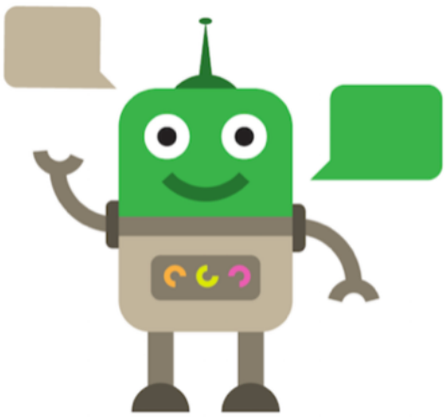
Can Machines Read *Jmulbed Senetcnes?*

COS597E Advanced NLP

Runzhe Yang, Zhongqiao Gao

2019.01.14



Motivation

"For emaxlpe, it deson't mttaer in waht oredr the ltteers in a wrod aepaprs, the olny iprmoatnt tihng is taht the frist and lsat ltteer are in the rghit pcale. The rset can be a toatl mses and you can sitll raed it wouthit pobelrm."

Can you read the above paragraph?

"For example, it dose't matter in what order the letter in a word appears, the only important thing is that the first and last letter are in the right place. The rest can be a total mess and you can still read it without problem."

Graham Ernest Rawlinson. 1976.
The significance of letter position in word recognition.
Ph.D. thesis, University of Nottingham

Motivation

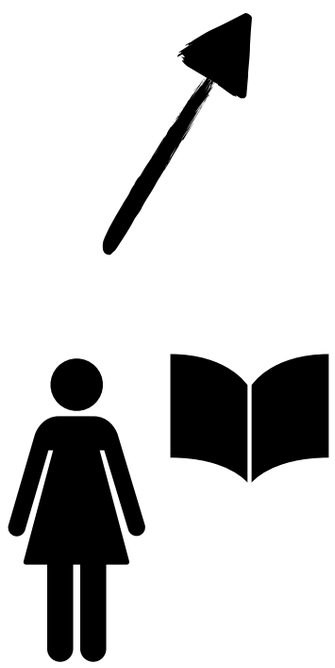
"For emaxlpe, it deson't mttaer in waht oredr the ltteers in a wrod aepaprs, the olny iprmoatnt tihng is taht the frist and lsat ltteer are in the rghit pcale. The rset can be a toatl mses and you can sitll raed it wouthit pobelrm."

"For example, it dose't matter in what order the letter in a word appears, the only important thing is that the first and last letter are in the right place. The rest can be a total mess and you can still read it without problem."

Graham Ernest Rawlinson. 1976.
The significance of letter position in word recognition.
Ph.D. thesis, University of Nottingham

Motivation

"For emaxlpe, … can sitll raed it wouthit pobelrm."



1. The strong language a priori

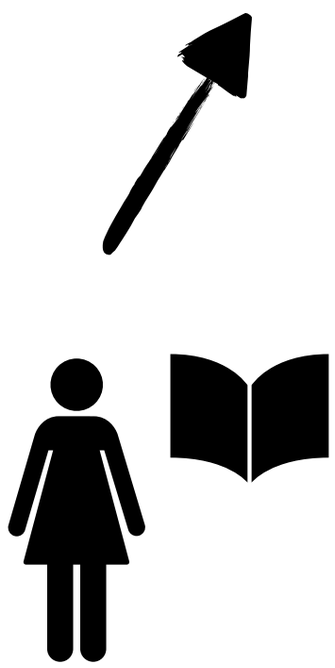
Pr[problem | pobelrm] ∝ Pr[pobelrm | problem] x Pr[problem]

Pr[pobelrm | pobelrm] ∝ Pr[pobelrm | pobelrm] x Pr[pobelrm]

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008.
Bayesian models of cognition.

Motivation

"For emaxlpe, … can sitll raed it wouthit pobelrm."



1. The strong language a priori

"you thought" "eyes read " "brains read"

Pr[problem | pobelrm] ∝ Pr[pobelrm | problem] x Pr[problem]

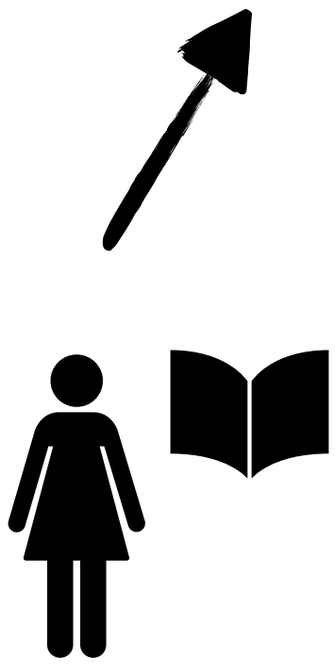
Pr[pobelrm | pobelrm] ∝ Pr[pobelrm | pobelrm] x Pr[pobelrm]

Pr[problem] >> Pr[pobelrm] *(strong a priori)*

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008.
Bayesian models of cognition.

Motivation

"For emaxlpe, … can sitll raed it wouthit pobelrm."



1. The strong language a priori

"you thought" "eyes read " "brains read"

Pr[problem | pobelrm] ∝ Pr[pobelrm | problem] x Pr[problem]

0.36 ≈ 0.4 x 0.9

Pr[pobelrm | pobelrm] ∝ Pr[pobelrm | pobelrm] x Pr[pobelrm]

0.01 ≈ 1.0 x 0.01

Pr[problem] >> Pr[pobelrm] *(strong a priori)*

pobelrm → problem

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008.
Bayesian models of cognition.

Motivation

"For emaxlpe, … can sitll raed it wouthit pobelrm."

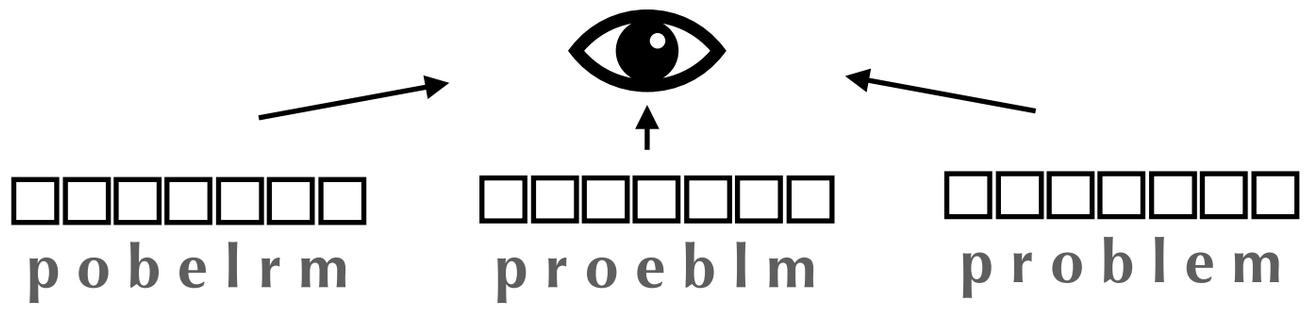
1. The strong language a priori

"you thought" "eyes read " "brains read"

Pr[problem | pobelrm] ∝ Pr[pobelrm | problem] x Pr[problem]

Pr[problem] >> Pr[pobelrm] *(strong a priori)*

2. The parallel processing hypothesis

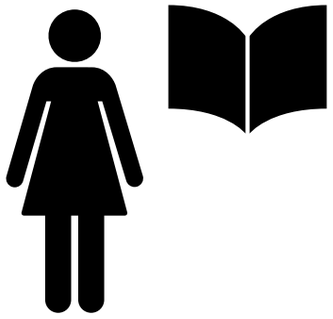


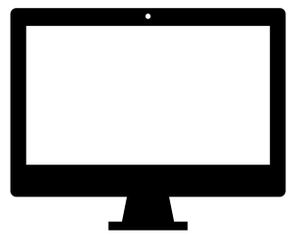
James T Townsend. 1990.
Serial vs. parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished.
Psychological Science, 1(1):46–54.

Motivation

"For emaxlpe, … can sitll raed it wouthit pobelrm."

