Re-contextualization and contextual Entity exploration

Sebastian Holzki
Authors: Joonseok Lee, Ariel Fuxman, Bo Zhao, and Yuanhua Lv

- PAPER PRESENTATION -
LEVERAGING KNOWLEDGE BASES FOR CONTEXTUAL ENTITY EXPLORATION
Information needs

Abstract

Learning to rank is a relatively new field of study, aiming at the selection from a set of training data with relevance labels are often evaluated using information retrieval measures, such as Discounted Cumulative Gain (NDCG) [1]. Until recently, most learning to rank algorithms were related to the above mentioned evaluation measures, optimization of these measures is that they depend on the numerical values output by the ranking function. A framework that addresses this challenge by optimizing over all the possible permutations of documents. A simple and approximate the average of NDCG over the space of rankings.
Contextual entity exploration problem

Removing the former disconnects “Silas Deane” from two context nodes, whereas removing the latter does not disconnect the selection node from any context node. This makes intuitive sense because while the Revolutionary War is very relevant to Silas Deane and the context of the document, the entity “Blacksmith” is irrelevant to the context (it is just mentioned on the Wikipedia page for Silas Deane because his father was a Blacksmith).
Using knowledge bases effectively

- Build a focused subgraph of the knowledge graph
  - User selection
  - Context around user selection
- Explore the graph for other relevant entities

Figure: Lee, Fuxman, Bo and Yuanhua [1]. (adapted)
Creating a reasonable ranking

Deliver a limited number of entities, ordered by relevance
How to measure relevance?

• Ability to connect user-selected entity and context entities  
  • Context-Selection Betweenness
• Overall relevance to user-selected entity  
  • Personalized Random Walk

Figure: Lee, Fuxman, Bo and Yuanhua [1]. (adapted)
Context-Selection betweenness (background)

Employing „Normalized Wikipedia Distance“ *

\[
NW D(u, v) = \frac{\log(\max(|I_u|, |I_v|)) - \log |I_u \cap I_v|}{\log |V| - \log(\min(|I_u|, |I_v|))}
\]

\(I_u\) and \(I_v\): Neighbours of \(u\) and \(v\)

- Measures the semantic distance of Wikipedia pages \((u\text{ and }v)\)
- More common neighbours \(\rightarrow\) less distant
- Normalized by corpus size

* Already introduced within this course to measure entity coherence, namely: \(mw-coh\).
Context-Selection betweenness

\[ CSB(v) = \frac{1}{Z} \sum_{c: v \in SP(s, c)} \frac{w(s, c)}{k \cdot l(s, c)} \]

\[ w(s, c) = \max(\theta - NWD(s, c), 0) \]

- \( \theta = 0.5 \): Emphasizes more relevant contexts
- Weight relative to path-length \( l \) and number of all shortest paths \( k \)
- Normalized by weight of all shortest paths \( Z \)
  (Not only those including node \( v \))

Result: Nodes \( v \) connecting \( s \) and \( c \) through short paths have higher Context-Selection Betweenness
Personalized Random Walk

Motivation:
Consider the overall relevance of an entity

• „Personalized“ means allowing arbitrary jumps:
  • Probability for user selected entity: $x_s$
  • Probabilities for context entities: $x_c/|C|$

• Result: emphasize on entities near user selection or context

Reminder:
„Random Walk“ assigns stationary probabilities to nodes in a graph, by randomly „walking“ through the graph.
Score Aggregation

\[
Relevance(v) = |V|R W(v) + \alpha \frac{|C|}{|V|} |C|C S B(v)
\]

1. Large graphs produce low probabilities
2. Context-Selection Betweenness is biased by number of context entities
3. Trust in Context-Selection Betweenness is affected by graph-size and number of context entities
Implementation details

- Wikipedia (2014) as knowledge base
- Entity linking through Wikipedia Hyperlinks

- Personalized Random Walk Parameters: $x_s = 0.05$; $x_c = 0$
- Aggregated score used as ranking

- User selection generated by crowd workers
- Evaluation of ranking also done by crowd workers
Results: comparison to baselines

Source: Lee, Fuxman, Bo and Yuanhua [1]
Results: scores and aggregation

Source: Lee, Fuxman, Bo and Yuanhua [1]
Summary

• Switching between primary and secondary sources
• Contextual entity exploration
• Relevance in knowledge graphs
• Context-Selection Betweenness + Personalized Random Walk
• Comparison to baseline

