Re-evaluating Word Mover's Distance

Ryoma Sato
Makoto Yamada
Hisashi Kashima

KYOTO UNIVERSITY

We carefully reran the experiments of WMD

- The Word Mover's Distance (WMD) [Kusner+ ICML 2015] is a fundamental method to compute text distances
  Word Mover's Embeddings [Wu+ EMNLP 2018], MoverScore [Zhao EMNLP 2019], Word Rotator's Distance [Yokoi + EMNLP 2020], etc. are based on it
- We carried out careful followup experiments
- We found classic baselines are competitive with WMD if we appropriate normalize them

Many pitfalls in ML → objective reeval is important

- Aims of this papers:

  To WMD users: we show objective evaluation results
  useful for choosing methods

  To ML community: we show common evaluation pitfalls
  please care them in your research

- We will see how many pitfalls ML researches have
- They tend to be advantageous for the proposed methods because of the publication bias and confirmation bias.
- It is import to carry out careful and objective re-evaluations by third-party groups

WMD measures similarity of texts

- Input: Two texts

  X: Obama speaks to the media in Illinois

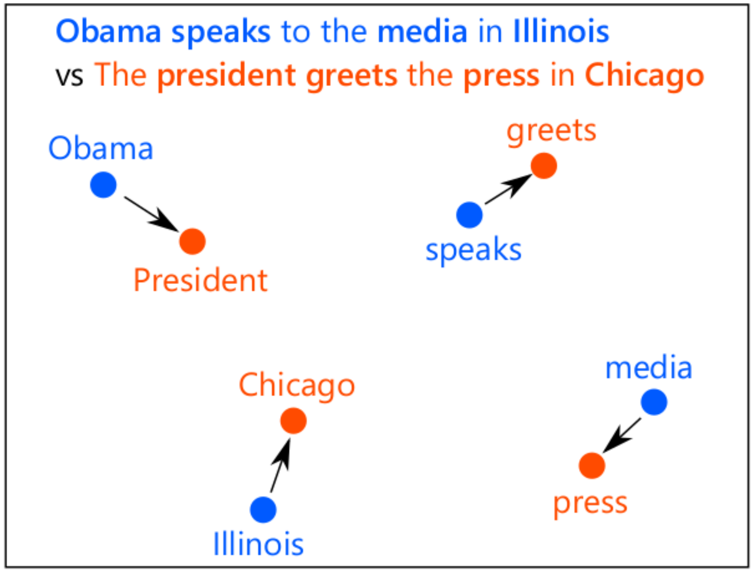
  Y: The president greets the press in Chicago

- Output: Distance between X and Y

- WMD:
  1. represent a text as a bag of word embeddings
  2. compute matching of embeddings
  3. compute the sum of distances of matched words

😆 versatile 😆 unsupervised

😆 effective 😫 heavy


Obama speaks to the media in Illinois
vs The president greets the press in Chicago
Obama
President
greets
speaks
Chicago
Illinois
media
press

👇 matching in the embedding space

The original paper used kNN classification

- The original paper [Kusner et al.] conducted kNN evaluations.
- The kNN documents are retrieved based on WMD.
- We follow this evaluation protocol in this paper.
- Datasets:

