Learning Probabilistic Sentence Representations from Paraphrases

Mingda Chen, Kevin Gimpel

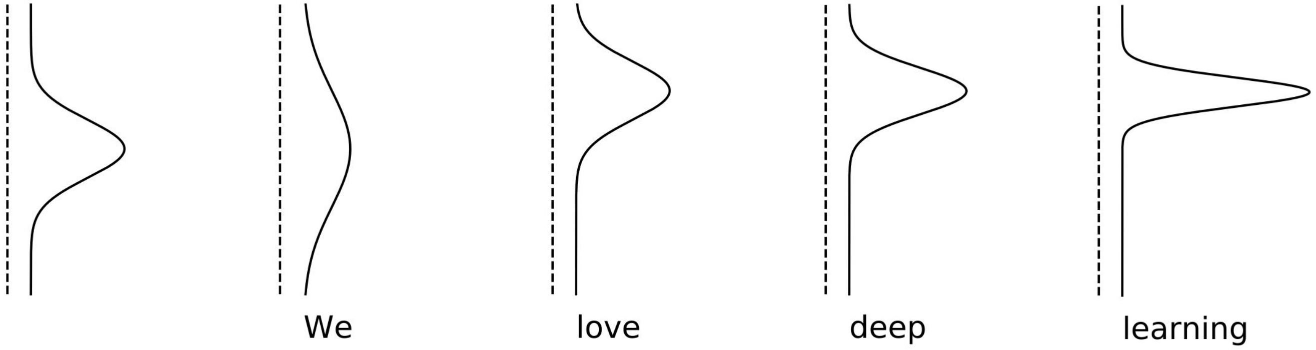
TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Motivation

- ❖ Probabilistic word representations have been shown to be useful for capturing notions of generality and entailment.
- ❖ Can we do the same thing with probabilistic sentence representations?

Proposed Approach

Word linear operator model (WLO) that treats each word as an “operator”.



1. The random variable for each sentence initially follows a standard multivariate Gaussian distribution.
2. Then, each word in the sentence transforms the random variable sequentially.
3. WLO leads to a random variable that encodes its semantic information.

Training

- Training uses paraphrases.
- A margin-based loss on paraphrase pairs (s_1, s_2)

$$\max(0, \delta - \boxed{d}(s_1, s_2) + d(s_1, n_1)) + \max(0, \delta - d(s_1, s_2) + d(s_2, n_2))$$

- Similarity function that outputs a scalar denoting the similarity of the input sentence pair.
- For probabilistic models, we use "Expected Inner Product of Gaussians" (Vilnis and McCallum, 2014).
- For other models, we use cosine similarity.

Evaluation

- Predictions:
  - based on the entropy of Gaussian distributions produced from probabilistic models.
  - based on the norm of vectors produced by other models.
- Datasets:
  - Sentence specificity: news, Twitter, Yelp reviews, and movie reviews.
    - For the news dataset, labels are either “general” or “specific”.
    - For the other datasets, labels are real values indicating specificity.
  - Stanford Natural Language Inference (SNLI) dataset.
    - Three categories: Entailment, Neutral, Contradiction.

Baselines

- Sentence representations trained on paraphrases
  - Word Sum: Summing word embeddings.
  - Word Avg: Averaging word embeddings.
- Pretrained representations from prior work
  - BERT: the representation for the “[CLS]” token.
  - ELMo Sum: summing the outputs from the last layer.
  - ELMo Avg: averaging the outputs from the last layer.

Results

	News	Twitter	Yelp	Movie
Prior work*	81.6	67.9	75.0	70.6
BERT	64.5	20.8	29.5	18.1
ELMo Avg	56.2	-9.4	-0.9	-22.5
ELMo Sum	65.4	46.2	72.7	59.3
Word Avg	54.6	-10.6	-32.3	-27.2
Word Sum	75.8	57.9	75.4	60.0
WLO	77.4	60.5	76.6	61.9

* trained on labeled sentence specificity data

Results

WLO achieves comparable performance to prior work, which was trained on labeled sentence specificity data

	News	Twitter	Yelp	Movie
Prior work*	81.6	67.9	75.0	70.6
BERT	64.5	20.8	29.5	18.1
ELMo Avg	56.2	-9.4	-0.9	-22.5
ELMo Sum	65.4	46.2	72.7	59.3
Word Avg	54.6	-10.6	-32.3	-27.2
Word Sum	75.8	57.9	75.4	60.0
WLO	77.4	60.5	76.6	61.9

* trained on labeled sentence specificity data

Results

Averaging-based models all failed on this task.

