A Preliminary Analysis of Syntax-Semantics Separation and Sense Selection in Word Embedding

QiPeng Guo [1] ShuaiChen Chang [1] Quan Gan [2] Xing Tian [3] XiangYang Xue [1] Zheng Zhang [3]

[1]Fudan University, China

[2]NYU

[3]NYU Shanghai

April 19, 2017

Goals and Their Intuition

- Separate syntax and semantics

Goals and Their Intuition

- Separate syntax and semantics
  - We will have a syntactic embedding (which we don't really care in qualitative analysis) and a semantic embedding for each word.
  - Syntax sends structural input to semantics and phonology (Chomsky, 1995)
  - Semantic-syntactic separation is also suggested by receptive aphasia (speaking grammatically without semantic significance) and expressive aphasia (in reverse).

Goals and Their Intuition

- Separate syntax and semantics
  - We will have a syntactic embedding (which we don't really care in qualitative analysis) and a semantic embedding for each word.
  - Syntax sends structural input to semantics and phonology (Chomsky, 1995)
  - Semantic-syntactic separation is also suggested by receptive aphasia (speaking grammatically without semantic significance) and expressive aphasia (in reverse).
- Select senses for a word given its context to deal with polysemous words.

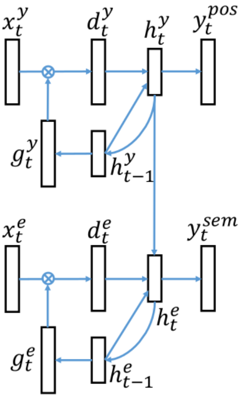
Goals and Their Intuition

- Separate syntax and semantics
  - We will have a syntactic embedding (which we don't really care in qualitative analysis) and a semantic embedding for each word.
  - Syntax sends structural input to semantics and phonology (Chomsky, 1995)
  - Semantic-syntactic separation is also suggested by receptive aphasia (speaking grammatically without semantic significance) and expressive aphasia (in reverse).
- Select senses for a word given its context to deal with polysemous words.
  - More favorably, use a single embedding to start with, and obtain the "context-dependent" embedding by *selecting* the senses from it.

Goals and Their Intuition

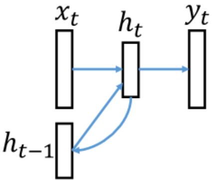
- Separate syntax and semantics
  - We will have a syntactic embedding (which we don't really care in qualitative analysis) and a semantic embedding for each word.
  - Syntax sends structural input to semantics and phonology (Chomsky, 1995)
  - Semantic-syntactic separation is also suggested by receptive aphasia (speaking grammatically without semantic significance) and expressive aphasia (in reverse).
- Select senses for a word given its context to deal with polysemous words.
  - More favorably, use a single embedding to start with, and obtain the "context-dependent" embedding by *selecting* the senses from it.
  - Contrary to one-sense-per-word approach.

Two-stream Model (Figure)



Baseline:

$$\boldsymbol{h}_t = \mathrm{GRU}(\boldsymbol{h}_{t-1}, \tanh(\boldsymbol{W}_x \boldsymbol{x}_t); \boldsymbol{\theta}_{\mathrm{GRU}})$$
$$\boldsymbol{y}_t = \mathrm{Softmax}(\boldsymbol{h}_t; \boldsymbol{\theta}_{\mathrm{Softmax}})$$



Two-stream Model

- Train a recurrent net to predict part-of-speech for the next word (syntactic stream).

$$
\begin{aligned}
\boldsymbol{d}_t^y &= \psi(\boldsymbol{h}_{t-1}^y, \boldsymbol{x}_t^y; \boldsymbol{W}_c^y, \boldsymbol{W}_x^y) \\
\boldsymbol{h}_t^y &= \mathrm{GRU}(\boldsymbol{h}_{t-1}^y, \boldsymbol{d}_t^y; \boldsymbol{\theta}_{\mathsf{GRU}}^y) \\
\boldsymbol{y}_t^{pos} &= \mathrm{Softmax}(\boldsymbol{h}_t^y; \boldsymbol{\theta}_{\mathsf{Softmax}}^y) \\
\psi(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t; \boldsymbol{W}_c, \boldsymbol{W}_x) &= g(\boldsymbol{W}_c \boldsymbol{h}_{t-1}) \odot \tanh(\boldsymbol{W}_x \boldsymbol{x}_t) \\
g(x) &= \begin{cases} 0 & x \leq 0 \\ 1/(1+e^{-x}) & x > 0 \end{cases}
\end{aligned}
$$

