Semantic Matching by Non-linear Word Transportation for Information Retrieval

Jiafeng Guo$^1$, Yixing Fan$^1$, Qingyao Ai$^2$, W. Bruce Croft$^2$

$^1$CAS Key Lab of Web Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences

$^2$Center for Intelligent Information Retrieval, University of Massachusetts Amherst
Outline

• Motivation
• Our Approach
• Experiments
• Conclusions
Outline

• Motivation
• Our Approach
• Experiments
• Conclusions
Motivation

Bag-of-words Representation

peace process in the Middle East

Exact (i.e., syntactic) Matching

<table>
<thead>
<tr>
<th>Arabian</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>predictions</td>
<td>2</td>
</tr>
<tr>
<td>peace</td>
<td>2</td>
</tr>
<tr>
<td>negotiations</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Bag-of-words Representation

As for the Arabian and Palestinian voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West Bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming Israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose?
Cons of BoW Representation

- BoW representation assumes strong independence between words

\[
\begin{align*}
\text{car} & \quad [0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,\ldots] \\
\text{automobile} & \quad [0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,\ldots] \\
\cosine(\text{car}, \text{automobile}) & = 0
\end{align*}
\]

\[
\begin{align*}
\text{car} & \quad [0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,\ldots] \\
\text{tie} & \quad [0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,\ldots] \\
\text{car} & \quad [0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,\ldots] \\
\text{sugar} & \quad [0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,\ldots] \\
\text{dist}(\text{car}, \text{tie}) & = \text{dist}(\text{car}, \text{sugar})
\end{align*}
\]

Vocabulary mismatch

Rich semantic meanings and relations in language are lost
Representative Techniques for Semantic Matching in IR

  - Pseudo-relevance feedback shows improvement
  - Prone to the problem of query drift

- **Latent models** [Deerwester et al. 1990, Wei and Croft 2006] *(Embed Query/Doc)*
  - Representing both documents and queries in a latent space
  - Do not improve the performance alone due to the loss of many detailed signals

  - A document is “translated” into a query by leveraging word dependency
  - Key difficulty is how to formalize and estimate the translation probability
Word Embedding

- Semantically meaningful representations of words can be efficiently acquired by distributional models [Mikolov et al. 2013, Pennington et al. 2014]
- Improve the performance of a variety of NLP tasks, such as POS tagging, named entity recognition, and sentiment analysis.

Word = vector of “features”

\[ \begin{align*}
\text{on} & \quad \text{the} \\
\text{of} & \quad \text{mat} \\
\text{dog} & \quad \text{cat} \\
\text{rat} & \quad \text{paws} \\
\text{milk} & \quad \text{cheese} \\
\text{eats} & \quad \text{drinks} \\
\text{sat} & \\
\end{align*} \]

Vector-Space cosine similarity between words \( w \) and \( v \)

\[
\cos(w, v) = \frac{Z_w^T Z_v}{||Z_w||_2 ||Z_v||_2}
\]
A more general representation for IR

How to compute relevance score?

| [1.0 2.5 8.3 4 ...] | 1 |
| [3.2 2 6.5 -1.0 ...] | 1 |
| [5.4 24 1 22 -1 ...] | 1 |
| [4.6 7.6 -1.2 3 ...] | 1 |

Bag-of-Word-Embeddings (BoWE) Representation

- BoWE provides a richer representation of queries and documents
- A good foundation for developing semantic matching based retrieval models
Existing BoWE based IR models

- Build compact vector representation for both query and document
  - Dual embedding space model [Nalisnick et al. 2016]
  - Weighted sum of word vectors [Vulić et al. 2015]
  - Fisher kernel framework [Clinchant et al. 2013]

- Suffer the same problem as latent models for IR
  - Detailed matching signals are lost
  - Cannot work well alone
Existing BoWE based IR models

- Translation models
  - Three transformations in GLM: direct term sampling, transformation via document sampling, and transformation via collection sampling [Ganguly et al 2015]
  - Neural language model + background language model model [Zuccon et al 2015]

- Linearly combination, marginal improvements

\[ p(q|w) = \frac{\text{sim}(q,w)}{\sum_{q'\in V} \text{sim}(q',w)} \]
Our Research Question

Is there a unified semantic matching framework based on BoWE representation that can lead to state-of-the-art retrieval performance?
Outline

• Motivation
• Our Approach
• Experiments
• Conclusions
A New View of the IR Problem

