Modeling Topics and Knowledge Bases with Embeddings

Dat Quoc Nguyen and Mark Johnson

Department of Computing
Macquarie University
Sydney, Australia

December 2016
Vector representations/embeddings of words

- One-hot representation: high-dimensional and sparse vector
  - Vocabulary size: N
  - N-dimensional vector: filled with 0s, except for a 1 at the position associated with word index
Vector representations/embeddings of words

- Deep learning evolution: most neural network toolkits do not play well with one-hot representations
  - Dense vector: low-dimensional distributed vector representation
  - Vector size $k \ll N$, for example: $k = 100$, Vocabulary size $N = 100000$
Vector representations/embeddings of words

- Unsupervised learning models are proposed to learn low-dimensional vectors of words efficiently, e.g. W2V Skip-gram
- Word embeddings learned from large external corpora capture various aspects of word meanings
  - Possibility to assign topics (i.e. labels) to clusters of “similar” meaning words
  - Topic 1 for \{banking, bank, transaction, finance, money laundering\}

⇒ Can we incorporate word embeddings to topic models?
• **Topic models** take a corpus of documents as input, and
  ▶ Learn a set of latent *topics* for the corpus
  ▶ Infer *document-to-topic* and *topic-to-word* distributions from co-occurrence of words within documents

If the corpus is small and/or the documents are short, the topics will be noisy due to the limited information of word co-occurrence

**IDEA:** Use the word embeddings learned on a large external corpus to improve the topic-word distributions in a topic model
Improving topic models with word embeddings

- Each word $w$ is associated with a pre-trained word embedding $\omega_w$
- Each topic $t$ is associated with a topic embedding $\tau_t$
- We define a latent feature topic-to-word distribution $\text{CatE}(w)$ over words:
  \[
  \text{CatE}(w \mid t) = \frac{\exp(\omega_w \cdot \tau_t)}{\sum_{w' \in V} \exp(\omega_{w'} \cdot \tau_t)}
  \]
  - Optimize the log-loss to learn $\tau_t$
- Our new topic models mix the CatE distribution with a multinomial distribution over words
  - Combine information from a large, general corpus (via the CatE distribution) and a smaller but more specific corpus (via the multinominal distribution)
Improving topic models with word embeddings

- Combine Latent Dirichlet Allocation (LDA) and Dirichlet Multinomial Mixture (DMM) with the word embeddings: LF-LDA & LF-DMM

\[ \alpha \rightarrow \theta \rightarrow z \rightarrow w_{Nd|D} \] (LDA)

\[ \alpha \rightarrow \theta \rightarrow z \rightarrow w_{Nd|D} \rightarrow \lambda \] (LF-LDA)

\[ \alpha \rightarrow \theta \rightarrow z \rightarrow w_{Nd|D} \] (DMM)

\[ \alpha \rightarrow \theta \rightarrow z \rightarrow w_{Nd|D} \rightarrow \lambda \] (LF-DMM)
## Improving topic models with word embeddings

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMM</td>
<td>LF-DMM</td>
<td>DMM</td>
</tr>
<tr>
<td>japan</td>
<td>japan</td>
<td>u.s.</td>
</tr>
<tr>
<td>nuclear</td>
<td>nuclear</td>
<td>oil</td>
</tr>
<tr>
<td>u.s.</td>
<td>u.s.</td>
<td>japan</td>
</tr>
<tr>
<td>crisis</td>
<td>plant</td>
<td>prices</td>
</tr>
<tr>
<td>plant</td>
<td>quake</td>
<td>stocks</td>
</tr>
<tr>
<td>china</td>
<td>radiation</td>
<td>sales</td>
</tr>
<tr>
<td>libya</td>
<td>earthquake</td>
<td>profit</td>
</tr>
<tr>
<td>radiation</td>
<td>tsunami</td>
<td>fed</td>
</tr>
<tr>
<td>u.n.</td>
<td>vote</td>
<td>growth</td>
</tr>
<tr>
<td>vote</td>
<td>korea</td>
<td>growth</td>
</tr>
<tr>
<td>korea</td>
<td>disaster</td>
<td>wall</td>
</tr>
<tr>
<td>europe</td>
<td>power</td>
<td>street</td>
</tr>
<tr>
<td>government</td>
<td>oil</td>
<td>china</td>
</tr>
<tr>
<td>election</td>
<td>japanese</td>
<td>fall</td>
</tr>
<tr>
<td>deal</td>
<td>plants</td>
<td>shares</td>
</tr>
</tbody>
</table>

(Topic coherence)

