# Distributed Representation of Documents with Explicit Explanatory Features: Background Research

## Kusner, Matt J., et al. "From Word Embeddings To Document Distances."

November 16th, 2015
SNU Data Mining Center
Han Kyul Kim

# Background

<u>Need for representing "distance" between documents</u>

- Accurately representing the distance between two documents is crucial in document retrieval, news categorization and clustering and multilingual document matching
- Suggested Methods:
  1. Bag of Words & TF-IDF
     - Due to high dimensionality of the vectors, near-orthogonality frequently occurs among the vector representations
     - Do not capture the distance between individual words
     - Example: "Obama speaks to the media in Illinois" vs. "The President greets the press in Chicago"
     - Variations of BOW models with different features exist
  2. Latent Semantic Indexing (LSI)
     - Eigendecomposes BOW feature space
  3. Latent Dirichlet Allocation (LDA)
     - Probabilistically groups similar words into topics and represent the documents as distribution over these topics
- Yet, no models improve the empirical performance of BOW on distance-based tasks

# Suggested Method

Word Mover's Distance (WMD)

- A new metric for the distance between text documents
- Utilizes word2vec embedded word vectors as semantic relationships are often preserved in vector operations
- Distance between two text documents A & B is the minimum cumulative distance that words from document A need to travel to match exactly to the words from document B
- Uses Earth Mover's Distance transportation problem to find the optimal solutions

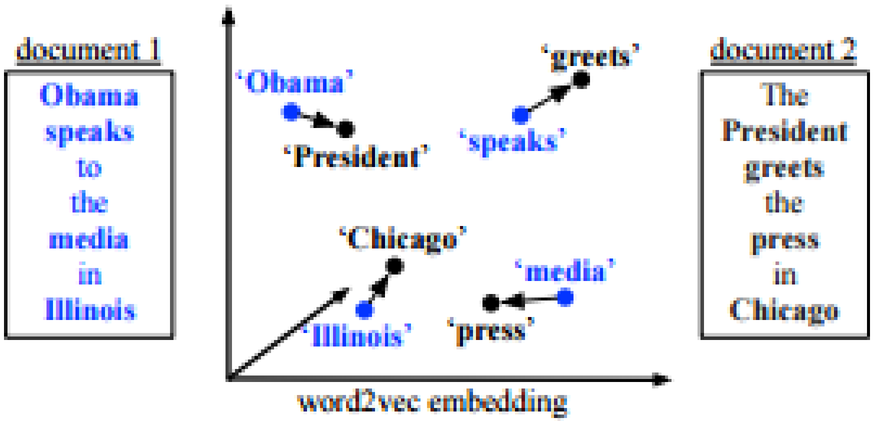
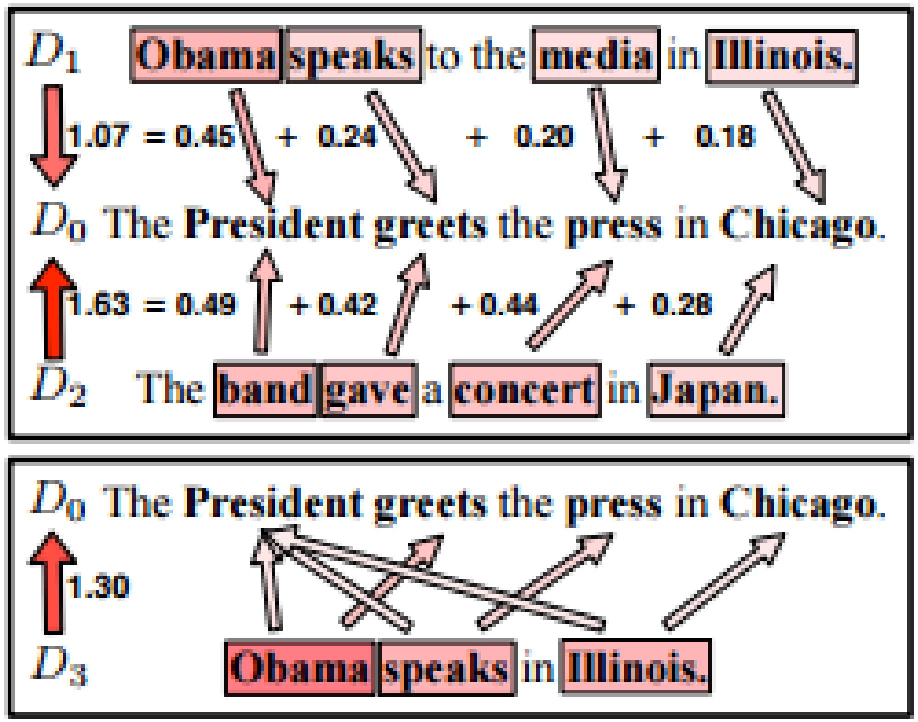



*Figure 1.* An illustration of the *word mover's distance.* All non-stop words (***bold***) of both documents are embedded into a *word2vec* space. The distance between the two documents is the minimum cumulative distance that all words in document 1 need to travel to exactly match document 2. (Best viewed in color.)

# Suggested Method

Properties and Advantages

1. Hyper-parameter free
2. Highly interpretable as the distance between two documents can be broken down and explained
3. High retrieval accuracy as it incorporates the effective knowledge encoding of word2vec



# Earth Mover's Distance

- A method to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features is given
- Intuitively, two distributions can be thought of as earth and hole, and EMD measures the least amount of work needed to fill the holes with earth
- Can be thought of as a transportation problem (suppliers supplying to several consumers)
- If two multidimensional data is given as:

$$P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\} \qquad Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$$

  - Can be thought of as (coordinate, weight)
- If $f_{ij}$ represents a flow from $p_i$ to $q_j$, following linear programming problem can be set up

$$\mathrm{WORK}(P, Q, \mathbf{F}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$$

Minimizing Objective Function

Constraints

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m,\ 1 \leq j \leq n$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \qquad 1 \leq i \leq m$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j} \qquad 1 \leq j \leq n$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right),$$

# WMD Formulation

- Difference between words (cost associated with travelling from one word to another): Euclidean distance difference in the word2vec embedding space

$$c(i,j) = \| x_i - x_j \|_2$$

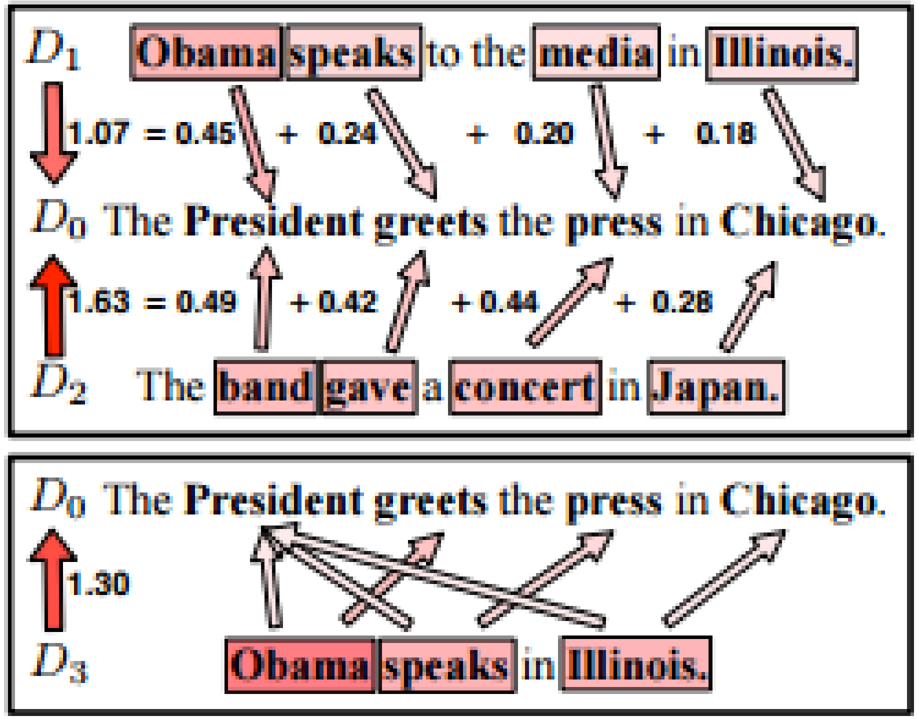
- Allow each word in document $d$ to be transformed into any word in document $d'$
- Let $\boldsymbol{T} \epsilon R^{n \times n}$ be a (sparse) matrix that denotes how much of word i in document $d$ travels to word j in $d'$
- $d_i$ : the number of word appearance in a document ($d_i = \frac{c_i}{\sum_{j=1}^{n} c_j}$)

