DE GRUYTER
OPEN

Research Paper

Using Machine Reading to Understand Alzheimer's and Related Diseases from the Literature

Satoshi Tsutsui[1], Yi Bu[1] & Ying Ding[†1,2]

[1]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA
[2]School of Information Management, Wuhan University, Wuhan 430072, China

Citation: Satoshi Tsutsui, Yi Bu & Ying Ding (2017). Using Machine Reading to Understand Alzheimer's and Related Diseases from the Literature.

Vol. 2 No. 4, 2017
pp 81–94
DOI: 10.1515/jdis-2017-0021

Received: Oct. 16, 2017
Revised: Nov. 4, 2017
Accepted: Nov. 12, 2017

Abstract

Purpose: This paper aims to better understand a large number of papers in the medical domain of Alzheimer's disease (AD) and related diseases using the machine reading approach.

Design/methodology/approach: The study uses the topic modeling method to obtain an overview of the field, and employs open information extraction to further comprehend the field at a specific fact level.

Findings: Several topics within the AD research field are identified, such as the Human Immunodeficiency Virus (HIV)/Acquired Immune Deficiency Syndrome (AIDS), which can help answer the question of how AIDS/HIV and AD are very different yet related diseases.

Research limitations: Some manual data cleaning could improve the study, such as removing incorrect facts found by open information extraction.

Practical implications: This study uses the literature to answer specific questions on a scientific domain, which can help domain experts find interesting and meaningful relations among entities in a similar manner, such as to discover relations between AD and AIDS/HIV.

Originality/value: Both the overview and specific information from the literature are obtained using two distinct methods in a complementary manner. This combination is novel because previous work has only focused on one of them, and thus provides a better way to understand an important scientific field using data-driven methods.

Keywords Machine reading; Alzheimer's disease; Knowledge discovery; Data mining

† Corresponding author: Ying Ding (E-mail: dingying@indiana.edu).

Research Paper

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder that causes dementia, where in its early stage people cannot remember recent events, and gradually have more difficulty managing daily life tasks or even recognizing family and friends. AD is to date incurable (Alzheimer's Association, 2015), although a great deal of resources are used to treat symptoms. The large number of sufferers from mild to severe cases creates a high social impact and tremendous costs for medical care that serves to manage symptoms instead of offering a cure. It is estimated that 4.7 million Americans have dementia, and the total number in 2050 is projected to be 13.8 million (Hebert et al., 2013). The economic costs of dementia were estimated to be $818 billion in 2015 (Prince et al., 2015). Due to its incurability, the costs for several sectors in the society, such as long-term care, home services, and non-professional caregivers, are greater than the cost of direct medical care (Bullock, 2004; Winblad et al., 2016; Yokoyama et al., 2016). Due to the severity and increasing number of people suffering from dementia, it is very important to promote further research on AD.

The current information available on the disease, however, is actually overwhelming. For example, simply searching "Alzheimer" in PubMed brings up 90,000+ articles[①]. Because no single researcher can read even a fraction of these papers, using computational techniques could be a partial solution. Toward this end, this paper presents a case study that uses machine reading to help gather data from the large amount of literature.

The machine reading technique of data mining is based on the idea that machines can integrate and summarize information for humans by reading and understanding large amounts of texts (Hirschberg & Manning, 2015). Previous work related to machine reading and AD used computational techniques such as topic modeling to read and take main points from a large number of papers, but its general purpose is to get overview information (Hughes et al., 2014; Lee et al., 2015; Song, Heo, & Lee, 2015; Sorensen, 2009; Sorensen, Seary, & Riopelle, 2010). Indeed, in order to understand the domain of AD, we first need an overview that will help to identify key details, such as what kind of major topics are in the literature. This will help determine what kind of specific information is needed. For instance, unless we are aware of a group or individuals that discuss AIDS within AD papers, we cannot understand how AIDS is related to AD (addressed in Section 4.2). This study first uses a topic modeling technique (Blei, Ng, & Jordan, 2003) to obtain major topics within AD literature.