Two-stream Model

- Train a recurrent net to predict part-of-speech for the next word (syntactic stream).

$$
\begin{aligned}
\boldsymbol{d}_t^e &= \psi(\boldsymbol{h}_{t-1}^e, \boldsymbol{x}_t^e; \boldsymbol{W}_c^e, \boldsymbol{W}_x^e) \\
\boldsymbol{h}_t^e &= \mathrm{GRU}(\boldsymbol{h}_{t-1}^e, [\boldsymbol{d}_t^e; \boldsymbol{h}_t^y]; \boldsymbol{\theta}_{\mathsf{GRU}}^e) \\
\boldsymbol{y}_t^{sem} &= \mathrm{Softmax}(\boldsymbol{h}_t^e; \boldsymbol{\theta}_{\mathsf{Softmax}}^e) \\
\psi(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t; \boldsymbol{W}_c, \boldsymbol{W}_x) &= g(\boldsymbol{W}_c \boldsymbol{h}_{t-1}) \odot \tanh(\boldsymbol{W}_x \boldsymbol{x}_t) \\
g(x) &= \begin{cases} 0 & x \leq 0 \\ 1/(1+e^{-x}) & x > 0 \end{cases}
\end{aligned}
$$

- Fix the syntactic stream and train another recurrent net to predict the next actual word, with syntactic information (semantic stream).
- Minimize the negative log-likelihood for both part-of-speech and words.

Two-stream Model

- Train a recurrent net to predict part-of-speech for the next word (syntactic stream).

$$
\begin{aligned}
\overrightarrow{\boldsymbol{d}}_t^{y} &= \psi(\overrightarrow{\boldsymbol{h}}_{t-1}^{y}, \boldsymbol{x}_t^{y}; \overrightarrow{\boldsymbol{W}}_c^{y}, \overrightarrow{\boldsymbol{W}}_x^{y}) \\
\overleftarrow{\boldsymbol{d}}_t^{y} &= \psi(\overleftarrow{\boldsymbol{h}}_{t-1}^{y}, \boldsymbol{x}_t^{y}; \overleftarrow{\boldsymbol{W}}_c^{y}, \overleftarrow{\boldsymbol{W}}_x^{y}) \\
\overrightarrow{\boldsymbol{h}}_t^{y} &= \mathrm{GRU}(\overrightarrow{\boldsymbol{h}}_{t-1}^{y}, \overrightarrow{\boldsymbol{d}}_t^{y}; \overrightarrow{\boldsymbol{\theta}}_{\mathsf{GRU}}^{y}) \\
\overleftarrow{\boldsymbol{h}}_t^{y} &= \mathrm{GRU}(\overleftarrow{\boldsymbol{h}}_{t+1}^{y}, \overleftarrow{\boldsymbol{d}}_t^{y}; \overleftarrow{\boldsymbol{\theta}}_{\mathsf{GRU}}^{y}) \\
\boldsymbol{y}_t^{pos} &= \mathrm{Softmax}([\overrightarrow{\boldsymbol{h}}_t^{y}; \overleftarrow{\boldsymbol{h}}_{t+2}^{y}]; \boldsymbol{\theta}_{\mathsf{Softmax}}^{y})
\end{aligned}
$$

- Fix the syntactic stream and train another recurrent net to predict the next actual word, with syntactic information (semantic stream).
- Minimize the negative log-likelihood for both part-of-speech and words.

Two-stream Model

- Train a recurrent net to predict part-of-speech for the next word (syntactic stream).

