained on a large corpus and can be plugged in a variety of downstream task models (sentimental analysis, classification, translation…) to automatically improve their performance by incorporating some general word/sentence representations learned on the larger dataset.

It's a form of *transfer learning*. Transfer learning has been recently shown to drastically increase the performance of NLP models on important tasks such as text classification. Go check the very nice work of Jeremy Howard and Sebastian Ruder (ULMFiT) to see it in action.

While unsupervised representation learning of sentences had been the norm for quite some time, the last few months have seen a shift toward supervised and multi-task learning schemes with a number of very interesting proposals in late 2017/early 2018.



Recent trend in Universal Word/Sentence Embeddings. In this post, we describe the models indicated in black. Reference papers for all indicated models are listed at the end of the post.

This post is thus a brief primer on the current state-of-the-art in Universal Word and Sentence Embeddings, detailing a few

- **strong/fast baselines**: *FastText, Bag-of-Words*
- **state-of-the-art models**: *ELMo, Skip-Thoughts, Quick-Thoughts, InferSent, MILA/MSR's General Purpose Sentence Representations & Google's Universal Sentence Encoder.*

Let’s start with word embeddings.

Recent Developments in Word Embeddings

A wealth of possible ways to embed words have been proposed over the last five years. The most commonly used models are word2vec and GloVe which are both unsupervised approaches based on the distributional hypothesis (*words that occur in the same contexts tend to have similar meanings*).

While several works augment these unsupervised approaches by incorporating the supervision of semantic or syntactic knowledge, purely unsupervised approaches have seen interesting developments in 2017–2018, the most notable being **FastText** (an extension of word2vec) and **ELMo** (state-of-the-art contextual word vectors).

F**astText** was developed by the team of Tomas Mikolov who proposed the word2vec framework in 2013, triggering the explosion of research on universal word embeddings.

The main improvement of FastText over the original word2vec vectors is the inclusion of character n-grams, which allows computing word representations for words *that did not appear in the training data* (“out-of-vocabulary” words).

FastText vectors are super-fast to train and are available in 157 languages trained on Wikipedia and Crawl. They are a great baseline.

The Deep Contextualized Word Representations (**ELMo**) have recently improved the state of the art in word embeddings by a noticeable amount. They were developed by the Allen institute for AI and will be presented at NAACL 2018 in early June.

Elmo knows quite a lot about words context

In ELMo, each word is assigned a representation which is a function of the entire corpus sentences to which they belong. The embeddings are computed from the *internal states of a two-layers bidirectional Language Model (LM)*, hence the name “ELMo”: ***E**mbeddings from **L**anguage **Mo**dels.*

Specificities of ELMo:

- **ELMo’s inputs are characters** rather than words. They can thus take advantage of sub-word units to compute meaningful representations even for out-of-vocabulary words (like FastText).
- **ELMo are concatenations of the activations on several layers of the biLMs.**Different layers of a language model encode different kind of information on a word (e.g. Part-Of-Speech tagging is well predicted by the lower level layers of a biLSTM while word-sense disambiguation is better encoded in higher-levels). Concatenating all layers allows to freely combine a variety of word representations for better performances on downstream tasks.

Now, let’s turn to universal sentence embeddings.

The Rise of Universal Sentence Embeddings

There are currently many competing schemes for learning sentence embeddings. While simple baselines like averaging word embeddings consistently give strong results, a few novel unsupervised and supervised approaches, as well as multi-task learning schemes, have emerged in late 2017-early 2018 and lead to interesting improvements.

Let's go quickly through the four types of approaches currently studied: from *simple word vector averaging baselines* to *unsupervised/supervised* approaches and *multi-task learning schemes (as illustrated above)*.

> *A good algorithm for computing such a baseline is detailed in the work of Arora et al. published last year at ICLR,* *[A Simple but Tough-to-Beat Baseline for Sentence Embeddings](): * use a popular word embeddings of your choice, encode a sentence in a linear weighted combination the word vectors and perform a common component removal (remove the projection of the vectors on their first principal component). *This general method has deeper and powerful theoretical motivations that rely on a generative model which uses a random walk on a discourse vector to generate text (we won't discuss the theoretical details here).*

A plot of HuggingFace's dialogs Bag-of-Words. Bag-of-Words approaches loose words ordering but keep a surprising amount of semantic and syntactic content. Interesting insights in Conneau et al. ACL 2018.

