Gating Mechanisms for Combining Character and Word-level Word Representations: An Empirical Study

Jorge A. Balazs, Yutaka Matsuo
The University of Tokyo, Graduate School of Engineering, Japan
{jorge, matsuo}@weblab.t.u-tokyo.ac.jp

東京大学 THE UNIVERSITY OF TOKYO

Paper Poster Code

Problem

Incorporating subword information into word representations has been shown to be beneficial, however there is no principled way for doing so.

Questions

- How does the method for combining character and word representations affect the quality of final word representations?
- What is effect these have in downstream performance?

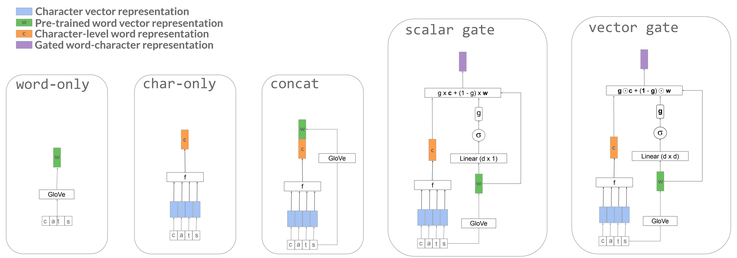
Summary

- A vector gate is the best at combining character and word representations, as measured by word similarity tasks.
- This mechanism learns that to properly model increasingly infrequent words, it has to increasingly rely on character-level information.
- Despite the increased expressivity of word representations it offers, it has no clear effect in sentence representations, as measured by sentence evaluation tasks.

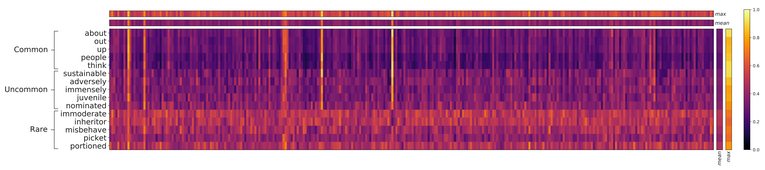
Model Description

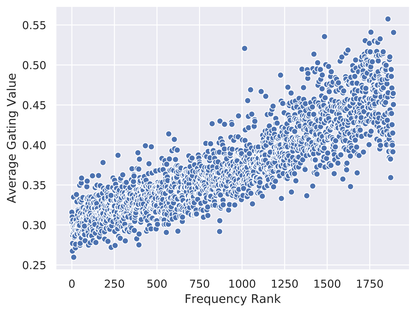
- We initialize word-level word representations $\boldsymbol{v}_i^{(w)}$ with GloVe, and create character-level word representations $\boldsymbol{v}_i^{(c)}$ with a BiLSTM over randomly-initialized character representations.
- We test 5 different ways of combining $\boldsymbol{v}_i^{(w)}$ and $\boldsymbol{v}_i^{(c)}$ into the final word representations $\boldsymbol{v}_i$:
  - **word-only**: $\boldsymbol{v}_i = \boldsymbol{v}_i^{(w)}$
  - **char-only**: $\boldsymbol{v}_i = \boldsymbol{v}_i^{(c)}$
  - **concat**: $\boldsymbol{v}_i = [\boldsymbol{v}_i^{(w)}; \boldsymbol{v}_i^{(c)}]$
  - **scalar gate**:
    $g_i = \sigma(\boldsymbol{w}^\top \boldsymbol{v}_i^{(w)} + b)$
    $\boldsymbol{v}_i = g_i \boldsymbol{v}_i^{(c)} + (1 - g_i) \boldsymbol{v}_i^{(w)}$
  - **vector gate**:
    $\boldsymbol{g}_i = \sigma(\boldsymbol{W} \boldsymbol{v}_i^{(w)} + \boldsymbol{b})$
    $\boldsymbol{v}_i = \boldsymbol{g}_i \odot \boldsymbol{v}_i^{(c)} + (1 - \boldsymbol{g}_i) \odot \boldsymbol{v}_i^{(w)}$
- We feed the final word representations $\boldsymbol{v}_i$ to a BiLSTM, and max-pool its output to obtain a sentence representation.
- Models are trained in the SNLI and MultiNLI (MNLI) datasets. The combined representation of the premise-hypothesis pair is defined as $\boldsymbol{s} = [\boldsymbol{s}_p; \boldsymbol{s}_h; |\boldsymbol{s}_p - \boldsymbol{s}_h|; \boldsymbol{s}_p \odot \boldsymbol{s}_h]$, where $\boldsymbol{s}_p$ is the sentence representation of the premise and $\boldsymbol{s}_h$ that of the hypothesis.
- $\boldsymbol{s}$ is finally mapped to label space through a fully-connected network.