- Formulate the semantic matching between documents and queries as a transportation problems
  - Document words: suppliers
  - Query words: consumers
  - Information: products

- The target of IR is to find the documents that can bring the maximum net returns for a given query
A New View of the IR Problem

- The transportation problem in IR
  1. Each document word has fixed information capacity based on its occurrences, while each query word has unlimited capacity to accommodate as much relevant information as possible from the document;
  2. The information gain of transporting (i.e., matching) a document word to a query word decides the transportation “profit”;
  3. The total profit on each query word should obey the law of diminishing marginal returns;
A Novel Semantic Matching Framework

- **Semantic Matching as Non-Linear Word Transportation**
  - Given a query and a document with BoWE representations, one aims to find a set of optimal flows \( F = \{f_{ij}\} \) that satisfy

\[
\begin{align*}
\max & \quad \sum_{j \in Q} \log \sum_{i \in D} f_{ij}r_{ij} \\
\text{subject to:} & \quad f_{ij} \geq 0 \quad \forall i \in D, \forall j \in Q \\
& \quad \sum_{j \in Q} f_{ij} = c_i \quad \forall i \in D
\end{align*}
\]

- \( f_{ij} \) denotes how much capacity of the i-th document word flows to the j-th query word
- \( r_{ij} \) denotes the corresponding transportation profit
- \( c_i \) denotes the information capacity of the i-th document word
Definition of Document Word Capacity

- Possible definitions:
  - Directly using term frequency $tf_i$ will favor long document
  - Simply applying length normalization will bias toward to short document

$$c_i = \frac{tf_i}{|D|}$$

- Adopt Bayesian smoothing with Dirichlet priors (as in LM)

$$c_i = \frac{tf_i + \mu c_i}{|D| + \mu}$$

- Each document can be viewed as a distribution over the entire vocabulary
Definition of Transportation Profit

- **Semantic closeness between words**: cosine similarity

  \[ r_{ij} = \max\left(\cos(w_i^{(d)}, w_j^{(q)}), 0\right) \quad \forall i \in D, \forall j \in Q \]

- **Matching risk** \( \alpha \): control the profit gap between exact and semantic matching

  \[ r_{ij} = \cos(w_i^{(d)}, w_j^{(q)})^\alpha \]

- A large \( \alpha \) would increase the profit gap between exact matching and semantical matching.

- Better to be query word dependent: the more discriminative a query word is, there is more risk for the semantic matching;
  - IDF is a strong signal of the discriminative power of a word

  \[ r_{ij} = \cos(w_i^{(d)}, w_j^{(q)})^{g(idf_j)} \quad \forall i \in D, \forall j \in Q \]

  \[ g(idf_j) = idf_j + b \]
Model Discussion

- What is the optimal solution of the model?
  - Without non-linear: relaxed transportation problem, match the document word to the most close query word (alignment)
  - With non-linear: decay the matching profit, encourage matching more distinct query words (explain the query as well as possible)

- A very general view of the semantic matching in IR
  - One may design various models under this view by defining different transportation problems
  - For instance
    - Diminishing return effect: Logarithm -> sigmoid functions
    - Document word capacity: Dirichlet smoothed weighting function -> other weighting and smoothing schemes appropriate for IR
    - Matching profit: modified cosine similarity -> Guassian kernel functions

\[
\text{max} \sum_{j \in Q} \sum_{i \in D} f_{ij} \log f_{ij} \\
\text{subject to: } f_{ij} \geq 0 \quad \forall i \in D, \forall j \in Q \\
\sum_{j \in Q} f_{ij} = c_i \quad \forall i \in D \\
\]
Relationship with Word Mover’s Distance

Word Mover’s Distance
[ICML 2015]

\[ \min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} c(i, j) \]

subject to:
\[ \sum_{j=1}^{n} T_{ij} = d_i \quad \forall i \in \{1, \ldots, n\} \]
\[ \sum_{i=1}^{n} T_{ij} = d_j' \quad \forall j \in \{1, \ldots, n\}. \]

- Dissimilarity between text documents
- Linear transportation problem (minimization type)
- Constraints on both sides
- Symmetric metric over a pair of objects with the same type

Our Model

\[ \max \sum_{j \in Q} \log \sum_{i \in D} f_{ij} r_{ij} \]

subject to:
\[ f_{ij} \geq 0 \quad \forall i \in D, \forall j \in Q \]
\[ \sum_{j \in Q} f_{ij} = c_i \quad \forall i \in D \]

