A New Approach for Summarizing SemRep Predications

Vahid Taslimitehrani *

Computer Science and Engineering Department, Kno.e.sis Center, Wright State University, Dayton, OH, 45324, USA

**E-mail: vahid@knoesis.org*

Olivier Bodenreider

Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, 20864, USA

E-mail: olivier@nlm.nih.gov

Semantic MEDLINE applies automatic summarization techniques to manage the semantic predications extracted from the biomedical literature by SemRep. It does so by selecting salient predications based on several criteria. In this study, we investigated a new technique to automatically summarize SemRep predications. Our technique leverages hierarchical relations from the UMLS Metathesaurus for aggregating the semantic predications. We also generated new inferences from the aggregated semantic predications. Several quantitative measures are defined to evaluate the system. We applied our method to summarize medications used to treat diseases and also adverse drug events reported in the biomedical literature. Our preliminary experimental results are promising in terms of summarization rate. They also indicate that less than half of the newly generated inferences correspond to existing relations. Further work is needed to evaluate the rest of the inferences.

1. Introduction and background

In the biomedical domain, PubMed provides access to million citations from some 6000 journals in the MEDLINE database. Complex biomedical information retrieval systems are necessary to help the user exploit this huge amount of data. Figure 1 represents parts of biomedical information retrieval system developed at National Library of Medicine (NLM). In the first part, SemRep,[1] a natural language processing application extracts semantic predications from the biomedical literature. For example, "`Xamoterol TREATS Congestive Heart Failure`" is a semantic predication extracted from a PubMed article by SemRep. Each semantic predication is also grounded by Unified Medical Language System (UMLS)[2] in SemRep. It means "Xamoterol" and "Congestive Heart Failure" are both Metathesaurus concepts, and the predicate, TREATS, is from the Semantic Network.

However, the number of citations returned by SemRep is usually huge and hard to manage by the users. For example, if you are looking for all medications used to treat congestive heart failure, SemRep returns more than 6000 semantic predications. To summarize the list of semantic predications, Fiszman et al. devised an abstraction summarization system[3] called Semantic MEDLINE for semantic predications from SemRep. Semantic MEDLINE takes a list of semantic predications from SemRep and returns a summarized list of semantic predications. Semantic MEDLINE summarizes the list of semantic predications based on the following four criteria:

(1) **Relevance**: Include predications on the topic of the summary that conform to the selected

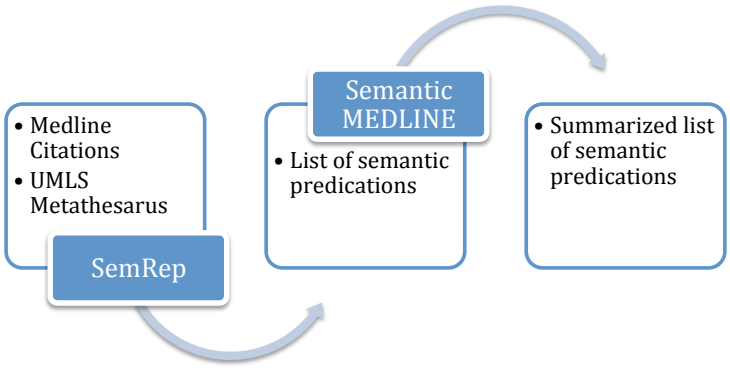

Fig. 1. A biomedical information retrieval system

summarization perspective.

(2) **Connectivity**: Include additional useful predications on the basis of having shared arguments with the relevant predications.
(3) **Novelty**: Eliminate, using UMLS hierarchical information, the predications the user already knows, identified as those having generic arguments, such as Pharmaceutical Preparations or Disease.
(4) **Saliency**: Eliminate predications with low frequency of occurrence.

In another word, Semantic MEDLINE removes some of the semantic predications in each step described above.

In this study, we propose a novel technique to summarize SemRep predications. The main idea of our technique is **aggregating semantic predications and making new inferences** from the aggregated predications. In another word, we group semantic predications together and make a new semantic predication. The new inferred semantic predication represents those aggregated predications.

The main difference between our method and Semantic MEDLINE is that Semantic MEDLINE removes semantic predications in order to summarize the list, but our technique aggregates semantic predications for the same purpose. We believe, our technique can be used in two places. We can use our technique as an alternative of Semantic MEDLINE or as complementary of Semantic MEDLINE. In the first scenario, our system takes a list of predications from SemRep and returns a summarized list and in the second scenario our system takes a summarized list of predications from Semantic MEDLINE and returns another summarized list.

1.1. *Example*

We used a simple example to motivate the reader and explain our summarization technique. Figure 2 represents a schematic view of our summarization technique. In this example, we summarize a short list of medications used to treat congestive heart failure including Dobutamine, Dopexamine, Dopexamine hydrochloride and Xamoterol. Each of these four medications is also UMLS Metathesaurus concepts. If you explore UMLS Metathesaurus, there

is a common characteristic between these medications. They have a similar parent in the UMLS Metathesaurus and the parent is "Selective beta-1 adrenoceptor stimulants". We used "IS_A" and "Inverse_ISA" relations in order to find the parents and children in the UMLS Metathesau

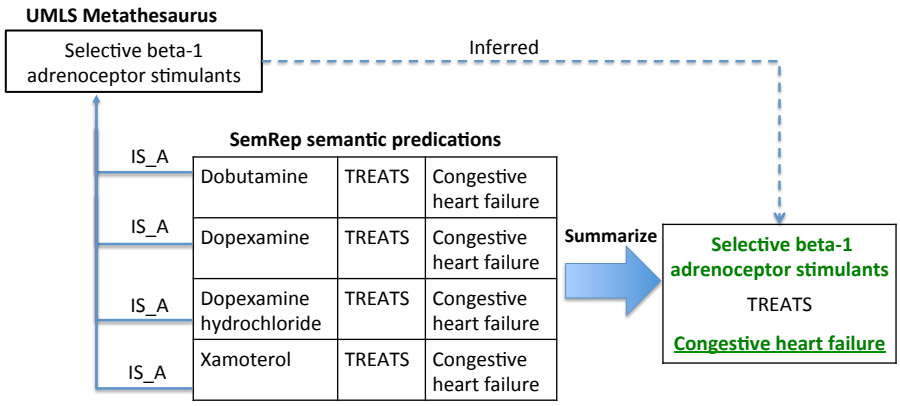

Fig. 2. A simple example of aggregating semantic predications

Since they have the same parent in the UMLS hierarchy, and they all treats congestive heart failure, we can easily aggregate them and make the following inference:

```
Selective beta-1 adrenoceptor stimulants, TREATS, Congestive Heart Failure
```

In this example, we showed four semantic predications are aggregated and a new inference is also generated. Our objective in this study is **leveraging hierarchical relations from the UMLS Metathesaurus for aggregating the semantic predications and generating new inferences** from the aggregated predications.

2. Methodology

3. Overview

In this section, we first give an overview of our summarization technique and then we will explain the steps of our technique. Figure 3 represent an schematic overview of the methodology.

Left side of the figure 3 is similar to the example investigated in the introduction. Let's discuss about a case that we are not allowed to aggregate. Looking at the right side of the figure 3, we found three medications *Med1*, *Med2*, and *Med3* have the same parent in the UMLS hierarchy and the parent is *Med*. However, in the UMLS hierarchy, *Med* has another child named *Med4* and there is no evidence from SemRep confirming *Med4* treats disease. It

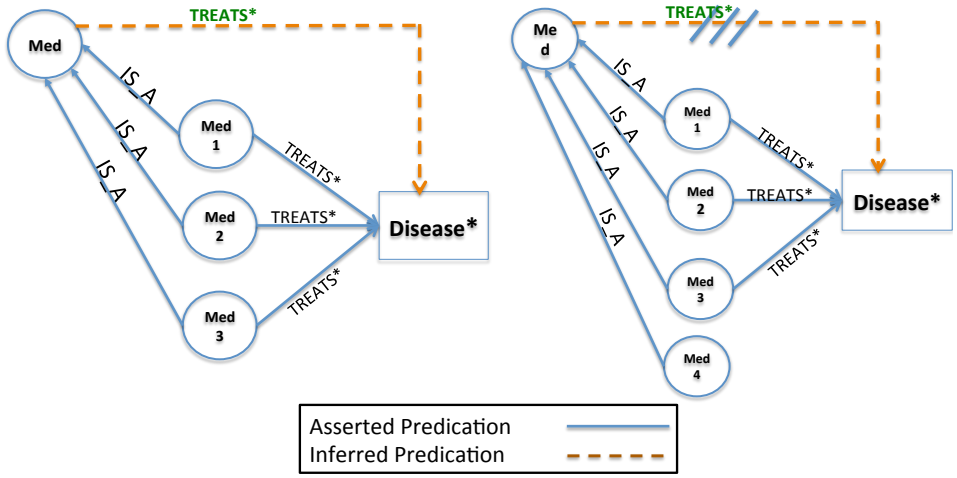

Fig. 3. Schematic overview of methodology

means there is no semantic predication in SemRep saying (`Med4`, `TREATS`, `disease`). In this case, we are not allowed to aggregate semantic predications and we leave them as they are.