	News	Twitter	Yelp	Movie
Prior work*	81.6	67.9	75.0	70.6
BERT	64.5	20.8	29.5	18.1
ELMo Avg	56.2	-9.4	-0.9	-22.5
ELMo Sum	65.4	46.2	72.7	59.3
Word Avg	54.6	-10.6	-32.3	-27.2
Word Sum	75.8	57.9	75.4	60.0
WLO	77.4	60.5	76.6	61.9

* trained on labeled sentence specificity data

Analysis

Equal-length sentence pairs in the SNLI test set.

	Entailment	Neutral	Contradiction
ELMo	78.3	58.3	63.4
BERT	65.0	55.7	56.3
Word Avg	77.5	50.0	57.2
Word Sum	75.0	54.7	57.7
WLO	75.8	54.7	57.2

The first sentence x entails the second sentence y if (1) entropy(x) > entropy(y), or (2) norm(x) < norm(y).

Analysis

Ideal performance: 100% 50% 50%

Equal-length sentence pairs in the SNLI test set.

	Entailment	Neutral	Contradiction
ELMo	78.3	58.3	63.4
BERT	65.0	55.7	56.3
Word Avg	77.5	50.0	57.2
Word Sum	75.0	54.7	57.7
WLO	75.8	54.7	57.2

The first sentence x entails the second sentence y if (1) entropy(x) > entropy(y), or (2) norm(x) < norm(y).

Analysis

Ideal performance: 100% 50% 50%

Equal-length sentence pairs in the SNLI test set.

	Entailment	Neutral	Contradiction
ELMo	**78.3**	58.3	**63.4**
BERT	65.0	55.7	56.3
Word Avg	77.5	50.0	57.2
Word Sum	75.0	54.7	57.7
WLO	75.8	54.7	57.2

ELMo gives the best performance in the entailment category, but it seems to conflate entailment with contradiction.

Analysis

Ideal performance: 100% 50% 50%

Equal-length sentence pairs in the SNLI test set.

	Entailment	Neutral	Contradiction
ELMo	78.3	58.3	63.4
BERT	65.0	55.7	56.3
Word Avg	**77.5**	**50.0**	**57.2**
Word Sum	**75.0**	**54.7**	**57.7**
WLO	**75.8**	**54.7**	**57.2**

Models trained on paraphrases perform best, achieving around 75% accuracy in the entailment category and around 50% accuracy in other categories.

Lexical Analysis

Small norm		Large norm	
small abs. ent.	small ent.	small abs. ent.	small ent.
,	addressing	staveb	cenelec
/	derived	jerusalem	ohim
by	decree	trent	placebo
an	fundamental	microwave	hydrocarbons
gon	beneficiaries	brussels	iec
as	tendency	synthetic	paras
having	detect	christians	allah
a	reservations	elephants	milan
on	remedy	seldon	madrid
for	eligibility	burger	±
from	film-coated	experimental	ukraine
’d	breach	alison	intravenous
—	exceed	63	electromagnetic
his	flashing	prophet	131
’	objectives	diego	electrons
upon	cue	mallory	northeast
under	commonly	ö	blister
towards	howling	natalie	http
’s	vegetable	hornblower	renal
with	bursting	korea	asteroid

Table 5: Examples showing top-20 lists of large-norm or small-norm words ranked based on small absolute entropy or small entropy in WLO.

- Words with small norm and small absolute entropy have little effect, both in terms of meaning and specificity.
- They are mostly function words.

Conclusion

- ❖ We trained sentence models on paraphrase pairs and showed that they naturally capture specificity and entailment.
- ❖ We benchmarked pretrained models using norm of the sentence vector, showing they can achieve reasonable performance.
- ❖ Our proposed WLO model, which treats each word as a linear transformation operator, achieves the best performance and lends itself to analysis.

Thanks!