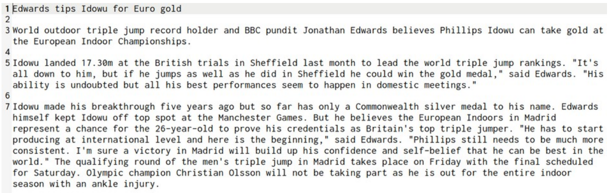
  bbcsports, twitter, recipe, ohsumed

  classic, reuters, amazon, 20news

Many duplicated samples exist

- **Misleading Fact 1**:
  These datasets contain many duplicated samples

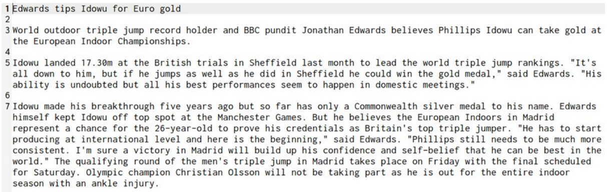
bbcsports



```
Edwards tips Idowu for Euro gold

World outdoor triple jump record holder and BBC pundit Jonathan Edwards believes Phillips Idowu can take gold at
the European Indoor Championships.

Idowu landed 17.30m at the British trials in Sheffield last month to lead the world triple jump rankings. "It's
all down to him, but if he jumps as well as he did in Sheffield he could win the gold medal," said Edwards. "His
ability is undoubted but all his best performances seem to happen in domestic meetings."

Idowu made his breakthrough five years ago but so far has only a Commonwealth silver medal to his name. Edwards
himself kept Idowu off top spot at the Manchester Games. But he believes the European Indoors in Madrid
represent a chance for the 26-year-old to prove his credentials as Britain's top triple jumper. "He has to start
producing at international level and here is the beginning," said Edwards. "Phillips still needs to be much more
consistent. I'm sure a victory in Madrid will build up his confidence and self-belief that he can be best in the
world." The qualifying round of the men's triple jump in Madrid takes place on Friday with the final scheduled
for Saturday. Olympic champion Christian Olsson will not be taking part as he is out for the entire indoor
season with an ankle injury.
```

athletics/012.txt in training

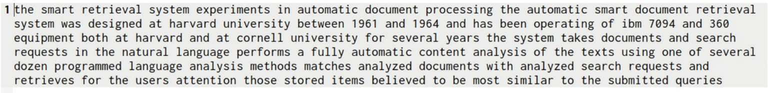
same



```
Edwards tips Idowu for Euro gold

World outdoor triple jump record holder and BBC pundit Jonathan Edwards believes Phillips Idowu can take gold at
the European Indoor Championships.

Idowu landed 17.30m at the British trials in Sheffield last month to lead the world triple jump rankings. "It's
all down to him, but if he jumps as well as he did in Sheffield he could win the gold medal," said Edwards. "His
ability is undoubted but all his best performances seem to happen in domestic meetings."

Idowu made his breakthrough five years ago but so far has only a Commonwealth silver medal to his name. Edwards
himself kept Idowu off top spot at the Manchester Games. But he believes the European Indoors in Madrid
represent a chance for the 26-year-old to prove his credentials as Britain's top triple jumper. "He has to start
producing at international level and here is the beginning," said Edwards. "Phillips still needs to be much more
consistent. I'm sure a victory in Madrid will build up his confidence and self-belief that he can be best in the
world." The qualifying round of the men's triple jump in Madrid takes place on Friday with the final scheduled
for Saturday. Olympic champion Christian Olsson will not be taking part as he is out for the entire indoor
season with an ankle injury.
```

athletics/020.txt in test

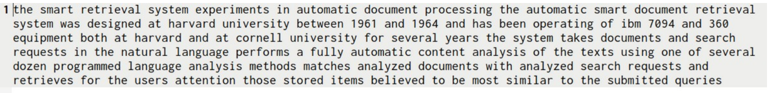
classic



```
the smart retrieval system experiments in automatic document processing the automatic smart document retrieval
system was designed at harvard university between 1961 and 1964 and has been operating of ibm 7094 and 360
equipment both at harvard and at cornell university for several years the system takes documents and search
requests in the natural language performs a fully automatic content analysis of the texts using one of several
dozen programmed language analysis methods matches analyzed documents with analyzed search requests and
retrieves for the users attention those stored items believed to be most similar to the submitted queries
```

class: CACM

same



```
the smart retrieval system experiments in automatic document processing the automatic smart document retrieval
system was designed at harvard university between 1961 and 1964 and has been operating of ibm 7094 and 360
equipment both at harvard and at cornell university for several years the system takes documents and search
requests in the natural language performs a fully automatic content analysis of the texts using one of several
dozen programmed language analysis methods matches analyzed documents with analyzed search requests and
retrieves for the users attention those stored items believed to be most similar to the submitted queries
```

class: CISI

Total: 4856 duplicated samples detected

We provide scripts to delete duplication

- **Misleading Fact 1**:

