Improving word mover's distance by leveraging self-attention matrix

Hiroaki Yamagiwa[1], Sho Yokoi[2,3], Hidetoshi Shimodaira[1,3]

[1]Kyoto University [2]Tohoku University [3]RIKEN AIP

EMNLP 2023 Findings

Housekeeping

Authors:
- Hiroaki Yamagiwa
- Sho Yokoi
- Hidetoshi Shimodaira

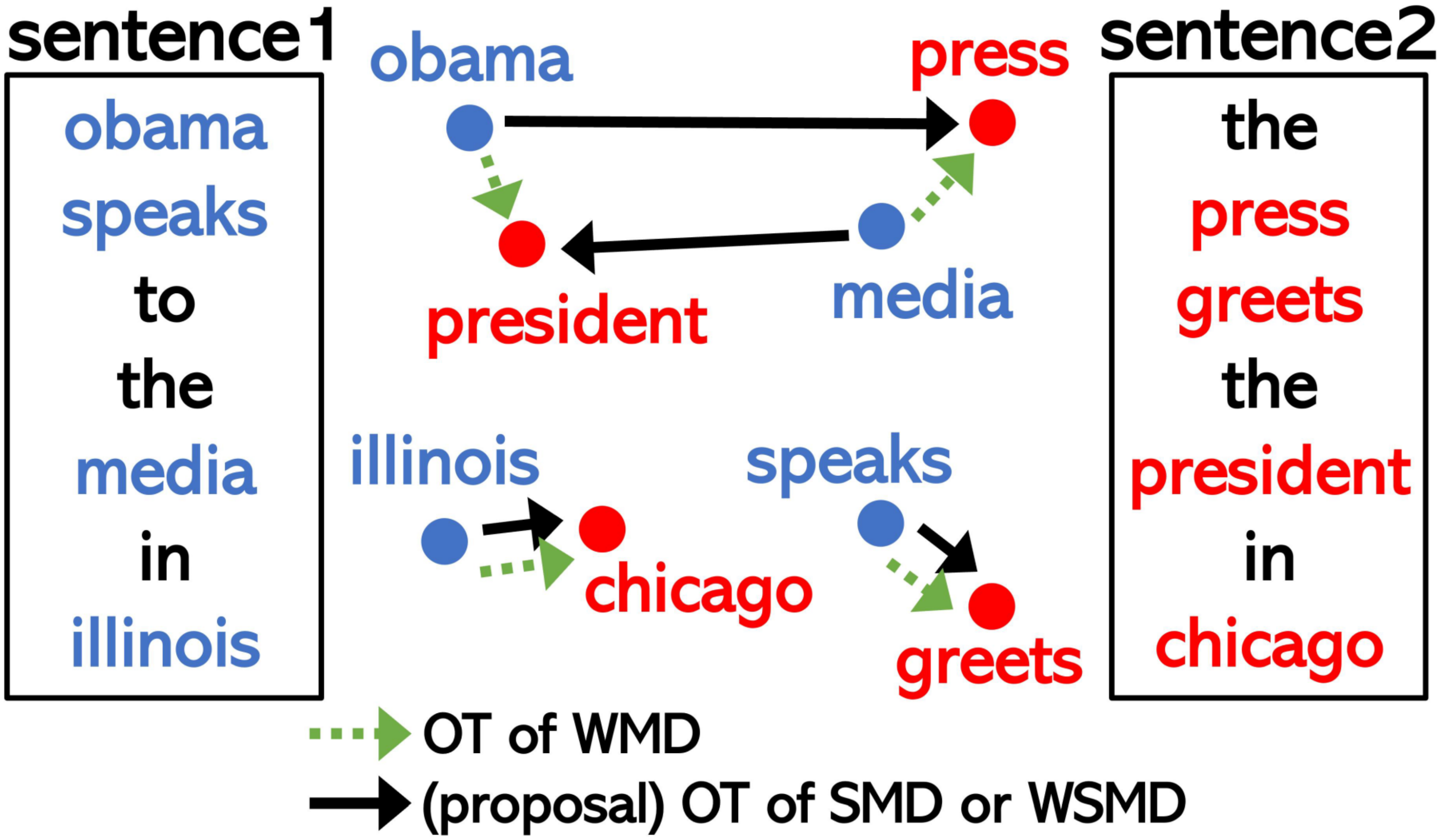
Materials:
- Paper: https://arxiv.org/abs/2211.06229
- Code: https://github.com/ymgw55/WSMD
- Slides: https://ymgw55.github.io/publication/wsmd/slides.pdf
- Poster: https://ymgw55.github.io/publication/wsmd/poster.pdf

Paper Code Slides Poster

Summary

- **Word Mover's Distance** (**WMD**) [1] uses the **Wasserstein distance** to measure semantic textual similarity.
- **WMD** **cannot** deal with **the order of words** within a sentence.
- We propose the **Word and sentence Structure Mover's Distance** (**WSMD**) that **can** address this limitation.