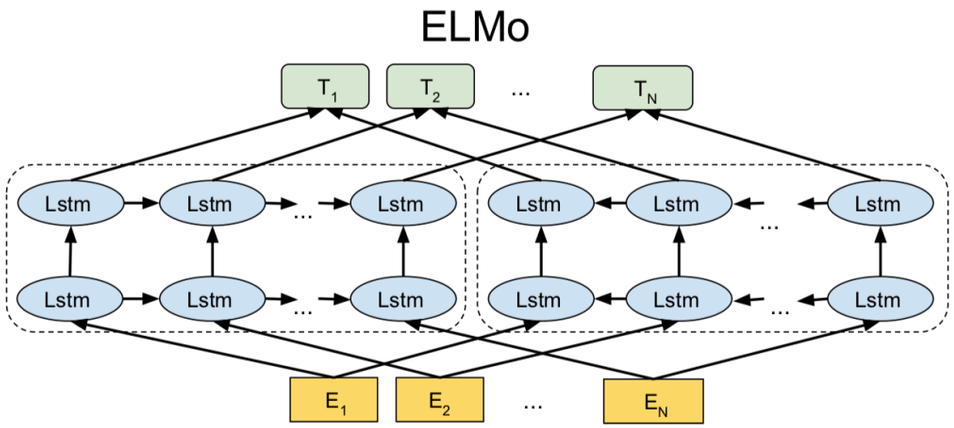
Humans can read jumbled sentences,
but can machines do?



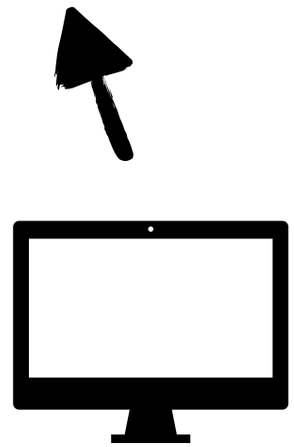


Background

1. The strong language a priori (human) LM using bidirectional context (machine)



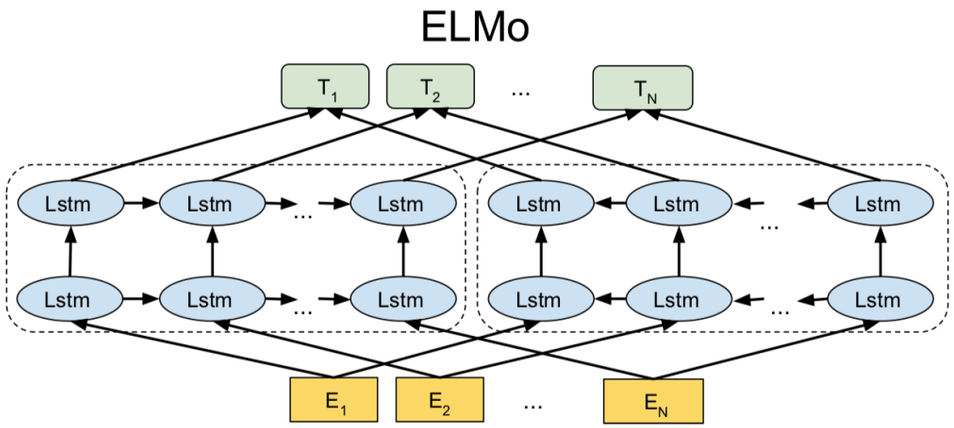
Humans can read jumbled sentences, but can machines do?



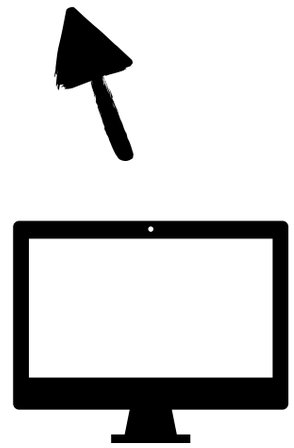
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. *arXiv preprint arXiv:1802.05365*.

Background

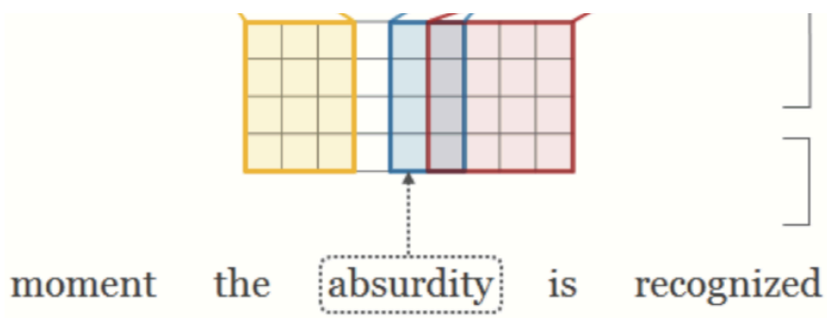
1. The strong language a priori (human)
LM using bidirectional context (machine)



Humans can read jumbled sentences, but can machines do?



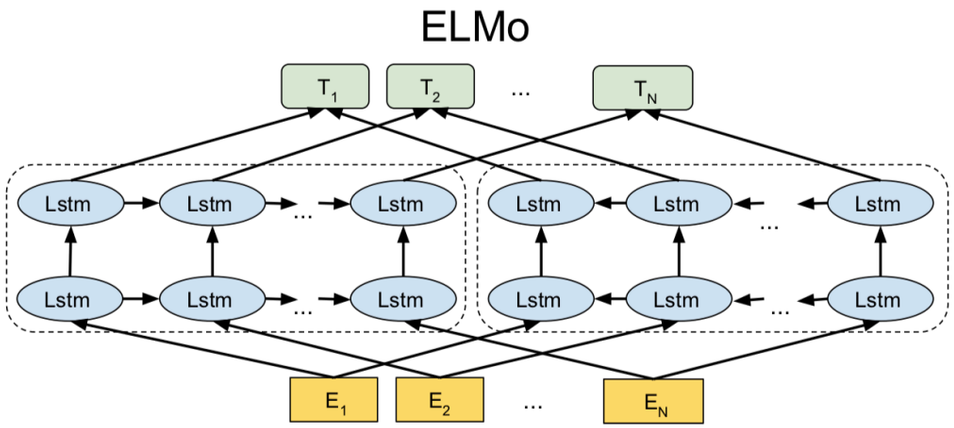
2. The parallel processing hypothesis (human)
Character-Level CNN / LSTM / Transformers



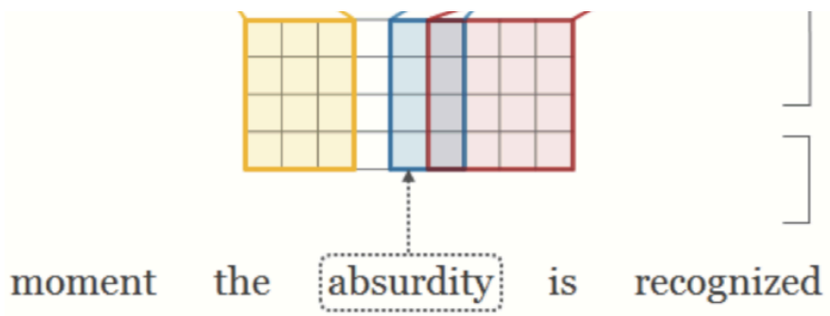
Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016, February).
Character-Aware Neural Language Models. In AAAI (pp. 2741-2749).

Background

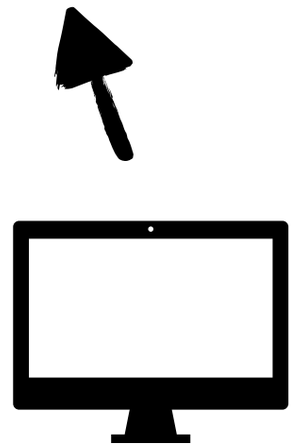
1. The strong language a priori (human)
LM using bidirectional context (machine)



2. The parallel processing hypothesis (human)
Character-Level CNN / LSTM / Transformers



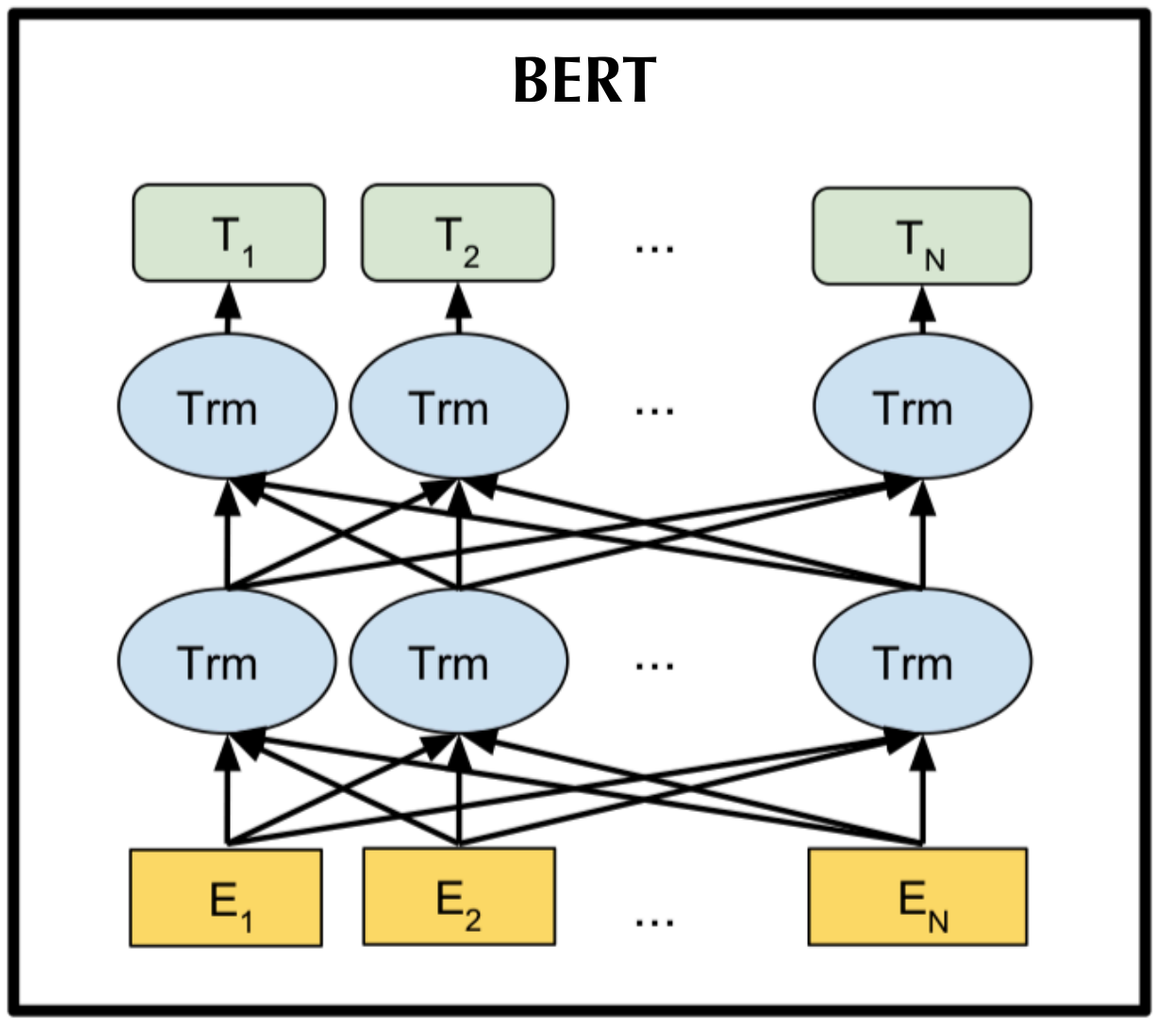
Humans can read jumbled sentences, but can machines do?



machine with noisy input:
Subramaniam et al., 2009,
Sakaguchi et al., 2016
Belinkov and Bisk, 2017
…

Methods

Pre-trained Deep Bidirectional Transformers



J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018.
***BERT**: Pre-training of Deep Bidirectional Transformers for Language Understanding*. ArXiv e-prints.

Methods

Jumbling Schemes

"What is considered the costliest disaster the insurance industry has ever faced ?"

Operation	Example (word-level jumbling)	Example (character-level jumbling)
Swap	What considered the is costliest disas-ter insurance the industry has ever ? faced	What is consdiered the csotliset dis-asetr the inuransce idnustry has ever faedc ?
Omit	What considered the costliest disater the insurance industry faced ?	What is consiered the cosliest saster the isrance industrindustry has ever faced ?
Add	What is considered the costliest disas-ter disaster the insurance industry has ever faced ?	Whapt is considerehd thes costl-ciessteb disasterb the insurancet indus-trydu has everyu faced ?

Examples of processed sentences under word-level and character-level jumbling schemes with probability 0.2.

Methods

Downstream Tasks

Customer Review (CR) (Wang and Manning, 2012)**, Text REtrieval (TREC)**(Li and Roth, 2002)**, and Semantic Text Similarity (STS)** (Cer et al., 2017)

Dataset	Task Description	Example Text	Label
CR	Sentiment analysis of reviews	*We tried it out Christmas night and it worked great*	Positive
TREC	Question Answering	*What are the twin cities?*	LOC:city
STS	Measuring the semantic similarity	{*Liquid ammonia leak kills 15 in Shanghai, Liquid ammonia leak kills at least 15 in Shanghai*}	4.6

Task descriptions for Custom Review dataset, Text REtrieval Conference, and Semantic Text Similarity with example input sentence and the corresponding ground truth label.

Methods

Downstream Tasks

(logistic regression) (logistic regression)

Customer Review (CR) (Wang and Manning, 2012)**, Text REtrieval (TREC)**(Li and Roth, 2002)**, and Semantic Text Similarity (STS)** (Cer et al., 2017)

Dataset	Task Description	Example Text	Label
CR	Sentiment analysis of reviews	*We tried it out Christmas night and it worked great*	Positive
TREC	Question Answering	*What are the twin cities?*	LOC:city
STS	Measuring the semantic similarity	{*Liquid ammonia leak kills 15 in Shanghai*, *Liquid ammonia leak kills at least 15 in Shanghai*}	4.6

Task descriptions for Custom Review dataset, Text REtrieval Conference, and Semantic Text Similarity with example input sentence and the corresponding ground truth label.

Methods

Downstream Tasks

(logistic regression) (logistic regression)

Customer Review (CR) (Wang and Manning, 2012)**, Text REtrieval (TREC)**(Li and Roth, 2002)**, and Semantic Text Similarity (STS)** (Cer et al., 2017)

(Pearson Coefficient)

Dataset	Task Description	Example Text	Label
CR	Sentiment analysis of reviews	*We tried it out Christmas night and it worked great*	Positive
TREC	Question Answering	*What are the twin cities?*	LOC:city
STS	Measuring the semantic similarity	{*Liquid ammonia leak kills 15 in Shanghai*, *Liquid ammonia leak kills at least 15 in Shanghai*}	4.6

Task descriptions for Custom Review dataset, Text REtrieval Conference, and Semantic Text Similarity with example input sentence and the corresponding ground truth label.

Experiments

Questions

(1) At which level of the jumbled text, can the neural NLP systems understand?

(2) What factors may influence neural NLP systems to understand jumbled sentences?

(3) How can we possibly make the neural NLP systems more robust?

Experiments

Effects of Jumbling Levels and Degrees

Operation	CR	TREC	STS
Random (baseline)	61.86	22.0	-0.014
Swap(char)	71.55	67.8	0.246
Omit(char)	69.03	68.4	0.286
Add(char)	69.91	66.4	0.293
Swap(word)	82.46	90.4	0.538
Omit(word)	82.68	90.4	0.530
Add(word)	82.70	87.0	0.536
Original	84.64	91.0	0.604

The test performance comparison between word-level and character-level jumbled sentence embeddings, original sentence embedding and random embedding, evaluated on three different downstream tasks with a jumbling probability 0.2.

Experiments

Effects of Jumbling Levels and Degrees

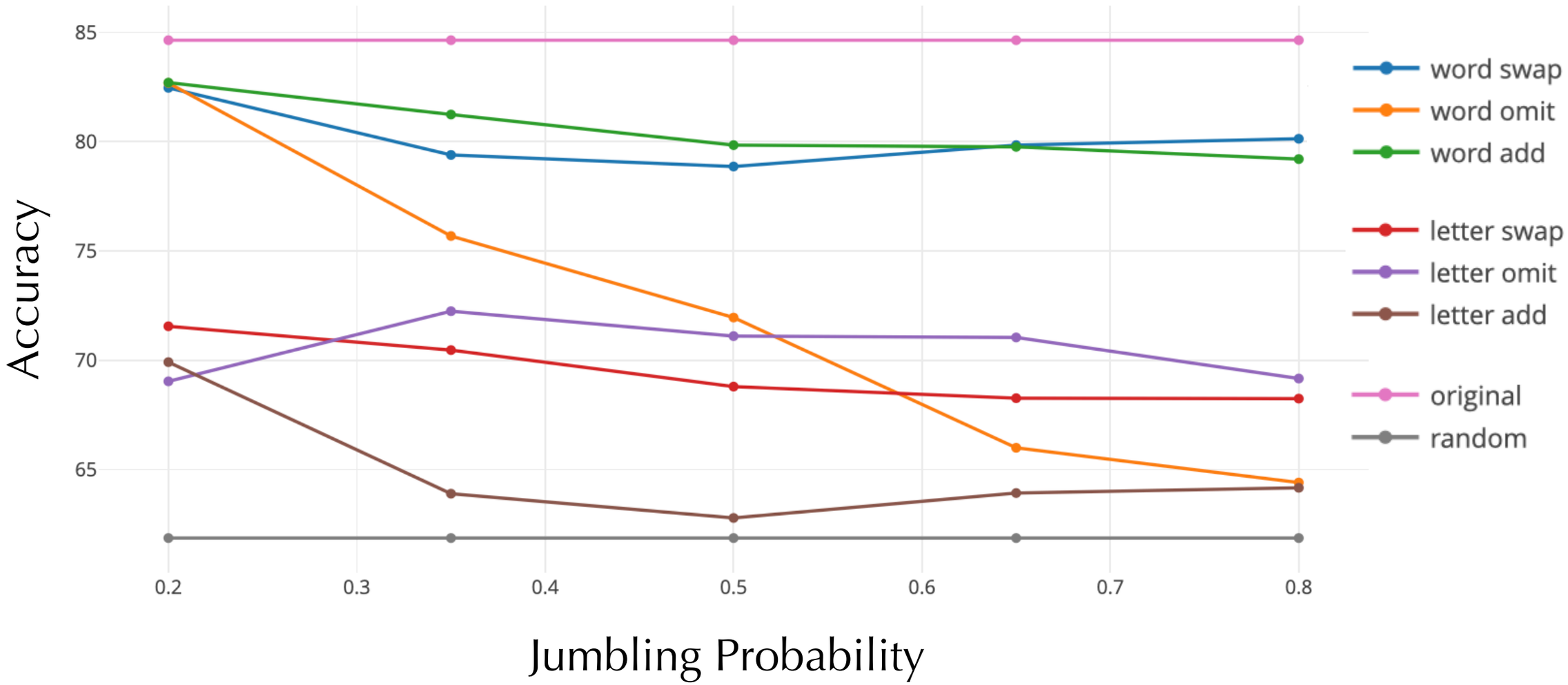
Operation	CR	TREC	STS
Random (baseline)	61.86	22.0	-0.014
Swap(char)	71.55	67.8	0.246
Omit(char)	69.03	68.4	0.286
Add(char)	69.91	66.4	0.293
Swap(word)	82.46	90.4	0.538
Omit(word)	82.68	90.4	0.530
Add(word)	82.70	87.0	0.536
Original	84.64	91.0	0.604

The test performance comparison between word-level and character-level jumbled sentence embeddings, original sentence embedding and random embedding, evaluated on three different downstream tasks with a jumbling probability 0.2.

Experiments

Effects of Jumbling Levels and Degrees

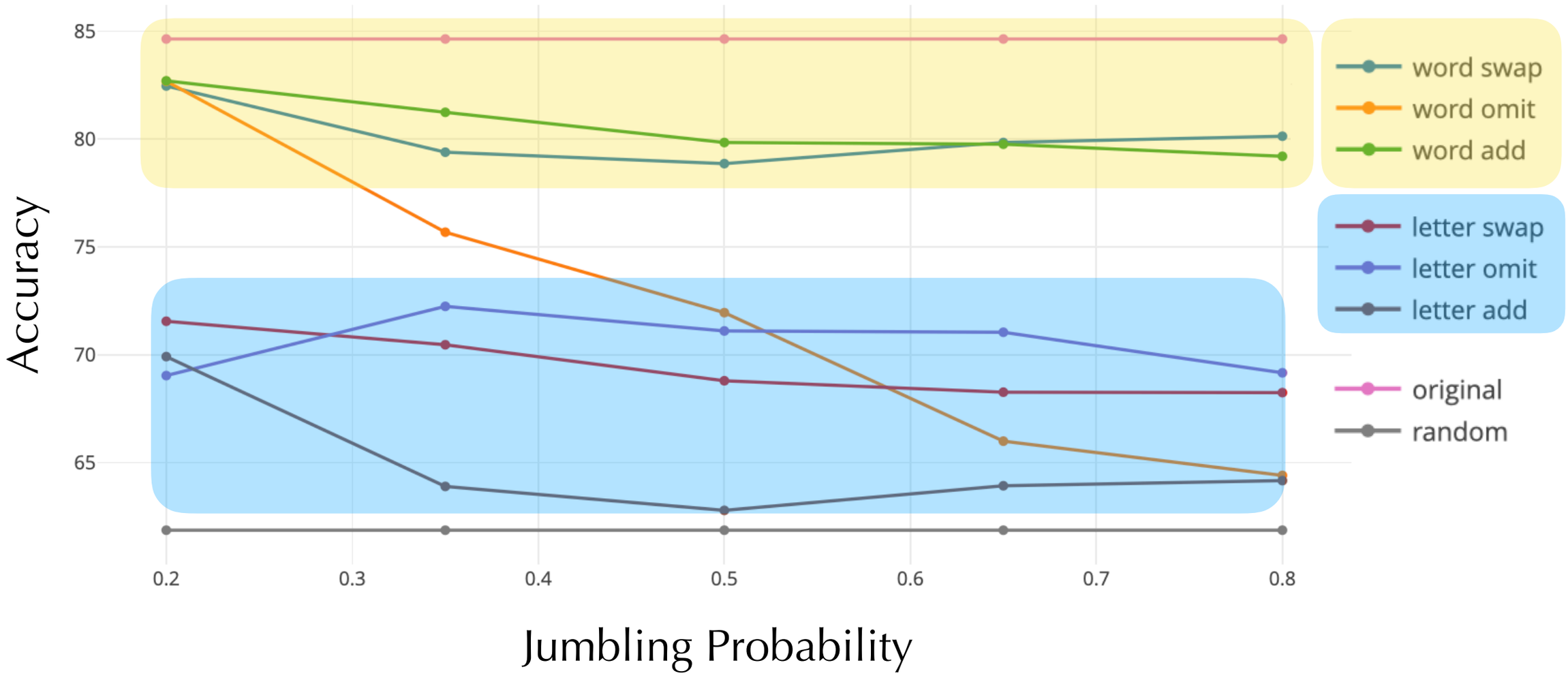
Customer Review



Experiments

Effects of Jumbling Levels and Degrees

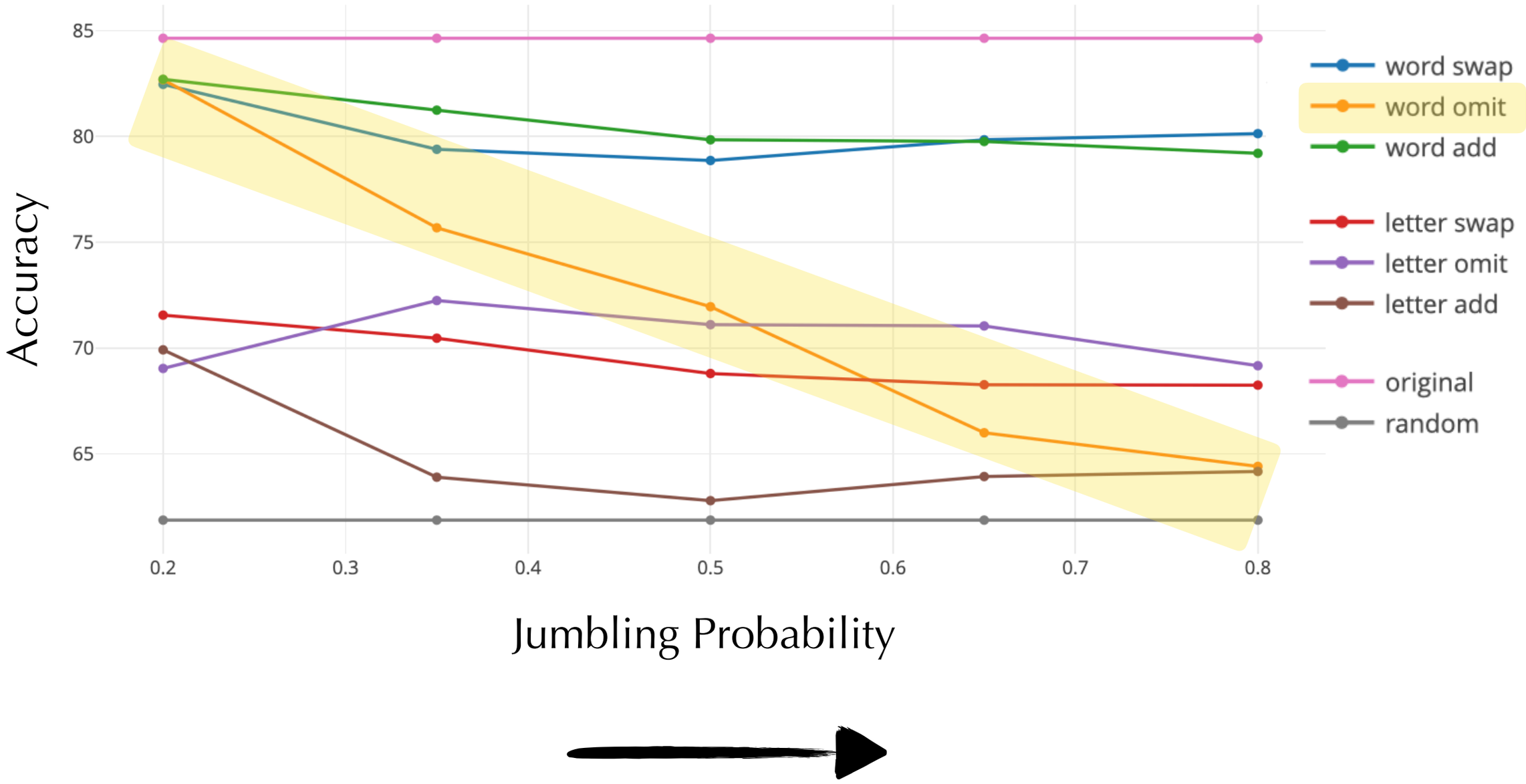
Customer Review



Experiments

Effects of Jumbling Levels and Degrees

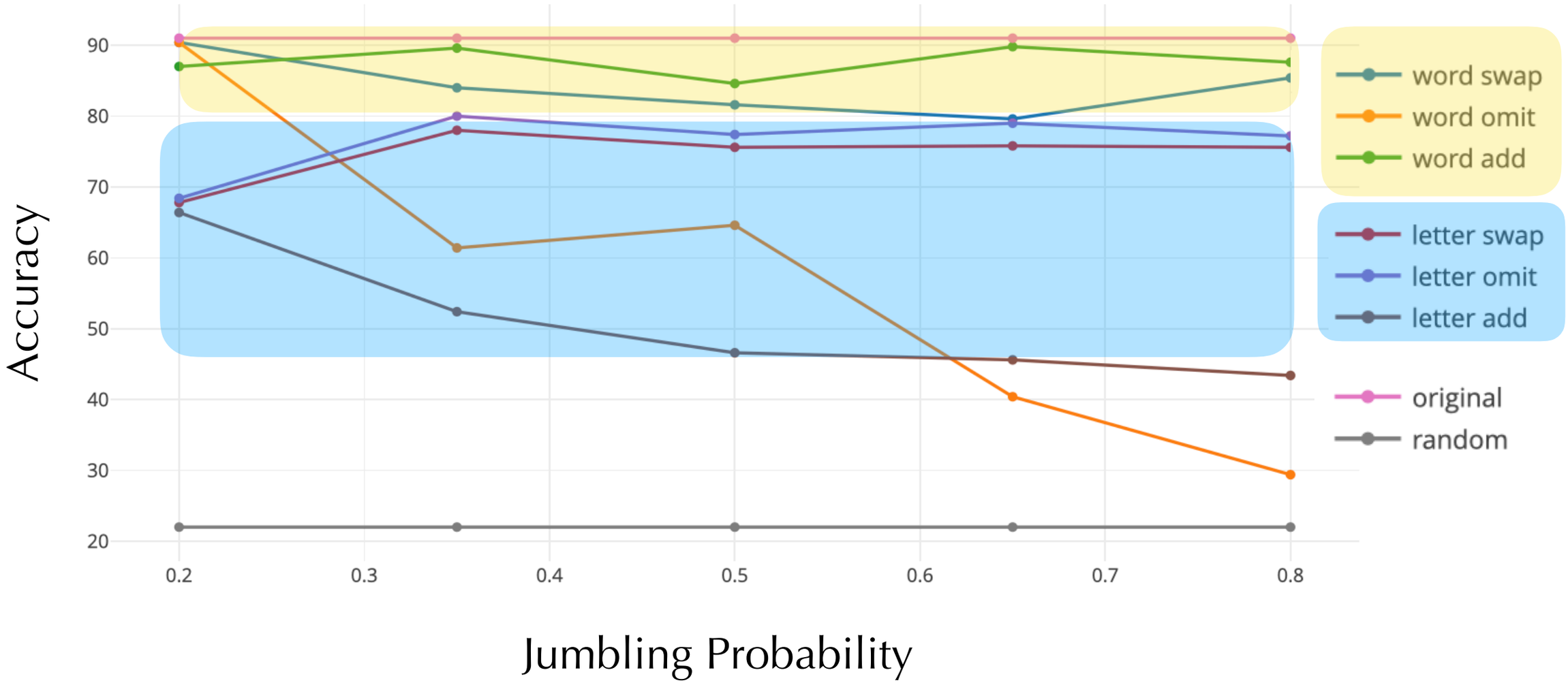
Customer Review



Experiments

Effects of Jumbling Levels and Degrees

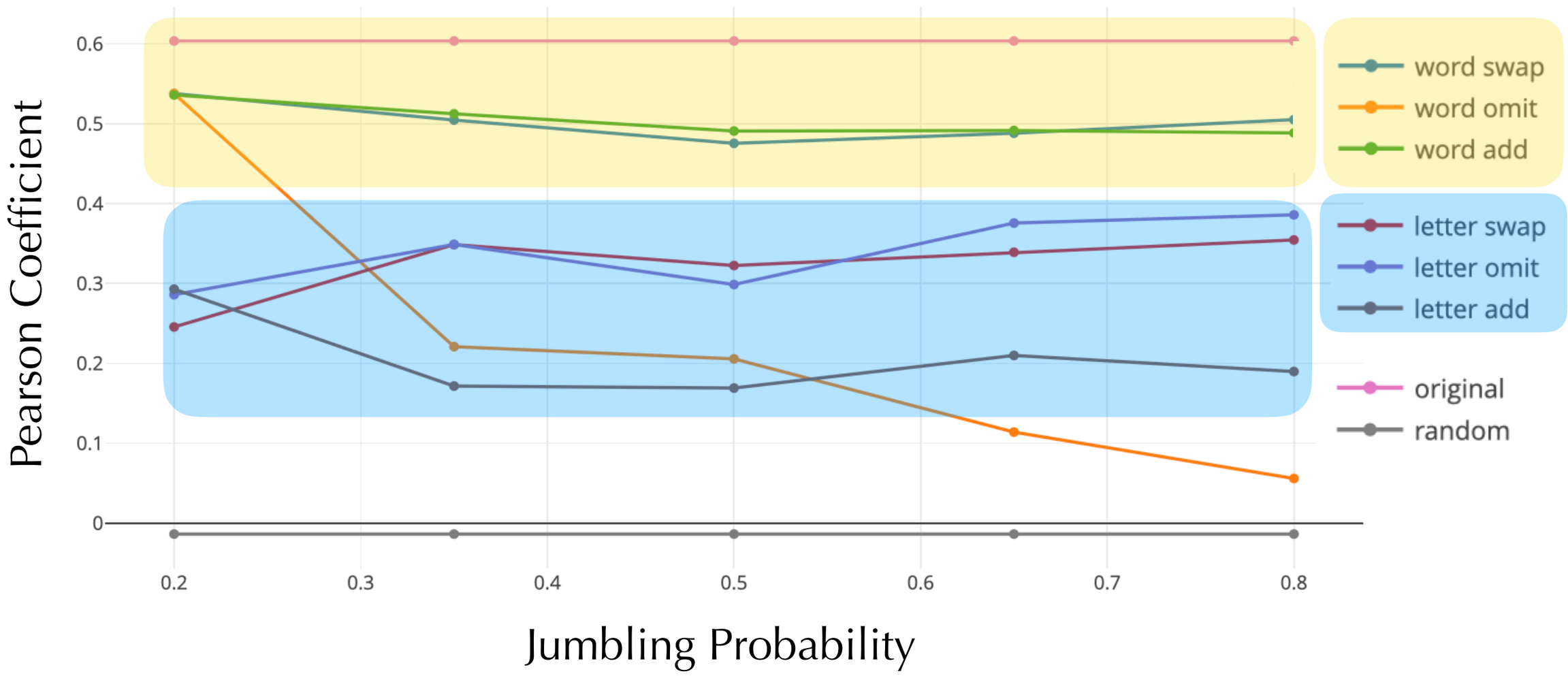
Text REtrieval



Experiments

Effects of Jumbling Levels and Degrees

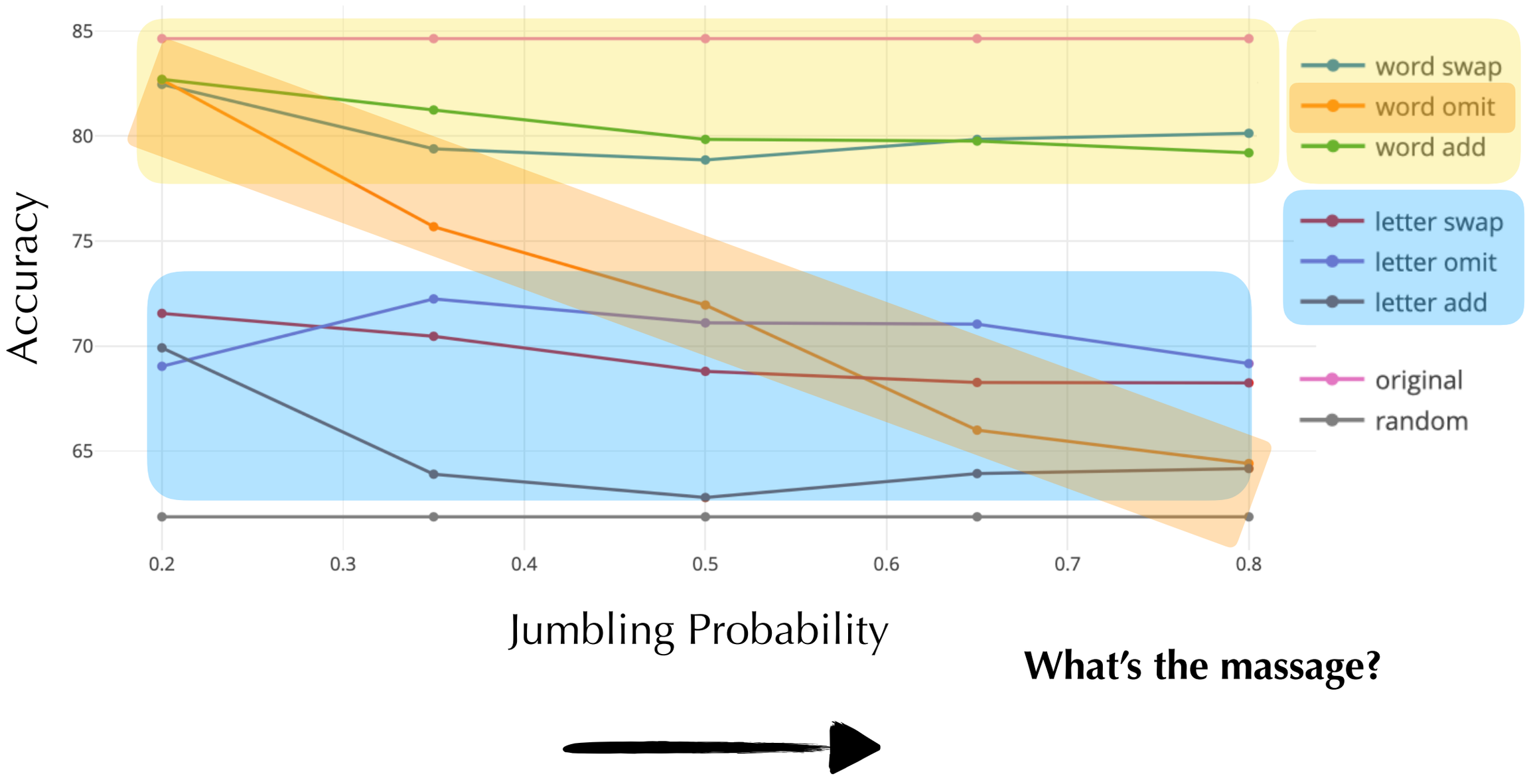
Semantic Text Similarity



Experiments

Effects of Jumbling Levels and Degrees

Customer Review



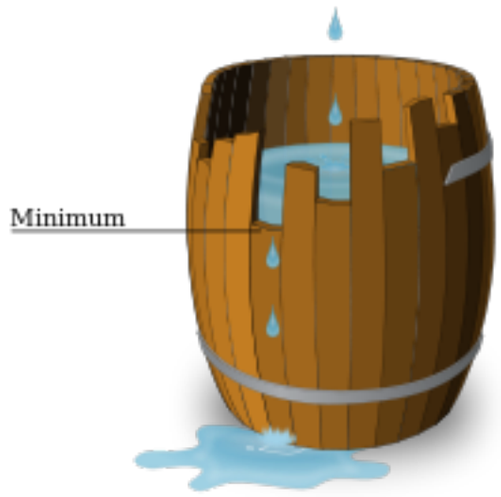
What's the massage?

Experiments

Effects of Jumbling Levels and Degrees

word-level omit: lose much information

other schemes: lose information but the under remaining information is still above machine's extraction capacity.



IMPLICATION: The main restriction of performance is the machine's capacity of capturing semantic information from jumbled sentence instead of competency or clearance of sentences.