- Significant improvement of topic coherence scores on all models and experimental corpora
- Obtain 5+\% absolute improvements in clustering and classification evaluation scores on the small or short datasets
- No reliable difference between pre-trained Word2Vec and Glove vectors
Vector representations/embeddings: “distance” results

\[ \mathbf{v}_{king} - \mathbf{v}_{queen} \approx \mathbf{v}_{man} - \mathbf{v}_{woman} \]

\[ \mathbf{v}_{Vietnam} - \mathbf{v}_{Hanoi} \approx \mathbf{v}_{Japan} - \mathbf{v}_{Tokyo} \]

\[ \approx \mathbf{v}_{Germany} - \mathbf{v}_{Berlin} \]

\[ \approx \mathbf{v}_{some\_relationship, \ saying: \ is\_capital\_of} \]

\[ \mathbf{v}_{Hanoi} + \mathbf{v}_{is\_capital\_of} \approx \mathbf{v}_{Vietnam} \]

\[ \mathbf{v}_{Tokyo} + \mathbf{v}_{is\_capital\_of} \approx \mathbf{v}_{Japan} \]

\[ \mathbf{v}_{Berlin} + \mathbf{v}_{is\_capital\_of} \approx \mathbf{v}_{Germany} \]

Triple (head entity, relation, tail entity)

\[ \mathbf{v}_h + \mathbf{v}_r \approx \mathbf{v}_t \]

\[ \implies \| \mathbf{v}_h + \mathbf{v}_r - \mathbf{v}_t \|_{1/2} \approx 0 \]

Link prediction in knowledge bases?
- **Knowledge bases (KBs)** of real-world triple facts (head entity, relation, tail entity) are useful resources for NLP tasks
- **Issue**: large KBs are still far from complete
- So it is useful to perform *link prediction in KBs* or *knowledge base completion*: predict which triples not in a knowledge base are likely to be true
- **Embedding models** for link prediction in KBs:
  - Associate entities and/or relations with dense feature vectors or matrices
  - Obtain SOTA performance and generalize to large KBs
The TransE model (Bordes et al., 2013) represents each relation $r$ by a translation vector $\mathbf{v}_r$, which is chosen so that $\|\mathbf{v}_h + \mathbf{v}_r - \mathbf{v}_t\|_{\ell_1/2} \approx 0$

- Good for 1-to-1 relationships, e.g. *is_capital_of*
- Not good for 1-to-Many, Many-to-1 and Many-to-Many, e.g. *gender*

STransE: a novel embedding model of entities and relationships in KBs

- Our STransE represents each entity as a low dimensional vector, and each relation by two matrices and a translation vector
- STransE choose matrices $\mathbf{W}_{r,1}$ and $\mathbf{W}_{r,2}$, and vector $\mathbf{v}_r$ so that: $\|\mathbf{W}_{r,1}\mathbf{v}_h + \mathbf{v}_r - \mathbf{W}_{r,2}\mathbf{v}_t\|_{\ell_1/2} \approx 0$
- Optimize a margin-based objective function to learn the vectors and matrices
We conducted experiments on two benchmark datasets WN18 and FB15k (Bordes et al., 2013).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#E</th>
<th>#R</th>
<th>#Train</th>
<th>#Valid</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN18</td>
<td>40,943</td>
<td>18</td>
<td>141,442</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td>FB15k</td>
<td>14,951</td>
<td>1,345</td>
<td>483,142</td>
<td>50,000</td>
<td>59,071</td>
</tr>
</tbody>
</table>

Link prediction task:

- Predict $h$ given $(?, r, t)$ or predict $t$ given $(h, r, ?)$ where ? denotes the missing element.
- Evaluation metrics: mean rank (MR) and Hits@10 (H10).

Find a new relation-path based embedding model in our CoNLL paper!
STransE results for search personalization

- Two users search using the same keywords, they are often looking for different information (i.e. difference due to the users’ interests)
- Personalized search customizes results based on user’s search history (i.e. submitted queries and clicked documents)
- Let \((q, u, d)\) represent a triple (query, user, document)
- Represent the user \(u\) by two matrices \(W_{u,1}\) and \(W_{u,2}\) and a vector \(v_u\), which represents the user’s topical interests, so that:
  \[\|W_{u,1}v_q + v_u - W_{u,2}v_d\|_{\ell_{1/2}} \approx 0\]
- \(v_d\) and \(v_q\) are pre-determined by employing the LDA topic model
- Optimize a margin-based objective function to learn the user embeddings and matrices

<table>
<thead>
<tr>
<th>Metric</th>
<th>Search Eng.</th>
<th>L2R SP</th>
<th>STransE</th>
<th>TransE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRR</td>
<td>0.559</td>
<td>0.631(+12.9)%</td>
<td>0.656(+17.3)%</td>
<td>0.645(+15.4)%</td>
</tr>
<tr>
<td>P@1</td>
<td>0.385</td>
<td>0.452(+17.4)%</td>
<td>0.501(+30.3)%</td>
<td>0.481(+24.9)%</td>
</tr>
</tbody>
</table>
Conclusions

- Latent feature vector representations induced from large external corpora can be used to improve topic modeling on smaller datasets.

- Our new embedding model STransE for link prediction in KBs.
  - Applying to a search personalization task, STransE helps to significantly improve the ranking quality.

- https://github.com/datquocnguyen
Thank you for your attention!