WMD vs. WSMD



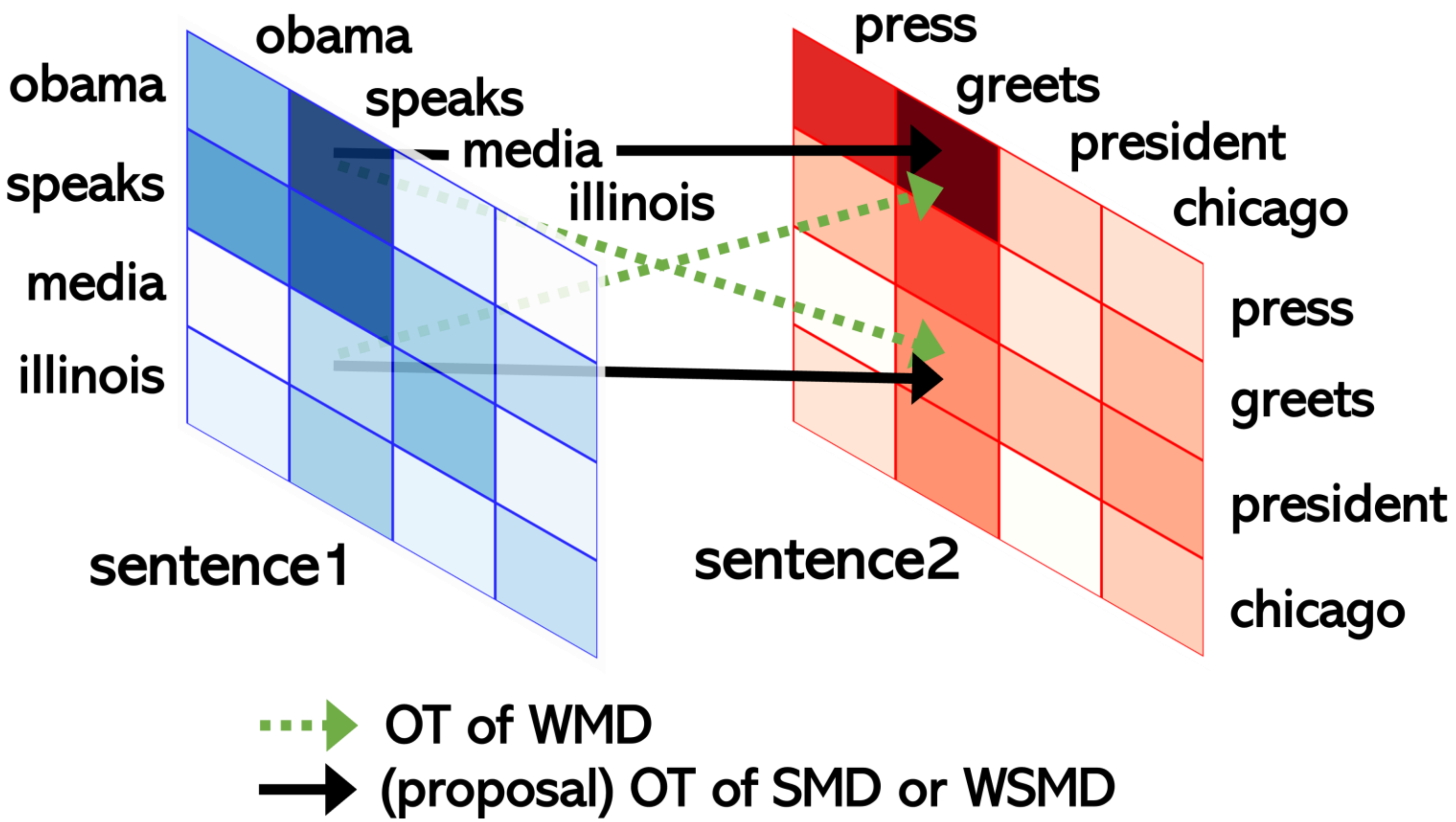
sentence1
obama
speaks
to
the
media
in
illinois
sentence2
the
press
greets
the
president
in
chicago
obama
press
president
media
illinois
chicago
speaks
greets
OT of WMD
(proposal) OT of SMD or WSMD

Approach

- Use the **Self-Attention Matrix** (**SAM**) from BERT-based models as **structure information**.
- Combine **WMD** and **SAM** using the **Fused Gromov-Wasserstein distance** [2].