① https://www.ncbi.nlm.nih.gov/pubmed/?term=alzheimer

Once the overview is obtained, specific questions are addressed with a focus on how AD is connected to AIDS/HIV. Answering these questions requires techniques to read the content of related papers, recognize key entities mentioned in the text, and identify the relations among these entities. Gathering entities and relations from texts is called information extraction (IE), which conventionally assumes pre-defined target information. For example, extracting gene-disease interactions information is a common IE task, but it requires that the target genes, diseases, and types of expected interactions among them are already identified. Yet it is not advisable to limit the types of information being sought in advance, as this weeds out potentially significant topics or details that can be helpful to the search. Moreover, even after the information being sought is fixed, it can change frequently depending on how we have understood the literature at that point, or due to shifts in perspectives on the topic. For example, we might be interested in which pathway contains a particular gene after a gene-disease extractor found that the gene is associated with AD. In this case, we need to build the pathway-gene extractor again if conventional IE techniques are employed. Therefore, an IE technique is used that does not require pre-defined targets, called open information extraction (Open IE) (Fader, Soderland, & Etzioni, 2011).

Open IE is an information extraction technique applied in natural language processing (Fader, Zettlemoyer, & Etzioni, 2014; Mausam, 2016) that gathers facts in the form of triples <*subject, predicate, object*>. For example, given the sentence, "Alzheimer is strongly correlated with the apoe genotype," it extracts <Alzheimer, is strongly correlated with, the apoe genotype>. These extracted facts are used to answer specific questions based on Latent Dirichlet Allocation (LDA) results (Blei et al., 2003). Note that it is always possible to trace the data back to specific sentences mentioning the answer. This provides an advantage because the information can be confirmed from reliable sources, that is, specific sentences in scientific papers.

The combination of the two methods of topic modeling and Open IE is complementary, in that Open IE answers key questions provided by topic modeling overviews. Topic modeling, specifically LDA, has been applied to a wide range of fields to reveal hidden topics from textual data (DiMaggio, Nag, & Blei, 2013; Hall, Jurafsky, & Manning, 2008; Hu et al., 2015). It is often difficult to make sense of each topic, however, even with substantial domain knowledge. This is because LDA just outputs topics as a distribution over terms, and does not provide information on how terms in the topic are linked together. On the contrary, Open IE can indicate how terms are specifically linked in texts. It is developed as a natural language processing (NLP) technique, and is applied to NLP tasks such as question answering. Yet when trying to understand a large collection of texts, searchers do not always

have specific questions to ask in advance. An overview is therefore needed to identify specific questions. Combining LDA and Open IE is complimentary, as LDA provides the overview, which is helpful to infer specific questions that are answered by Open IE.

The rest of this paper is organized as follows. Section 2 briefly summarizes the key related work. Section 3 describes the proposed machine reading approach using the two distinct methods of LDA and Open IE to better comprehend large collections of literature both at the overview and specific levels. Section 4 presents the results of this approach for the medical domain of Alzheimer's disease, whose related papers are far beyond what a single researcher can read. Due to the high social and fiscal impact of the disease, the need for further research is urgent. Finally, Section 5 concludes the paper.

2 Related Work

Computational techniques have been extensively used to understand a scientific domain, and applications for the topic of AD has also gathered a great deal of attention (Hughes et al., 2014; Lee et al., 2015; Song, Heo, & Lee, 2015; Sorensen, 2009; Sorensen et al., 2010). Sorensen (2009), for instance, investigated the productivity and impact of the top 100 AD researchers using citation analysis, and identified the role of AD within the field of neurodegenerative diseases. In defining an AD-specific *h*-index ranking to obtain an overview of the whole domain, Sorensen et al. (2010) focused on co-authorship networks and revealed major author communities. Chen et al. (2014) studied cholinesterase inhibitors within AD research, and analyzed research trends. Hughes et al. (2014) explored collaboration networks using papers from the Alzheimer Disease Center to reveal the impact of this organization. Lee et al. (2015) analyzed networks of medical entities to provide an overview of AD research, but did not investigate specific relations between entities. Song et al. (2015) studied AD literature both at the topic and entity levels; To our knowledge, this research is the only one that considers relations between entities, yet they only used 54 pre-defined relations defined in the Unified Medical Language System[②]. Overall, the related studies introduced so far mainly address overview information of the field of AD, while this paper has another purpose of obtaining specific information without specifying specific information searches in advance.