Future work (Source: [1])
• Different contextualization systems have low degree of overlap
• Idea: combine them and re-rank
[Backup] Normalized Wikipedia Distance Example

\[
NWD(u, v) = \frac{\log(\max(|I_u|, |I_v|)) - \log |I_u \cap I_v|}{\log |V| - \log(\min(|I_u|, |I_v|))}
\]

\[
|I_u| = 4 \\
|I_v| = 3 \\
|I_u \cap I_v| = 2 \\
|V| = 9
\]

NWD = (log(4) - log(2)) / (log(9) - log(3)) ≈ 0.631
Results: influence of entity linking

Source: Lee, Fuxman, Bo and Yuanhua [1]
Authors: Nam Khanh Tran, Andrea Ceroni, Nattiya Kanhabua and Claudia Niederée

- PAPER PRESENTATION -
BACK TO THE PAST: SUPPORTING INTERPRETATIONS OF FORGOTTEN STORIES BY TIME-AWARE RE-CONTEXTUALIZATION
Context of documents from the past

Prior to 1964, many of the cigarette companies advertised their brand by falsely claiming that their product did not have serious health risks. A couple of examples would be "Play safe with Philip Morris" and "More doctors smoke Camels". Such claims were made both to increase the sales of their product and to combat the increasing public knowledge of smoking's negative health effects.

Advertisement poster from the 1950s

Source: Course slides of Nam Khanh Tran [3]
Four goals of contextualizing „the past“

1. **Relevance**: Context $c$ has to be relevant to information item $d$
2. **Complementarity**: Avoid displaying duplicate information
3. **Time-awareness**: Consider creation-time of $d$
4. **Conciseness**: Show brief context, not whole pages
Prerequisites

• Document with contextualization hooks
  • Contextualization hook:
    „... parts of document, that require contextualization.“ [2]
    *Entities, concept mentions, ...*

• Knowledge base

• **Facility to provide proper context for hooks**
  (Proper: with regard to the four goals)
Main concept of the paper

- Document
  - Hook identification
  - Query formulation
  - Context ranking
  - Context retrieval
  - Ranked context
    1 ...
    2 ...
    3 ...

Source: Course slides of Nam Khanh Tran [3] and original paper [2]
Query formulation

1. Document-based queries
   (title, lead paragraph)

2. Hook-based queries
   (all hooks, each hook on its own)

3. Learning to select promising queries out of power-set of hooks
   - Trained on Recall@k
   - Some prediction-features:
     - Linguistic: entities, document length, number of nouns, …
     - Temporal scope (±2 years), temporal similarity, temporal document frequency
     - …

Query Performance Prediction (QPP)
Context retrieval

• Query-likelihood language model:

\[ P(c|q) \propto P(c) \prod_{w \in q} P(w|c)^{n(w,q)} \]

• Rank contextualization units \( c \) according to the probability \( P(c|q) \) => How likely is it, that \( c \) generates the given query \( q \) at random

• Assuming independent query terms

• Serves for relevance goal
Learning to re-rank context

- Contextualization units initially ranked by retrieval model

- Goal of reranking: Achieve high precision-measure
- Also consider a variety of features

- Trained by examples: [proper context => document]
- Incorporated features:
  - Topic diversity
  - Entity difference, text difference (complementarity)
  - Temporal similarity of context and document
Evaluation: query formulation methods

Highlighted: variations of machine-learned approach (Query Performance Prediction)
Evaluation: re-ranking context

<table>
<thead>
<tr>
<th>Baselines</th>
<th>qpp_@100</th>
<th>P@1</th>
<th>P@3</th>
<th>P@10</th>
<th>MAP</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM-T</td>
<td>0.5882</td>
<td>0.4967</td>
<td>0.4176</td>
<td>0.5446</td>
<td>0.2796</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>0.5882</td>
<td>0.5490</td>
<td>0.4529</td>
<td>0.5802</td>
<td>0.3097*</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>0.8054*</td>
<td>0.6993*</td>
<td>0.5140*</td>
<td>0.6498*</td>
<td>0.3951*</td>
<td></td>
</tr>
</tbody>
</table>

RF: learning to rank approach using RankForest
LM: standard query-likelihood language model
LM-T: modification of LM including time-awareness

- Same query formulation method QPP@100
- Learning to rank performs significantly better
- Complementarity plays an important role
Summary

• Contextualization need for documents from the past
• Usual entity contextualization falls short
• Time-awareness and complementarity improve context quality

• Query formulation optimized for recall
• Context retrieval and ranking optimized for precision

• Results show significant improvements
Topic: Contextualization – overall summary

• Focus on user-selection/hooks
• Well structured and large knowledge bases are mandatory
• Some goals require more effort (time-awareness)

• Automated contextualization makes life a lot easier

Thanks for your attention!
Sources

