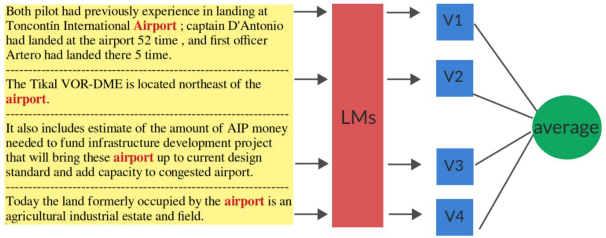
Sentence selection strategies for distilling word embeddings from BERT

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 diaigms:



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 attampts 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 Puntarenas Province , Lagos ' parent decided to move to Limón where Cristhian 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 w

A **banana** is an elongated, edible fruit – botanically a berry – produced by several kinds of large herbaceous flowering plants in the genus Musa

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

Bananas, 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 Multual Information) neighbors are mentioned

One day Mitchell posted a photo of herself on Twitter next to a bruised **banana** in response to trolls who had compared her freckles to the overripe **fruit**

DEF: definition of w from English fragment of wikitionary

GENERIC: sentences with high confidence score from GenericsKB

Evaluation Results

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

		McR				CSLB				WNSS				BND			
		1	5	10	20	1	5	10	20	1	5	10	20	1	5	10	20
MASK	RAND	44.8	57.0	59.8	61.5	31.0	47.4	51.7	53.8	39.3	53.6	56.0	59.1	28.0	36.2	38.0	40.0
	INTRO	44.0	57.9	58.7	60.7	34.4	47.3	50.8	53.8	41.6	54.2	56.5	57.6	28.3	36.6	38.5	40.0
	HOME	55.3	59.2	61.9	60.0	42.0	50.4	53.2	54.5	45.9	55.2	58.2	58.7	28.9	35.8	37.4	39.1
	POS	42.1	51.6	54.7	56.8	29.5	44.8	47.4	50.1	38.1	51.0	54.9	56.4	28.3	35.7	38.3	39.9
	ENUM	42.9	53.9	55.5	57.1	29.9	43.3	47.8	44.6	41.0	52.5	54.8	56.2	28.3	36.1	39.3	40.0
	PMI	56.8	57.0	59.2	61.6	48.9	46.1	**54.0**	54.4	43.1	55.1	58.3	58.6	29.5	37.6	39.7	41.0
	GENERIC	46.7	52.5	55.4	57.0	33.9	45.7	47.8	50.4	36.4	51.3	55.3	57.8	26.0	34.4	36.7	38.8
	DEF+HOME	56.9	59.9	62.2	**64.1**	49.6	50.4	53.2	**56.0**	49.6	55.2	58.5	59.3	29.2	35.7	37.4	39.1
	DEF+RAND	55.6	58.2	59.2	62.6	48.8	49.8	51.8	55.5	50.3	55.2	57.1	58.6	29.3	35.7	37.9	39.4
LAST	RAND	55.5	59.0	62.3	61.6	46.1	48.3	53.9	53.5	49.4	56.5	58.0	59.0	35.4	42.9	44.7	45.7
	INTRO	53.4	58.7	61.5	59.8	43.3	48.8	50.1	51.8	50.2	**58.3**	58.0	59.1	35.8	42.8	**44.8**	45.6
	HOME	58.3	61.8	62.6	63.0	47.8	48.7	51.8	51.0	52.0	**58.3**	59.1	**59.6**	35.7	42.3	43.9	44.9
	POS	53.7	59.5	58.9	59.8	43.4	52.8	53.2	55.3	45.4	55.4	57.5	58.6	32.6	38.7	40.5	41.5
	ENUM	47.4	59.8	58.1	60.0	41.9	47.8	47.3	52.5	49.5	55.3	57.0	57.4	35.3	42.7	43.7	45.4
	PMI	55.2	60.0	61.8	63.4	43.9	**53.0**	**54.0**	54.4	49.7	57.0	59.1	**59.6**	36.6	42.2	44.6	**45.8**
	GENERIC	54.3	60.7	59.8	61.1	45.1	49.2	51.2	51.3	50.3	57.3	57.9	58.9	36.1	42.3	43.2	44.3
	DEF+HOME	57.0	60.4	61.6	63.0	**50.1**	48.4	52.5	51.7	55.2	**58.3**	**59.6**	59.4	37.2	42.4	44.0	45.1
	DEF+RAND	57.6	60.5	58.8	61.9	**50.1**	49.1	51.5	52.9	55.2	58.0	59.4	59.1	37.2	41.7	44.1	45.5

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

	BERT-LARGE				RoBERTa-BASE				RoBERTa-LARGE			
	McR	CSLB	WNSS	BND	McR	CSLB	WNSS	BND	McR	CSLB	WNSS	BND
RAND	62.2	55.6	59.4	39.6	59.8	51.6	57.9	39.0	61.3	55.0	59.5	40.3
HOME	63.2	54.8	59.8	39.0	59.3	48.2	58.0	38.4	61.4	53.4	60.3	40.0
PMI	**65.0**	55.4	59.7	**41.3**	**63.6**	**54.0**	**58.7**	**39.8**	62.7	**56.0**	**60.1**	**44.1**
DEF+HOME	62.9	**56.8**	**59.9**	39.1	61.2	50.5	58.5	39.0	**63.1**	53.4	60.0	39.8

The above tabel report F1 scores of static word vectors from other LMs