  These datasets contain many duplicated samples

- The data and train/test split by WMD paper have been used in many following researches

  [Huan et al. NeurIPS 2016, Yurochkin et al. NeurIPS 2019, Le et al. NeurIPS 2019, Takezawa et al. ICML 2021, Wu et al. EMNLP 2018, Mollaysa et al. ICML 2017, Gupta et al. AAAI 2020, Skianis et al. AISTATS 2020]

- We suspect many readers and researchers are not aware of it

- We release code to detect & delete duplication

https://github.com/joisino/reeval-wmd

Baseline methods were not normalized

- The original paper used bag-of-words (BoW) and TF-IDF as baseline methods

- **Misleading Fact 2**:
  Baseline methods were not normalized

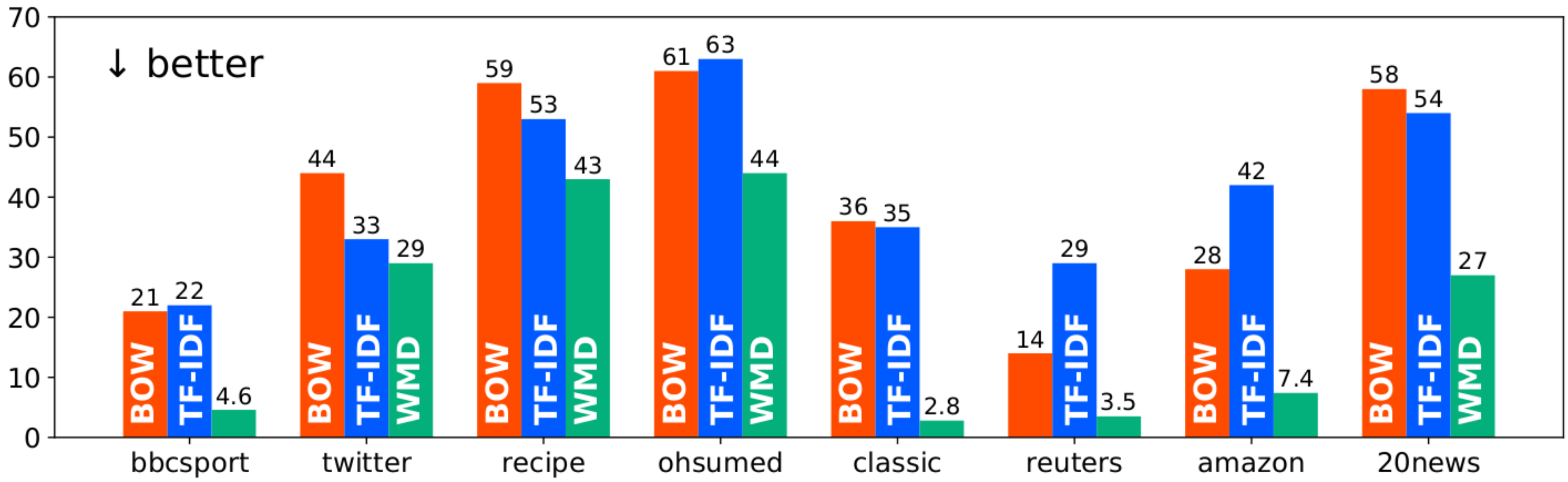
- Lengths of documents vary
  → Lengths of raw BoW vectors vary
  → Even if two documents share the topic,
  they are detected distant if their lengths are different

  ↔ WMD were normalized in the original evaluation

- We ran re-evaluation with normalization to BoW $x'_{\mathrm{BoW}} \leftarrow \frac{x_{\mathrm{BoW}}}{\|x_{\mathrm{BoW}}\|}$

