

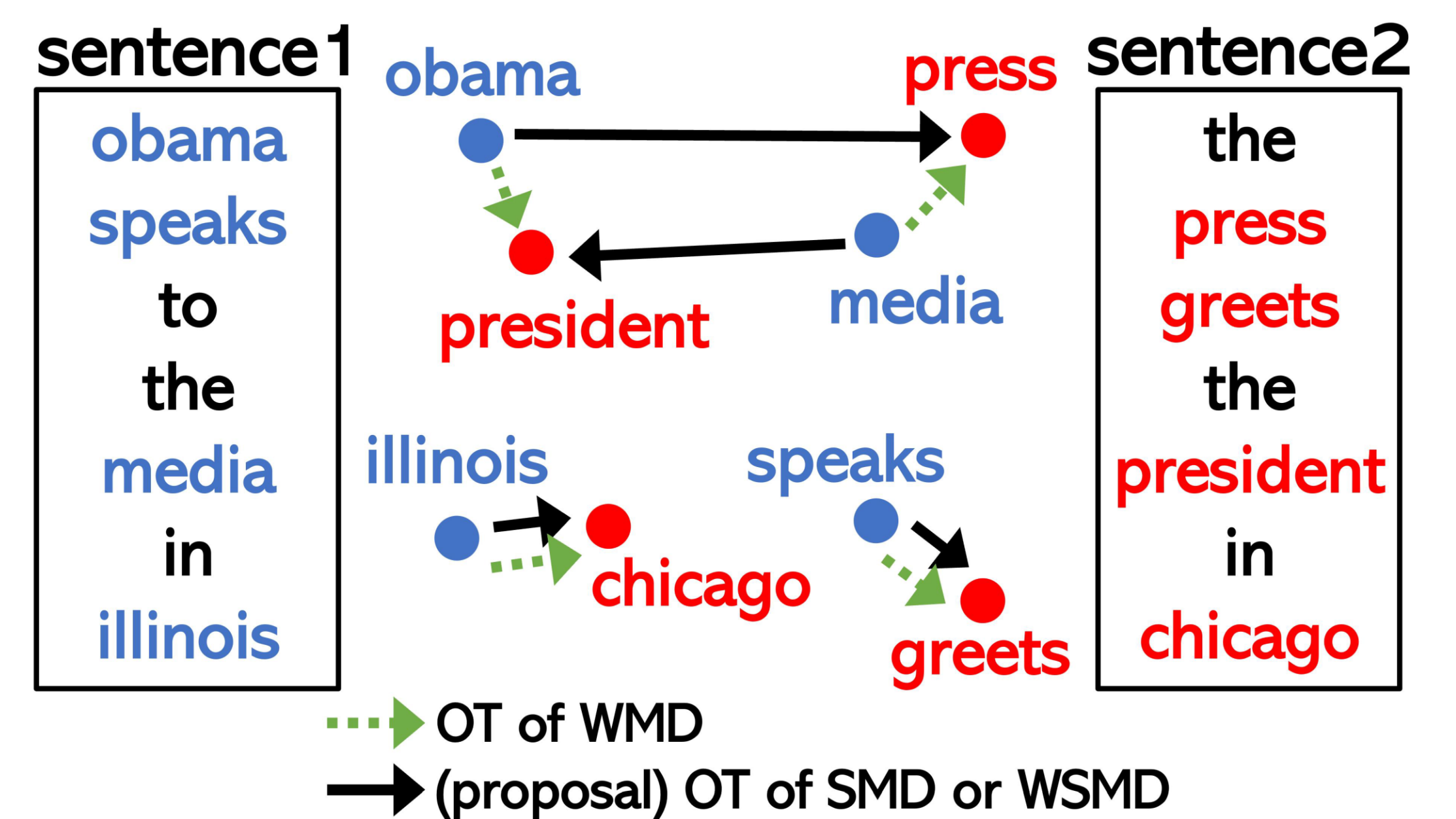


Background

- **Word Mover's Distance (WMD)** [1] uses the **Wasserstein distance** to measure semantic textual similarity.
- WMD cannot address the order of words within a sentence.