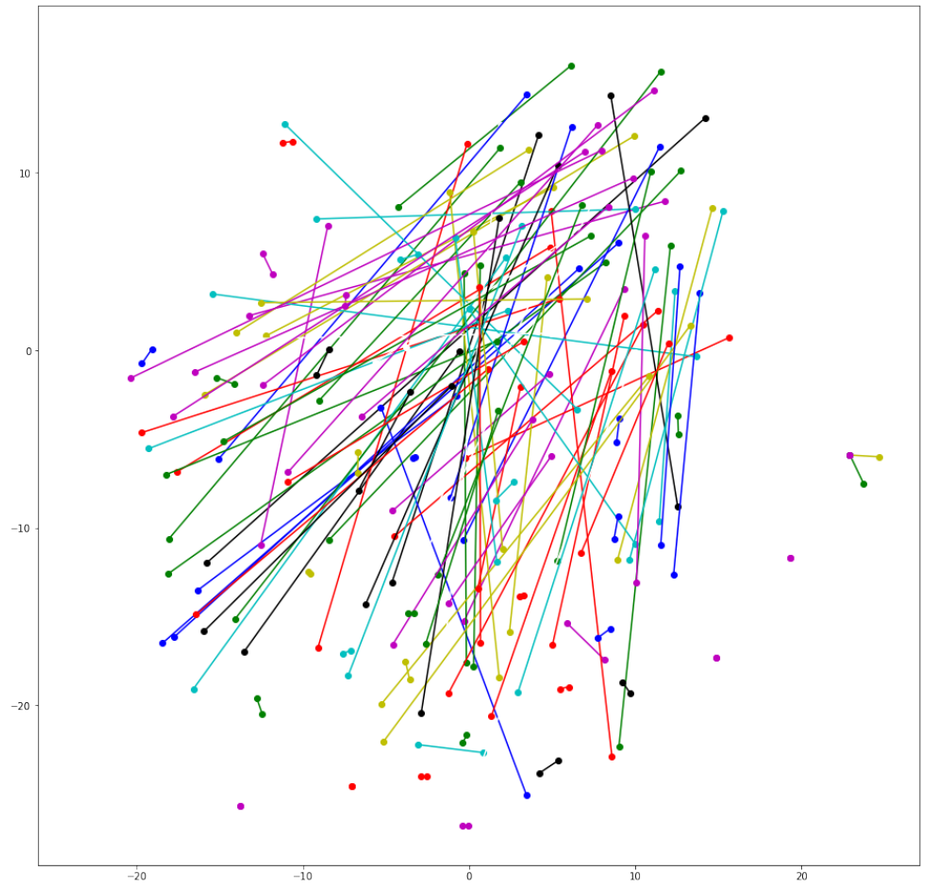
Experiments

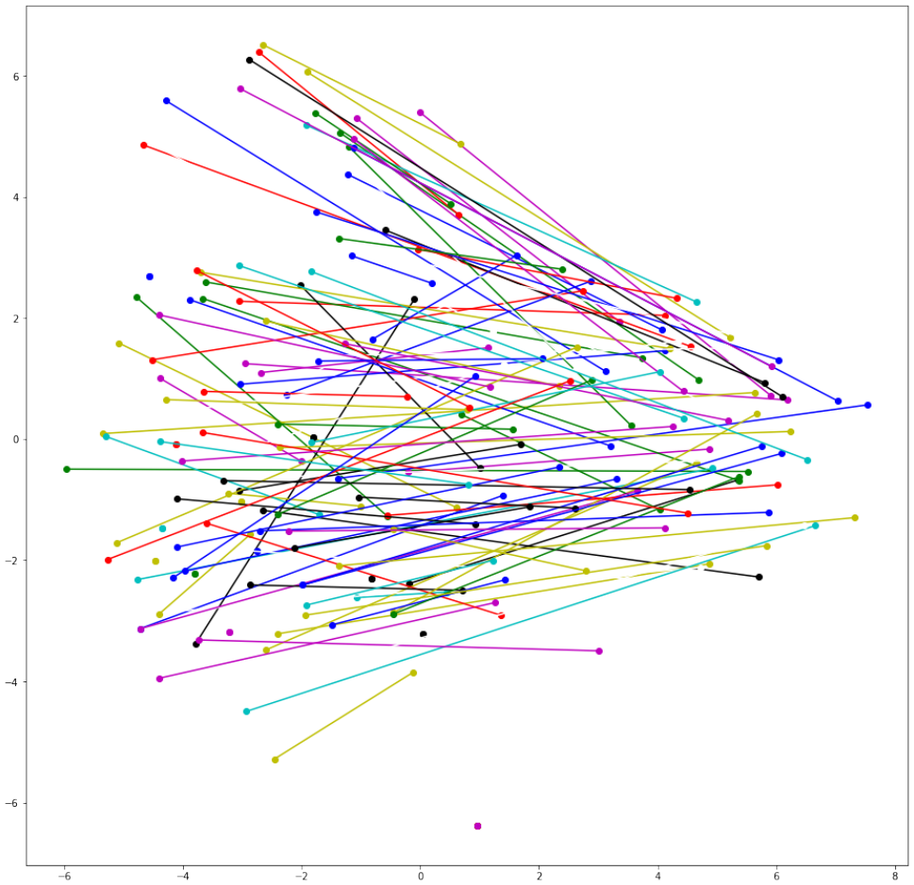
Visualization of Jumbled Sentence Embedding

t-SNE or PCA

↙ preserve local geometry

↘ preserve distances / angles





Character-level “add”, CR dataset

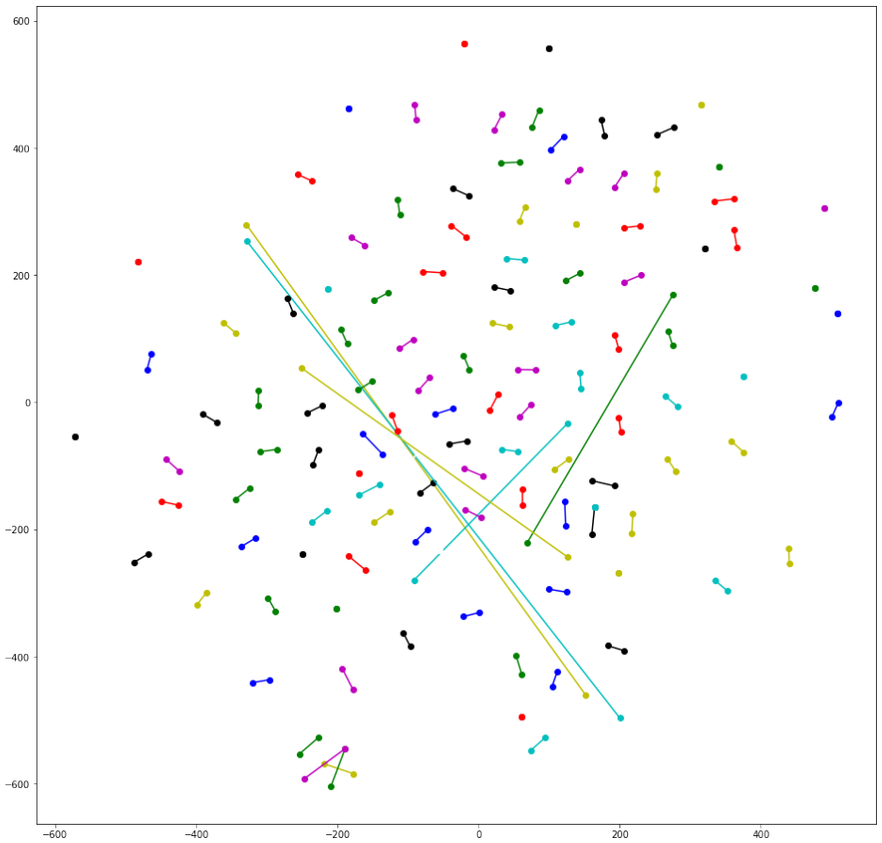
Experiments

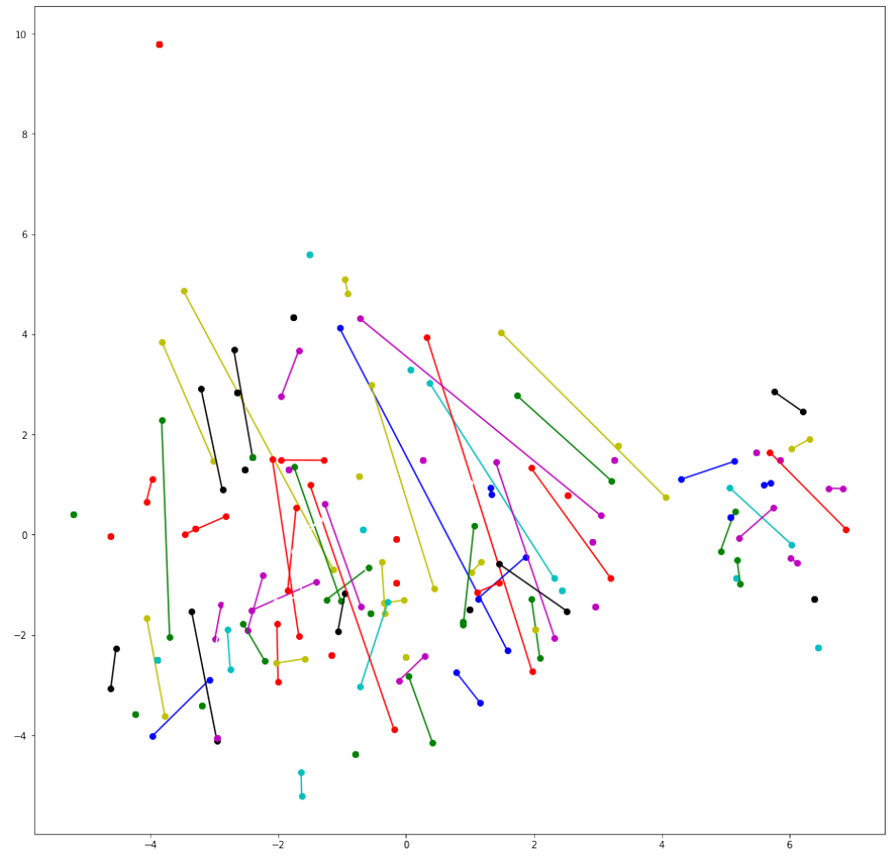
Visualization of Jumbled Sentence Embedding

