Finding the Differences between the Perceptions of Experts and the Public in the Field of Diabetes

Dahee Lee, Won Chul Kim, Min Song
Department of Library and Information Science, Yonsei University
Seoul, Korea
{leedahee, krevas, min.song}@yonsei.ac.kr

ABSTRACT
Automatic information extraction techniques such as named entity recognition and relation extraction have been developed but it is yet rare to apply them to various document types. In this paper, we applied them to academic literature and social media's contents in the field of diabetes to find distinctions between the perceptions of biomedical experts and the public. We analyzed and compared the experts' and the public's networks constituted by the extracted entities and relations. The results confirmed that there are some differences in their views, i.e., biomedical entities that interest them and relations within their knowledge range.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services—Web-based services; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms
Experimentation, Measurement

Keywords
Diabetes; social media; named entity recognition; relation extraction; degree centrality; semantic relatedness

1. INTRODUCTION
As the vast volume of unstructured data has become freely available, a number of researchers have developed the methods or techniques to automatically handle such data. The examples include named entity recognition (NER) and relation extraction (RE). However, most of previous studies tried to achieve better performance of NER or RE, and relatively few ones made use of them for the purpose of content analysis.

In this study, we aim to observe how differently the experts and the general public possess biomedical perspectives in the field of diabetes. We applied NER and RE to documents written by them, namely academic articles and social networking sites' comments. We then comparatively analyzed the extracted biomedical entities and relations with two approaches. Specifically, the entity types addressed in our study were disease (DS), drug (DR), body/organ (BD), food (FD), and nutrient (NT) while the relation type was limited to biomedical verbs. Our study would eventually broaden the application scope of those automatic information extraction techniques and validate their usefulness and effectiveness.

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742773

2. METHODOLOGY
Our method consists of three stages: data collection, data processing, and data analysis, as shown in Figure 1.

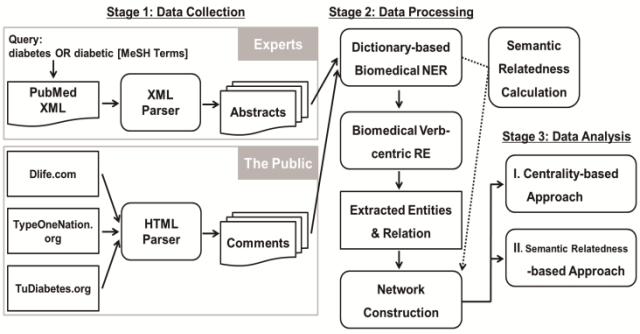

Figure 1: Three stages of the methodology.

In the first stage, we collected documents written by experts and the public. From three social networking sites for diabetes (DLife.com, TypeOneNation.org, TuDiabetes.org), we gathered all the comments and their upload time information from the sites by Jsoup, a HTML parser. A total of 246,334 comments uploaded from March, 2005 to December, 2014 were obtained, and only 245,655 among them had actual contents. For the experts' data, we retrieved 127,969 XML-formatted articles on the subject of diabetes from PubMed (http://www.ncbi.nlm.nih.gov/pubmed). The search query was "diabetes OR diabetic" in the Medical Subject Heading (MeSH) term field, and the publication year was restricted to the same period with the public's data. After we extracted abstracts from each article using a SAX parser, the final number of articles with abstracts was 107,418. We intentionally collected a bigger size of the public's documents since they were likely to contain less biomedical entities and their relations.

The next stage involves pair generation and network construction. Pair generation was carried out by the engine which extends Stanford CoreNLP [1] and is still under development for further performance improvement. It enables us to perform dictionary-based NER with abbreviation resolution, sentence splitting, and lemmatization, and also RE. We integrated the records of multiple dictionaries, ontologies or databases to acquire plentiful entity names for NER. All the synonyms are mapped into each representative entity name. For RE, we used the list of 398 biomedical verbs provided by [2] and the simplest feature to extract biomedical verb located between entities on dependency tree. We then established two networks, each for the experts and the public with the extracted entities and relations. We made the edge weight to stand for semantic relatedness value of two entities, which was calculated using S-space [3].

We comparatively analyzed two networks with two approaches at the final stage. Centrality-based approach is to detect core entities from each network on the basis of degree centrality [4],

and semantic-relatedness-based approach is to discover relations with a big perception gap among overlapping pairs using the weight difference. We assumed that extracted information on the networks reflects each community's view.

3. RESULTS AND DISCUSSIONS

3.1 Centrality-based Approach

Table 1 demonstrates top five entities based on degree centrality per entity type. About half (46%) of those mentioned by experts do not coincide with those by the public while the other half do, implying that they have different ranges of interest and knowledge to some extent. For example, experts have explored biomedical entities associated with type 2 diabetes (centrality of 0.2286) far more than type 1 diabetes (0.1254). The public however deals with type 1 diabetes (0.0734) as much as type 2 (0.0586). Among drug names, streptozotocin is ranked high on the experts' network (0.2101) because it is normally used in the lab to induce diabetes on experimental animals. Furthermore, the public frequently talks about fiber's effect on diabetes (0.1298), whereas experts substantially emphasize the importance of cholesterol control (0.1274).

Table 1: Top 5 entities by degree centrality per entity type.

	Disease	
Rank	Experts	The Public
1	diabetes	diabetes
2	type 2 diabetes	type 1 diabetes
3	type 1 diabetes	type 2 diabetes
4	leukocyte adhesion deficiency	tropical spastic paraparesis
5	gestational diabetes	cancer

	Body/Organ		Drug	
Rank	Experts	The Public	Experts	The Public
1	blood	blood	insulin	insulin
2	kidney	heart	streptozotocin	LSD*
3	liver	liver	LSD*	palmitic acid
4	heart	stomach	palmitic acid	metformin
5	muscle	skin	nitric oxide	GHB*

	Food		Nutrient	
Rank	Experts	The Public	Experts	The Public
1	water	water	glucose	glucose
2	fruit	cheese	protein	protein
3	salt	fruit	cholesterol	fiber
4	vegetable	milk	triglyceride	carbohydrate
5	bread	snack	fatty acid	sodium

*LSD and GHB refer to lysergic acid diethylamide and gamma-hydroxybutyric acid respectively.

3.2 Semantic-relatedness-based Approach

Table 2 shows top five entities that have not only biomedical relationships with diabetes but a large difference between the semantic relatedness scores (edge weights) from the experts' and the public's networks. It indicates a clear distinction on specific biomedical understanding between them. Overall, the public is less aware of how much those five related entities are semantically related with diabetes (lower weights) as well as how diversely their biomedical relations can be expressed in comparison with experts (smaller number of types). It can be inferred that the public generally doesn't recognize the strong association between palmitic acid (PA) and type 2 diabetes, even though it was already proved that PA acts as a reliable biomarker for type 2 diabetes [5]. In the case of metformin, the public only knows that it is likely to be prescribed for type 2 diabetics, whereas the experts so far have studied how it can reduce the risk of type 2 diabetes. Meanwhile, the semantic relatedness between glucose and diabetes is higher on the public's network as their relationship is widely known.

Table 2: Top 5 entities related to diabetes and their relations by semantic relatedness difference.

	Entity Pair			Relation	
Rank	Related Entity (Type)	Diabetic Entity (Type)	Weight Difference	Experts' Weight (Type)	The Public's Weight (Types)
1	cancer (DS)	diabetes (DS)	0.0595	0.1823 (increase, assemble and 32 others)	0.1228 (increase, diagnose and 13 others)
2	palmitic acid (DR)	type 2 diabetes (DS)	0.0496	0.0560 (increase, model and 49 others)	0.0064 (list, control and 9 others)
3	insulin (DR)	type 2 diabetes (DS)	0.0441	0.0562 (increase, control and 98 others)	0.0121 (control, slow and 20 others)
4	glucose (NT)	diabetes (DS)	0.0418	0.0178 (control, group and 60 others)	0.0596 (control, diagnose and 65 others)
5	met-formin (DR)	type 2 diabetes (DS)	0.0402	0.0544 (control, reduce and 33 others)	0.0142 (diagnose, control and 3 others)

4. CONCLUSION

We figured out the dissimilar perceptions of the experts and the public in the area of diabetes by applying NER and RE techniques to documents written by them and analyzing the results with two approaches. To summarize, they present a certain contrast in terms of biomedical entities of great interest and perceived relationships with regard to the corresponding types and degrees. We plan to adopt other centrality measures and analyze the existence of the time delay for knowledge transfer between two subject groups. The study contributes to the deeper understanding of the biomedical knowledge gap between experts and the public.

5. ACKNOWLEDGEMENTS

This work was supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

6. REFERENCES

[1] Manning, Christopher D., et al. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstration 2014*, 55-60.

[2] Sun, L., and Korhonen, A. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2:638-647.

[3] Jurgens, D., and Stevens, K. The S-Space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations*, 30-35.

[4] Wasserman, S., and Faust, K. Social network analysis: Methods and applications. Cambridge university press. 1994.

[5] Trombetta, Antonella, et al. Increase of Palmitic Acid Concentration Impairs Endothelial Progenitor Cell and Bone Marrow–Derived Progenitor Cell Bioavailability Role of the STAT5/PPARγ Transcriptional Complex. *Diabetes,* 62:1245-1257. 2013.