Obtaining specific information from texts has been well studied as information extraction in natural language processing. Information extraction (IE) is a task that

② https://www.nlm.nih.gov/research/umls/

automatically extracts structured information from texts. For example, many IE systems can extract entities such as genes, diseases, drugs, and relations between them from the medical literature (e.g. Song et al., 2015). However, because predefined relations are required, they are not effective when no relations are extracted in advance. Open IE systems (Mausam, 2016) overcome this issue and use raw textual phrases as relations, which has been applied to several NLP tasks such as question answering (Fader et al., 2014). To the best of our knowledge, this paper is the first to use Open IE to understand large amounts of literature in a specific medical domain in combination with topic modeling methods.

3 Methodology

The methodology of this paper is visually summarized in Figure 1. First, we collected literature with a focus that can be specified by key terms, specific periods, and/or target journals. After consulting domain experts, we collected a set of PubMed papers relevant to AD[3] using search terms *Alzheimer, Mild cognitive impairment, Dementia, Significant memory concern,* and *Subjective memory complaint* without any constraint of publication year. The earliest paper obtained was published in 1945. The data set collected includes 160,091 papers with MeSH terms, which are the index terms given to the PubMed articles and abstracts. Second, we applied LDA to keywords. Two types of keywords were employed: MeSH terms and genes mentioned in the abstracts. These genes were extracted from a dictionary-based extractor using LingPipe[4] with the NCBI human gene list[5]. In addition, we applied LDA to subsets of the papers with a narrower focus by year range, hoping to obtain more interesting observations than by looking at all papers. Specifically, we made two subsets of papers published in the last decade (2005–2014) and the previous decade (1995–2004). Having referred to some previous papers such as those of Zhang et al. (2017), we set the number of topics as five when using LDA because the purpose is to get a macro-level overview. Next, we applied Open IE, specifically Reverb (Fader et al., 2011), to textual content (e.g. abstracts or full text) of the papers. Each extracted triple is linked to a specific sentence in a paper. We then examined the topic modeling results and obtained some questions based on the results. Finally, we answered these questions using extracted triples. This part can be iterative because an answer to a question could raise new questions.

[3] The query was performed in October, 2015.
[4] http://alias-i.com/lingpipe/index.html
[5] https://www.ncbi.nlm.nih.gov/gene

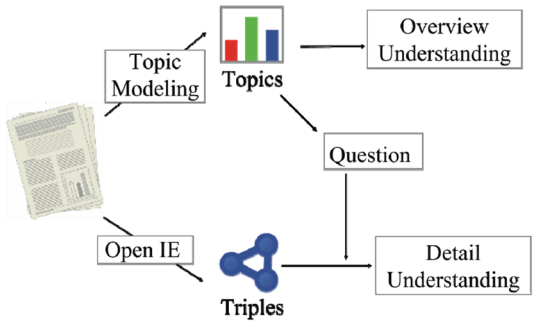

Figure 1. Conceptual sketch of methodology.

4 Results and Discussion

4.1 Overview

Based on the methods shown above, we obtained 1,469,008 triples and organized them in a relational database so that they can be traced back to a specific sentence in a paper. The extracted triples[6] and a simple demo website[7] were then available. The topic modeling results for MeSH terms and genes are shown in Tables 1 and 2, respectively. For each period and each topic, top five terms are shown, where topics are ranked by their popularity[8]. The easily noticeable observation is only that the fifth topic in 1995–2004 from the MeSH terms is obviously related to AIDS/HIV. However, the results can provide more relevant observations if the characteristics of LDA are considered.

A basic characteristic of LDA is that it provides each topic a distribution of terms. This means the first term in a topic is its most representative term. Moreover, LDA can represent each paper as a distribution of topics, which enables the ranking of topics by popularity. From these characteristics, the popularity observations from MeSH terms and genes respectively are presented.