Going beyond simple averaging, the first major proposals were using **unsupervised** training objectives, starting with the *Skip-thoughts vectors* proposed by Jamie Kiros and co-workers in 2015.

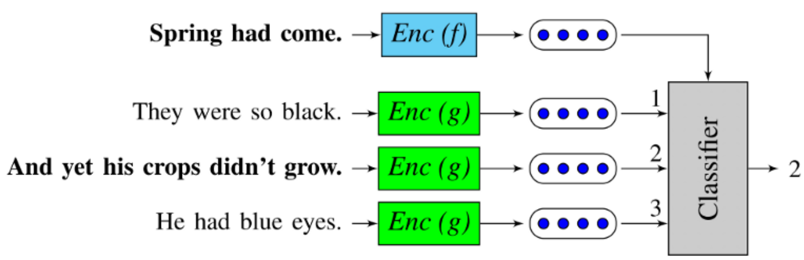
Unsupervised schemes learn sentence embeddings as a byproduct of learning to predict a coherent succession of sentences or a coherent succession of clauses inside a sentence. These approaches can (in theory) make use of any text dataset as long as it contains sentences/clauses juxtaposed in a coherent way.

Skip-thoughts vectors is the archetypical example of learning unsupervised sentence embeddings. It can be though as the equivalent for sentences of the skip-gram model developed for word embeddings: *rather than predicting the words surrounding a word, we try to predict the surroundings sentences of a given sentence*. The model consists in an RNN-based encoder-decoder which is trained to reconstruct the surrounding sentences from the current sentence.

One interesting insight in the Skip-Thought paper was a *vocabulary expansion scheme*: Kiros et al. handled words not seen during training by learning a linear transformation between their RNN word embedding space and a larger word embedding such as word2vec.

Quick-thoughts vectors are a recent development of the Skip-thoughts vectors, presented this year at ICLR. In this work, the task of predicting the next sentence given the previous one is reformulated as a classification task: *the decoder is replaced by a classifier which has to choose the next sentence among a set of candidates*. It can be interpreted as a discriminative approximation to the generation problem.

One strength of this model is its speed of training (an order of magnitude compared to Skip-thoughts model) making it a competitive solution to exploit massive dataset.

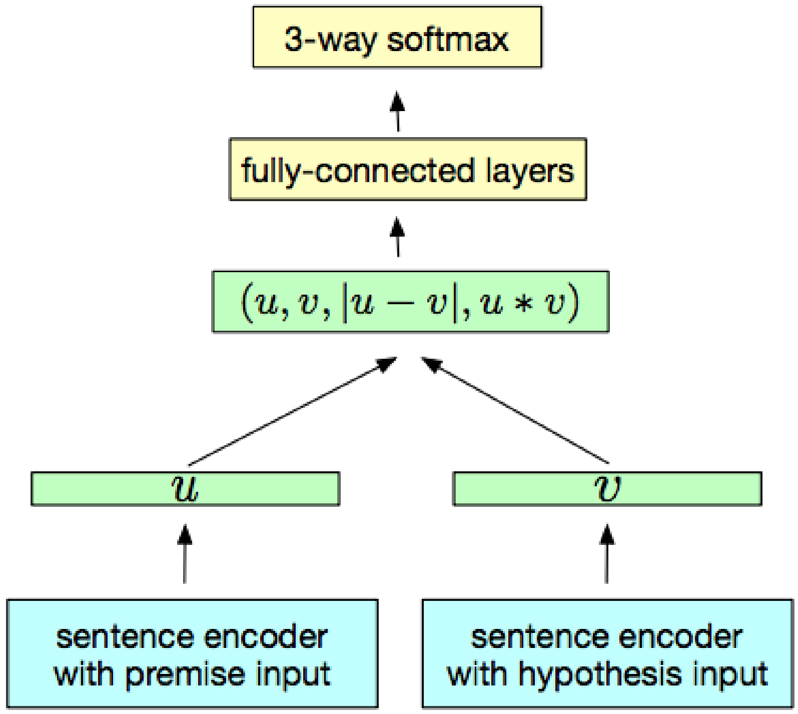


Quick-thoughts classification task. The classifier has to chose the following sentence from a set of sentence embeddings. Source: "An efficient framework for learning sentence representations" by Logeswaran et al.