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\} \qquad (1)$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}.$$

# WMD Formulation




<u>Observations</u>

- "Moves" the words to semantically similar words
- (D1, D2) = both have same TF-IDF distance from D0
- Still valid when the number of words in documents vary

# Complexity and Model Relaxation

- Complexity for solving WMD optimization problem: $O(p^3 \log p)$

  * p = number of unique words in the entire documents
- Can overcome the high complexity of the model via

  1. <u>Word Centroid Distance</u>
     - Represent each document by its weighted average vector and use that centroid vector to find the distance between the documents
     - Centroid distance serves as lower bound on WMD
     - Scales to $O(dp)$
     - Can incorporate this method to narrow down the search space in calculating exact WMD

# Complexity and Model Relaxation

2. <u>Relaxed Word Moving Distance</u>
   - To provide tighter bound, remove the two constraint from the original WMD formulation consecutively and take the maximum distance between the two
   - Need to find only the most similar word vector $x_j$ in document $d'$

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1,\dots,n\}$$

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \quad (1)$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1,\dots,n\}.$$

**Take the maximum distance between these two relaxed models**

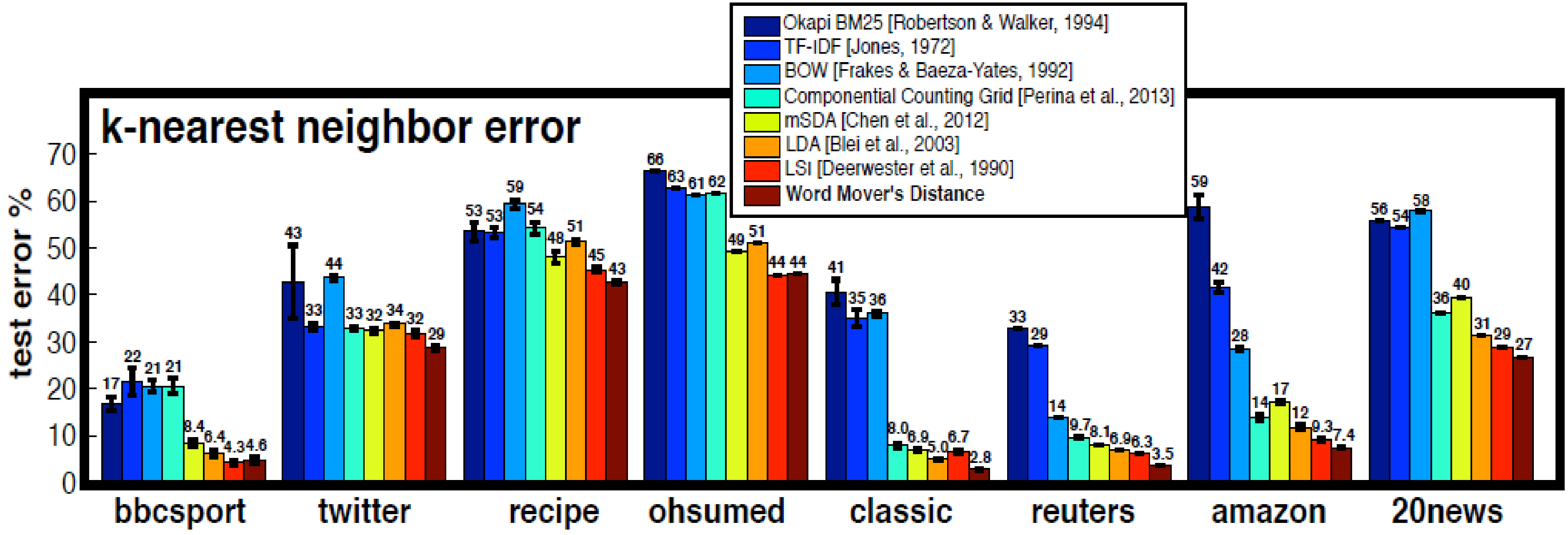
# Data

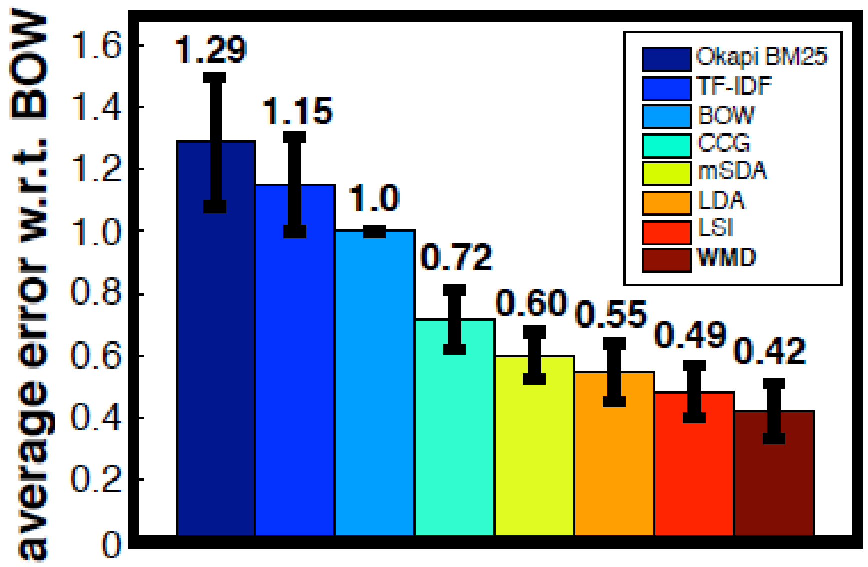
*Table 1.* Dataset characteristics, used in evaluation.

| NAME | $n$ | BOW DIM. | UNIQUE WORDS (AVG) | $\|\mathcal{Y}\|$ |
|---|---|---|---|---|
| BBCSPORT | 517 | 13243 | 117 | 5 |
| TWITTER | 2176 | 6344 | 9.9 | 3 |
| RECIPE | 3059 | 5708 | 48.5 | 15 |
| OHSUMED | 3999 | 31789 | 59.2 | 10 |
| CLASSIC | 4965 | 24277 | 38.6 | 4 |
| REUTERS | 5485 | 22425 | 37.1 | 8 |
| AMAZON | 5600 | 42063 | 45.0 | 4 |
| 20NEWS | 11293 | 29671 | 72 | 20 |

- Used the trained word2vec model from Google
- Words that are not present in the trained word2vec model is dropped during computing WMD metric

# K-NN Result


k-nearest neighbor error
test error %
Okapi BM25 [Robertson & Walker, 1994]
TF-IDF [Jones, 1972]
BOW [Frakes & Baeza-Yates, 1992]
Componential Counting Grid [Perina et al., 2013]
mSDA [Chen et al., 2012]
LDA [Blei et al., 2003]
LSI [Deerwester et al., 1990]
Word Mover's Distance
bbcsport
twitter
recipe
ohsumed
classic
reuters
amazon
20news

