A new approach for summarizing SemRep predications

Vahid Taslimitehrani
Dr. Olivier Bodenreider
We want to propose an alternative/complementary summarization technique.
Background

- Semantic MEDLINE summarization system
  - Semantic MEDLINE (2008)
    - Relevance
    - Connectivity
    - Novelty
    - Saliency
  - Degree Centrality (2011)
  - Clustering Cliques (2013)
Motivation – an example

UMLS Metathesaurus

Selective beta-1 adrenoceptor stimulants

SemRep semantic predications

<table>
<thead>
<tr>
<th>IS_A</th>
<th>Dobutamine</th>
<th>TREATS</th>
<th>Congestive heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_A</td>
<td>Dopexamine</td>
<td>TREATS</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>IS_A</td>
<td>Dopexamine hydrochloride</td>
<td>TREATS</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>IS_A</td>
<td>Xamoterol</td>
<td>TREATS</td>
<td>Congestive heart failure</td>
</tr>
</tbody>
</table>

Inferred

Summarize

Selective beta-1 adrenoceptor stimulants TREATS Congestive heart failure
Motivation - an example

Based on the example, we observe:

1. 4 semantic predications are aggregated.
2. A new inference is made.

Objective:
Our technique leverages hierarchical relations from the UMLS Metathesaurus for aggregating the semantic predications and generating new inferences from the aggregated predications.
Methodology-an overview

Asserted Predication

Inferred Predication
Methodology

- Let’s discuss about the details of methodology using an example. We’re interested to summarize the semantic predications returned in response to the following question:

What are the medications used to TREAT Congestive heart failure (C0018802)?

- SemRep returns 6013 semantic predications.
- SemRep returns 684 unique treatment options.
Methodology

- **Step 1**: Retrieve unique semantic predications from SemRep when
  - The predicate is TREATS or any descendant.
  - The object is a disease or any descendant.

- **Step 2**: Extract semantic groups of each subject and remove those predications with procedures as semantic group (T061, ...)

- **Step 3**: Retrieve all parents of each medication returned from step 2.

Xamoterol is one of the medications from step 2 (Xamoterol TREATS Congestive heart failure)
Xamoterol has two parents:
  1. Selective beta-1 adrenoceptor stimulants (Xamoterol IS_A Selective beta-1 adrenoceptor stimulants)
  2. Sympathomimetics (Xamoterol IS_A Sympathomimetics)
Methodology

- **Step 4**: Retrieve all children of the parents returned from step 3.

**Selective beta-1 adrenoceptor stimulants** has 4 children:

1. Dobutamine (Selective beta-1 adrenoceptor stimulants INVERSE_ISA Dobutamine)
2. Dopexamine (Selective beta-1 adrenoceptor stimulants INVERSE_ISA Dopexamine)
3. Dopexamine hydrochloride (Selective beta-1 adrenoceptor stimulants INVERSE_ISA Dopexamine hydrochloride)
4. Xamoterol (Selective beta-1 adrenoceptor stimulants INVERSE_ISA Xamoterol)

**Sympathomimetics** has many children such as:

1. Adrenergic alpha-agonists
2. Dopamine
3. Ephedrine
4. Etilefrine
5. Xamoterol

- **Step 5**: Some of the parents are too generic such as C0993159 (Oral product). If an ancestor has too many descendants, then it is not used.
Methodology

- **Step 6**: For each child of a parent returned from step 5, we need to verify the child TREATS the disease or not. If all children TREAT the disease, we aggregate semantic predcitions and make a new inference.

<table>
<thead>
<tr>
<th>Selective beta-1 adrenoceptor stimulants</th>
<th>Sympathomimetics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dobutamine TREATS CHF</td>
<td>Adrenergic TREATS CHF</td>
</tr>
<tr>
<td>Dopexamine TREATS CHF</td>
<td>Dopamine TREATS CHF</td>
</tr>
<tr>
<td>Dopexamine hydrochloride TREATS CHF</td>
<td>Ephedrine TREATS CHF</td>
</tr>
<tr>
<td>Xamoterol TREATS CHF</td>
<td>Xamoterol TREATS CHF</td>
</tr>
</tbody>
</table>

- Aggregate

- Do not aggregate

**Selective beta-1 adrenoceptor stimulants TREATS Congestive heart failure**

- Adrenergic TREATS CHF ✔
- Dopamine TREATS CHF ✔
- Ephedrine TREATS CHF ✗
- Xamoterol TREATS CHF ✔

- Do not aggregate
Methodology

**Step 7:** If we can aggregate semantic predications in step 6, we continue to aggregate into the higher levels. If not, it is the highest level of summarization.

- **Selective beta-1 adrenocceptor stimulants** has just one parent:
  1. Adrenergic beta agonist

- and **adrenergic beta agonist** has 6 children:
  1. Selective beta-1 adrenocceptor stimulants
  2. Selective beta-2 adrenocceptor stimulants
  3. Dobutamine
  4. Dobutamine hydrochloride
  5. Isoproterenol
  6. Mirabegon
Implementation

- We used Biomedical Knowledge Repository (BKR) to implement our technique.
  - UMLS in RDF
  - SemRep predcations in RDF
- Created at NLM by Dr. Olivier Bodenreider and Dr. Thomas Rindflesch.
- We used 2013 version that includes more than 27 million predications from 13 million articles by SemRep.
- Technologies
  - Semantic Web
    - RDF, SPARQL standards
    - Virtuoso triple store
  - Programming
    - Java
We defined 4 quantitative measures to evaluate the performance of our technique.