Approach

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

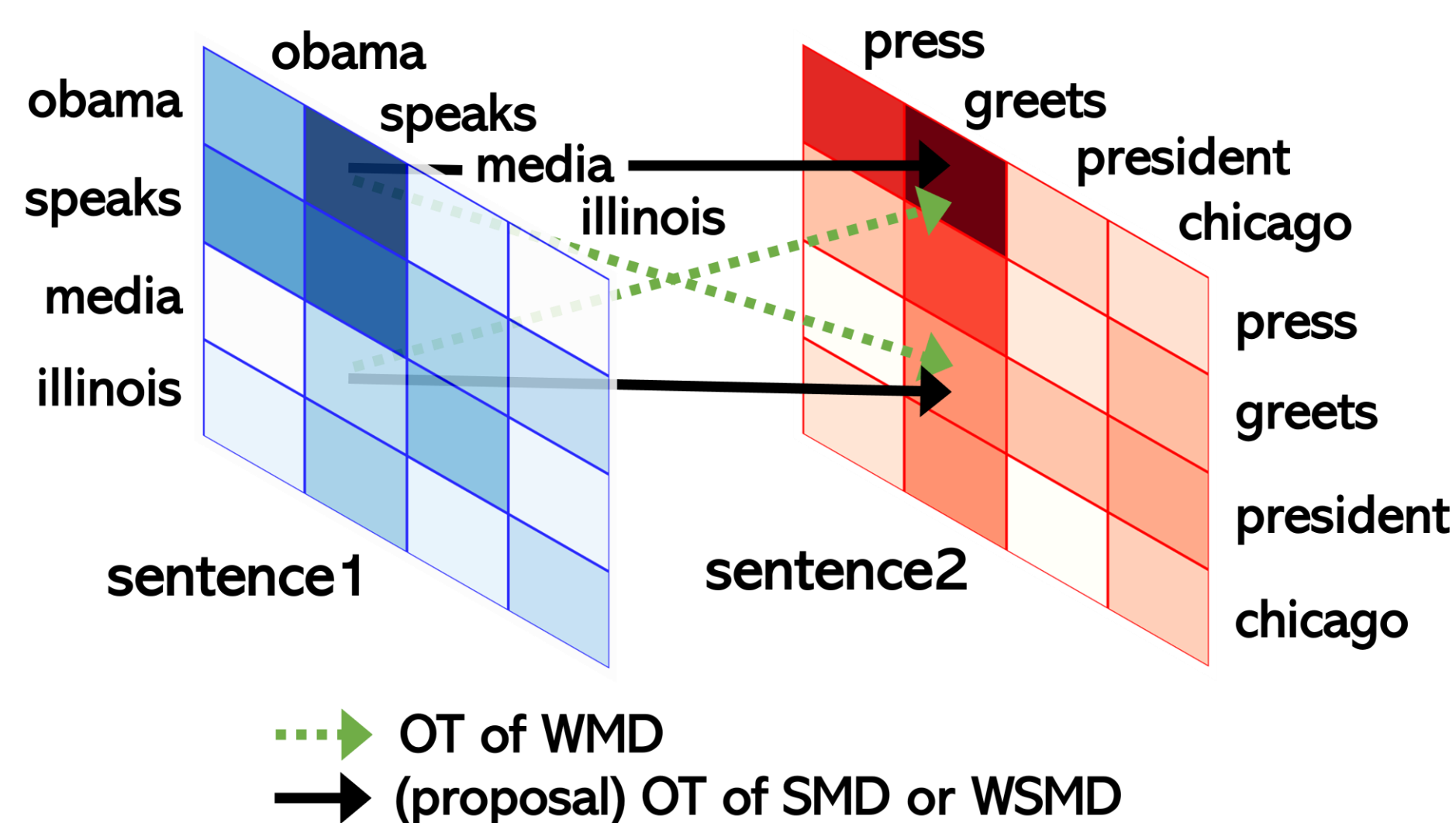


Results

- The method improved the performance of WMD-like methods in a paraphrase identification task [3].