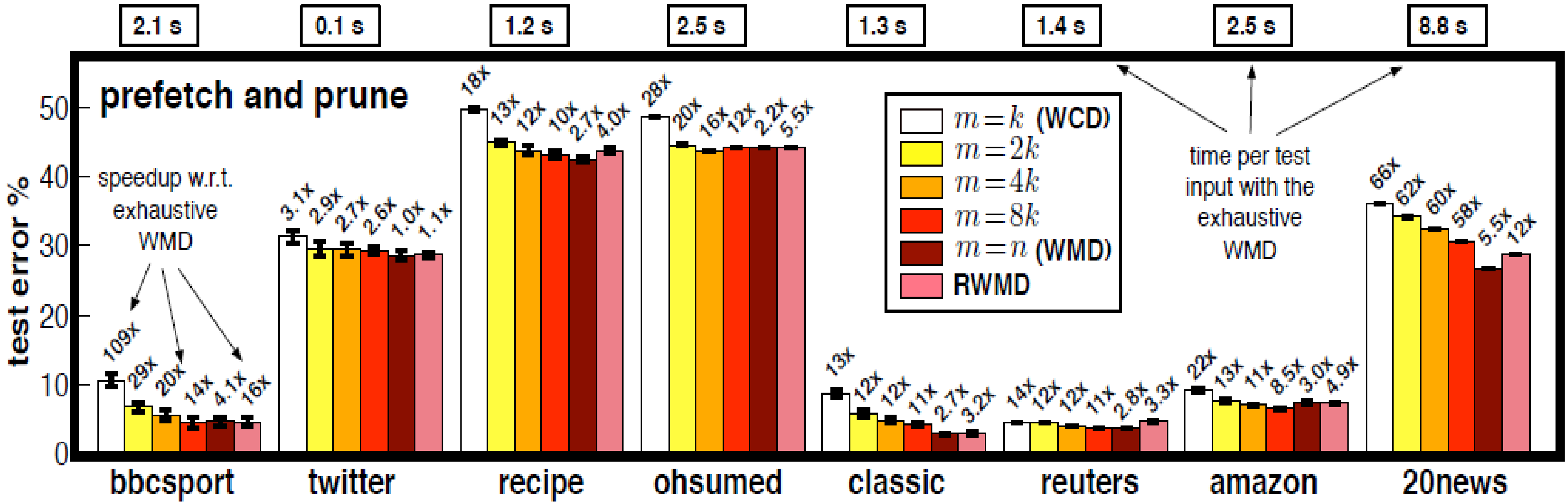

average error w.r.t. BOW
1.29
1.15
1.0
0.72
0.60
0.55
0.49
0.42
Okapi BM25
TF-IDF
BOW
CCG
mSDA
LDA
LSI
WMD

# Using Different Embedding Methods

| DOCUMENT $k$-NEAREST NEIGHBOR RESULTS | | | | | |
|---|---|---|---|---|---|
| DATASET | HLBL | CW | NIPS (W2V) | AMZ (W2V) | NEWS (W2V) |
| BBCSPORT | 4.5 | 8.2 | 9.5 | 4.1 | 5.0 |
| TWITTER | 33.3 | 33.7 | 29.3 | 28.1 | 28.3 |
| RECIPE | 47.0 | 51.6 | 52.7 | 47.4 | 45.1 |
| OHSUMED | 52.0 | 56.2 | 55.6 | 50.4 | 44.5 |
| CLASSIC | 5.3 | 5.5 | 4.0 | 3.8 | 3.0 |
| REUTERS | 4.2 | 4.6 | 7.1 | 9.1 | 3.5 |
| AMAZON | 12.3 | 13.3 | 13.9 | 7.8 | 7.2 |

- Used different embedding methods (Hierarchical log0bilinear model & Collobert Weston Model) and word2vec trained on different datasets to observe the change in k-nn performance
- WMD method seems to be very sensitive to the training set used in word2vec method
- Perhaps explains its reason behind very low k-nn error

# Using Different Embedding Methods




- When m = k, WCD metric for classification
- For all other results pre-fetch m instances via WCD, use RWMD to check if a document can be pruned and only if not compute the exact WMD distance until k documents are selected
- RWMD omits all WMD computations
- Relaxed models significantly decrease the computational time without the loss of accuracy

# Conclusion

- Compared to word2vec clustering, this method doesn't provide direct representation of documents
- Space generated by word2vec embedding space is effective in capturing semantic information, and can be applied to capturing the semantics of documents
- Although it didn't elaborate or list any examples to show the explanatory power of WMD method, it was enough to suggest a glimpse of such effect
- Yet, the experiment seemed biased, giving unfair advantage to WMD

# Reference

Kusner, Matt J., et al. "From Word Embeddings To Document Distances."

Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems. 2013.

R. Liu, D. Wang, and C. Xing, “Document classification based on word vectors.” ISCSLP, 2014