t-SNE or PCA

↙ preserve local geometry

↘ preserve distances / angles





Word-level “add”, CR dataset

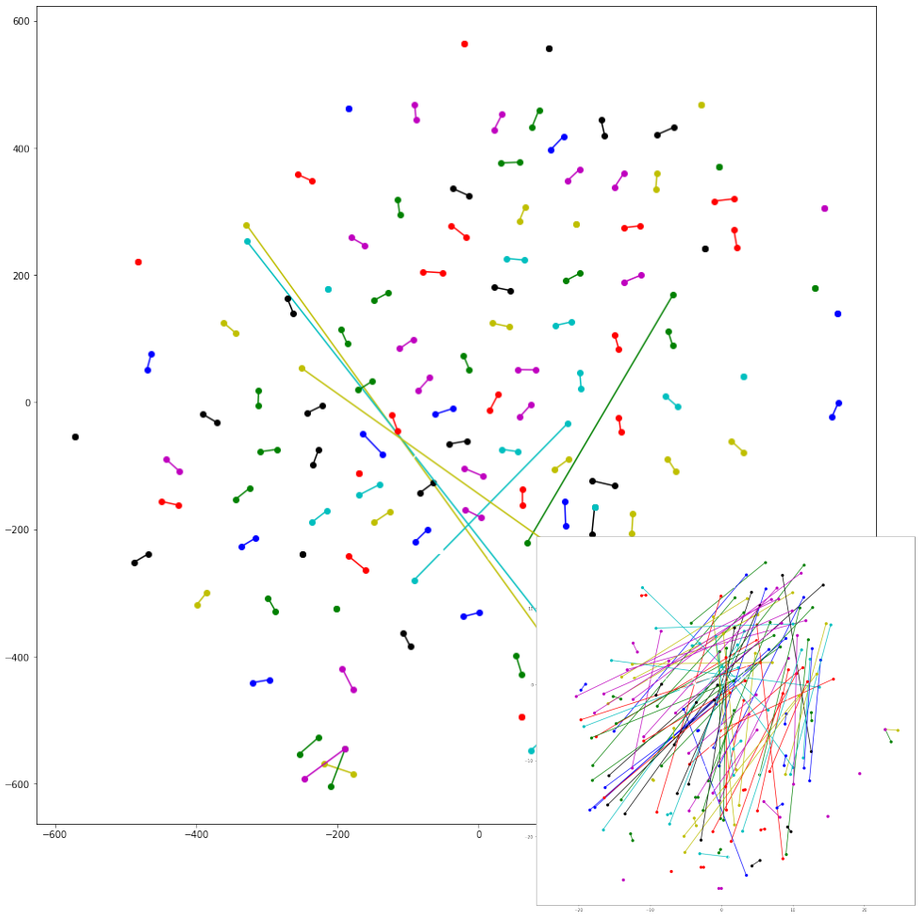
Experiments

Visualization of Jumbled Sentence Embedding

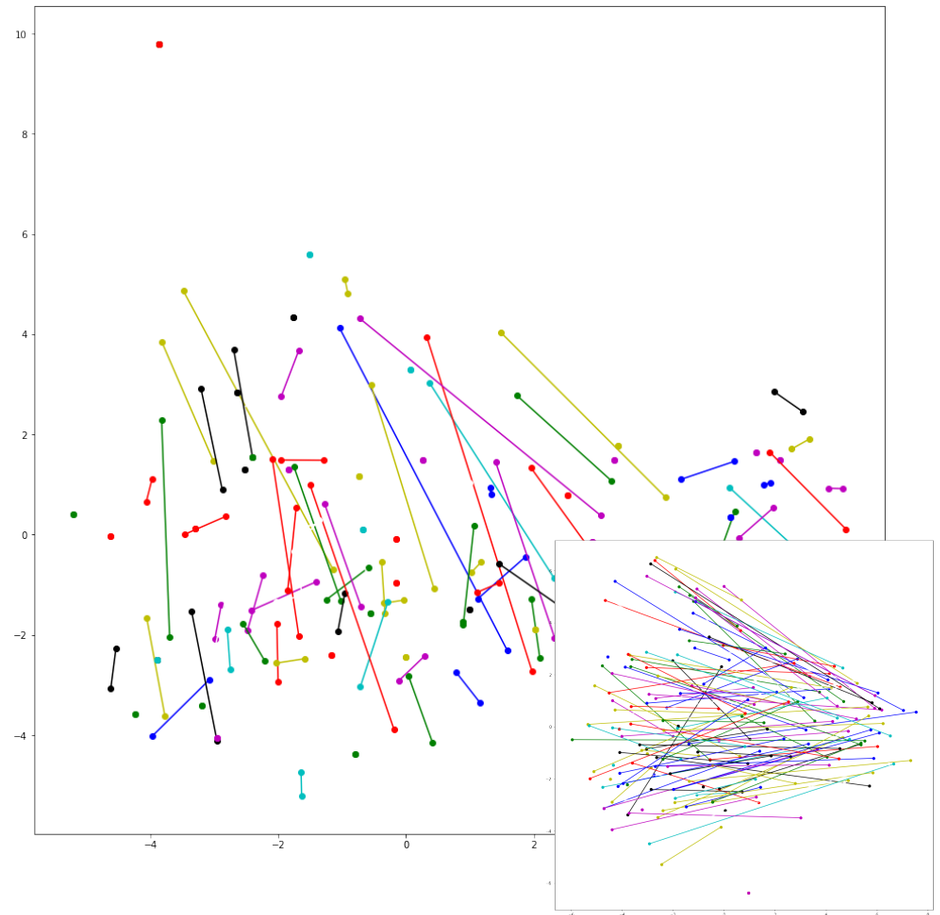
t-SNE or PCA

↙ preserve local geometry

↘ preserve distances / angles



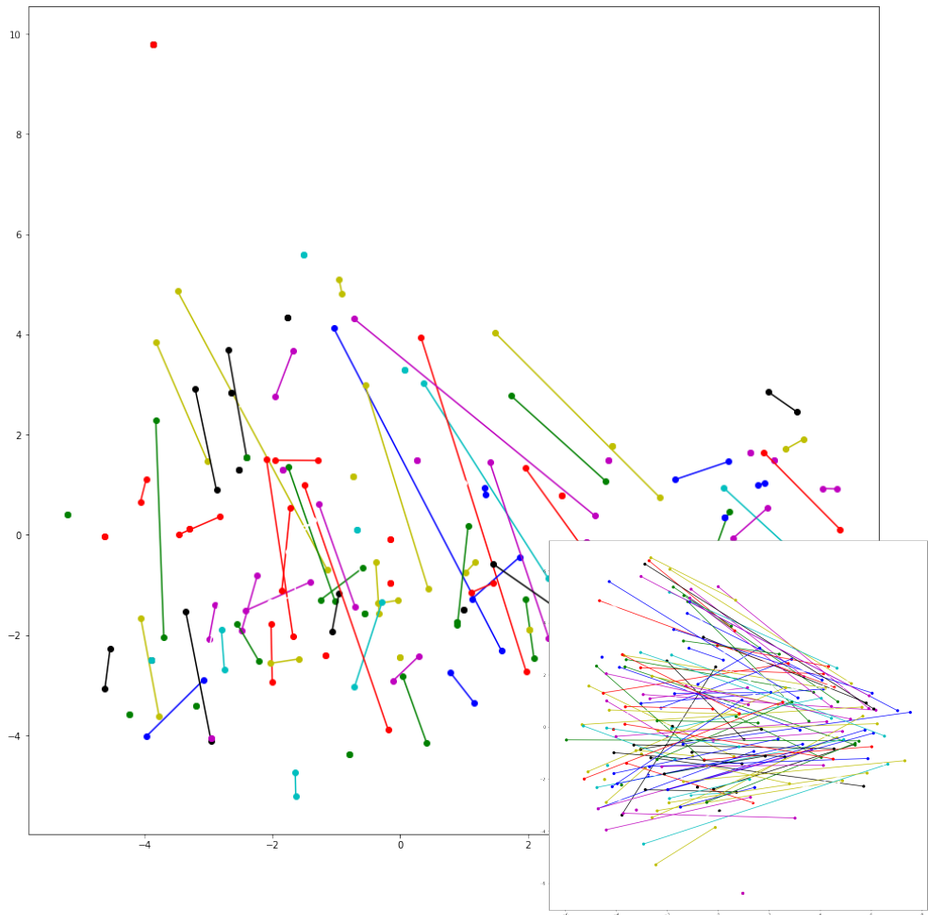
Character-level jumbling is more violent than word-level jumbling.



There are some induced biases in jumbled sentence embedding.

Experiments

Effects of Induced Biases: A Simple Cure



There are some induced biases in jumbled sentence embedding.

Can we remove these **induced biased** to improve the model's **robustness**?

Experiments

Effects of Induced Biases: A Simple Cure

	CR	CR-bias	TREC	TREC-bias	STS	STS-bias
Original	84.93		90.8		0.6036	
Swap (w)	84.43	62.28	89.2	89.6	0.537	0.475
Omit (w)	82.62	82.33	89.2	89.6	0.523	0.325
Add (w)	81.74	82.44	91.2	91.4	0.523	0.514
Swap (c)	69.04	**81.86**	69.6	**86.8**	0.219	**0.463**
Omit (c)	68.74	**79.44**	70.8	**79.8**	0.223	**0.369**
Add (c)	70.52	**83.36**	67.4	**90.0**	0.206	**0.537**

Can we remove these **induced biased** to
improve the model's **robustness**?

Experiments

Effects of Induced Biases: A Simple Cure

	CR	CR-bias	TREC	TREC-bias	STS	STS-bias
Original	84.93		90.8		0.6036	
Swap (w)	84.43	62.28	89.2	89.6	0.537	0.475
Omit (w)	82.62	82.33	89.2	89.6	0.523	0.325
Add (w)	81.74	82.44	91.2	91.4	0.523	0.514
Swap (c)	69.04	**81.86**	69.6	**86.8**	0.219	**0.463**
Omit (c)	68.74	**79.44**	70.8	**79.8**	0.223	**0.369**
Add (c)	70.52	**83.36**	67.4	**90.0**	0.206	**0.537**

Can we remove these **induced biased** to improve the model's **robustness**? **—Yes!**

Conclusion

Our Contributions

(1) We evaluated the state-of-the-art model's ability to process jumbled sentences on **three classic downstream tasks mimicking human cognitive abilities**, including sentiment classification, information retrieval, and semantic similarity.

(2) We found the machine's ability of reading jumbled sentences is more **sensitive to the types of jumbling than degrees of jumbling**, and the induced biases of jumbled embeddings greatly impair performance.

(3) The removal of these induced biases significantly **improves the machine's robustness** of reading **character-level jumbled sentences** on all three tasks.