$$\begin{aligned}
\overrightarrow{\boldsymbol{d}}^{e}_{t} &= \psi(\overrightarrow{\boldsymbol{h}}^{e}_{t-1}, \boldsymbol{x}^{e}_{t}; \overrightarrow{\boldsymbol{W}}^{e}_{c}, \overrightarrow{\boldsymbol{W}}^{e}_{x}) \\
\overleftarrow{\boldsymbol{d}}^{e}_{t} &= \psi(\overleftarrow{\boldsymbol{h}}^{e}_{t-1}, \boldsymbol{x}^{e}_{t}; \overleftarrow{\boldsymbol{W}}^{e}_{c}, \overleftarrow{\boldsymbol{W}}^{e}_{x}) \\
\overrightarrow{\boldsymbol{h}}^{e}_{t} &= \mathrm{GRU}(\overrightarrow{\boldsymbol{h}}^{e}_{t-1}, \left[\overrightarrow{\boldsymbol{d}}^{e}_{t}; \overrightarrow{\boldsymbol{h}}^{y}_{t}\right]; \overrightarrow{\boldsymbol{\theta}}^{e}_{\mathrm{GRU}}) \\
\overleftarrow{\boldsymbol{h}}^{e}_{t} &= \mathrm{GRU}(\overleftarrow{\boldsymbol{h}}^{e}_{t+1}, \left[\overleftarrow{\boldsymbol{d}}^{e}_{t}; \overleftarrow{\boldsymbol{h}}^{y}_{t}\right]; \overleftarrow{\boldsymbol{\theta}}^{e}_{\mathrm{GRU}}) \\
\boldsymbol{y}^{sem}_{t} &= \mathrm{Softmax}([\overrightarrow{\boldsymbol{h}}^{e}_{t}; \overleftarrow{\boldsymbol{h}}^{e}_{t+2}]; \boldsymbol{\theta}^{e}_{\mathrm{Softmax}})
\end{aligned}$$

- Fix the syntactic stream and train another recurrent net to predict the next actual word, with syntactic information (semantic stream).
- Minimize the negative log-likelihood for both part-of-speech and words.

Sense Selection

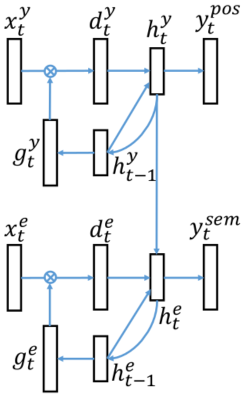
- Recall that we want to select only a subset of the dynamic embeddings:
  $\psi(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t; \boldsymbol{W}_c, \boldsymbol{W}_x) = g(\boldsymbol{W}_c \boldsymbol{h}_{t-1}) \odot \tanh(\boldsymbol{W}_x \boldsymbol{x}_t)$
- We want $\boldsymbol{g} = g(\boldsymbol{W}_c \boldsymbol{h}_{t-1})$ to be sparse.
- Additional penalty for each gating function output:

$$\mathcal{L}_s = \rho \log(\rho/\bar{g}) + (1-\rho)\log(1-\rho/\bar{g})$$

$$\bar{g} = \sum_{i=1}^{|\boldsymbol{g}|} g_i$$

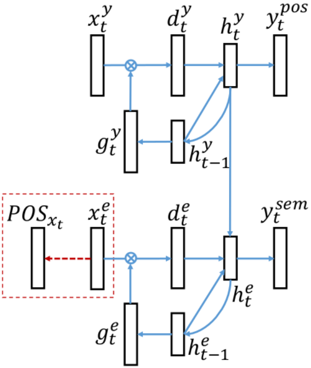
Purging Syntactic Information

- How to ensure that there is no syntactic information left in semantic embedding $\boldsymbol{x}^e$?



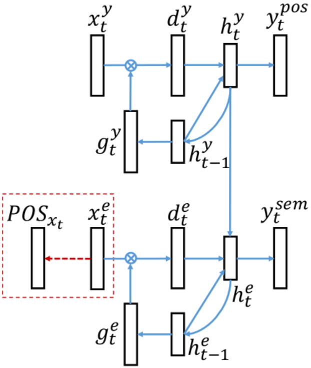
Purging Syntactic Information

- How to ensure that there is no syntactic information left in semantic embedding $\boldsymbol{x}^e$?
  - Idea: use an *adversary* to predict the part-of-speech directly from $\boldsymbol{x}^e$;
  - Then the model tries to fool the adversary .