Original evaluation

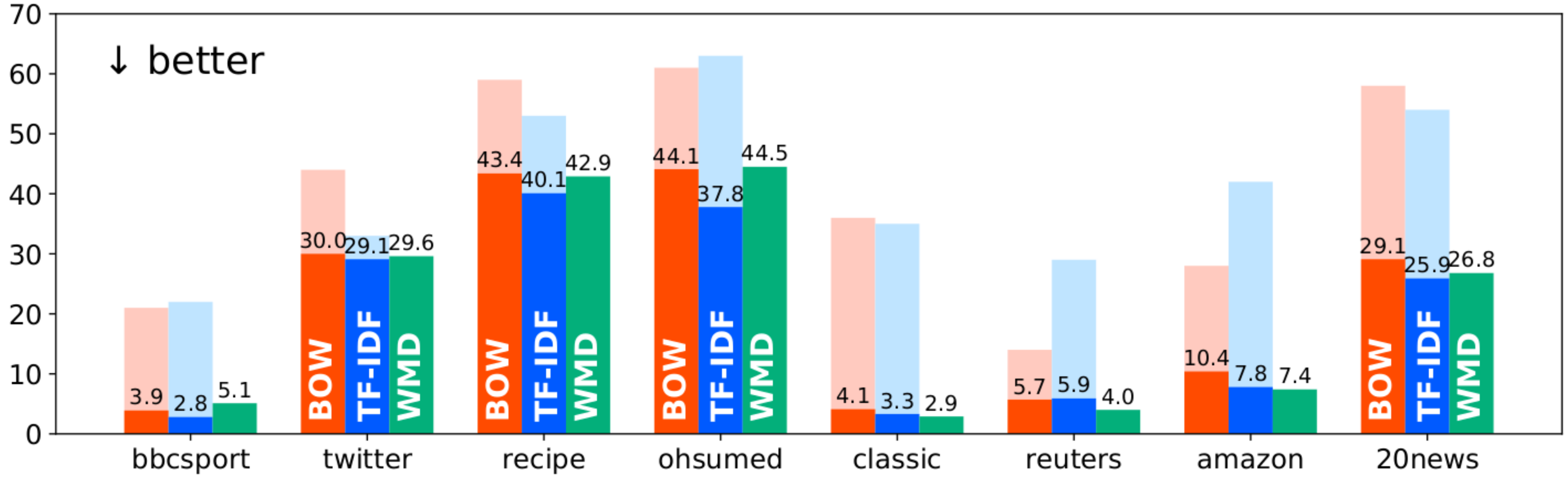
- What was reported in the original paper
- BoW and TF-IDF are much worse than WMD



👆 k-NN classification errors. **Lower is better**.

After normalization

- After normalization (Our re-evaluation)
- BoW and TF-IDF become much better

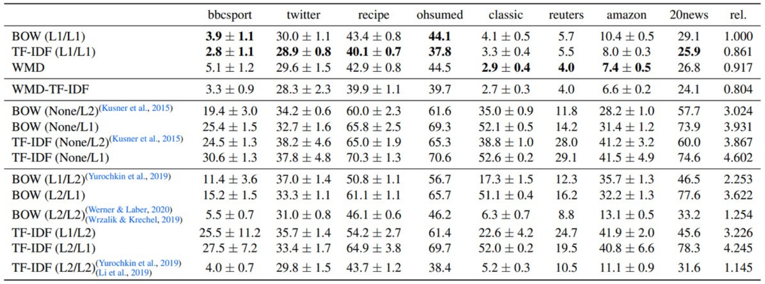


k-NN classification errors. **Lower is better**.