In MeSH term topic modeling, Huntington's disease (HD) always appears in the first topic regardless of the rank all years or that of the last and second to last decades (Table 1). This means HD has consistently held certain popularity within the AD literature. Moreover, Creutzfeldt-Jakob syndrome (CJS), which is also a neurodegenerative disease like AD, was found to be popular in the period of 1995–2004 but not recently. This is because the disease is the first word in the first topic of that period, but later in 2005–2014, it only appears in the fourth topic.

[6] http://homes.soic.indiana.edu/stsutsui/machine_reading/data/index.html
[7] http://homes.soic.indiana.edu/stsutsui/machine_reading
[8] For the details on calculating topic popularity, see Chen et al. (2017).

Table 1. MeSH LDA results ranked by popularity.

Year	1st topic	2nd topic	3rd topic	4th topic	5th topic
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 coretex	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

Research Paper

Table 2. Gene LDA results ranked by popularity.

Year	1st topic	2nd topic	3rd topic	4th topic	5th topic
All	**APP**	**APOE**	INS	MS	MAPT
(1945–2015)	TNF	BCHE	**HTT**	PRNP	SDS
	BDNF	CAT	PSEN1	GFAP	SST
	BACE1	CA3	CA1	ALB	GRN
	NGF	CA1	**APP**	MDD1	PSD
1995–2004	**APP**	**APOE**	PRNP	TNF	**HTT**
	MS	BDNF	INS	PSEN1	CA1
	CD4	ACT	NGF	GFAP	SDS
	SPY	A2M	TF	MAPT	ALB
	PSEN1	LDLR	TTR	BCHE	CA3
2005–2014	**APOE**	**APP**	**HTT**	MS	PSEN1
	INS	GFAP	PRNP	BDNF	**APP**
	TNF	CD4	CA1	MAPT	**APOE**
	APP	CAT	BACE1	BCHE	ALB
	ACE	NOTCH3	NGF	SYP	PSEN2

In gene topic modeling, APP and APOE are always the most popular genes, as they appear as top words in either the first or second topic, regardless of time periods (Table 2). Other than the first or second topic, HTT became popular in recent periods as it appears as the top term in the third topic in 2005–2014, while it appears as the top term in the fifth topic in 1995–2004.

Another characteristic of LDA is that highly co-occurring terms constitute a topic. Moreover, it sometimes distinguishes the term co-occurrences in different contexts by having the same term in multiple topics. From these characteristics, observations of a term in different contexts for MeSH terms and genes are presented, respectively.

In MeSH terms, CJS appears in three different topics: the first topic in the period 1995–2004, the fourth topic in the period 2005–2014, and the fifth topic in all years. We examined co-occurring terms, where magnetic resonance imaging (MRI) co-occurs in the topic list in 1995–2004, but the term caregivers co-occurs in 2005–2014. MRI is a brain-imaging technique while caregiver is the person who takes care of the patient or person suffering from the disease. This observation indicates that CJS can be studied both from two contexts of brain imaging research and patient care. Interestingly, when investigating the topic in all years, we find that the two contexts are merged into one topic because it has both MRI and nursing homes, where patients are given care.

The research also found that APP and APOE appear in multiple topics and multiple ranks in genes. For example, APP/APOE is always the top gene in the first and second topics, but also appears as the fifth gene at the third topic in all years, and the second and third genes as the fifth topic in 2005–2014 (Table 2). This

observation indicates that APP and APOE can be studied in a context where each gene itself is the key to the topic, but also in a context where it is secondary to the topic.

4.2 Question Answering Examples

This section demonstrates the power of Open IE and answers questions specific to AD that are inferred by the LDA results and their interpretations. For example, the topic related to AIDS/HIV is found within the AD literature. Open IE can tell how AIDS/HIV is actually related to AD, which is answered by Open IE later in this section. In fact, Open IE can answer more basic questions such as the definition of terms, which could be helpful for researchers with limited knowledge of AD (e.g. information scientists or scientometrians who expect to study AD from the literature) to better understand a domain only from the literature. For example, you cannot interpret the results discussed in Section 4.1 if you do not know HD, CJS, MRI, APP, and APOE. Therefore, we first show how Open IE can answer these simple questions in order to understand these basic terms.