4. Algorithm

In this section, we explain different steps of our proposed algorithm. To facilitate understanding the process, we use a simple example after explaining each step. Assuming the questions is "What are the medications used to treat congestive heart failure?". SemRep extract 6013 semantic predications from the biomedical literature with *TREATS* as predicate and *Congestive Heart Failure* as object. We are going to apply the following steps to summarizing the list of semantic predications.

Step 1: In the first step, we retrieve a list of unique semantic predications from SemRep when

- The predicate is `TREATS` or any descendant.
- The object is a disease or any descendant.

For example, the object can be congestive heart failure or any descendant like chronic congestive heart failure. Since a medical evidence can be repeated in many articles, we just keep a unique list of semantic predications. After removing duplicated predications from our list, 684 unique semantic predications are remained on the list.

Step 2: Some of the treatment options are procedures. Since we are summarizing the list of medications, we used Semantic Groups to recognize procedures and remove them from the list of semantic predications. In this step, 184 predications are removed from the list.

Step 3: In the third step, we use UMLS Metathesaurus to retrieve all parents of each medication returned from step 2. For example, Xamoterol is one of the medications used to treat congestive heart failure. Xamoterol has two parents in the UMLS hierarchy:

- Selective beta-1 adrenoceptor stimulants (`Xamoterol`, `IS_A`, `Selective beta-1`

adrenoceptor stimulants)
- Sympathomimetics (Xamoterol, IS_A, Sympathomimetics)

Step 4: In this step, we use UMLS Metathesaurus again to retrieve all children of the parents returned from the previous step. As we said earlier, Xamoterol has two parents, and each one has the following children:

- Selective beta-1 adrenoceptor stimulants
  (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
  (1) Adrenergic alpha-agonists
  (2) Dopamine
  (3) Ephedrine
  (4) Xamoterol
  (5) ....

Sympathomimetics has many more children but we just mention some of them here.
Step 5: Some of the parents of the medications are too general, and they have too many children. For example, there is a concept in UMLS named "Oral Tablet" and it has more than 500 children. We used a predefined threshold on the number of children to recognize these concepts. If the number of children exceeds the threshold, the concept is not used.
Step 6: In this step, for each child of a parent returned from step 5, we need to verify the child TREATS the disease or not. If all children TREATS the disease, we are allowed to aggregate semantic predications and make a new inference.

As we said earlier, selective beta-1 adrenoceptor stimulants has four children, and they are medications used to treat congestive heart failure. In other words, the following semantic predications are in SemRep:

(1) Dobutamine, TREATS, Congestive Heart Failure
(2) Dopexamine, TREATS, Congestive Heart Failure
(3) Dopexamine hydrochloride, TREATS, Congestive Heart Failure
(4) Xamoterol, TREATS, Congestive Heart Failure

Since all children of selective beta-1 adrenoceptor stimulants TREATS congestive heart failure, we are allowed to aggregate semantic predications and generate the following inference:

Selective beta-1 adrenoceptor stimulants, TREATS, Congestive Heart Failure

Since we cannot verify all children of sympathomimetics treats congestive heart failure in SemRep, then we are not allowed to aggregate, and we leave those predications as they are.
Step 6: In the last step of our algorithm, if there is no success in the aggregation in the previous step, we stop the process and it is the highest level of aggregation. If we aggregated predications in the previous step, we have to add the generated inference to the list of predications, return to step 3, and repeat the process again for further aggregation.

5. Implementation

We used Biomedical Knowledge Repository (BKR) to implement our technique.[4] The BKR contains relations extracted from PubMed articles by SemRep and normalizes entities to concepts in the UMLS. It includes approximately 27 million semantic predications extracted from 13 million articles in PubMed. These semantic predications are transformed into RDF format together with the provenance information about the article where the predication is extracted.

We also used semantic web technologies to implement the system. Semantic predications are stored in the Virtuoso triple store, and SPARQL queries are used to extract semantic predications. Java programming language is also used to develop the prototype.

6. Results and Discussion

In this section, we report a systematic evaluation of our proposed summarization system. We investigate the following two questions:

- What are the medications used to treat disease X?
- What are the medications caused disease X? (adverse drug events)

To evaluate our system, we first need to choose a list of diseases. We designed two scenarios: In the first scenario, we select five diseases among the diseases with high number of semantic predications (more than 400 unique predications) and in the second scenario, we select five diseases among the diseases with the medium number of semantic predications (between 100 and 400 unique predications).

We also define a set of quantitative measures to evaluate the performance of our method. Our proposed measures are the following:

(1) **Summarization rate**: It measures the percentage of reduced predications after applying our technique. We have to mention again that we do not remove any predication in our system, and we just aggregate them and then substitute with the new generated inferences.

$$Summarization\ Rate = 1 - \frac{\#\ of\ semantic\ predications\ after}{\#\ of\ semantic\ predications\ before} \tag{1}$$

(2) **Inference ratio**: It measures the average number of semantic predications grouped together.

$$Inference\ Ratio = 1 - \frac{\#\ of\ predications\ can\ be\ aggregated}{\#\ of\ inferences} \tag{2}$$

(3) **Number of generated inferences**

(4) **Ratio of validated inferences**: A new inference is validated, if it is already extracted by the SemRep. When a new inference is validated, it means someone talked about it in another PubMed article.

Table 1 represents the performance of our method for the first question on five diseases with the high number of semantic predications. For example, if a user queries about the medications used to treat hypertensive disorder, our technique aggregates 26% of the predications and at the same time generates 63 new inferences. 24 new generated inferences are already validated and confirmed to be correct by the literature.

Table 1. Question # 1 - Performance of our summarization technique on 5 diseases with the high number of semantic predications

Disease	**Number of predications**	**Summarization rate**	**# generated inferences**	**Rate of validated inferences**	**Inference ratio**
Hypertensive disorder	1122	26%	63	38%	5.6
Congestive heart failure	499	29%	24	21%	7
Depression	400	27%	39	23%	3.7
Myocardial infarction	419	29%	24	16%	6
Schizophrenia	401	30%	26	38%	5.6

Table 2 represents the performance of our method for the first question on five diseases with the medium number of semantic predications. Summarization rate does not change too much comparing to Table 1 but the number of inferences decreased a bit.

Table 2. Question # 1 - Performance of our summarization technique on 5 diseases with the medium number of semantic predications

Disease	**Number of rate**	**Summarization inferences**	**# generated predications**	**Rate of validated inferences**	**Inference ratio**
Hypercholesterolemia	240	21%	18	33%	3.8
Pruritus	179	47%	12	50%	8
Burn injury	316	45%	11	55%	13.9
Pseudomonas infection	201	26%	23	43%	3.3
Glaucoma	343	29%	21	62%	5.7

In the second part of our experiments, we worked on the second question and tried to summarize the list of semantic predications when the predicate is `CAUSES`. Table 3 represents the performance of our method on five diseases with the high number of semantic predications. One important observation is that the rate of validated inferences dropped dramatically from the first question. One explanation is that physicians usually do not report adverse events about the general medications.

There are some observations from Tables 1 to 4. First, summarization rate is almost stable on different experiments. It shows our approach is not just applicable to diseases with the high number of semantic predications and works well on diseases with the medium number of semantic predications. Second, the number of generated inferences has a linear relation with the number of predications. If the number of semantic predications drops, the chance of aggregating and making new inferences will decrease.

Table 3. Question # 2 - Performance of our summarization technique on 5 diseases with the high number of semantic predications

Disease	Number of rate	Summarization inferences	# generated predications	Rate of validated inferences	Inference ratio
Traumatic Injury	1045	40%	83	17%	6
Ischemia	622	44%	32	3%	9.5
Cerebrovascular accident	401	34.5%	16	6%	9.6
Obstruction	1815	37%	75	16%	10
Septicemia	748	41%	42	14%	11.5

Table 4. Question # 2 - Performance of our summarization technique on 5 diseases with the medium number of semantic predications

Disease	Number of rate	Summarization inferences	# generated predications	Rate of validated inferences	Inference ratio
Wounds & injuries	139	44%	6	50%	8.4
Cardiovascular disease	170	37%	8	25%	6.7
Pulmonary embolism	152	35%	7	28%	6.6
Asthma	192	41%	11	36%	11.3
Gastroesophageal reflux disease	101	32%	5	60%	14.8

7. Conclusion

In this study, we propose a new summarization system for SemRep. Our technique is based on aggregating semantic predications using UMLS Metathesaurus and generating new inferences from the aggregated predications. We believe our system can be complementary to Semantic MEDLINE. Our preliminary results are promising.

8. Acknowledgments

This research was supported in part by an appointment to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

1. T. Rindflesch, M. Fiszman, B. Libbus. Semantic interpretation for the biomedical research literature. *Medical informatics*, 399-422, 2005.
2. O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32: 267D–270, 2005.
3. M. Fiszman, T. Rindflesch, H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. *Proceedings of the HLT-NAACL workshop on computational lexical semantics. Association for Computational Linguistics*, 76-83, 2004.
4. O. Bodenreider, T. Rindflesch. Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications. *Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland*, 2006.