Distillation of Knowledge from the Research Literature on Alzheimer's Dementia

Wutthipong Kongburan
King Mongkut's University of Technology Thonburi
Thailand
58130800102@st.sit.kmutt.ac.th

Mark Chignell
University of Toronto
Canada
chignell@mie.utoronto.ca

Jonathan Chan
King Mongkut's University of Technology Thonburi
Thailand
jonathan@sit.kmutt.ac.th

ABSTRACT

Many countries are aging societies. Since abilities generally deteriorate with age, technologies can assist older adults in their daily life. Loss of cognitive status is particularly severe in cases of dementia, with around 70% (according to Alzheimers.net) of dementia cases involving Alzheimer's Dementia (AD), a progressive and currently incurable disease. There is considerable research on AD with thousands of relevant publications being added to the PubMed online database every year. The knowledge incorporated in this large body of work is spread across hundreds of thousands of pages of text, making it difficult to distill and mobilize that knowledge in terms of treatments and guidelines. Text mining technology may assist in distilling knowledge from the vast corpus of research literature on Alzheimer's dementia. In this paper, we apply the Named Entity Recognition (NER) system, a text mining (TM) method used to group words into classes, in order to extract useful information from free texts. We present findings concerning how well NER can extract information from a corpus of AD research publications.

CCS CONCEPTS

Applied computing → Life and medical sciences → Health care information systems

KEYWORDS

Aging society; Alzheimer intervention; Named entity recognition; PubMed; Quality of life;

1. INTRODUCTION

An estimated 10% of the world population was aged 65 or older as of this writing, and in many countries in Europe and Japan that proportion is over 20% and climbing. In one example of this demographic trend, in 2015 Statistics Canada reported that, for the first time there were more people aged 65 or over than there were under 15[1]. Meanwhile, in Japan, the proportion of elderly (over the age of 65) citizens reached 26% in 2015[2]. An increasing number of older adults is associated with an increased burden of health problems, because many physical and cognitive functions decline even with healthy aging, and declines are typically more pronounced in the case of disease. Alzheimer disease (AD) is one of the most prevalent chronic medical conditions affecting older people and is a major cause of severe decline in cognition and loss of the ability to live independently. As of this writing there are close to 50 million cases of AD or related dementias worldwide[3]. As many as 50 to 70 percent of all dementia cases are AD, according to Alzheimers.net. In addition, 1-in-9 Americans over 65 has AD[4]. Behavioral symptoms associated with dementia include repetitive speech, wandering, and sleep disturbances, along with loss of memory and an increase in risk of conditions such as depression and delirium. As of this writing there are no effective treatments for AD and the clinical focus has been on managing the symptoms of dementia. Since many types of treatment have been proposed, information about what works when dealing with behavioral problems associated with people at different stages of AD can enhance quality of life not only for those with AD but also for their caregivers.

The main aim of the research reported in this paper is to demonstrate how Text Mining (TM) can extract useful information about AD treatments from the scientific literature on AD. First we describe the construction of a training dataset (corpus) from the abstracts of scientific papers with a focus on AD. We then used Named Entity Recognition (NER), trained using the training data set, to label entities of interest within a sample set of real-world test cases. The results demonstrate that NER can be used to classify relevant entities within the AD literature.

2. BACKGROUND

NER is a key approach to TM that identifies keywords in text streams and classifies them into predefined relevant categories such as gene, or protein. Various techniques have been proposed to develop NER systems. They can be categorized as rule-based, dictionary-based and Machine Learning (ML)-based (see more information in [2]). As can be seen in [2, 4, 5], when the appropriate resources are available, the ML-based solutions present several advantages, and perform better than dictionary-based and rule-based approaches. In this paper, we use ML-based TM to deal with the problem of NER.

We used the NER classifier developed at Stanford University. Stanford NER is a Java implementation of NER labelled sequences of words in a text which include names of people, locations, and company names. NER used the Conditional Random Fields (CRFs) technique [8] to train the classifier based on a training set of labeled entities within a corpus of documents. Other projects that have used CRFs in NER include Gimli [1] and BANNER [9]. These two open source tools automatically tagging genes, proteins and other entity names in

[1] http://www.statcan.gc.ca/daily-quotidien/150929/dq150929b-eng.pdf
[2] http://www8.cao.go.jp/kourei/english/annualreport/2014/pdf/c1-1.pdf

[3] https://www.alz.org/documents_custom/2016-facts-and-figures.pdf
[4] http://www.alzheimers.net/resources/alzheimers-statistics/

WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
DOI: http://dx.doi.org/10.1145/3041021.3054934

text. Relevant tools also include ChemSpot [10], a named entity recognition tool for identifying mentions of chemicals in natural language texts.