We first use an example of Huntington's disease (HD), which in the previous section, was observed to consistently hold certain popularity within the AD literature. The immediate question "What is Huntington's disease?" can be answered by searching triples with a pattern <Huntington disease, is, ?x > where ?x means some words are identified. In this case, 446 distinct triples were found. The top two frequent answers and other three randomly sampled results are shown in Table 3. The number inside the parenthesis represents the number of triples that matched the pattern. Now it can be discovered that HD is *an inherited neurodegenerative disorder*.

Table 3. Question answering example: What is Huntington's disease?

Question	Answer example
What is Huntington's 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*)

Extraction from texts does not always result in correct data, so some manual inspections are required. Limitations of this approach include the wrong answer of *complex, single gene*. One possible method to reduce error is to only use highly frequent triples. However, low frequency triples also carry interesting and important facts. For example, *an incurable genetic neurodegenerative disorder* has been mentioned only once in our corpus but this answer carries an important fact that

HD is incurable. Therefore, we decided not to rely on the frequency of triples, but instead use manually inspected results. Another issue is that Open IE does not provide normalization, so term variants are treated distinctly. This could be an issue when triples are searched with a pattern. For example, *HD* needs to be added as a term variant to Huntington's disease searches if we want to get complete results. Yet these term variants can be obtained by querying terms that have the same definition. For example, checking other subjects defined as *an inherited neurodegenerative disorder* leads us to another term variant, *Huntington's disease*.

The same approach is also applied herein to find definitions of CJS, MRI, APP, and APOE (Table 4). These definitions are complementary in allowing more LDA results to be interpreted. It can be confirmed that, for example, APP and APOE are strongly related to AD. This fact coincides with the observation in Section 4.1, that they are always the most popular genes.

Table 4. What are CJS, MRI, APP, and APOE?

Term	Definition
AD	a neurodegenerative disorder
AD	a genetically complex and heterogeneous disorder
CJS	a rare neurodegenerative disease
CJS	a fatal neurodegenerative illness
CJS	an incurable disease
MRI	a useful diagnostic marker
MRI	a promising AD biomarker
MRI	the most widely used and less invasive medical imaging technique
APP	a transmembrane glycoprotein
APP	an extremely complex molecule
APP	**the key player in AD pathogenesis**
APOE	the major apolipoprotein
APOE	**the only confirmed susceptibility gene for AD**
APOE	the most prevalent and best established genetic risk factor for late-onset AD.

This study also uses more interesting questions than term definitions. Suppose we know that HD is also a neurodegenerative disease like AD. The next natural question is, "How is AD related to HD?" We first queried <AD, ?x, HD >, but could not find relevant relations. After extending the pattern to <AD, ?r1, ?x> & <?x, ?r2, HD>, one finds an equivalent of finding two-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 Figure 2. It can be observed that AD and HD share some symptoms such as *cognitive impairment, depression,* and *vascular dysfunction*. *Neuronal death* is also common in both AD and HD. Moreover, Figure 2 indicates that *apoe genotype* does not affect HD while it is strongly correlated with AD. This negative relation indicates the advantage of using Open IE in addition to LDA,

which relies on co-occurrences, because co-occurrences cannot distinguish positive or negative relations. One could regard relations that are not found as negative relations, however, assuming that extracted knowledge is complete, but negative relations that are mentioned explicitly in papers could provide more important facts than others.

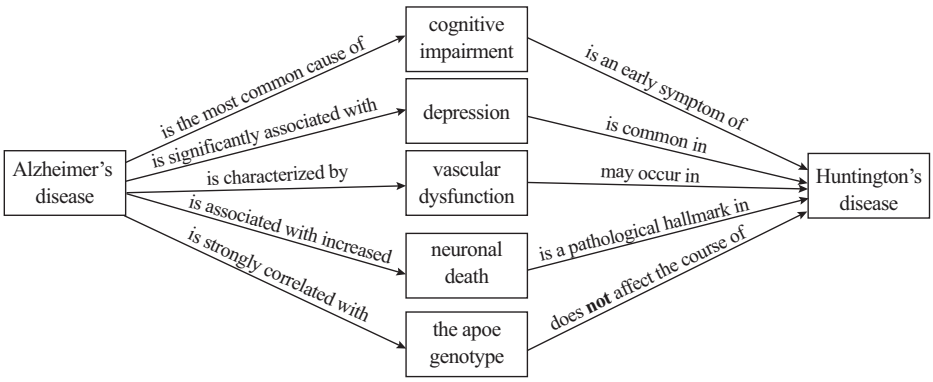

Figure 2. Two-step paths from AD to HD.