- **Summarization rate:**
  \[
  1 - \frac{\text{# semantic predications after}}{\text{# semantic predications before}} = 1 - \frac{300}{400} = 0.25
  \]

- **Inference ratio:**
  \[
  \frac{\text{# predications can be aggregated}}{\text{# inferences}} = \frac{130}{30} \approx 4.3
  \]

- Number of generated inferences
- Ratio of validated inferences
Experimental results

- We investigate two questions:

1. **What are the medications used to treat disease X?**
2. **What are the medications caused disease X? (adverse drug events)**

- For each question, we select five diseases with
  - high numbers of predications (more than 400 unique predications)
  - medium numbers of predications (between 100 and 400 predications)
### Experimental results-question #1

<table>
<thead>
<tr>
<th>Disease</th>
<th># pred.</th>
<th>Summarization rate</th>
<th># generated inferences</th>
<th>Ratio of validated inferences</th>
<th>Inference ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertensive disorder</td>
<td>1122</td>
<td>26%</td>
<td>63</td>
<td>38%</td>
<td>5.6</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>499</td>
<td>29%</td>
<td>24</td>
<td>21%</td>
<td>7</td>
</tr>
<tr>
<td>Depression</td>
<td>400</td>
<td>27%</td>
<td>39</td>
<td>23%</td>
<td>3.7</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>419</td>
<td>29%</td>
<td>24</td>
<td>16%</td>
<td>6</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>401</td>
<td>30%</td>
<td>26</td>
<td>38%</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Diseases with high numbers of semantic predications (more than 400)
### Experimental results-question #1

<table>
<thead>
<tr>
<th>Disease</th>
<th># pred.</th>
<th>Summarization rate</th>
<th># generated inferences</th>
<th>Ratio of validated inferences</th>
<th>Inference ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypercholesterolemia</td>
<td>240</td>
<td>21%</td>
<td>18</td>
<td>33%</td>
<td>3.8</td>
</tr>
<tr>
<td>Pruitus</td>
<td>179</td>
<td>47%</td>
<td>12</td>
<td>50%</td>
<td>8</td>
</tr>
<tr>
<td>Burn injury</td>
<td>316</td>
<td>45%</td>
<td>11</td>
<td>55%</td>
<td>13.9</td>
</tr>
<tr>
<td>Pseudomonas Infections</td>
<td>201</td>
<td>26%</td>
<td>23</td>
<td>43%</td>
<td>3.3</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>343</td>
<td>29%</td>
<td>21</td>
<td>62%</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Diseases with the medium numbers of semantic predications
## Experimental results-question #2

<table>
<thead>
<tr>
<th>Disease</th>
<th># pred.</th>
<th>Summarization rate</th>
<th># generated inferences</th>
<th>Ratio of validated inferences</th>
<th>Inference ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traumatic Injury</td>
<td>1045</td>
<td>40%</td>
<td>83</td>
<td>17%</td>
<td>6</td>
</tr>
<tr>
<td>Ischemia</td>
<td>622</td>
<td>44%</td>
<td>32</td>
<td>3%</td>
<td>9.5</td>
</tr>
<tr>
<td>Cerebrovascular accident</td>
<td>401</td>
<td>34.5%</td>
<td>16</td>
<td>6%</td>
<td>9.6</td>
</tr>
<tr>
<td>Obstruction</td>
<td>1815</td>
<td>37%</td>
<td>75</td>
<td>16%</td>
<td>10</td>
</tr>
<tr>
<td>Septicemia</td>
<td>748</td>
<td>41%</td>
<td>42</td>
<td>14%</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Diseases with high numbers of semantic predications (more than 400)
# Experimental results-question #2

<table>
<thead>
<tr>
<th>Disease</th>
<th># pred.</th>
<th>Summarization rate</th>
<th># generated inferences</th>
<th>Ratio of validated inferences</th>
<th>Inference ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wounds &amp; injuries</td>
<td>139</td>
<td>44%</td>
<td>6</td>
<td>50%</td>
<td>8.4</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>170</td>
<td>37%</td>
<td>8</td>
<td>25%</td>
<td>6.7</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>152</td>
<td>35%</td>
<td>7</td>
<td>28%</td>
<td>6.6</td>
</tr>
<tr>
<td>Asthma</td>
<td>192</td>
<td>41%</td>
<td>11</td>
<td>36%</td>
<td>11.3</td>
</tr>
<tr>
<td>Gastroesophageal reflux disease</td>
<td>101</td>
<td>32%</td>
<td>5</td>
<td>60%</td>
<td>14.8</td>
</tr>
</tbody>
</table>

Diseases with the medium numbers of semantic predications
Limitations

- We need to validate the rest of new inferences by the experts.
- In the current experiments, we’re using SNOMED_CT and MEDCIN hierarchy and IS_A relations are used to explore the hierarchy. The code has the ability to generalize to any other hierarchies.
Conclusion

- We propose a new technique to summarize SemRep predications.
- Our technique is based on aggregating semantic predications and in the same time making new inferences from the aggregated inferences.
- We also designed four measures to evaluate the performance of our technique.
- Preliminary experimental results are promising.
- We can use this technique as complementary to the semantic MEDLINE summarization.
Acknowledgment

I would like to thank

- Dr. Olivier Bodenreider
- Dr. Paul Fontelo
- Dr. Marcelo Fizman
- Dr. Thomas Rindflesch
- Dr. McDonald
- Lister Hill Center