NER requires a corpus for training the classifier. Typically this requires laborious manual annotation of entities. Previous, we discussed development of a thyroid cancer intervention corpus in order to label relevant treatment entities for thyroid cancer [7]. In this work, we develop an AD corpus in order to recognize disease names and interventions, relating to AD, in texts.

3. PROPOSED METHOD

The main aim of this work is to applying text mining approach to automatically identify references to diseases and interventions in research documents relevant to AD. We first describe the information retrieval process used to gather relevant documents. We then describe how entities are defined and labeled. Initial results using NER on research documents about AD are then reported.

3.1 Corpus Construction

All document abstracts used in our training corpus were retrieved from PubMed. We performed three different searches to gather a total of 50 abstracts. First 20 abstracts were retrieved from a PubMed search with a query that captured the overall topic of Alzheimer's treatment appearing in the abstract: *Alzheimer treatment AND hasabstract[text] AND Humans[Mesh] AND English[lang]*. It should be noted that the results were sorted by best match, and the query limits the results to abstracts of human studies in English (accessed on 9/2/2017). We then used *Alzheimer treatment* as a keyword in searching Google Scholar on February 9, 2017 (https://scholar.google.com). The first ten (i.e., top ten ranked) hits listed by Google Scholar were added to the corpus. Finally, the last 20 abstracts were retrieved from PubMed with the more specialized query focused on nonpharmacologic treatment: *Nonpharmacologic treatment for Alzheimer AND hasabstract[text] AND Humans[Mesh] AND English[lang]*. After preprocessing, each document in the corpus consisted of a title and a content abstract.

3.2 Annotation Procedure

The procedure used for annotating disease and intervention entities is summarized in this section.

All entities to be annotated were either nouns or noun phrases. The three labels assigned (mutually exclusive) were *Disease*, *Intervention*, and *Other*. For example, *"*An exercise program is recommended for patients with mild to moderate Alzheimer disease" would be labeled as *exercise program– Intervention*, *Alzheimers disease* – Disease, *patients – Other* etc.

In the sentence: *"*We treated 17 patients who had moderate to severe Alzheimer's disease with oral tetrahydroaminoacridine, a centrally active anticholinesterase", both the specific term *oral tetrahydroaminoacridine*, and the more general term *anticholinesterase*, were labeled as *Intervention*.

In contrast, for the following sentence: *"*The aim of this research was the assessment of the long-term benefit of non-drug therapies in Alzheimer's disease", *non-drug therapies* was judged to be too general a term to be included.

Pronouns and co-references were excluded. For example, *"*Cholinesterase inhibitors represent first-line therapy for patients with mild to moderate AD, and it is also used in the treatment of dementia". In this case *it* was unlabeled.

Special characters (e.g., quote, dash, or parenthesis) at the beginning or end of entities were not considered. For example, "The effect of angiotensin converting enzyme (ACE) inhibitors on Alzheimer disease (AD) remains unclear". The label was assigned as: *angiotensin converting enzyme ACE inhibitors– Intervention*, and *AD–Disease*. The parenthesis was not labeled.

3.3 Entity Definition

As AD progresses, impairment of memory, judgment, attention span, and problem solving skills are followed by severe apraxias and a global loss of cognitive abilities[5]. In order to build a corpus that was consistent and well-formed terminology entities were defined using the criteria described in the following paragraphs.

Disease included both synonymous mentions and abbreviated forms of AD. Although AD is the most common type of dementia, there are other dementias such as frontotemporal dementia, and other conditions can also cause dementia, such as Parkinson's disease, vascular disease and Creutzfeldt-Jakob disease[6]. Symptoms or consequences of dementia are broad and include cognitive impairment, memory loss, cognitive decline, depression, anxiety, psychosis, and agitation.

Intervention, as defined in developing the corpus, is an object or action that is used or performed by doctor or other clinician targeted in order to help the AD patient. Other aspects of interventions include cause, risk factor, and diagnosis tool. However, in this work we focus only on treatment and preventive care intervention. AD interventions include non-pharmacologic therapy: e.g., exercise, cognitive activity, gardening, word games, listening to music and cooking, and pharmacologic therapy. The main FDA-approved drugs for AD include cholinesterase inhibitors, and memantine. Gene targeting therapy interventions for AD were not considered in developing the corpus.

3.4 Study Design

The main objective of the study was to demonstrate the use of text mining to automatically identify knowledge constructed from research publications in a medical domain, using the particular example of dementia. In the case study reported here, we used three entity labels (*Intervention*, *Disease*, *Other*). We used training data gathered from a specially constructed corpus of 50 medical research paper abstracts, and we used a set of test cases comprising the first page of text from 10 prominent webpages that were relevant to Alzheimer's research (listed below).