For a long time, **supervised** learning of sentence embeddings was thought to give lower-quality embeddings than unsupervised approaches but this assumption has recently been overturned, in part following the publication of the *InferSent* results.

Unlike the *unsupervised* approaches detailed before, *supervised* learning requires a labelled dataset annotated for some task like Natural Language Inference (e.g. with pairs of entailed sentences) or Machine Translation (with pairs of translated sentences) which poses the question of the specific task to choose and the related question of the size of the dataset required for good quality embeddings. We talk more about these questions in the next and last section on Multi-task learning but before that, let's see what's behind the InferSent breakthrough that was published in 2017.

InferSent is an interesting approach by the simplicity of its architecture. It uses the *Stanford Natural Language Inference (SNLI) Corpus* (a set of of 570k pairs of sentences labelled with 3 categories: neutral, contradiction and entailment) to train a classifier on top of a sentence encoder. Both sentences are encoded using the same encoder while the classifier is trained on a pair representation constructed from the two sentence embeddings. Conneau et al. adopt a bi-directional LSTM completed with a max-pooling operator as sentence encoder.



A supervised sentence embeddings model (InferSent) to learn from a NLI dataset. Source: "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data" by A. Conneau et al.

The success of InferSent lead poses the following question in addition to the usual quest for selecting the best neural net model:

> *Which supervised training task would learn sentence embeddings that better generalize on downstream tasks?*

Multi-task learning can be seen as a generalization of Skip-Thoughts, InferSent, and the related unsupervised/supervised learning schemes, that answer this question by trying to combine several training objectives in one training scheme.

Several recent proposals for multi-task learning were published in early 2018. Let's quickly go through MILA/MSR's **General Purpose Sentence Representation** and Google's **Universal Sentence Encoder.**

In the paper describing MILA & Microsoft Montreal's work and presented at ICLR 2018 (Learning General Purpose Distributed Sentence Representation via Large Scale Multi-Task Learning), Subramanian et al observe that to be able to generalize over a wide range of diverse tasks, it is necessary to encode multiple aspects of the same sentence.

The authors thus leverage a *one-to-many multi-tasking learning framework* to learn a universal sentence embedding by switching between several tasks. The 6 tasks chosen (Skip-thoughts prediction of the next/previous sentence, neural machine translation, constituency parsing and natural language inference) share the same sentence embedding obtained by a bi-directional GRU. Experiments suggest that syntactic properties are better learned when adding a multi-language neural machine translation task, length and word order are learned with a parsing task and training a natural language inference encodes syntax information.

Google's Universal Sentence Encoder, published in early 2018, follows the same approach. Their encoder uses a transformer-network that is trained on a variety of data sources and a variety of tasks with the aim of dynamically accommodating a wide variety of natural language understanding tasks. A pre-trained version has been made available for TensorFlow.

This concludes our short summary on the current state of Universal Words and Sentence Embeddings.

The domain has seen a lot of interesting developments in the last few months together with great progresses in the ways we assess and probe the performance of these embeddings and their inherent bias/fairness (a real issue when you talk about Universal Embeddings). We didn't have time to talk about these latest topics but you can find a few links in the references.

I hope you enjoyed this brief!

Some references

- Very recently, C. Perone and co-workers published a nice and extensive comparison between ELMo, InferSent, Google Universal Sentence Encoder, p-mean, Skip-thought, etc. Here is a link to the paper: https://arxiv.org/abs/1806.06259
- If you're interested in the way we evaluate sentence embeddings, you should definitely check the recent work of Facebook on SentEval and its probing tasks as well as the recently published GLUE benchmark by NYU, UW and DeepMind researchers.