- Semantic relevance between queries and documents
- Non-linear transportation problem (maximization type)
- Relax the constraints on the query side due to the vague intent
- No identity of indiscernibles property
Efficient Solution

Efficient pruning and indexing strategy

Original problem

Top-K pruning

Directly solved by convex optimization approaches
Outline

• Motivation
• Our Approach
• Experiments
• Conclusions
Experimental Settings

- **Dataset:**
  - Robust04: news collection
  - GOV2, ClueWeb09-Cat-B: Web collection

- **Pre-processing:**
  - Word stemming using Krovetz stemmer
  - Stop word remove in queries using INQUERY stop list

- **Word Embedding:**
  - Corpus specific: Word2Vec
  - Corpus independent: Google vectors, Glove vectors

- **Evaluation Methodology:**
  - 5-fold cross validation
  - Tuned towards MAP
  - Evaluated by MAP, nDCG@20, P@20

<table>
<thead>
<tr>
<th></th>
<th>Robust04</th>
<th>GOV2</th>
<th>ClueWeb-09-Cat-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>0.6M</td>
<td>35M</td>
<td>38M</td>
</tr>
<tr>
<td>Document Count</td>
<td>0.5M</td>
<td>25M</td>
<td>34M</td>
</tr>
<tr>
<td>Collection Length</td>
<td>252M</td>
<td>22B</td>
<td>26B</td>
</tr>
<tr>
<td>Query Count</td>
<td>250</td>
<td>150</td>
<td>150</td>
</tr>
</tbody>
</table>
Baseline Methods

- Models based on exact matching
  - **QL**: Query likelihood model based on Dirichlet smoothing
  - **BM25**: The classical probabilistic retrieval model
  - **SDM**: SDM is a state-of-the-art language model addressing term dependence using Markov random fields

- Models using semantic matching
  - **RM3**: One of the representative PRF models
  - **LM+LDA**: The LDA-based document model within the language modeling framework
  - **LM+WE-VS**: A linear combination of word embedding based retrieval model and unigram language
  - **WE-GLM**: A word embedding based translation model

- Our model: **NWT**
Experimental Results

Our model can significantly outperform basic exact matching baselines;

Our model even performs better than the state-of-the-art n-gram based model SDM;

Our model can significantly outperform existing latent models and word embedding based models;

Our model’s performance is comparable with PRF methods;

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic titles</th>
<th></th>
<th>Topic descriptions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
<td>P@20</td>
<td>MAP</td>
</tr>
<tr>
<td>Exact Matching Baselines</td>
<td>QL</td>
<td>0.253−</td>
<td>0.415−</td>
<td>0.369−</td>
<td>0.246−</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.255−</td>
<td>0.418−</td>
<td>0.370</td>
<td>0.241−</td>
</tr>
<tr>
<td></td>
<td>SDM</td>
<td>0.263</td>
<td>0.423</td>
<td>0.375</td>
<td>0.261</td>
</tr>
<tr>
<td>Semantic Matching Baselines</td>
<td>RM3</td>
<td>0.295+</td>
<td>0.423</td>
<td>0.375</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>LM+LDA</td>
<td>0.258−</td>
<td>0.421</td>
<td>0.374</td>
<td>0.247−</td>
</tr>
<tr>
<td></td>
<td>LM+WE-VS</td>
<td>0.255−</td>
<td>0.417−</td>
<td>0.370−</td>
<td>0.253−</td>
</tr>
<tr>
<td></td>
<td>WE-GLM</td>
<td>0.255−</td>
<td>0.417</td>
<td>0.371</td>
<td>0.252−</td>
</tr>
<tr>
<td>Our Approach</td>
<td>NWT</td>
<td>0.274</td>
<td>0.426</td>
<td>0.380</td>
<td>0.268</td>
</tr>
</tbody>
</table>

Significant improvement or degradation with respect to NWT is indicated (+/-) (p-value<0.05)
### Experimental Results

