Sentence selection strategies for distilling word embeddings from BERT

Yixiao Wang, Zied Bouraoui, Luis Espinosa-Anke, and Steven Schockaert
Cardiff university, UK; CRIL-CNRS, Universite d’Artois, France.

Distilling static word vectors from language models is useful when word meaning has to be modelled in the absence of context.

Method

The most common strategy for deriving static word vectors of a word that we want to model can be illustrated by the following diags:

Both plot had previously experience in landing at Tocumon International Airport, captain D’Antonio had landed at the airport 52 time, and first officer Artes had landed there 5 time. The Tikal VOR-DME is located northeast of the airport.

It also includes estimate of the amount of AIP money needed to fund infrastructure development project that will bring these airport up to current design standard and add capacity to congested airport.

Today the land formerly occupied by the airport is an agricultural estate and field.

Given an sentence that mention the target word, here is an example of how do we get contextualized representation for the target word from a LM.

| No scheduled flight operate at this airport |
| [CLS] + [token1] + [token2] + … + [token_n] + [SEP] |
| _v_CLS_ + _v_1_ + _v_2_ + … + _v_n_ + _v_SEP_ |

From here, we have three alternative ways to get contextualized word representation:

LAST: keep the original sentences and get last layer outputs

MASK: replace the word w by a [MASK] and get last layer outputs

AVG: get the average of all hidden layers and output layer as outputs

However, randomly selecting sentence that mention the target word is far from optimal. This paper attempts to distill high quality static word vectors in an efficient way by strategically selecting a few high quality sentences for each word.

Sentence selection strategies

RANDOM (baseline):

Born in Pantaneras Province, Lagoons’ parent decided to move to Limón where Christian went to school and worked in banana plantation.

INTRO: sampling sentences that occur in the introductory section of a wikipedia article.

The work, created in an edition of three, consists of a fresh banana taped to a wall with a piece of duct tape.

HOME: sampling sentences from wikipedia article about banana.

POS: sentences that start with word w in plural form.

BANANA, grown mainly for domestic consumption, amount to a steady annual average crop of 70,000 tons.

ENUM: sentences in which w is preceded or succeeded by a comma or the word “and”

These have included: bacon maple ale and chocolate, peanut butter, and banana ale

PMI: sentences in which both w and its top PMI (Pointwise Mutual Information) neighbors are mentioned

The above table report F1 scores of static word vectors from BERT-base.

| Evaluation Results |

| Stategic word vectors are evaluated on the tasks of predicting semantic properties of words. |

| No scheduled flight operate at this airport |
| [CLS] + [token1] + [token2] + … + [token_n] + [SEP] |
| _v_CLS_ + _v_1_ + _v_2_ + … + _v_n_ + _v_SEP_ |

From here, we have three alternative ways to get contextualized word representation:

LAST: keep the original sentences and get last layer outputs

MASK: replace the word w by a [MASK] and get last layer outputs

AVG: get the average of all hidden layers and output layer as outputs

However, randomly selecting sentence that mention the target word is far from optimal. This paper attempts to distill high quality static word vectors in an efficient way by strategically selecting a few high quality sentences for each word.

Sentence selection strategies

RANDOM (baseline):

Born in Pantaneras Province, Lagoons’ parent decided to move to Limón where Christian went to school and worked in banana plantation.

INTRO: sampling sentences that occur in the introductory section of a wikipedia article.

The work, created in an edition of three, consists of a fresh banana taped to a wall with a piece of duct tape.

HOME: sampling sentences from wikipedia article about banana.

POS: sentences that start with word w in plural form.

BANANA, grown mainly for domestic consumption, amount to a steady annual average crop of 70,000 tons.

ENUM: sentences in which w is preceded or succeeded by a comma or the word “and”

These have included: bacon maple ale and chocolate, peanut butter, and banana ale

PMI: sentences in which both w and its top PMI (Pointwise Mutual Information) neighbors are mentioned

The above table report F1 scores of static word vectors from BERT-base.

| Evaluation Results |

| Stategic word vectors are evaluated on the tasks of predicting semantic properties of words. |

| No scheduled flight operate at this airport |
| [CLS] + [token1] + [token2] + … + [token_n] + [SEP] |
| _v_CLS_ + _v_1_ + _v_2_ + … + _v_n_ + _v_SEP_ |

From here, we have three alternative ways to get contextualized word representation: