NATURAL LANGUAGE PROCESSING FOR MINING NEUROANATOMICAL RELATIONS AMONG BRAIN REGIONS

by

Erinç Gökdeniz

B.S., in Computer Engineering, Ege University, 2004

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2016

ACKNOWLEDGEMENTS

I'd like to express my deepest gratitude to my thesis supervisors Assist. Prof. Arzucan Özgür and Prof. Reşit Canbeyli. It has been a privilege working with them. They have been supportive all the time with their positive and very kind attitude. They encouraged me on hard times to give my best. It has been a great learning period for me and their approach made me see the problems from different angles. I'm really thankful to them for accepting me as their thesis student.

I'm very grateful to my thesis committee members; Assoc. Prof. Haluk Bingöl, Assoc. Prof. Burak Güçlü and Dr. Suzan Üsküdarlı for their times and sharing their invaluable knowledge. I genuinely believe that this study will evolve to its next stage with their feedback.

I thank Leon French and his co-workers for creating and sharing the manually annotated WhiteText corpus.

I would like to thank Assist. Prof. Pınar Öz for her support during the pattern definition/selection phases and sharing her time and knowledge on PVT.

I am also thankful to my dear friend Çağrı Polat for his endless support and encouragement.

Last but not the least, I'd like to thank my family. Words are not sufficient to explain my gratitude to them. I thank my beatiful wife Özlem and our lovely kids Efe and Arda for being in my life and for their love, patience and time... Without them, I'd never be able to complete this journey. I thank my mother Nihal Boçkay and my brother İlker Gökdeniz who have always been with me. They are the reason who I am and they are part of everything I succeed. Thank you very much for supporting me all the time.

I'd like to dedicate my study to my father, Murat Gökdeniz...

ABSTRACT

NATURAL LANGUAGE PROCESSING FOR MINING NEUROANATOMICAL RELATIONS AMONG BRAIN REGIONS

Identifying the relations among different regions of the brain is vital for a better understanding of how the brain functions. While a large number of studies have investigated the neuroanatomical and neurochemical connections among brain structures, their specific findings are found in publications scattered over a large number of years and different types of publications. Text mining techniques have provided the means to extract specific types of information from a large number of publications with the aim of presenting a larger, if not necessarily an exhaustive picture. By using natural language processing techniques, the present study aims to identify relations among brain regions in general and relations relevant to the paraventricular nucleus of the thalamus (PVT) in particular.

We introduce a linguistically motivated approach based on patterns defined over the constituency and dependency parse trees of sentences. Besides the presence of a relation between a pair of brain regions, the proposed method also identifies the directionality of the relation, which enables the creation and analysis of a directional brain region connectivity graph. The approach is evaluated over the manually annotated data sets of the WhiteText Project. In addition, as a case study, the method is applied to extract and analyze the connectivity graph of PVT, which is an important brain region that is considered to influence many functions ranging from arousal, motivation, and drug-seeking behavior to attention. The results of the PVT connectivity graph show that PVT may be a new target of research in mood assessment.

ÖZET

Beynin çalışma şeklini daha iyi anlayabilmek için beynin bölümleri arasındaki ilişkileri anlamak çok önemlidir. Beynin her bir bölümü birbiri ile kimyasal veya fonksiyonel etkileşim halindedir ve bu etkileşimleri inceleyen çok fazla sayıda çalışma bulunmaktadır. Bu çalışmalarda yer alan beyin bölümleri arasındaki ilişkiler, çevrimiçi erişilebilir yayınlanmış makalelerde yer almaktadır. Metin madenciliği teknikleri kullanılarak özellikli ilişkilerin çıkartılması bize ilişkiler hakkındaki büyük resmi görmemiz konusunda yardımcı olmaktadır. Biz de bu çalışmamızda, doğal dil işleme (NLP) teknikleri kullanarak beynin bölümleri arasındaki ilişkileri yayınlanmış makalelerden çıkartmayı hedeflemekteyiz. Çalışmamızda "Paraventricular Thalamic Nucleus (PVT)" adı verilen beyin bölümünün ilişkileri üzerinde yoğunlaşıyoruz.

Dilbilimsel bir yaklaşımla, örüntülere bağlı olarak ilişkilerin yer aldığı cümleler seçilerek, daha sonrasında bu cümleler üzerinde bağlılık ayrıştırıcı ve öğe ayrıştırıcı kullanılarak ilgili beyin bölümleri ve birbirleriyle ilişkileri çıkartılmıştır. Çalışmamızda, ilişkilerin yanısıra bu ilişkilerin yönü de tayin edilerek beyin bölümlerinin birbiriyle bağlantılarını gösteren bir bağlantı grafiği sunulmaktadır. Geliştirdiğimiz sistem, Whitetext projesinin derlemi üzerinde değerlendirildikten sonra aynı metodlar PVT beyin bölümünün bağlantı grafiğini çıkartma ve analiz etme konularında kullanılmaktadır. PVT, uyarılma, isteklendirme, ilaç arama davranışı ve dikkat gibi çok sayıda işlev üzerinde etkisi olduğu inanılan önemli bir beyin bölümüdür. Çalışmamızın sonuçlarının göstereceği üzere PVT beyin bölümü davranış değerlendirmesi konusunda yeni bir araştırma odağı olabilir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$(?i)$	case insensitive sign in regular expression
\s	a whitespace character in regular expression
\w	a word sign in regular expression

LIST OF ACRONYMS/ABBREVIATIONS

AC	Amygdaloid Nucleus
AL	Lateral Amygdaloid Nucleus
AMG	Amygdala
amod	Adjectival Modifier
API	Application Programming Interface
BAMS	Brain Architecture Management System
BioNLP	Biomedical Natural Language Processing
BNST	The Bed Nucleus of Stria Terminalis
BSTvl	Ventrolateral Bed Nucleus of Stria Terminalis
CB	calbindin
CE	Central Nucleus of Amygdala
CgG	Cingulate Gyrus
CRF	Conditional Random Field
CTb	Cholera Toxin b
DAB	Diaminobenzidine
det	Determiner
dobj	Direct Object
DR	Dorsal Raphe
GATE	General Architecture for Text Engineering
ID	Identifier
JCN	Journal of Comparative Neurology
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings
NAc	Nucleus Accumbens
NAS	Nucleus Accumbens
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
nn	Noun Compound Modifier

NP	Noun Phrase
nsubj	Nominal Subject
nsubjpass	Passive Nominal Subject
Pa	Paraventricular Nucleus of Thalamus
PBN	Parabrachial Nucleus
PBS	Phosphate-buffered Saline
PFC	Prefrontal Cortex
PL	Parietal Lobe
PP	Prepositional Phrase
pPVT	Posterior Paraventricular Nucleus of Thalamus
prep	Prepositional Modifier
Pt	Parataenial Nucleus
PV	parvalbumin
PVT	Paraventricular Nucleus of Thalamus
SAX	Simple API for XML
SCN	Suprachiasmatic Nucleus
SD	Stanford Dependencies
SLK	Shallow Linguistic Kernel
UIMA	Unstructured Information Management Architecture
US	United States
vmod	Reduced Non-finite Verbal Modifier
xcomp	Clausal Complement
XML	Extensible Markup Language
xsubj	Controlling Subject

1. INTRODUCTION

The brain is the most complex organ in a vertebrate's body and contains numerous interconnected structures. The interactions between these specialized structures form the substrate for different functions such as arousal, motivation, attention, etc. These interactions are classified as neuroanatomical, chemical (type of neurotransmitters) and functional (connection method, or attributed cognitive function) types. Many studies have been conducted to identify the relations among brain regions in various species and this information is already available in the free text of the biomedical literature, albeit scattered in a large number of studies published over a sizable time period. Our aim is to propose a linguistically empowered approach by using natural language processing (NLP) techniques to automatically extract the relations among the brain regions from the publications. By doing so, we first target obtaining the neuroanatomical relations among the brain regions, and then extend these with the neurochemical and functional relations. After generating a map of connections, we will be in a position to automatically extract a brain region's relations and its roles on many functions.

Extraction of neuroanatomical relations is important to understand the existing studies on brain regions and it can be used for future guidance in new researches. For example, it is possible to observe a relation between two brain regions and the impact of this relation on a specific function. Considering the amount of publications in neuroscience domain, extracting relations manually from the related publications and creating a map of brain regions is not an easy task. With a simple search on just two neuroscience journals on PubMed it can be seen that there are 16,924 papers in "The Journal of Comparative Neurology" and 32,745 papers in "The Journal of Neuroscience", which would need automation for relation extraction.

In this study, we propose a natural language processing based approach for neuroanatomical relation extraction from neuroscience publications. Unlike previous studies that used supervised machine learning methods originally proposed for protein-

protein interaction extraction, we target developing a high-precision knowledge-based linguistically motivated approach specifically designed for the neuroscience domain. Our main motivation with this study is to build the first fully linguistic approach for extracting directed neuroanatomical relations. Our approach is based on using predefined patterns for selecting the potential neuroanatomical relation describing sentences and leveraging the deeper syntactic analysis for identifying the related brain region entities. We use specifically the constituency and dependency parse trees for syntactic analysis of the sentences. Previous studies identified only the relations among the brain region mentions, which were missing two important aspects: the directionality of the relation and unique brain region entities to obtain the overall interaction map of a brain region. We believe it is an important asset to identify the inputs and outputs of a brain region to better understand the impact of that relation on any functionality. For this purpose, we obtain the directionality of the relations directly from the pattern keywords and secondly we generate a brain region dictionary to be able to match the brain regions in the publications. The brain region entities are identified and normalized by utilizing this dictionary. With the directionality information and the relations of each brain region entity, we create a connectivity graph to show the connections and directions among the brain regions.

Consider the following sample: It reveals a relation between "dorsal midline thalamus" and "hypothalamus" brain regions with the pattern keyword of "projection from". We extract these brain regions using NLP techniques which will be shown in detail in the following sections. We also identify the direction of the relation directly from the pattern keyword. This sentence shows the direction of the relation in the pattern as "from Brain Region A to Brain Region B", therefore the projection is from dorsal midline thalamus to hypothalamus for this relation.

"There is a substantial projection from the dorsal midline thalamus to the hypothalamus, which appears from the retrograde tracer labeling to originate primarily in Pa." [1]

As a case study, we focus on a specific brain region, the paraventricular nucleus of the thalamus (PVT), which belongs to midline and intralaminar group of thalamic nuclei and has long been considered to have a non-specific effect on cortical arousal. PVT is one of the core components in the circadian timing system and receives input from all major components of this system. There are several studies which investigated the anatomical and chemical relations of the PVT. We would like to better understand the role of PVT in the circadian timing system and on a variety of psychological functions like fear, drug addiction, arousal, attention, motivation, etc. With this motivation, we aim to extract the anatomical, chemical and functional connections of the PVT brain region. As a starting point, we focus on the neuroanatomical relations and try to generate a big picture of the relations among the brain regions. Our main reason for choosing the PVT in particular is due to recent studies that attribute more specific functions to this group of thalamic nuclei because of their rich neuroanatomical and neurochemical projections [1–3].

For system development and evaluation of our study, we use the WhiteText corpus [4–6] consisting of abstracts from the neuroscience domain. We present our results on a manually annotated corpus of 14 full text articles relevant to a specific brain region (PVT). Finally, as a case study, we apply our method to extract neuroanatomical relations from articles relevant to PVT in PubMed.

This research has made the following contributions to the neuroanatomical relation extraction domain:

- The first fully linguistic based automated relation extraction system for neuroanatomical relation extraction. Previous studies in this domain were based on machine learning methods with some additional simple linguistic rules.
- A brain region dictionary which is used to recognize the brain region entities in the publications and is made publicly available for future neuroscience text mining studies. Since there is no standard use of brain region names in publications, we gathered the synonyms and the acronyms of the brain regions into a dictionary. Even though there are ontologies that contain all this information, our dictionary

is more text-mining oriented as it will be shared in more detail in Discussions section.

- Directionality of the relations between the brain regions.
- PVT connectivity graph: We automatically extracted the relations from the existing PVT-specific publications in PubMed and displayed the interactions of the brain regions in a connectivity graph.
- We have evaluated our system on PVT-specific annotated corpus and obtained 75.78% precision and 37.89% recall with 50.52% F-Measure values. When we measure the same methodology from the dictionary by matching the extracted relations directly from the dictionary instead of the annotated corpus, the precision level increased to 87.58% and recall reached 43.79%.
- We examined full text publications in addition to abstracts and obtained higher recall values compared to abstracts.
- Manually annotated sentences in the full-text PVT corpus are also made publicly available for future neuroscience text mining studies.

The rest of the study is organized as follows: Section 2 gives information about the related work on relation extraction and NLP domains. Details of the our methods that are used for relation extraction and preparation of the data are shared in Section 3. Within this section, the dictionary creation, the pattern selection are explained in detail and at the end of the section, the system development steps are introduced briefly. In Section 4, the evaluation of our linguistic relation extraction system is given and also compared with the previous studies. PVT Case study results are also shown in this section. Finally, we conclude with our findings and discussion points in Section 5. In this section we also identify future direction for this work.