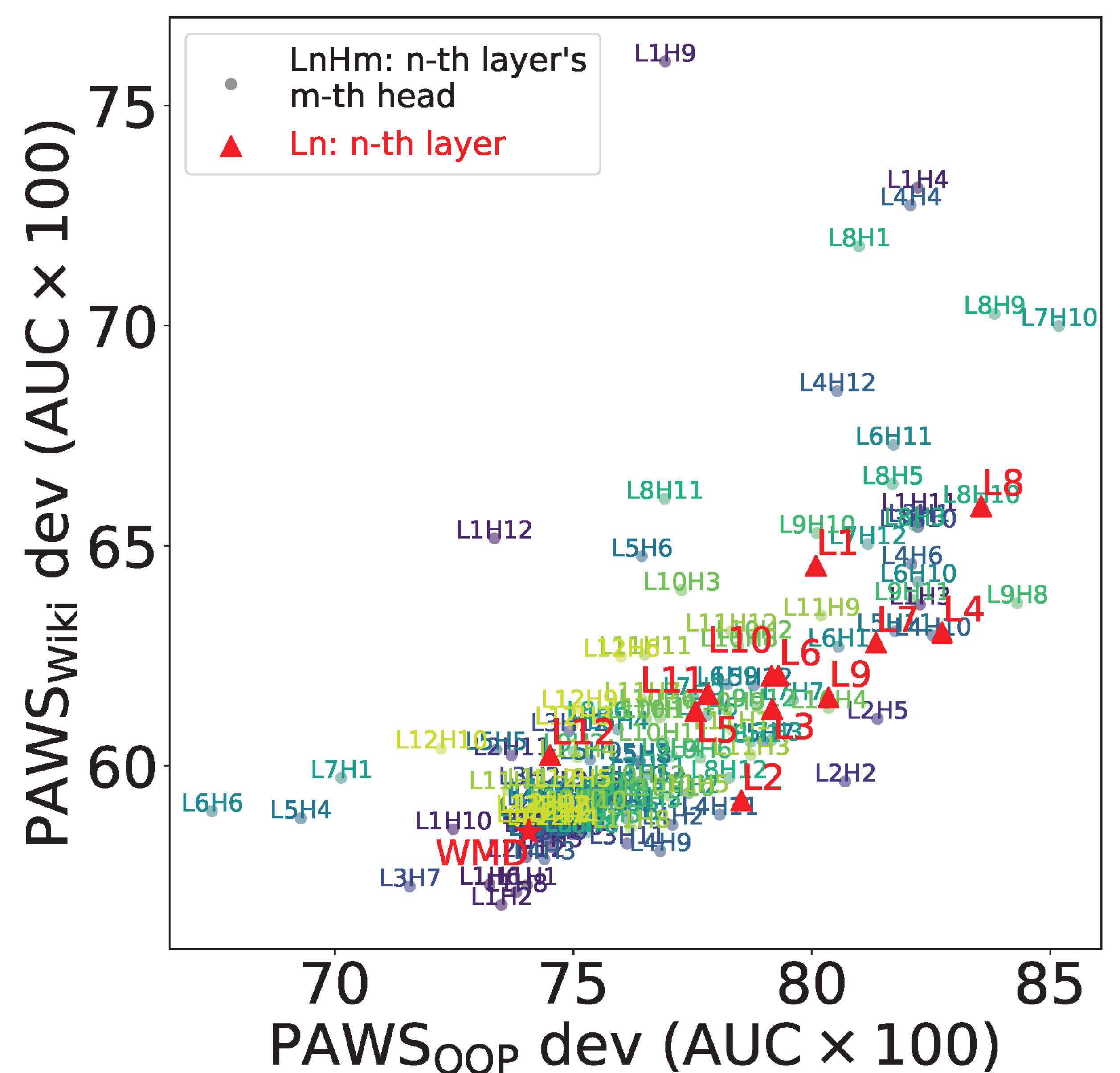
Methods

Self-Attention Matrix (SAM)



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

Head Selection



- The performance varies significantly among different attention heads.
- By selecting layers and using the heads within them, we average the **WSMD** to evaluate the performance.