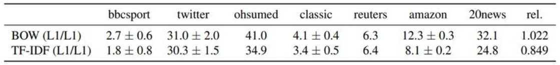
Improvements are actually five percents

- The original paper claimed 60% improvements over BoW

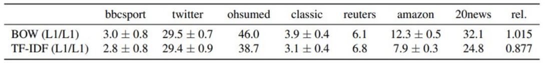
- We conducted several careful experiments and found that the improvement was actually 5%



	bbcsport	twitter	recipe	ohsumed	classic	reuters	amazon	20news	rel.
BOW (L1/L1)	**3.9** ± **1.1**	30.0 ± 1.1	43.4 ± 0.8	**44.1**	4.1 ± 0.5	5.7	10.4 ± 0.5	29.1	1.000
TF-IDF (L1/L1)	**2.8** ± **1.1**	**28.9** ± **0.8**	**40.1** ± **0.7**	**37.8**	3.3 ± 0.4	5.5	8.0 ± 0.3	**25.9**	0.861
WMD	5.1 ± 1.2	29.6 ± 1.5	42.9 ± 0.8	44.5	**2.9** ± **0.4**	**4.0**	**7.4** ± **0.5**	26.8	0.917
WMD-TF-IDF	3.3 ± 0.9	28.3 ± 2.3	39.9 ± 1.1	39.7	2.7 ± 0.3	4.0	6.6 ± 0.2	24.1	0.804
BOW (None/L2)(Kusner et al., 2015)	19.4 ± 3.0	34.2 ± 0.6	60.0 ± 2.3	61.6	35.0 ± 0.9	11.8	28.2 ± 1.0	57.7	3.024
BOW (None/L1)	25.4 ± 1.5	32.7 ± 1.6	65.8 ± 2.5	69.3	52.1 ± 0.5	14.2	31.4 ± 1.2	73.9	3.931
TF-IDF (None/L2)(Kusner et al., 2015)	24.5 ± 1.3	38.2 ± 4.6	65.0 ± 1.9	65.3	38.8 ± 1.0	28.0	41.2 ± 3.2	60.0	3.867
TF-IDF (None/L1)	30.6 ± 1.3	37.8 ± 4.8	70.3 ± 1.3	70.6	52.6 ± 0.2	29.1	41.5 ± 4.9	74.6	4.602
BOW (L1/L2)(Yurochkin et al., 2019)	11.4 ± 3.6	37.0 ± 1.4	50.8 ± 1.1	56.7	17.3 ± 1.5	12.3	35.7 ± 1.3	46.5	2.253
BOW (L2/L1)	15.2 ± 1.5	33.3 ± 1.1	61.1 ± 1.1	65.7	51.1 ± 0.4	16.2	32.2 ± 1.3	77.6	3.622
BOW (L2/L2)(Werner & Laber, 2020)(Wrzalik & Krechel, 2019)	5.5 ± 0.7	31.0 ± 0.8	46.1 ± 0.6	46.2	6.3 ± 0.7	8.8	13.1 ± 0.5	33.2	1.254
TF-IDF (L1/L2)	25.5 ± 11.2	35.7 ± 1.4	54.2 ± 2.7	61.4	22.6 ± 4.2	24.7	41.9 ± 2.0	45.6	3.226
TF-IDF (L2/L1)	27.5 ± 7.2	33.4 ± 1.7	64.9 ± 3.8	69.7	52.0 ± 0.2	19.5	40.8 ± 6.6	78.3	4.245
TF-IDF (L2/L2)(Yurochkin et al., 2019)(Li et al., 2019)	4.0 ± 0.7	29.8 ± 1.5	43.7 ± 1.2	38.4	5.2 ± 0.3	10.5	11.1 ± 0.9	31.6	1.145