2. RELATED WORK

Most previous studies on text mining in the biomedical domain have focused on extracting information about proteins and genes from scientific publications. Shared tasks such as BioCreative [7, 8] and BioNLP [9–11] have boosted research in this area. Both rule-based [12, 13] and machine learning based methods [14, 15] have been proposed for identifying names of proteins/genes in scientific texts. Several approaches ranging from entity co-occurrence [16, 17] and pattern matching based methods [18] to more complex natural language processing and/or machine learning based methods have been proposed for extracting the relations among proteins [19–24].

Developing text mining methods in the neuroinformatics domain for identifying brain region entities and mining the neuroanatomical relations among them is a relatively new research topic, compared to the more widely studied areas of biomedical text mining focusing on genes, proteins, and diseases. Only a handful of studies have been conducted in neuroscience text mining, most of which adapt and extend the methods proposed in the well-studied area of protein-protein interaction extraction. In the context of the Neuroscholar system, which is one of the first studies tackling the use of advanced natural language processing methods for neuroscience data mining, Burns et al. [25] extracted neuroanatomical information from tract-tracing experiments with an F-Measure of 79% on identifying the mentions of five types of neuroscience named entities related to tract-tracing-experiments. They used conditional random fields (CRF) with a feature set utilizing morphological, lexical, syntactic and semantic information on a manually annotated corpus of 1047 sentences from 21 documents. French et al. [4] had a similar CRF based approach with a richer feature set and reported 92% precision and 86% recall on the task of identifying brain region mentions in text. Particularly for this purpose, they have created the Whitetext corpus which includes 1,377 abstracts with the annotated 18,242 brain regions.

Even though Neuroscholar was one of the first attempts to extract neuroanatomical relations, the evaluation results for the connectivities were not reported. The

first study with the evaluation results in this domain was conducted by French et al. (2012) [5]. They have focused on the connectivity between the entities and applied both co-occurrence based methods and kernel-based supervised machine learning methods, which have originally been proposed for extracting protein-protein interactions. With the co-occurrence based methods, they checked the existence of two brain regions in a sentence, and also in the abstract. Additionally, the presence of one of the connectivity related keywords (projection, efferent, pathway, etc.) was second checkpoint in their studies for co-occurrence. French et al. (2012) [5] obtained high recall and low precision with the co-occurrence based methods. At the abstract level, they obtained 100% precision and 2.2% recall values. On the other hand, the precision increased to 13.3% and the recall decreased to 72.4% at the sentence level. For kernel based methodologies, they used the framework of Tikk et al. [23] which evaluated nine different kernel based methods on the task of protein-protein interaction extraction. The framework of Tikk et al. [23] later on became the base evaluation framework in different tasks such as drug-drug interaction extraction [26]. These kernels are the similarity functions that are used for pattern analysis by comparing two entities and computing their similarity. French et al. [5] applied seven of these kernel methods and they obtained their best results with the shallow linguistic kernel by recalling 70.1% of the connectivities with 50.3% precision. Shallow Linguistic Kernel (SLK) is mainly based on the shallow linguistic processing of the sentences such as tokenization, part of speech tagging and lemmatization. It takes two different entities (brain region entities) and decides whether there is a relation among them by using the shallow linguistic information at the local (neighbouring words) and global context (sentence level). They have also compared their extracted connectivity results with the Brain Architecture Management System (BAMS) and found that 63.5% of the extracted connectivities exist in BAMS.

Recently, within the scope of WhiteText project, French et al. shared an enhanced corpus with 1,828 newly annotated abstracts. French et al. [6] tested their approach on this corpus and obtained similar findings to their previous studies with a precision of 51% and recall of 67%. Similarly, Richardet et al. [27] built their research on this approach by using the kernels. In addition to kernels they applied some filters and lexical rules that are developed according to the sentence structures. The proposed

filters are mostly applied in order to remove the unlikely brain region connections. The sentences were skipped if they had more than seven brain region mentions or the count of characters were more than 500, etc.. On the rule-based extension, they defined 9 basic rules using the Apache UIMA Ruta workbench [28]. These rules are basic patterns (i.e., projection from Brain Region A to Brain Region B) and mainly depend on the surface structures of the sentences such as the locations of the brain regions in the sentences. They have improved the precision values with the filter and the rules with the cost of reduced recall. Since there are several evaluation values for this study, the details can be found in the Results section. Finally, Vasques et al. [29] extended this work to find the targets of a seed in tractography projects. They have selected 3 brain regions, the internal globus pallidus, the subthalamic nucleus and the nucleus accumbens and made systematic connectivity review in the literature and compared this with their automated text mining study.