Another important observation from LDA results relates to AIDS and HIV, which are very different diseases (HIV is common to AIDS, in that all people with AIDS generally have HIV, but is not the full-blown AIDS disease in terms of symptoms and treatment). Also, a domain expert working on AD brain imaging was queried, but he did not know how HIV/AIDS are related to AD. Similar to the HD example, triples were thus queried with <AD, ?r1, ?x> & <?x, ?r2, HIV>. The results shown in Figure 3 confirm meaningful facts: HIV actually has similar symptoms to AD such as *delirium* and *dementia*, but does not infect neurons nor endothelial cells while AD affects them.

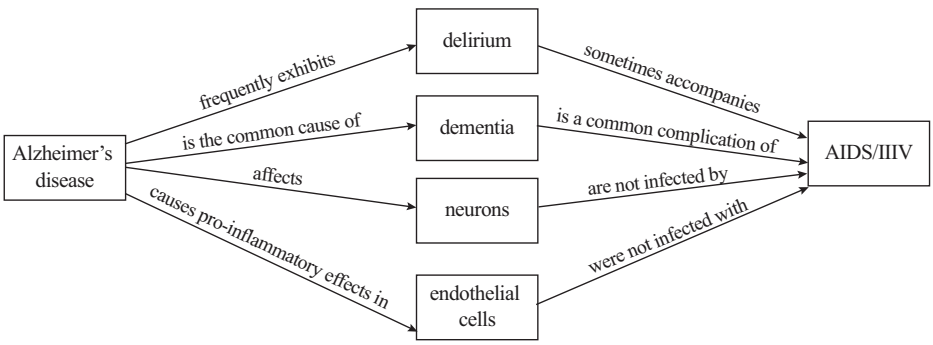

Figure 3. Two-step paths from AD to HIV.

Some questions need additional resources to answer. For example, this study was able to confirm that the apoe gene is strongly correlated with AD (as seen in the previous section), yet now we are interested in other genes that also have high correlation with AD. The natural way to find answers is to search triples with the pattern <?x, correlated with AD> & <?x, is, gene>. But this query gave no results because it is rare for researchers to write a sentence that contains "[a gene name] is a gene." To solve this issue, we used an additional resource of NCBI human genes® to restrict ?x in the pattern. In other words, the final query is <?x, correlated with AD> where ?x is in the gene list. This query found 62 answers including APP, BDNF, and CR2.

Some questions from topic modeling results cannot be answered by Open IE. These are "why" questions because Open IE simply extracts facts mentioned in the paper. For example, we observed that CJS was popular in the period of 1995–2004, but not 2005–2015. Open IE, however, cannot answer why this change happened. A doctoral student working on AD who was queried inferred that the reason might be that Stanley B. Prusiner was awarded the Nobel Prize in 1997 for his discovery of Prions, the pathogen of CJS, which created an upsurge in interest. However, this cause and effect relation cannot be verified, as Open IE can by no means infer that.

5 Conclusions

This paper provides a case study of using the machine reading method to understand the domain of Alzheimer's disease (AD), and its relation to other diseases such as HIV and AIDS. AD is a field whose number of the related papers is overwhelmingly high, although there is a vital need for further research that may actually help find the causes of the disease as well as a cure. We demonstrate that machine reading helps identify specific information that offers a better understanding via overviews provided by topic modeling. The use of both methods of LDA and Open IE in a mutually complementary way reveals how the topic modeling technique connects AD and HIV/AIDS. Based on this observation, when querying the Open IE extractions, the two diseases are found to have different mechanisms but share some symptoms such as dementia.

