A Multi-Task Approach for Disentangling Syntax and Semantics in Sentence Representations

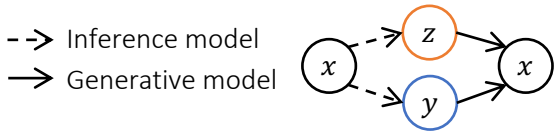
Mingda Chen, Qingming Tang, Sam Wiseman, Kevin Gimpel
Toyota Technological Institute at Chicago

Learning Sentence Representations

Sentence → (Neural Networks) → Fixed-length vector

*Can we encode semantics and syntax into **separate** representations?*

vMF-Gaussian VAE

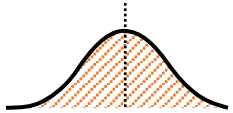


- Syntactic variable, Gaussian distribution
- Semantic variable, vMF distribution

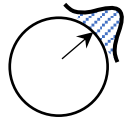
$$p_\theta(x, y, z) = p_\theta(y)p_\theta(z)p_\theta(x|y, z)$$
$$q_\phi(y, z|x) = q_\phi(y|x)q_\phi(z|x)$$

Background

Gaussian distribution



vMF distribution



Neural Architecture

$q_\phi(y|x)$ Word averaging encoder

$q_\phi(z|x)$ (1) Word averaging encoder
(2) Bidirectional LSTM encoder

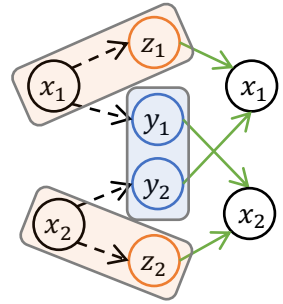
$p_\theta(x|y, z)$

(1) Bag-of-words decoder (2) LSTM decoder

$$\mathop{\mathbb{E}}_{\substack{y\sim q_\phi(y|x)\\ z\sim q_\phi(z|x)}}\left[\sum_{t=1}^{T}\log\frac{\exp g_\theta([y;z])_{x_t}}{\sum_{j=1}^{V}\exp g_\theta([y;z])_j}\right] \mathop{\mathbb{E}}_{\substack{y\sim q_\phi(y|x)\\ z\sim q_\phi(z|x)}}\left[\sum_{t=1}^{T}\log p_\theta(x_t|y,z,x_{1:t-1})\right]$$

Multi-Task Training

Aligned paraphrase x_1, x_2 → ParaNMT-50M: 50 million paraphrases



Word Position Loss (WPL)

Paraphrase Reconstruction Loss (PRL)

Discriminative Paraphrase Loss (DPL)

WPL $\mathop{\mathbb{E}}_{z\sim q_\phi(z|x)}\left[-\sum_i \log \mathrm{softmax}(f([e_i; z]))_i\right]$ → Three-layer FFNN

PRL $\mathop{\mathbb{E}}_{\substack{y_2\sim q_\phi(y|x_2)\\ z_1\sim q_\phi(z|x_1)}}[-\log p_\theta(x_1|y_2, z_1)]$ → Swap semantic variables, Keep syntactic variables

DPL $\max(0, \delta - d(x_1, x_2) + d(x_2, n_2))$ → Cosine similarity based on mean directions

Final Loss: ELBO + αWPL + βPRL + γDPL

Evaluation

Semantic similarity *Human annotated*

What is the best way to repair a cracked bathtub?
What is the best way to fix this garage floor? Score: 0

Syntactic similarity *Human annotated and Automatically parsed/tagged*

Query sentence → Find nearest neighbor based on syntactic representations → Score the parse trees/tags between these two sentences

- Labeled F1 score for constituency parsing
- Tree edit distance for constituency parsing
- Accuracy for part-of-speech tagging

Experiment Results

	Semantic similarity		Syntactic similarity	
InferSent	67.8		28.0	
SkipThought	42.1		30.9	
ELMo	57.7		30.4	
BERT	4.5		28.6	
	Sem. var. (↑)	Syn. var. (↓)	Sem. var. (↓)	Syn. var. (↑)
base	45.5	40.8	25.2	25.0
WPL	51.5	28.1	**24.1**	28.2
DPL	68.4	37.8	25.1	26.1
PRL	67.9	29.6	24.7	26.9
PRL+WPL	69.8	23.2	24.4	28.1
PRL+DPL	71.2	31.7	25.0	26.2
DPL+WPL	71.0	24.1	25.1	28.8
ALL	72.3	20.1	25.4	29.3
ALL+LSTM	**72.9**	**11.3**	25.3	**38.8**

Table 1. Pearson correlations (%) for STS benchmark and Labeled F1 scores for constituent parsing.

Nearest Neighbors

Query	Syn. similar	Sem. similar
starting	trying sharing chasing	rising wake initial forward
times	officer plan gang liar	twice later thousand once
jokes	photos finding baby	funny humor prize stars
area	bottle lesson suit bags	sector location zone fields rooms
considered	stable limited odd scary	thought assumed regard reasons
we've got to get a move on.	you'll have to get in there.	come on, we gotta move.
and that was usually the highlight of my day.	and yet that was not the strangest aspect of the painting.	i really enjoyed it when i did it.
you're gonna save her life.	you're gonna give a speech.	you will save her.
this is just such a surprise.	this is just a little gain.	oh. this is a surprise.

See our follow-up work on controlled generation
Controllable Paraphrase Generation with a Syntactic Exemplar
M. Chen, Q. Tang, S. Wiseman, K. Gimpel. ACL 2019.