Purging Syntactic Information

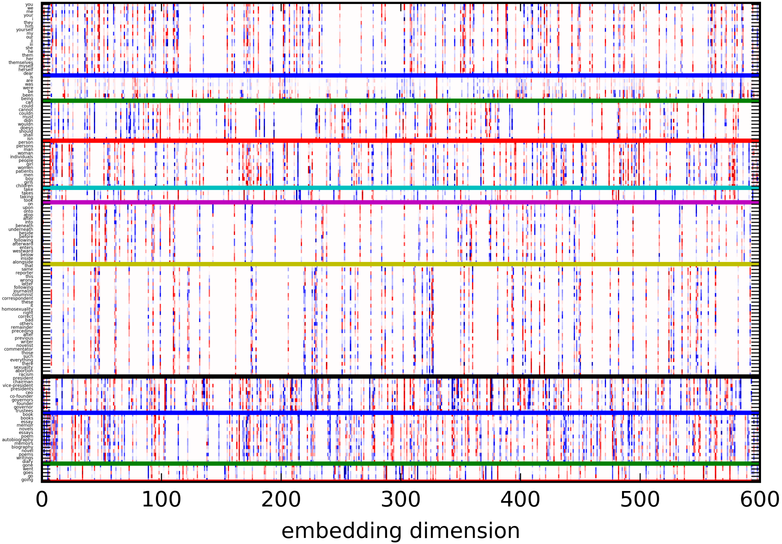
- How to ensure that there is no syntactic information left in semantic embedding $\boldsymbol{x}^e$?
  - Idea: use an *adversary* to minimize $\sum_t -\log P(\mathrm{POS}_t^e \mid \boldsymbol{x}_t^e; \theta_{\mathrm{adv}})$ w.r.t. θ_{adv} ;
  - Then the model tries to additionally minimize the penalty $\mathcal{L}_{\mathrm{adv}} = -H(\mathrm{pos} \mid \boldsymbol{x}_t^e; \theta_{\mathrm{adv}})$ w.r.t. $\boldsymbol{x}_t^e$, freezing everything else .



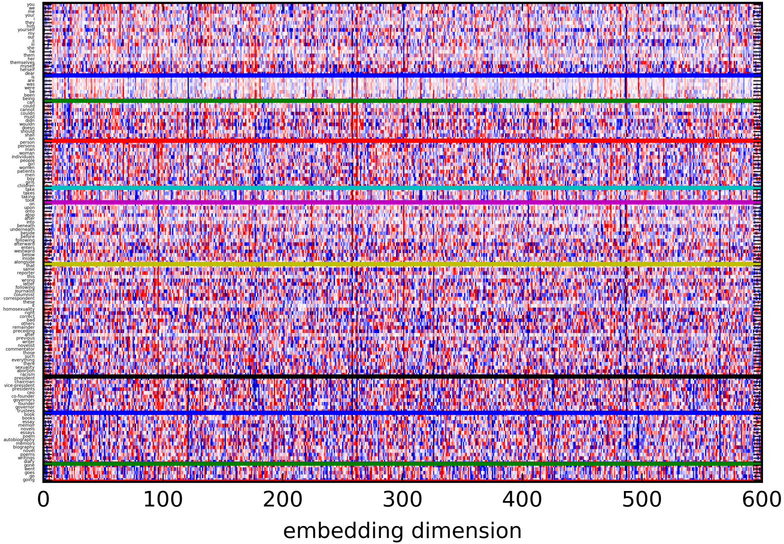
Sparsity (Results)

- Visualization (next frame): Pick a sentence and a word there, replace the word with every word in the vocabulary, and search the nearest neighbor of $\boldsymbol{d}_t^e$:

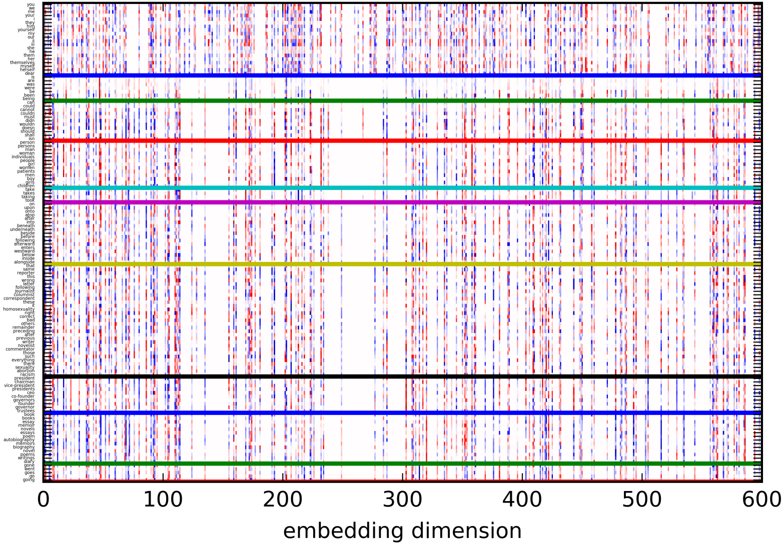
Sparsity (Results)



