FacetE: Exploiting Web Tables for Domain-Specific Word Embedding Evaluation

Michael Günther, Paul Sikorski, Maik Thiele, and Wolfgang Lehner

DBTest ’20 Workshop at SIGMOD 2020

19.06.2020
NLP Systems Workflow

Data Storage with textual data

Extracted text data

Language Model

Numerical Representation (Vectors)

5.02, 43.07, ...

Numerical Representation (Vectors)
NLP Systems Workflow

State-of-the-art Language Models: Word Embeddings

Data Storage with textual data

Extracted text data

Language Model

Numerical Representation (Vectors)

Deep Neuronal Network

Training on Dummy Task

Extract Weights as Pre-Trained Language Model

Large Text corpora in natural language

5.02, 43.07, …..
NLP Systems Workflow

Data Storage with textual data

Extracted text data → Language Model → Numerical Representation (Vectors)

Classification and Regression Tasks

Similarity Search Tasks
Word Embedding for Systems

**ML Systems**
- Utilize implicitly encoded knowledge from large text corpora
- Capture semantic similarities of text values

**Database Systems**
- Semantic text similarity queries
- Data exploration
- Data integration

**Information Retrieval Systems**
- Semantic search
- Query Expansion
- Multi-lingual search

Choice of the word embedding model is crucial for the performance!
Evaluation of Word Embedding Models

Word Similarity
- Similar Words by cosine similarity of word vectors
  \[ \text{sim}_{\text{cos}}(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} \]
- Example: most similar to “king”?
  → prince, man, and queen

Analogy Queries
- Retrieve Similar Relations
  \[ a - b \approx c - ? \]
- 3CosAdd: \( \arg \max_{d \in V \{a, b, c\}} \text{sim}_{\text{cos}}(d, c - a + b) \)
- Example: man – woman ≈ king - ?
  → queen

Schematic Representation of Word Vectors
Evaluation of Word Embedding Models

Common Similarity Datasets
- WS-353: 353 word pairs of general domain knowledge quantifying semantic relatedness
- SimLex-999: 999 word pairs of general domain knowledge quantifying semantic similarity

Depend on human notion of similarity → Require human labeling effort

Common Analogy Query Datasets
- Google Analogy: 550 semantic and syntactic relations, mostly city-country relations
- MSR: 8,000 analogies of 800 syntactic relations

Facts of general domain knowledge → Automatic extraction possible

| Embedding Model | WS353 | RW | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CBoW</td>
<td>57.2</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>SkipGram</td>
<td>62.8</td>
<td>37.2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Embedding Model</th>
<th>Semantic</th>
<th>Syntactic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBoW</td>
<td>57.3</td>
<td>68.9</td>
<td>63.7</td>
</tr>
<tr>
<td>SkipGram</td>
<td>66.1</td>
<td>65.1</td>
<td>65.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>


Limitations:
- Only small datasets
- Return a single value only
- Only general domain
Evaluation of Word Embedding Models

Common Similarity Datasets
- **WS-353** 353 word pairs of general domain knowledge quantifying semantic relatedness
- **SimLex-999** 999 word pairs of general domain knowledge quantifying semantic similarity

**Limitations:**
- Only small datasets
- Return a single value only
- Only general domain

**Design Goals:**
- Large number of relations
- Flexible structure
- Multiple categories

**Design Strategies:**
- Extraction from millions of web tables
- Organization in facets
- Definition of categories

Common Analogy Query Datasets
- **Google Analogy** 550 semantic and syntactic relations, mostly city-country relations
- **MSR** 8,000 analogies of 800 syntactic relations

**Facts of general domain knowledge**
- Automatic extraction possible

**Limitations:**
- Depend on human notion of similarity
  - Require human labeling effort

**Design Goals:**
- Flexible structure
- Multiple categories

**Design Strategies:**
- Extraction from millions of web tables
- Organization in facets
- Definition of categories
Dataset Design

Data Source: Web Tables
- Large amount of knowledge
- General enough to be expected in pre-trained word embedding models
- Redundancy allows to exclude temporary facts (e.g. time dependent facts like home soccer team to visiting team)

Target Design: Facets
- Each Facet $F:O \rightarrow V$ assigns objects (e.g. Soccer Player) to values (e.g. Teams)
- Allows flexible construction of application specific evaluation datasets
- More flexible than hierarchical categorization
Extraction Pipeline

1. **Pre Filtering**: Frequency and Regex Filter, Facet Creation
2. **Soft Functional Dependencies**: Check contradiction of most frequent relation
3. **Post Filtering**: Filter by Pooling, Blacklist, ...
4. **Categorization**: Assign facets to 8 broader categories

250 Facets / 600K Values

Word Embeddings ➔ Analogy Evaluation

125M Web Tables
Extraction Pipeline

1) Pre-Filtering
- Filters infrequent and non-textual data of English tables

- Remove non-textual data
- Remove infrequent columns

2) Soft Functional Dependencies: Check contradiction of most frequent relation

3) Post Filtering: Filter by Pooling, Blacklist, ...