Self-Attention Matrix

- Both pairs, (obama, speaks) and (press, greets), show a high attention weight.



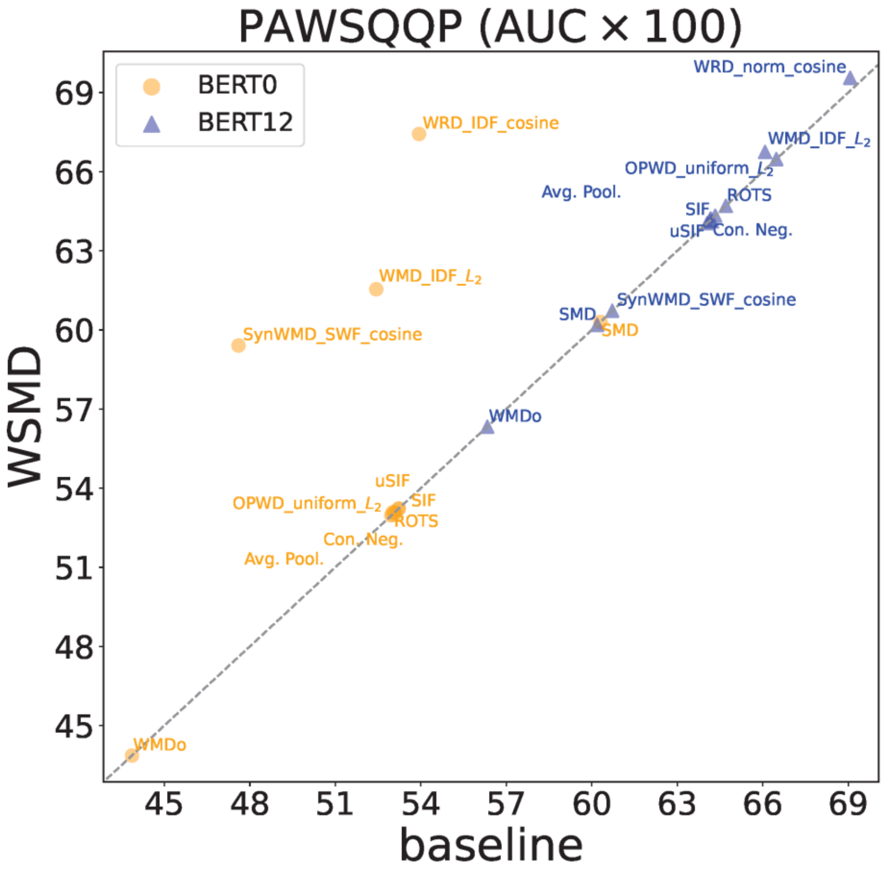
Proposed Method

- Let A and A' be the **SAM**s for sentences s and s'.
- Define the **Word and sentence Structure Mover's Distance** (**WSMD**) as follows:

$$\mathrm{WSMD}\big((s, A), (s', A')\big) = \min_{\boldsymbol{P} \in \boldsymbol{\Pi}(\mathrm{u},\mathrm{u}')} \underbrace{\sum_{i,j,i',j'} \left\{ (1-\lambda) \underbrace{\boxed{C_{ij}}}_{\text{Wasserstein}} + \lambda \underbrace{\boxed{k}}_{\text{Normalization parameter}} \underbrace{\boxed{\left| A_{ii'} - A'_{jj'} \right|^2}}_{\text{Gromov-Wasserstein}} \right\} \boxed{\boxed{P_{ij}} P_{i'j'}}}_{\text{Fused Gromov-Wasserstein}}$$

Results

- We used the PAWS [3] dataset for paraphrase identification.
- **WSMD** was effective for **WMD-like** methods such as **WMD**, **WRD**, and **SynWMD**.



The values above the diagonal line show the performance improvement by WSMD.

References

1. Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. *From word embeddings to document distances*. ICML.
2. Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. 2019. *Optimal transport for structured data with application on graphs*. PMLR.
3. Yuan Zhang, Jason Baldridge, and Luheng He. 2019. *PAWS: Paraphrase adversaries from word scrambling*. NAACL.