Proposed Method

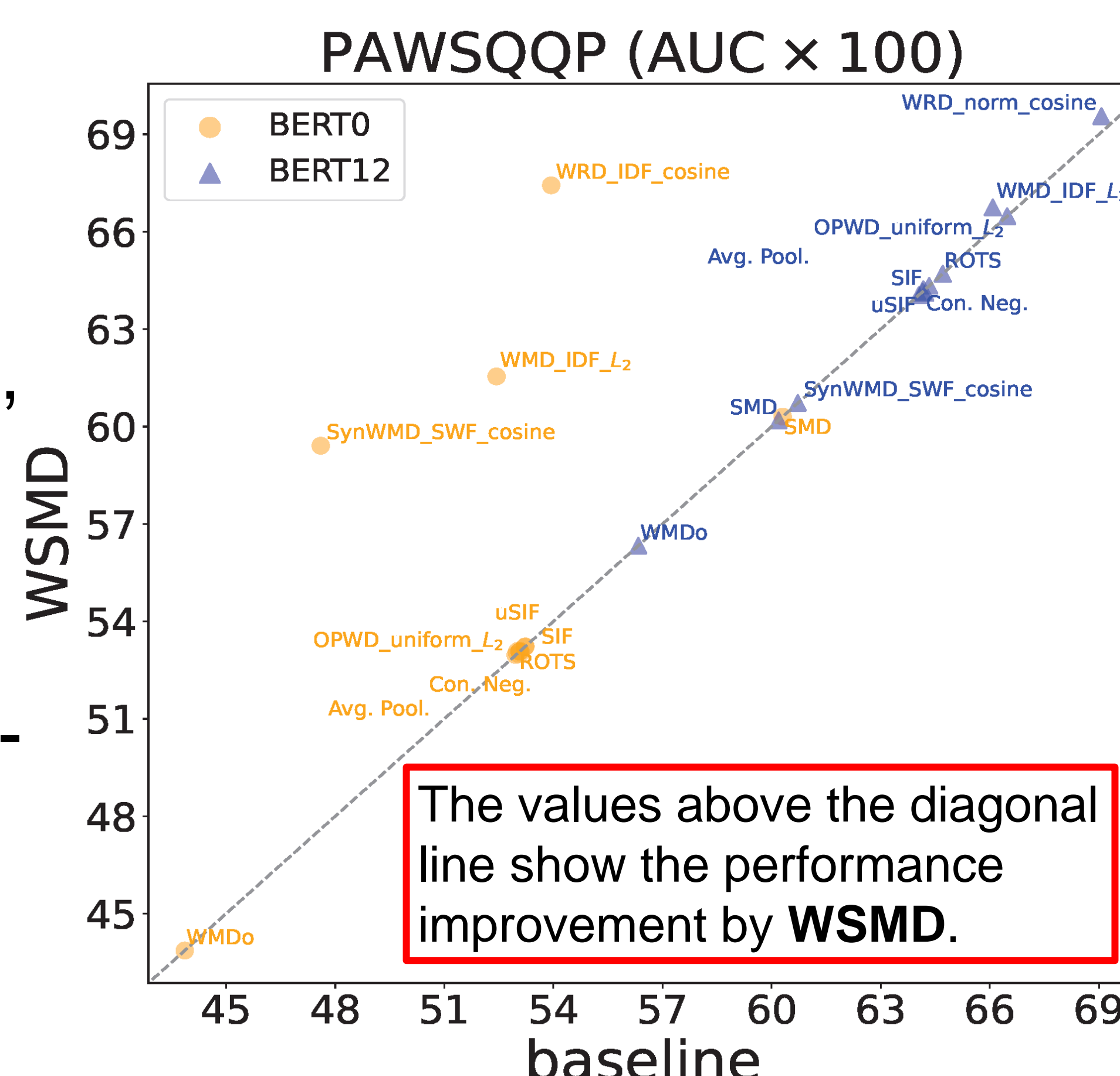
Let A and A' be the **SAMs** for sentences s and s' . Define the **Word and sentence Structure Mover's Distance (WSMD)** as follows:

$$WSMD((s, A), (s', A')) = \min_{P \in \Pi(u, u')} \sum_{i,j,i',j'} \left\{ (1-\lambda) C_{ij} + \lambda k |A_{ii'} - A'_{jj'}|^2 \right\} P_{ij} P_{i'j'}$$

Wasserstein Gromov-Wasserstein Fused Gromov-Wasserstein

Experiments

- For paraphrase identification, we used the PAWS [3] dataset, which contains sentence pairs with high word overlap.
- **WSMD** was effective for WMD-like methods such as WMD, WRD, and SynWMD.



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.