Using Frame Embeddings to Identify Semantically Related Software Requirements

Waad Alhoshan, Riza Batista-Navarro and Liping Zhao
Semantic Relatedness in RE

- Detecting related NL requirements is **tricky** sometimes!
  - Inherent problems due to NL, e.g. ambiguity and incompleteness
  - Writing in NL does not adhere to any formalism

- Req-1: The transaction records are kept into a central database of the Bank and only authorised users are able to view the documents.
- Req-2: The Bank’s reports are stored and restricted i.e. accessing the logs should be allowed to specific users.
- Req-3: The Bank’s clients are requested to confirm their personal information regularly.
- Req-4: Every year the bank control system shall ask the clients to verify their contact information.
Semantic Frames

- **Semantic Frame** is defined as a coherent structure of concepts.
- **FrameNet** is an implementation of that theory:
  - More than 1200 frames.
  - Curated by language experts.
  - Frame contents: Definition, Core and non-core frame elements, lexical units and semantic relations with other frames (if any).
Semantic Frames Cont.

Req-1: The transaction records are kept into a central database of the Bank and only authorised users are able to view the documents.

FN-Req-1: The transaction records [Records] are kept [Storing] into a central database of the Bank and only authorised [Deny_or_grant_permission] users are able [Capability] to view [Perception_active] the documents [Text].

Req-2: The Bank’s reports are stored and restricted i.e. accessing the logs should be allowed to specific users.

FN-Req-2: The Bank’s reports [Text] are stored [Storing] and restricted [Deny_or_grant_permission] i.e. accessing the logs [Records] should be allowed [Preventing_or_letting] to specific [Specific_individual] users.

Req-3: The Bank’s clients are requested to confirm their personal information regularly.

FN-Req-3: The Bank’s clients are requested [Request] to confirm [Verification] their personal information [Information] regularly [Frequency].

Req-4: Every year the bank control systems shall ask the clients to verify their contact information.

Word Embedding

Background
NLP4RE

“Customizing general NLP techniques to make them applicable for solving the problems requirements engineers face in their daily practice.” [1]
NLP4RE Cont.

• We published papers based on Corpus-based investigation research for using FrameNet in RE (FN-RE corpus) [2] [3].
• Also, to investigate ways for measuring Semantic Relatedness between Frames from RE perspective.
  o Knowledge-based measures (WUP and Path) [4]
  o Context-based measure(pre-trained word embedding for SE) [4] [5].
Proposed Approach

Frame-to-Frame Method

1. Train Word Embeddings for RE domain
2. Generate Frame Embeddings
3. Measure Semantic Relatedness
Proposed Approach

Step 1: Train Word Embedding

1.0 GB of app reviews (> 3m reviews)
Proposed Approach

Step 2: Generate Frame Embedding

- Retrieve all n LUs of Frame X along with their POS tags.
- Search to match every LU of Frame X in the training word embedding.
- Calculate the sum and obtain the average of the total word embedding pertaining to the LUs of Frame X.
Step 2: Generate Frame Embedding Cont.
Step 3: Measure Semantic Relatedness

\[ Ra = (F_{a1}, F_{a2}, ..., F_{an}) \text{ and } Rb = (F_{b1}, F_{b2}, ..., F_{bm}) \]

\[ M = [Ra, Rb] = 
\begin{bmatrix}
FR(F_{a1}, F_{b1}) & \cdots & FR(F_{an}, F_{b1}) \\
\vdots & \ddots & \vdots \\
FR(F_{a1}, F_{bm}) & \cdots & FR(F_{an}, F_{bm})
\end{bmatrix} \]

\[ \vec{M} = \begin{bmatrix}
(FR(F_{a1}, F_{b1}) + \cdots + FR(F_{an}, F_{b1}))/n \\
\vdots \\
(FR(F_{a1}, F_{bm}) + \cdots + FR(F_{an}, F_{bm}))/n
\end{bmatrix} = \begin{bmatrix}
\overrightarrow{FR_1} \\
\vdots \\
\overrightarrow{FR_n}
\end{bmatrix} \]

\[ \downarrow M = \begin{bmatrix}
(FR(F_{a1}, F_{b1}) + \cdots + FR(F_{an}, F_{b1}))/m \\
\vdots \\
(FR(F_{a1}, F_{bm}) + \cdots + FR(F_{an}, F_{bm}))/m
\end{bmatrix} = \begin{bmatrix}
\downarrow FR_1 \\
\vdots \\
\downarrow FR_m
\end{bmatrix} \]

\[ SR (Ra, Rb) = \cos(\vec{M}, \downarrow M) = \frac{\vec{M} \cdot \downarrow M}{||\vec{M}|| ||\downarrow M||} \]
Evaluation Plan