Gating Mechanisms



Gating Values





Word-level Results

		MEN	MTurk287	MTurk771	RG65	RW	SimLex999	SimVerb3500	WS353	WS353R	WS353S
SNLI	w	71.78	35.40	49.05	61.80	18.43	19.17	10.32	39.27	28.01	53.42
	c	9.85	-5.65	0.82	-5.28	17.81	0.86	2.76	-2.20	0.20	-3.87
	cat	71.91	**35.52**	48.84	62.12	18.46	19.10	10.21	39.35	28.16	53.40
	sg	70.49	34.49	46.15	59.75	18.24	17.20	8.73	35.86	23.48	50.83
	vg	**80.00**	32.54	**62.09**	**68.90**	**20.76**	**37.70**	**20.45**	**54.72**	**47.24**	**65.60**
MNLI	w	68.76	50.15	68.81	65.83	18.43	42.21	25.18	61.10	58.21	70.17
	c	4.84	0.06	1.95	-0.06	12.18	3.01	1.52	-4.68	-3.63	-3.65
	cat	68.77	50.40	68.77	65.92	18.35	42.22	25.12	61.15	58.26	70.21
	sg	67.66	49.58	68.29	64.84	18.36	41.81	24.57	60.13	57.09	69.41
	vg	**76.69**	**56.06**	**70.13**	**69.00**	**25.35**	**48.40**	**35.12**	**68.91**	**64.70**	**77.23**

Sentence-level Results

		Classification							Entailment	Relatedness	Semantic Textual Similarity	
		CR	MPQA	MR	SST2	SST5	SUBJ	TREC	SICKE	SICKR†	STS16†	STSB†
SNLI	w	80.50	84.59	74.18	78.86	42.33	**90.38**	**86.83**	86.37	88.52	59.90*	71.29*
	c	74.90*	78.86*	65.93*	69.42*	35.56*	82.97*	83.31*	84.13*	83.89*	59.33*	67.20*
	cat	80.44	84.66	74.31	78.37	41.34*	90.28	85.80*	**86.40**	88.44	59.90*	71.24*
	sg	**80.59**	84.60	**74.49**	**79.04**	41.63*	90.16	86.00	86.10*	**88.57**	60.05*	71.34*
	vg	80.42	**84.66**	74.26	78.87	**42.38**	90.07	85.97	85.67	88.31*	**60.92**	**71.99**
MNLI	w	83.80	**89.13**	79.05	83.38	45.21	91.79	89.23	84.92	86.33	66.08	71.96*
	c	70.23*	72.19*	62.83*	64.55*	32.47*	79.49*	74.74*	81.53*	75.92*	51.47*	61.74*
	cat	**83.96**	89.12	**79.23**	83.70	45.08*	**91.92**	**90.03**	**85.06**	86.45	**66.17**	71.82*
	sg	83.88	89.06	79.22	83.71	45.26	91.66*	88.83*	84.96	86.40	65.49*	71.87*
	vg	83.45*	89.05	79.13	**83.87**	**45.88**	91.55*	89.49	84.82	**86.50**	65.75	**72.82**

Correlations Between Sentence and Word-level Tasks