4) Categorization: Assign facets to 8 broader categories

- 250 Facets / 600K Values
- Word Embeddings
- Analogy Evaluation

125M Web Tables

Country | Date | Team
-------|------|------
Country | Team | Nick-name
Team | Country

Column-Tuples → Basis for Facets
2) Soft-Functional Dependencies

- Determine static facts

1) Determine most frequent relation pairs

<table>
<thead>
<tr>
<th>Team</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenal</td>
<td>England</td>
</tr>
<tr>
<td>AC Milan</td>
<td>Italy</td>
</tr>
<tr>
<td>Juventus</td>
<td>Italy</td>
</tr>
</tbody>
</table>

2) Check on contradictions

\[
SFD(o, v) = \frac{\text{count}(o, v)}{\sum_{v':(o, v')} \text{count}(o, v')}
\]

\[
SFD(\text{Arsenal, England}) = \frac{2}{3}
\]

Most frequent for “Arsenal”

Pre Filtering: Frequency and Regex Filter, Facet Creation

Soft Functional Dependencies: Check contradiction of most frequent relation

Post Filtering: Filter by Pooling, Blacklist, ...

Categorization: Assign facets to 8 broader categories

125M Web Tables

Word Embeddings

Analogy Evaluation

One Contradiction
3) **Post-Filtering**

- **Blacklists**
  Remove too generic facets

- **Word Embedding Pooling**
  Retain only facets modeled by at least one word embedding model

---

**Name** | **Description**
--- | ---

---

**Pre Filtering:** Frequency and Regex Filter, Facet Creation

**Soft Functional Dependencies:** Check contradiction of most frequent relation

**Post Filtering:** Filter by Pooling, Blacklist, ...

**Categorization:** Assign facets to 8 broader categories

250 Facets / 600K Values

Word Embeddings

Analogy Evaluation

125M Web Tables
**Extraction Pipeline**

4) **Categorization**

- Assign each of the 250 facets on of 8 broader categories (e.g. geographic, music, sports, ...)

<table>
<thead>
<tr>
<th>Team</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC Milan</td>
<td>Italy</td>
</tr>
<tr>
<td>Juventus</td>
<td>Italy</td>
</tr>
<tr>
<td>Arsenal</td>
<td>England</td>
</tr>
</tbody>
</table>

**Keywords for categories**

**Similarity to Keywords**

<table>
<thead>
<tr>
<th>Cat.</th>
<th>Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music</td>
<td>0.15</td>
</tr>
<tr>
<td>Sports</td>
<td>0.53</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Word Embedding Model**

**Pre Filtering:** Frequency and Regex Filter, Facet Creation

**Soft Functional Dependencies:** Check contradiction of most frequent relation

**Post Filtering:** Filter by Pooling, Blacklist, ...

**Categorization:** Assign facets to 8 broader categories

125M Web Tables

250 Facets / 600K Values

**Word Embeddings**

**Analogy Evaluation**
Evaluation

Evaluation of Categories

Setup
- 4 Pre-trained word embedding models: GloVe, Word2Vec-SkipGram, fastText, SentenceBert
- Selection of 4 FacetE categories

Calculation
- Select facets $F: O \rightarrow V$ from the categories
- Determine the value $V$ for each object $O$ with 3CosAdd analogy method
- Calculate amount of correctly assigned values
- Calculate average in each category

Coverage: For some text values word embedding models can not determine a vector

Evaluation of 4 Categories
Evaluation of Categories

Setup
- 4 Pre-trained word embedding models: GloVe, Word2Vec, SkipGram, fastText
- Selection of 4 FacetE categories

Calculation
- Select facets $F: O \rightarrow V$ from the categories
- Determine the value $V$ for each object $O$ with 3CosAdd analogy method
- Calculate amount of correctly assigned values
- Calculate average in each category

Coverage: For some text values word embedding models can not determine a vector

Observation
- No single best model
- High Coverage
Evaluation

Evaluation of a Single Object Set

Setup

- 4 Pre-trained word embedding models: GloVe, Word2Vec-SkipGram, fastText, SentenceBert
- Selection of all facets for cities

Calculation

- Determine the value $V$ for each object $O$ with 3CosAdd analogy method
- Calculate amount of correctly assigned values for each city name
- Calculate average across all objects

![Evaluation of a Single Object Set - Cities](image-url)
Evaluation

Evaluation of a Single Object Set

Setup
- 4 Pre-trained word embedding models: GloVe, Word2Vec-SkipGram, fastText, SentenceBert
- Selection of all facets for cities

Calculation
- Determine the value $V$ for each object $O$ with 3CosAdd analogy method
- Calculate amount of correctly assigned values for each city name
- Calculate average across all objects

Observation
Word2Vec performs better on geographic data, however GloVe has a better representation of cities

Evaluation of a Single Object Set - Cities
Conclusion

Web Table Extraction Pipeline

- Web Tables are a good resource for structured relations of general common knowledge
- Pipeline is able to process millions of tables → Reusable for other table corpora

Facet Structure

- Enables flexible construction of evaluation datasets
- Evaluation of different granularity Single Facts (e.g. City → Country), Objects (e.g. Cities) or Domains (e.g. Geographic)

Evaluation of Common Word Embedding Models

- Large differences in accuracy values on different domains
- No best model for all cases

FacetE Dataset: https://www.kaggle.com/guenthermi/facete