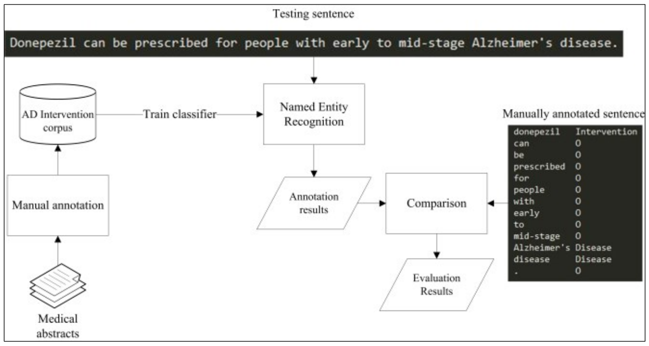
1) http://www.alz.org/alzheimers_disease_alternative_treatments.asp
2) http://www.alz.org/research/science/alzheimers_disease_treatments.asp
3) http://www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/treatment/txc-20167132
4) http://www.nhs.uk/Conditions/Alzheimers-disease/Pages/Treatment.aspx
5) https://www.nia.nih.gov/alzheimers/topics/treatment
6) https://www.fightdementia.org.au/about-dementia/health-professionals/clinical-resources/non-pharmacological-treatments
7) http://www.everydayhealth.com/alzheimers/non-medical-alzheimers-therapy.aspx

[5] https://www.ncbi.nlm.nih.gov/mesh/68000544

[6] http://www.alz.org/dementia/types-of-dementia.asp

8) http://www.memantine.com/en/patients_and_caregivers/treatment/index.php
9) http://patient.info/doctor/alzheimers-disease
10) http://www.alzheimersresearchuk.org/about-dementia/helpful-information/treatments-available

These webpages were retrieved using *Alzheimer treatment* as a keyword search in the Google search engine. Moreover, with the same keyword search, we filtered the search results by time (past week, past month, past year) in order to focus on new AD interventions mentioned in webpages at different periods of time within the recent past (as of this writing). In order to automatically assign the entities in the test set, we used an existing text classifier, Stanford NER [6], otherwise known as CRF classifier. It uses many features such as term appearance (e.g., capitalization, prefixes and suffixes) and orthographic features (e.g., alphanumeric characters, dashes, and Roman numeral characters) to build a classifier. We used our prototype corpus, in word token format, to train the classifier. We then used the resulting classifier to automatically identify entities in the testing data. We evaluated the performance of the NER by comparing the entities that it did (and didn't) label with the results obtained when the same test cases were hand-labeled by a human. A summary of the training process and experiment are shown in Fig. 1.



```
Donepezil can be prescribed for people with early to mid-stage Alzheimer's disease.
```

```
donepezil    Intervention
can          O
be           O
prescribed   O
for          O
people       O
with         O
early        O
to           O
mid-stage    O
Alzheimer's Disease
disease      Disease
.            O
```

Figure 1: Training process and experimental flow

4. RESULTS AND DISCUSSION

The AD intervention corpus consisted of 569 sentences, with 274 cases of *Disease*, and 265 cases of *Intervention* as determined by manual labeling. These entities comprised 4 unique *Diseases*, and 138 unique *Interventions*. The most frequent entity was *AD*, and *cholinesterase inhibitor*, for *Disease*, and *Intervention*, respectively.

Table 1 shows the results that were obtained. It can be seen that precision was very high (i.e., the entities that were identified by the NER classifier were labeled correctly). However, recall was low likely due to the small size of the training corpus and the fact that it did not contain a large enough sample of AD interventions that could be recognized in test case documents.

We also examined the confusion matrix, and the number of false positive and false negative errors. Our NER classifier was more likely to classify intervention entities as *Other* than as *Disease*. Most errors involving mislabeling of *Interventions* involved terms that were not in the training corpus (e.g., *Aromatherapy, relaxing song, art therapy, coconut oil, pet therapy, Rivastigmine or Exelon, Galantamine or Razadyne*). In addition, most erroneous labeling of *Interventions* as *Diseases* were due to a disease name being embedded within the name of the intervention (e.g., *Disease Alzheimer's drugs*). As another example of the challenges faced, Proper nouns can be difficult to handle. In general, excluding proper nouns is beneficial e.g., *Alzheimer's Healthcare Center*. However, removing proper nouns from consideration is far from foolproof. For instance, drug names typically have an initial capital letter and appear to be proper nouns; but, they are also *Interventions*.

Table 1: Intervention identification performance

Measure	Performance
Precision	0.987
Recall	0.397
F1-score	0.566
Accuracy	0.982

The results obtained show that many AD interventions can be inferred as shown in Fig. 2 (note that only the *Intervention* labels are visualized in the figure). The two most frequently occurring entities are *cholinesterase inhibitors* and *memantine*, as was also found in the training corpus. Meanwhile, the most frequent unseen (i.e., not appearing in the training set) entities observed in the test web pages were *Alzheimer's drugs*, and *AChE inhibitors*.