This study has several implications. First of all, it shows that the literature on a topic can answer specific questions relating to it, which has not been attempted in the literature to date. From the perspective of Alzheimer's disease, the approach provided in this article could help domain experts find important relations between entities in a similar manner as this study identified relations between AD and

® http://www.ncbi.nlm.nih.gov/gene/

HIV/AIDS. Methodologically, this approach can serve as a preliminary knowledge extraction step for literature-based knowledge discovery if future researchers hope to construct a curated knowledge base for a specific purpose.

One limitation of this approach is that we need to manually clean the data, such as remove false extractions. Moreover, this study is not able to answer abstract questions when these answers are not written explicitly in texts. In the future, it would be helpful to develop a method to automate the process to detect false extractions. We could also integrate existing medical knowledge bases to answer more complex or nuanced questions.

Author Contributions

S. Tsutsui (stsutsui@indiana.edu) designed the research framework and the methods, analyzed the data, and wrote the manuscript. Y. Bu (buyi@iu.edu) wrote the manuscript. Y. Ding (dingying@indiana.edu, corresponding author) instructed the research team and revised the manuscript.

References

Alzheimer's Association. (2015). 2015 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 11(3), 332–384.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.

Bullock, R. (2004). The needs of the caregiver in the long-term treatment of Alzheimer disease. Alzheimer Disease & Associated Disorders, 18 Suppl 1, S17–S23.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. Journal of Informetrics, 11(4), 1175–1189.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics, 41(6), 570–606.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535–1545). Stroudsburg, PA: Association for Computational Linguistics.

Fader, A., Zettlemoyer, L., & Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 1156–1165). New York: ACM.

Hall, D., Jurafsky, D., & Manning, C.D. (2008). Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 363–371). Stroudsburg, PA: Association for Computational Linguistics.

Hebert, L.E., Weuve, J., Scherr, P.A., & Evans, D.A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology, 8 (19), 1778–1783.

Hirschberg, J., & Manning, C.D. (2015). Advances in natural language processing. Science, 349(6245), 261–266.

Hu, B., Dong, X., Zhang, C., Bowman, T.D., Ding, Y., Milojević, S., . . . & Larivière, V. (2015). A lead-lag analysis of the topic evolution patterns for preprints and publications. Journal of the Association for Information Science and Technology, 66(12), 2643–2656.

Hughes, M.E., Peeler, J., Hogenesch, J.B., & Trojanowski, J.Q. (2014). The growth and impact of Alzheimer disease centers as measured by social network analysis. JAMA Neurology, 71(4), 412–420.

Lee, D., Kim, W.C., Charidimou, A., & Song, M. (2015). 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.

Mausam, M. (2016). Open information extraction systems and downstream applications. In Proceedings of the 25th International Joint Conference on Artificial Intelligence. Retrieved on November 4, 2017, from http://www.cse.iitd.ac.in/~mausam/papers/publications.html.

Prince, M., Wimo, A., Guerchet, M., Ali, G., Wu, Y., & Prina, M. (2015). World Alzheimer Report 2015: The global impact of dementia: An analysis of prevalence, incidence, cost and trends. London: Alzheimer's Disease International.

Song, M., Heo, G.E., & Lee, D. (2015). Identifying the landscape of Alzheimer's disease research with network and content analysis. Scientometrics, 102(1), 905–927.

Song, M., Kim, W.C., Lee, D., Heo, G.E., & Kang, K.Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. Journal of Biomedical Informatics, 57, 320–332.

Sorensen, A.A. (2009). 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.

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

Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., . . . & Zetter- berg, H. (2016). Defeating Alzheimer's disease and other dementias: A priority for European science and society. The Lancet Neurology, 15(5), 455–532.

Yokoyama, J.S., Wang, Y., Schork, A.J., Thompson, W.K., Karch, C.M., Cruchaga, C., . . . & Desikan, R.S. (2016). Association between genetic traits for immune-mediated diseases and Alzheimer disease. JAMA Neurology, 73(6), 691–697.

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2017). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. Journal of the Association for Information Science and Technology. Retrieved on November 4, 2017, from http://onlinelibrary.wiley.com/doi/10.1002/asi.23916/pdf.