Machine Reading Approach to Understand Alzheimer's Disease Literature

Satoshi Tsutsui
School of Informatics and Computing
Indiana University,
Bloomington, IN, USA
stsutsui@indiana.edu

Ying Ding
School of Informatics and Computing
Indiana University,
Bloomington, IN, USA
dingying@indiana.edu

Guilin Meng
Department of Neurology,
Tenth People's Hospital
Tongji University, Shanghai, China
15216758879@163.com

ABSTRACT

Alzheimer's disease (AD) is an irremediable disease that has a high social impact and requires further research. However, the current amount of literature is already overwhelming. This paper shows a case study that applies a machine reading approach to understand the overwhelming amount of papers without much prior knowledge on AD. Machine reading enables us to understand the vast amount of AD literature both at an overview and specific level.

CCS Concepts

•**Information systems** → *Data mining;*

Keywords

Alzheimer's disease, Machine Reading, Bibliometrics

1. INTRODUCTION

Alzheimer disease (AD) is a neurodegenerative disorder that causes dementia. While it has a high social impact with many patients and high costs for medical care, AD still remains to be incurable [1]. Therefore, it is necessary to promote further research. However, the current information available is actually overwhelming. For example, searching Alzheimer in PubMed leads to 80,000 articles. Because no single researcher can read all the papers, using computational techniques to read these articles could be a solution. Towards this end, this paper presents a case study that uses machine reading to help us understand the overwhelming amount of AD literature assuming minimal prior knowledge on AD.

The computational analysis of publication data is often called bibliometrics. However, this methodology usually aims to obtain a big picture of a domain assuming prior knowledge, and does not aim to understand the textual content of papers without domain knowledge. For example, bibliometrics can show which topics are popular in AD, but does not show diseases that have a similar symptom to AD. In order to claim that we understand the literature, we need to obtain concrete information as well as an overview of the field.

This paper proposes a machine reading approach rather than using traditional methods from bibliometrics. Machine reading is formally defined as "*the idea that machines could become intelligent, and could usefully integrate and summarize information for humans, by reading and understanding the vast quantities of text that are available*" [5]. In this sense, bibliometrics can be a part of machine reading because it does summarize information and provides overviews for humans, but only overviews and not enough details to understand the contents of papers deeply.

Indeed, in order to understand AD literature without domain expertise, we first need an overview before diving into the content. In other words, we need an overview to determine what kind of specific information is interesting. For instance, unless we are aware of a group that discusses AIDS within AD papers, we cannot be interested in how AIDS is related to AD (this is the actual example that will be addressed later). Therefore, we first use statistical techniques, Latent Dirichlet Allocation (LDA) [2] in particular, to obtain the overview of AD.

Once we know the overview, we can be interested in specific questions (e.g. how AD is connected to AIDS). Answering these questions requires techniques to read the content of related papers both at an entity and and at a relation level. Extracting entities and relations from texts is called information extraction (IE). However, conventional IE assumes predefined target information. For example, extracting gene-disease interaction is a common IE task, but that means we already know the target genes, diseases, and expected interactions between them. However, our goal is to understand literature with minimal prior knowledge, so we do not have a specific target in advance. Moreover, even after we obtained a target, it can be changed frequently depending on how we have understood the literature so far. For example, we might be interested in which pathway contains a particular gene after gene-disease extractor found that the gene is associated with AD. In this case, we need to build the pathway-gene extractor again. Therefore, we resort to an IE technique that does not require pre-defined targets, which is called open information extraction (Open IE).

Open IE extracts facts in a form of triples <*subject, predicate, object*>. For example, it extracts <Alzheimer, is strongly correlated with, the apoe genotype> given a sentence "Alzheimer

CIKM Indianapolis USA

DOI: 10.475/123_4

is strongly correlated with the apoe genotype." We use this to answer specific questions based on LDA results. Note that we can always go back to specific sentences mentioning the answer. This makes it possible to build a paper search engine based on triples. For example, we can find a paper that discusses the cause of AD, not just containing the keywords of *cause* and *AD* without context. We made the demo search system available [1] as well as data [2].

Our combination of two methods (LDA and Open IE) is complementary. LDA has been applied to a wide range of fields to reveal hidden topics from textual data. However, it is often difficult to make sense of each topic even with domain knowledge. This is because LDA just outputs topics as a distribution over words, and does not provide how terms in a topic are linked together. Meanwhile, Open IE can tell how terms are specifically linked in sentences. It is developed as a natural language processing (NLP) technique, and is applied to NLP tasks such as question answering. However, when it is used to understand a large amount of literature without prior knowledge, asking specific questions is already difficult. Therefore, it is a complementary combination to use LDA and Open IE. LDA provides the overview of AD, LDA results are used to infer specific questions, and these questions are answered by Open IE.

The contribution of this paper is summarized in three folds. First, we present a case study to apply machine reading framework to understand overwhelming amount of AD papers without assuming much prior knowledge. Second, we propose a way to use two distinct methods (LDA and OpenIE) in a mutually complementary way. Third, we made a paper search system available so that AD researchers can enjoy the idea of machine reading.

2. RELATED WORK

2.1 Bibliometrics on AD

Several previous works [11, 12, 6, 7, 9] applied bibliometrics to the field of AD using statistical techniques. Most study accompany content analysis to interpret the results, which is performed manually, meaning that they assume prior knowledge on AD. [11] investigates the productivity and impact of the top 100 AD researchers. It also identifies the role of AD within the fields of neurodegenerative disease. [12] focuses on co-author network to revel major author communities. [3] particularly focuses on cholinesterase inhibitors within AD research, and analyzes the trend of the research. [6] analyzed collaboration network using papers from Alzheimer Disease Center in order to reveal the impact of the center. [7] analyzes network of medical entities to provide the overview of the AD research. However, it does not investigate specific relations between entities. [9] investigates AD literature both at the topic level and entity with relation level. To our knowledge, this is the only study that considers relations between entities. However, they only use 54 predefined relations defined in Unified Medical Language System [3].

2.2 Information Extraction

Many IE systems are developed to extract entities such as genes, diseases, drugs, and interactions between them (e.g. [10]). However, because they require predefined relations, they are not effective when when we do not have relations that we want to extract in advance. Open IE systems (see [8] for comprehensive summary) overcome this issue and use raw textual phrases as relations. Open IE has been applied to several NLP tasks (e.g. general question answering). However, to our knowledge, this paper is the first study to use Open IE to understand large amounts of literature in a specific medical domain.

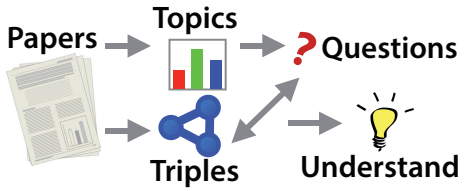

Figure 1: Conceptual sketch of our framework

3. GENERAL FRAMEWORK

Our approach discussed in Introduction is summarized in Figure 1. First, collect literature with a focus. The focus is specified by key terms, specific periods, and/or target journals. Second, apply LDA to keywords. The keywords can be author-provided, indexer-annotated (e.g. mesh terms), or automatically extracted. Note that the purpose is to obtain an overview, so the number of topics should be small. Next, apply Open IE to textual content (e.g. abstracts or full text) of the papers. Each extracted triple is linked to a specific sentence in a paper. Finaly, examine LDA generated topics to find interesting questions. Try to answer these questions by Open IE triples. This part can be iterative because an answer to a question could raise new questions.

4. CASE STUDY ON AD LITERATURE

This section reports how we actually applied the framework above into the domain of Alzheimer's Disease. Note that the purpose of this paper is not to propose a new method with quantitative evaluation. Rather, we aim to show a case study to use machine reading to understand AD literature without prior knowledge, and discuss both merits and limitations.

4.1 Text Processing Approach

First, we collected a set of PubMed papers relevant to AD[4]. We used search terms of *Alzheimer, Mild cognitive impairment, Dementia, Significant memory concern*, and *Subjective memory complaint.* These terms were determined by consulting a domain expert. This is the only part we assumed the domain knowledge. We collected 160,091 papers with title, author, journal, index terms, and abstracts.

Second, we applied LDA to the collected papers with MeSH terms, which are the index terms given to the PubMed articles. In addition, we applied LDA to subsets of the papers with a narrower focus by year range, hoping to obtain more interesting questions than by just looking at all papers. Specifically, we made two subsets of papers published in the last decade (2005 - 2014) and the second last decade (1995 - 2004) In LDA, we fixed the number of topics into five. We

[1]http://homes.soic.indiana.edu/stsutsui/machine_reading
[2]http://homes.soic.indiana.edu/stsutsui/machine_reading/data
[3]https://www.nlm.nih.gov/research/umls/
[4]The querying was performed 2015 October

Table 1: LDA Results (Topics are ranked by the popularity)

year	topic1	topic2	topic3	topic4	topic 5
all 1945 - 2015	**Huntington Disease**	Mental Disorders	tau Proteins	Aging	Creutzfeldt-Jakob Syndrome
	Parkinson Disease	Caregivers	Amyloid beta-Protein Precursor	Cognition	**Apolipoproteins E**
	Neurons	Dementia, Vascular	Neurodegenerative Diseases	Cholinesterase Inhibitors	Magnetic Resonance Imaging
	Cerebral Cortex	**AIDS** Dementia Complex	Brain Diseases	Memory Disorders	Nursing Homes
	Nerve Tissue Proteins	Schizophrenia	Amyloid	Neuropsychological Tests	Genetic Predisposition to Disease
1995 - 2004	Creutzfeldt-Jakob Syndrome	Amyloid beta-Protein Precursor	Parkinson Disease	Cholinesterase Inhibitors	**AIDS** Dementia Complex
	Apolipoproteins E	Neurons	Caregivers	Dementia, Vascular	Aging
	Huntington Disease	Membrane Proteins	Neuropsychological Tests	Memory Disorders	Peptide Fragments
	tau Proteins	Nerve Tissue Proteins	Cognition	Nootropic Agents	**HIV**-1
	Magnetic Resonance Imaging	Neurodegenerative Diseases	Memory	Schizophrenia	**HIV** Infections
2005 - 2014	Neurons	Aging	Cognition	Parkinson Disease	Cholinesterase Inhibitors
	Peptide Fragments	Neuropsychological Tests	Neurodegenerative Diseases	Caregivers	Neuroprotective Agents
	tau Proteins	Magnetic Resonance Imaging	Mental Disorders	Amyloid	Dementia, Vascular
	Amyloid beta-Protein Precursor	Memory Disorders	Nursing Homes	Creutzfeldt-Jakob Syndrome	Frontotemporal Dementia
	Huntington Disease	Memory	Amyotrophic Lateral Sclerosis	Depression	Amyloid Precursor Protein Secretases

filtered out terms that appeared in more than 20% of the papers and terms that is only used in less than five papers. This resulted in 6528 unique terms. The results of LDA is shown in Table 1. The topics are ranked by popularity.

Third, we applied Open IE, specifically Reverb [4], to all the sentences in the abstracts. We obtained 1,469,008 triples, and organized them in a relational database so that we can trace back to a specific sentence in a paper.

Lastly, we manually examined the LDA results and obtained some questions based on the results. Then, we tried to answer these questions using extracted triples. The details with real questions are presented in the next section. Note that these questions are just examples and our approach is capable of answering more. We also discuss some limitations.

4.2 Answerable Questions

The first term we focused on is *Huntington Disease* (HD) because it consistently appears in the first topic in the whole corpus and subsets (recall that the topics are ranked). The immediate question is "What is Huntington Disease ?" This question can be answered by searching triples with a pattern <Huntington Disease, is , ?x > where ?x means some words. We found 446 distinct triples. The top two frequent answers and other three randomly sampled results are shown in Table 2. The number inside the parenthesis represents the number of the triples that matched the pattern. Now we answer that HD is *an inherited neurodegenerative disorder*, which is similar to AD.

Table 2: What is Huntington Disease?

an inherited neurodegenerative disorder (44)
a neurodegenerative disorder (28)
a hereditary brain disease (2)
an incurable genetic neurodegenerative disorder (1)
a complex , single-gene (1) (wrong extraction)

We can also see limitations of our approach such as a wrong answer of *complex, single gene*. Extraction from text is not always correct so we need to exclude the noise. One possible approach is to use only high frequent triples. However, we observed that low frequent triples also carry interesting and important facts. For example, *an incurable genetic neurodegenerative disorder* is mentioned only once in the collection of papers we retrieved, but this answer carries important fact that HD is incurable. Therefore, we decided not to rely on the frequency. Later in this paper, we omit the number of times mentioned. Another issue is that Open IE does not provide normalizations, so similar terms are treated distinctly. This is a challenge when we search triples with a pattern. For example, we need to add *HD* as a synonym to Huntington's disease. These synonyms can also be obtained by querying extracted triples. In this example, we manually checked other subjects described as *an inherited neurodegenerative disorder*, and added other synonyms such as *huntington's disease*. Although providing evaluation on cleaning effort is not the focus of this paper, we note that it typically takes five to ten minutes to remove noise and add synonyms per question. We omit this cleaning process in the rest of the paper due to space limitations.

Now we know that HD is also a neurodegenerative disease like AD, the next natural question is, "How AD is related to HD ?." We first queried <AD, ?x , HD >, but could not find interesting relations. Then we extended the pattern to <AD, ?r1 , ?x> & <?x, ?r2, HD>. This is equivalent to finding 2-step paths from node AD to node HD on a directed graph. It is also possible to find a path the other way from HD to AD, which gives similar results. The query resulted in 408 paths with 61 distinct middle nodes. Part of them are shown in Table 3. We can see AD and HD shares some symptoms such as *cognitive impairment, depression, vascular dysfunction. Neuronal death* is also common in both AD and HD. Moreover, we note that the fifth row in the upper half of Table 3 indicates that *apoe genotype* is not affect the HD while strongly correlated with AD. This is one of the merits to use Open IE, not just co-occurrence in a sentence, because co-occurrence cannot distinguish these negative ones and other positive ones. Of course, we could regard relations not being found is always negative relations assuming that extracted knowledge is complete, but negative relations that is mentioned explicitly in papers could be more interesting facts than others.

The following natural question is,"What is apoe?", which is mentioned to be the genotype that is strongly correlated with AD. Actually, apoe is the only officially confirmed susceptibility gene for AD. We discovered this fact by searching the triples with query <apoe, is, ?x>. We also see that apolipoproteins e, which is the full name for apoe, appears on the LDA results shown in Table 1. The next question could be what other gene also correlate with AD. This question is discussed in the next section.

Another interesting observation from LDA results is AIDS and HIV because they are very different disease to us non-domain experts. We also asked a domain expert who is working on brain imaging about AD, but he did not know how these diseases are related. Therefore, similar to the HD example, we queried triples with <AD, ?r1 , ?x> & <?x, ?r2, HIV>. The result is shown in the latter half of Table 3. It shows interesting facts that HIV actually has similar symptoms with AD such as *delirium* and *dementia*, but does not infect neurons nor endothelial cells while AD affects them.

Table 3: Two step paths from AD to HD or HIV

AD	is the most common cause of	cognitive impairment	is an early symptom of	HD
	are significantly associated with	depression	is common in	
	is characterized by	vascular dysfunction	may occur in	
	is associated with increased	neuronal death	is also a pathological hallmark in	
	is strongly correlated with	the apoe genotype	does **not** affect the course of	
AD	frequently exhibit	delirium	sometimes accompany	HIV
	is the common cause of	dementia	is a common complication of	
	affect	neurons	are **not** infected by	
	causes pro-inflammatory effects in	endothelial cells	were **not** infected with	

4.3 Questions that need additional resources

Some questions needs additional resources to answer. We confirmed that apoe gene is strongly correlated with AD in the last section, and now are interested in other genes that also have high correlation with AD. The natural way to find answers is to search triples with the patten <?x, correlate with , AD> & <?x, is, gene>. However, this query gave no results because it is rare for researchers to write a sentence that contains *gene name* is a gene . To solve this issue, we used an additional resource of NCBI human genes [5] to restrict ?x in the pattern. In other words, the final query is <?x, correlate with , AD> where ?x is in the gene list. This query found 62 answers including APP, BDNF or CR2.

4.4 Unanswerable questions

There are some questions from LDA results that Open IE cannot answer. In the LDA results showed in Table 1, we see that Apolipoproteins E and AIDS/HIV appear as top terms in topics in the second last decade (1995 - 2004), but disappear in the last decade (2005 - 2014) while Huntington Disease consistently appear at the first topics. This indicates that Apolipoproteins E and AIDS/HIV catch less attention from researchers while Huntington Disease keeps the attention. To confirm this fact, we plotted the proportion of publications that have these terms over all the publications in Figure 2. The figure shows the same trend. Then, the natural question is, why did these changes happen. However, Open IE triples cannot explain the changes because it is usually not mentioned in the papers while Open IE extracts facts mentioned in the paper.

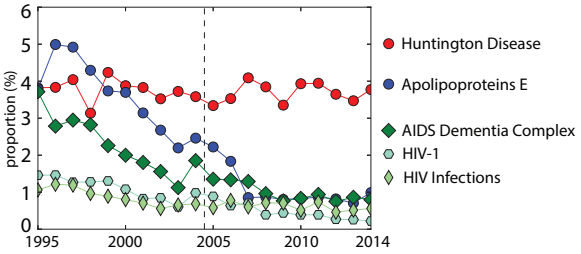

Figure 2: Ratio of papers with a term over all papers

5. CONCLUSIONS

This paper focuses on the problem of overwhelming AD literature despite the needs for further research. We provide a case study to use machine reading approach. The case study indicates both advantages and limitations of machine reading. One of the advantages, unlike previous bibliometric approaches, is that machine reading can understand specific information as well as an overview without requiring domain expertise. Also, we are able to find interesting connections between AD-related entities such as AD and HIV by using two distinct methods in a mutually complementary way. Limitations includes the manual clearing effort (i.e. removing noise and normalization), and disability to answer abstract questions whose answer is not written explicitly in text. The future direction should be how to use machine reading to discover novel knowledge that even domain experts have not yet discovered.

[5]http://www.ncbi.nlm.nih.gov/gene/

6. REFERENCES

[1] A. Association. 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 11(3):332–384, 2015.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, mar 2003.

[3] H. Chen, Y. Wan, S. Jiang, and Y. Cheng. Alzheimer's disease research in the future: bibliometric analysis of cholinesterase inhibitors from 1993 to 2012. *Scientometrics*, 98(3):1865–1877, 2014.

[4] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011.

[5] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[6] M. E. Hughes, J. Peeler, J. B. Hogenesch, and J. Q. Trojanowski. The growth and impact of alzheimer disease centers as measured by social network analysis. *JAMA neurology*, 71(4):412–420, 2014.

[7] D. Lee, W. C. Kim, A. Charidimou, and M. Song. A bird's-eye view of alzheimer's disease research: reflecting different perspectives of indexers, authors, or citers in mapping the field. *Journal of Alzheimer's Disease*, 45(4):1207–1222, 2015.

[8] Mausam. Open Information Extraction Systems and Downstream Applications. In *IJCAI*, 2016.

[9] M. Song, G. E. Heo, and D. Lee. Identifying the landscape of alzheimer‘s disease research with network and content analysis. *Scientometrics*, 102(1):905–927, 2015.

[10] M. Song, W. C. Kim, D. Lee, G. E. Heo, and K. Y. Kang. Pkde4j: Entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics*, 57:320 – 332, 2015.

[11] A. A. Sorensen. Alzheimer's disease research: scientific productivity and impact of the top 100 investigators in the field. *Journal of Alzheimer's Disease*, 16(3):451–465, 2009.

[12] A. A. Sorensen, A. Seary, and K. Riopelle. Alzheimer's disease research: A coin study using co-authorship network analytics. *Procedia-Social and Behavioral Sciences*, 2(4):6582–6586, 2010.