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic titles</th>
<th>Topic descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td><strong>GOV2 collection</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact Matching Baselines</td>
<td>QL</td>
<td>0.295−</td>
<td>0.409−</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.295</td>
<td>0.421−</td>
</tr>
<tr>
<td></td>
<td>SDM</td>
<td>0.319+</td>
<td>0.441+</td>
</tr>
<tr>
<td>Semantic Matching Baselines</td>
<td>RM3</td>
<td>0.301</td>
<td>0.395−</td>
</tr>
<tr>
<td></td>
<td>LM+WE-VS</td>
<td>0.295−</td>
<td>0.408−</td>
</tr>
<tr>
<td></td>
<td>WE-GLM</td>
<td>0.299−</td>
<td>0.411−</td>
</tr>
<tr>
<td>Our Approach</td>
<td>NWT</td>
<td>0.304</td>
<td>0.422</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model Name</th>
<th>Topic titles</th>
<th>Topic descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td><strong>Clueweb-09-Cat-B collection</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact Matching Baselines</td>
<td>QL</td>
<td>0.100−</td>
<td>0.224−</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.101</td>
<td>0.225−</td>
</tr>
<tr>
<td></td>
<td>SDM</td>
<td>0.109</td>
<td>0.242−</td>
</tr>
<tr>
<td>Semantic Matching Baselines</td>
<td>RM3</td>
<td>0.103</td>
<td>0.224−</td>
</tr>
<tr>
<td></td>
<td>LM+WE-VS</td>
<td>0.101−</td>
<td>0.225−</td>
</tr>
<tr>
<td></td>
<td>WE-GLM</td>
<td>0.102−</td>
<td>0.228−</td>
</tr>
<tr>
<td>Our Approach</td>
<td>NWT</td>
<td>0.107</td>
<td>0.236</td>
</tr>
</tbody>
</table>

Similar observations can be found on the other two large Web collections.
Model Analysis

- **Impact of word embeddings:**
  - Corpus-specific word embeddings are better than corpus-independent word embeddings
  - CBOW works better than SG
Impact of embedding dimension:
- With lower dimensionality, the similarity between word embeddings might be coarse and hurt the semantic matching performance
- With larger dimensionality, one may need more data to train reliable word embeddings

Impact of indexed neighbor:
- With a small neighbor size, the performance improvement is limited since very few semantic matching signals can be leveraged
- When the neighbor size is large enough, we can obtain quite stable results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Embedding</th>
<th>MAP</th>
<th>NDCG@20</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles</td>
<td>CBOW-50d</td>
<td>0.255</td>
<td>0.409</td>
<td>0.360</td>
</tr>
<tr>
<td></td>
<td>CBOW-100d</td>
<td>0.265</td>
<td>0.420</td>
<td>0.371</td>
</tr>
<tr>
<td></td>
<td>CBOW-300d</td>
<td>0.274</td>
<td>0.426</td>
<td>0.380</td>
</tr>
<tr>
<td></td>
<td>CBOW-500d</td>
<td>0.272</td>
<td>0.426</td>
<td>0.379</td>
</tr>
<tr>
<td>Descriptions</td>
<td>CBOW-50d</td>
<td>0.246</td>
<td>0.394</td>
<td>0.336</td>
</tr>
<tr>
<td></td>
<td>CBOW-100d</td>
<td>0.256</td>
<td>0.400</td>
<td>0.342</td>
</tr>
<tr>
<td></td>
<td>CBOW-300d</td>
<td>0.268</td>
<td>0.413</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>CBOW-500d</td>
<td>0.265</td>
<td>0.411</td>
<td>0.352</td>
</tr>
</tbody>
</table>
Outline

• Motivation
• Our Approach
• Experiments
• Conclusions
Conclusions

- A new view of semantic matching for IR
  - Non-linear word transportation framework

- Three important advantages of our model
  - Transportation based on the BoWE can capture detailed semantic matching signals between words;
  - The word alignment effect and the marginal diminishing effect can better model the semantic matching process;
  - The flexibility in model definition enables the design of models dedicated to the IR task.

- Future Work
  - explore different model variations within the transportation framework
Thanks!

Email: guojiafeng@ict.ac.cn
Data and Codes can be found at
http://www.bigdatalab.ac.cn/benchmark/
Relationship with Semantic Matching based Retrieval models

- **Statistical Translation models**
  - The transportation bears a similarity to the translation process;
  - Translation models are usually formalized under the probabilistic framework;
  - Our model uses a non-linear objective, no strict requirement for the transportation profit to be a probability

- **Query Expansion**
  - Share similar ideas with global methods, but different in how to obtain the word association relationships and the usage of the relationships;
  - Orthogonal to query expansion with local methods;

- **Latent Models**
  - Latent models use compact vector for the document
  - Our model uses BoWE for the document, so both exact and semantic matching signals can be captured
Inspiration

• “From Word Embeddings To Document Distances” [Kusner et al. ICML 2015]

• Word Mover’s Distance (WMD)
  • Measure the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to “travel” to reach the embedded words of another document