Sparsity (Results)



Sparsity (Results)



Measuring Cleanness

- Assumption: if separated successfully, given a word w and its embedding, the word of the nearby embedding should have more varied syntactic roles different from that of w.

$$\frac{1}{K} \sum_{w' \in NN_K(w)} 1_{POS(w)=POS(w')}$$

- Result of this quantity:

Model	**Top-5**	**Top-10**	**Top-15**	**Top-20**
Baseline	64.9%	69.8%	72.1%	73.1%
SynSem	60.7%	64.9%	66.7%	67.5%
SynSem + adv	**48.4%**	**52.9%**	**54.8%**	**55.7%**

Measuring Cleanness

- Assumption: if separated successfully, given a word w and its embedding, the word of the nearby embedding should have more varied syntactic roles different from that of w.

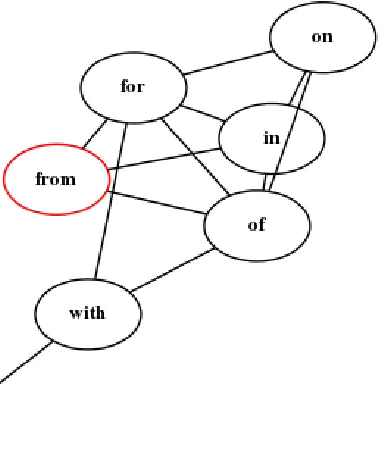
$$\frac{1}{|V|}\sum_{w\in V}\frac{1}{K}\sum_{w'\in NN_K(w)}1_{POS(w)=POS(w')}$$

- Result of this quantity:

Model	Top-5	Top-10	Top-15	Top-20
Baseline	64.9%	69.8%	72.1%	73.1%
SynSem	60.7%	64.9%	66.7%	67.5%
SynSem + adv	**48.4%**	**52.9%**	**54.8%**	**55.7%**

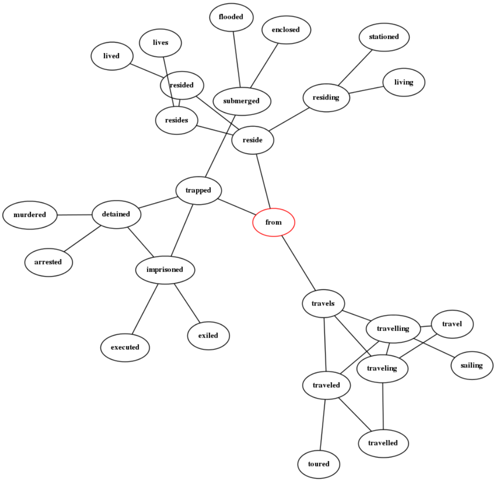
Word Association

- Start from a word embedding, find the K-nearest neighbors and repeat:



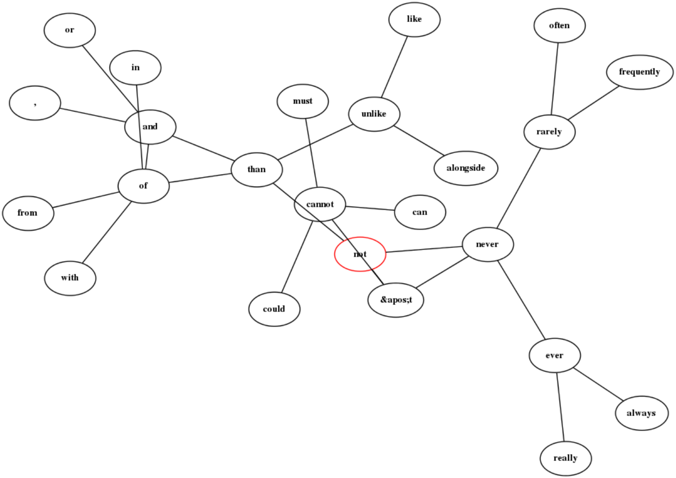
Word Association

- Start from a word embedding, find the K-nearest neighbors and repeat:



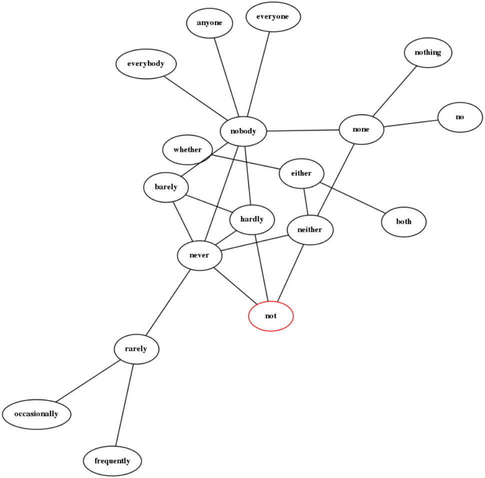
Word Association

- Start from a word embedding, find the K-nearest neighbors and repeat:



Word Association

- Start from a word embedding, find the K-nearest neighbors and repeat:



More Results: Language Modeling

Model	**Dataset**	**PPL**	$\|\theta\|$	**POS PPL**
Zaremba et al., 2014	PTB	78.4	66M	-
Merity et al., 2016	PTB	70.9	21M	-
Zilly et al., 2016	PTB	66.0	24M	-
Baseline Uni	PTB	74.4	6.7M	-
Baseline Uni + POS	PTB	71.7		-
SynSem Uni	PTB	73.1	7.1M	6.35
Baseline Uni	Wiki	70.3	98M	-
SynSem Uni	Wiki	67.0	100M	5.47
Baseline Bi	Wiki	39.0	149M	-
SynSem Bi	Wiki	**25.5**	150M	3.15

More Results: SCWS

- Stanford Contextual Word Similarities (SCWS) Dataset:
  - `... majority decisions are not binding on the` **`minority`** `...`
    `... the signatories was subject to` **`minor`** `reservations ...`
    10 humans rated an average score of 5.3

More Results: SCWS

- Stanford Contextual Word Similarities (SCWS) Dataset:
  - `...Polymers with microcrystalline regions are generally tougher ( can be` **`bent`** `more without breaking )...`
    `...Sigismund was` **`bent`** `on strengthening the power of the monarchy...`
    10 humans rated an average score of 1.7

More Results: SCWS

- Stanford Contextual Word Similarities (SCWS) Dataset:
  - `...Polymers with microcrystalline regions are generally tougher ( can be` **`bent`** `more without breaking )...`

    `...Sigismund was` **`bent`** `on strengthening the power of the monarchy...`

    10 humans rated an average score of 1.7
- We can also take a subset where each pair of queried words are identical.
  - But apparently nobody has done this before (to our best knowledge).

More Results: SCWS

- Stanford Contextual Word Similarities (SCWS) Dataset:
  - `...Polymers with microcrystalline regions are generally tougher ( can be` **`bent`** `more without breaking )...`

    `...Sigismund was` **`bent`** `on strengthening the power of the monarchy...`

    10 humans rated an average score of 1.7
- We can also take a subset where each pair of queried words are identical.
  - But apparently nobody has done this before (to our best knowledge).
- We give score by scaling the embedding (either $[\boldsymbol{d}^y; \boldsymbol{d}^e]$ or $\boldsymbol{d}^e$ only) to norm 1 and computing the cosine distance.

More Results: SCWS

Model	w/o Ctxt	with Ctxt	Polysemous
Skip-gram-300d	65.2	-	-
Huang	62.8	65.7	-
Chen	66.2	68.9	-
Neelakantan-50d	64.0	66.1	-
Neelakantan-300d	67.3	69.3	*17.2*[1]
Li et al	-	69.7	-
Qiu et al	-	66.1	-
Baseline	62.0	-	-
SynSem(400)	64.3	61.9	22.4
SynSem(600)	66.2	63.1	-
SynSem(800)	66.6	66.2	-

[1] Our reproduction

Discussion & Future Work

- In terms of performance in language modeling and SCWS, it does not seem worth such hassle...
  - Do we really need to explicitly separate & select for downstream tasks (e.g. language modeling, machine translation)?
  - Likely not, but still needs investigation.