SEM-REQ manually labelled dataset

Baseline system based on Google Word2Vec Model
I: SEM-REQ Dataset

1. **Annotating 1770** requirements pairs from FN-RE corpus [2][3] by 3 annotators independently.
2. **Validating** the dataset
   1. with an average F-score of 77.5%
3. **Harmonisation** to produce final dataset of SEM-REQ.
II: Baseline System

- Using pre-trained word embeddings, i.e., Google’s Word2Vec model.
- Applying same procedure of measuring semantic relatedness (i.e., cosine metric and embeddings averaging)
Preparing the Results

• We compared the F2F method with the baseline system by applying each of them to SEM-REQ.
Results

F1 Scores Comparison

Performance (F1 Score) vs. Threshold (t)

- F2F Method
- Baseline System
### Examples

<table>
<thead>
<tr>
<th>ID_A</th>
<th>Sentence_A</th>
<th>ID_B</th>
<th>Sentence_B</th>
<th>F2F Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN-REQ-005-2</td>
<td>He <strong>ACCESS</strong>hsaving or lacking access the website, <strong>CREATE</strong>sCreating a profile and <strong>PROVIDE</strong>sSupply his educational professional and personal information</td>
<td>FN-REQ-007-2</td>
<td>On registration, they <strong>NEED</strong>Have as requirement to <strong>PROVIDE</strong>Supply name and address, payment details (credit card, etc), shoe sizes, gender, and any special details</td>
<td>0.528667032</td>
</tr>
<tr>
<td>FN-REQ-007-2</td>
<td>On registration, they <strong>NEED</strong>Have as requirement to <strong>PROVIDE</strong>Supply name and address, payment details (credit card, etc), shoe sizes, gender, and any special details</td>
<td>FN-REQ-022-3</td>
<td><strong>WHEN</strong>temporal collocation all items have been <strong>CHOSEN</strong>Choosing, the shopper <strong>PROVIDE</strong>supply a delivery address.</td>
<td>0.420046491</td>
</tr>
<tr>
<td>FN-REQ-030-5</td>
<td>John <strong>INDICATE</strong>sIndicating that he <strong>WISH</strong>sDesiring to <strong>WITHDRAW</strong>Removing $50 dollars.</td>
<td>FN-REQ-030-8</td>
<td>The ATM <strong>VERIFY</strong>sVerification that the amount may be <strong>WITHDRAW</strong>Removing from his account.</td>
<td>0.581840709</td>
</tr>
<tr>
<td>FN-REQ-015-9</td>
<td>After the Account Manager <strong>APPROVE</strong>sDeny or grant permission the purchase, an authorisation signature <strong>MAY</strong>Possibility be <strong>REQUIRE</strong>sHave as requirement.</td>
<td>FN-REQ-022-1</td>
<td>The Pizza Ordering <strong>SYSTEM</strong>Gizmo <strong>ALLOW</strong>sPreventing or letting the user of a web browser to <strong>ORDER</strong>Request entity pizza for home delivery.</td>
<td>0.400741491</td>
</tr>
<tr>
<td>FN-REQ-007-1</td>
<td>Customers will <strong>NEED</strong>Have as requirement to <strong>REGISTER</strong>Recording with the Odd Shoe Company to <strong>MAKE</strong>Intentionally create orders.</td>
<td>FN-REQ-015-9</td>
<td>After the Account Manager <strong>APPROVE</strong>sDeny or grant permission the purchase, an authorisation signature <strong>MAY</strong>Possibility be <strong>REQUIRE</strong>sHave as requirement.</td>
<td>0.509795616</td>
</tr>
</tbody>
</table>
Ongoing Work
References


Thank you

Waad Alhoshan
PhD Candidate, Software & Systems Research Group
School of Computer Science, University of Manchester
Email: Waad.alhoshan@postgrad.manchester.ac.uk