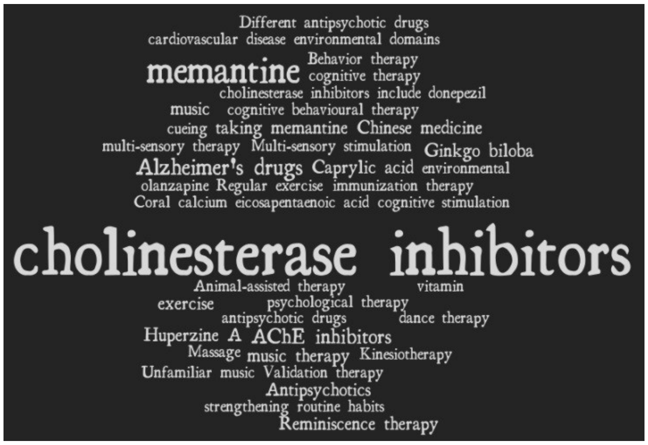


Figure 2: Word cloud of extracted interventions for Alzheimer's Disease

There were also a number of new entities identified in the test set, not seen earlier in the training set, as exemplified in the entities shown in Table 2.

5. CONCLUSIONS

Alzheimer's Disease (AD) is one of the most costly and damaging diseases associated with aging. Although a vast amount of medical research has been published on AD in recent decades, and many experiments have been carried out, the results are typically provided as unstructured text. Consequently, TM is needed to quickly scan through documents and extract specific terms and concepts. In this paper, we present an applying NER, text mining approach, to discover non-pharmacologic management and general pharmacologic treatment of AD from a corpus of free text documents representing scientific research publications. In addition, we showed how the NER approach, applied to a topic specific corpus, can extract statements relating to the use of various outcomes in treating AD. The work reported here demonstrates the value of a TM approach in distilling knowledge from research findings related to AD that are reported in text. Better knowledge of the availability and impact of treatments should eventually lead to treatment and care regiments that achieve the goal of improved quality of life in older adults and caregivers [3].

The present results show that a carefully labeled training set corpus can form a good basis for subsequently automated entity recognition within medical research publications. While the present research focused only on AD it seems likely that similar results should be obtained with other medical syndromes and contexts.

In future research, it would be good to use a larger corpus, a larger set of entity types, and more extensive test data in showing that the methods introduced here can be scaled up to more extensive distillations of knowledge within the research literature. NER would then need to be followed by additional analyses to identify useful treatment guidelines and options, so that useful clinical evidence can be synthesized from large volumes of research literature.

Table 2: Example of unseen Interventions

Search criteria	Intervention
Normal search	unfamiliar music, dance therapy, massage, environmental cueing, Chinese medicine, coral calcium, animal-assisted therapy, multi-sensory therapy,
Past week	physical therapy, intranasal insulin therapy
Past month	physical therapy, intranasal insulin therapy, antiepileptic drugs, light therapy, flashing light therapy, mouse's spine
Past year	flashing light therapy, routine screening, biogen therapy, nilvadipine, BACE inhibitors

6. ACKNOWLEDGMENTS

This work was partially supported by the Petchra Pra Jom Klao Doctoral Scholarship of King Mongkut's University of Technology Thonburi.

7. REFERENCES

[1] Campos, D., Matos, S., and Oliveira, J. L. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, *14*(1), 54.

[2] Campos, D., Oliveira, J. L., and Matos, S. 2012. Biomedical named entity recognition: a survey of machine-learning tools. INTECH Open Access Publisher.

[3] Chan, J. H. Digital information and communication in the ageing society. In *Proceedings of the 6th International Conference on Applications of Digital Information and Web Technologies* 2015. DOI= http://dx.doi.org/10.3233/978-1-61499-503-6-3.

[4] Cohen, A. M., and Hersh, W. R. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

[5] Eltyeb, S., and Salim, N. 2014. Chemical named entities recognition: a review on approaches and applications. *Journal of cheminformatics*, *6*(1), 17.

[6] Finkel, J. R., Grenager, T., and Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. (June. 2005), 363-370. Association for Computational Linguistics.

[7] Kongburan, W., Padungweang, P., Krathu, W., and Chan, J. H. Semi-automatic construction of thyroid cancer intervention corpus from biomedical abstracts. In *Eighth International Conference on Advanced Computational Intelligence,* ICACI, (February. 2016), 150-157.

[8] Lafferty, J., McCallum, A., and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning,* ICML, Vol. 1, (June. 2001), 282-289.

[9] Leaman, R., and Gonzalez, G. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on Biocomputing*, Vol. 13, (January. 2008), 652-663.

[10] Rocktäschel, T., Weidlich, M., and Leser, U. 2012. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, *28*(12), 1633-1640.