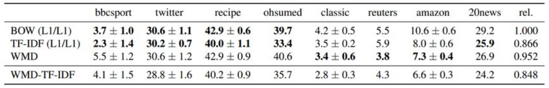


	bbcsport	twitter	ohsumed	classic	reuters	amazon	20news	rel.
BOW (L1/L1)	2.7 ± 0.6	31.0 ± 2.0	41.0	4.1 ± 0.4	6.3	12.3 ± 0.3	32.1	1.022
TF-IDF (L1/L1)	1.8 ± 0.8	30.3 ± 1.5	34.9	3.4 ± 0.5	6.4	8.1 ± 0.2	24.8	0.849



	bbcsport	twitter	ohsumed	classic	reuters	amazon	20news	rel.
BOW (L1/L1)	3.0 ± 0.8	29.5 ± 0.7	46.0	3.9 ± 0.4	6.1	12.3 ± 0.5	32.1	1.015
TF-IDF (L1/L1)	2.8 ± 0.8	29.4 ± 0.9	38.7	3.1 ± 0.4	6.8	7.9 ± 0.3	24.8	0.877

Table 3. kNN classification errors with clean data. Lower is better. The same notations as in Table 2.



	bbcsport	twitter	recipe	ohsumed	classic	reuters	amazon	20news	rel.
BOW (L1/L1)	**3.7** ± **1.0**	**30.6** ± **1.1**	**42.9** ± **0.6**	**39.7**	4.2 ± 0.5	5.5	10.6 ± 0.6	29.2	1.000
TF-IDF (L1/L1)	**2.3** ± **1.4**	**30.2** ± **0.7**	**40.0** ± **1.1**	**33.4**	3.5 ± 0.2	5.9	8.0 ± 0.6	**25.9**	0.866
WMD	5.5 ± 1.2	30.6 ± 1.2	42.9 ± 0.9	40.6	**3.4** ± **0.6**	**3.8**	**7.3** ± **0.4**	26.9	0.952
WMD-TF-IDF	4.1 ± 1.5	28.8 ± 1.6	40.2 ± 0.9	35.7	2.8 ± 0.3	4.3	6.6 ± 0.3	24.2	0.848

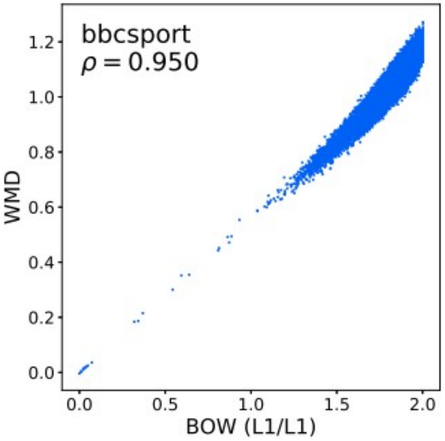
← careful evaluations
(see our paper)

- Inadvertent baselines lead misunderstanding claims

- 5% improvement is genuine
  If speed is important, WMD may not be worth trying
  Otherwise, WMD may be a worth candidate over BoW

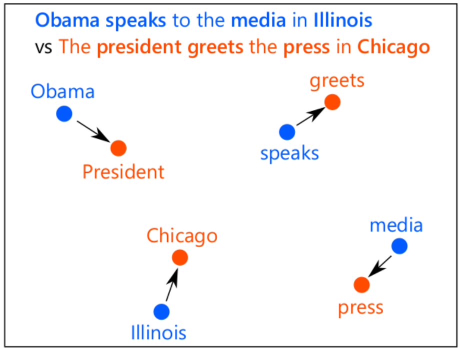
Distances are similar due to high dimensionality

- We found not only performance but also values of WMD are similar to those of BoW



- This is because two embeddings are almost orthogonal in a high dimensional space

- For example, d(Obama, Obama) = 0
  d(Obama, President) = 1.17
  d(Obama, band) = 1.34


Obama speaks to the media in Illinois
vs The president greets the press in Chicago
Obama
President
greets
speaks
Chicago
Illinois
media
press

← Two dimensional illustration does not reflect this "almost equidistant" property

Objective re-evaluation is important

- We found several misleading facts on the original evaluations of WMD paper

  Other facts & experiments are available in our paper

- Lessons 🎁

  Inadvertent baselines lead misunderstanding claims

  It is difficult to design perfect experiments

  It is import to carry out careful and objective re-evaluations by third-party groups

https://github.com/joisino/reeval-wmd