With our knowledge-based approach, we focus on automated neuroanatomical relation extraction by using NLP techniques. Differently from the previous studies, it's fully linguistic based. We leverage constituency and dependency parse trees of the sentences and extract the relations by using a brain region dictionary. On machine learning based methodologies, the main approach is to extract the relations, whereas on our linguistically motivated approach, we also get the directions of the relations by using pre-defined patterns. The directionality of the relations are used in the connectivity graph to display the circuits of the brain regions to better understand the interactions and their impacts on body functions. We also evaluated our results with the Whitetext corpus to compare with the supervised and semi-supervised previous studies.

3. MATERIALS AND METHODS

3.1. Data Preparation

3.1.1. Corpus

Two different corpora are used in this research. The first is a corpus of PVT related publications which contains 558 publications retrieved from PubMed with a specific query on this brain region. This corpus is mainly used for the PVT case study to test the system that we have developed. Second corpus contains the articles from Journal of Comparative Neurology and includes 3,205 abstracts that are manually annotated for Whitetext Project [4–6]. This corpus includes brain region mentions and connectivity information among these brain regions.

3.1.1.1. PVT Corpus. The PVT corpus is used in two different ways during the evaluation. Abstracts of the 451 publications which are not publicly available and 107 publicly available full text publications constituted the first data set and provided the basis for our application on PVT Case Study. Secondly, 14 of these full text papers were selected by domain experts and fully annotated with brain region mentions and connectivity statements. The evaluation for these two datasets can be found under Results section. PubMed IDs of the publications, with abstracts and publicly available full text publications can be found as supplementary data[1]

The PVT corpus is derived from the publications that are fetched from PubMed. PubMed[2] is a database of citations and abstracts from MEDLINE (Medical Literature Analysis and Retrieval System Online)[3] and life science journals for biomedical literature. This database is maintained by the US National Library of Medicine. PubMed also provides an E-Utilies API [30] that allows users to query the database with dif-

[1] *https://github.com/erincgokdeniz/relation_extraction*
[2] *http://www.ncbi.nlm.nih.gov/pubmed*
[3] *https://www.nlm.nih.gov/pubs/factsheets/medline.html*

ferent parameters. This API is mainly a structured interface to Entrez [30] query and database system which is a search engine on top of health sciences databases at the National Center for Biotechnology Information (NCBI). In our research we have leveraged E-utilies API by using several queries to retrieve the list of PubMed IDs, abstracts of the publications in xml or json formats.

In our research, we needed to retrieve PVT related publications from PubMed to be used in the PVT case study. For this purpose, we focused on some keywords for our search on PubMed. "paraventricular", "thalamic", "thalamus" words are the main keywords on our query and we decided that the results should not include publications about hypothalamus. Therefore we filtered out "hypothalamus" or "hypothalamic" keywords. The final query that we used for publication retrieveal on PubMed was:

("paraventricular"[All Fields] AND ("thalamic"[All Fields] OR "thalamus"[All Fields]) AND ("nucleus"[All Fields] OR "nuclei"[All Fields]) NOT "hypothalamus"[All Fields] NOT "hypothalamic"[All Fields]) OR ("Paraventricular Thalamic Nucleus"[All Fields] OR "paraventricular nucleus of thalamus"[All Fields] OR "paraventricular nucleus of the thalamus"[All Fields] OR "paraventricular thalamus"[All Fields])

PubMed advanced search builder allows users to query particular metadata information about the publications, such as author, MeSH (Medical Subject Headings), etc. In our search, we used "All Fields" since we were interested in all publications that might be related with PVT. In this query, following terms are searched explicitly in the metadata of the publications:

- "paraventricular" AND "thalamic/thalamus" AND "nuclei/nucleus" which does not include any "hypothalamic/hypothalamus" field as metadata
- "paraventricular thalamic nucleus"
- "paraventricular nucleus of thalamus"
- "paraventricular nucleus of the thalamus"

By 14th of August 2015, this query on PubMed retrieved 558 publications and

the PVT case study was built on top of this result set. During the documentation period of this research, the same query resulted in 569 PVT related publications on 14th of December 2015.

As the first standalone application of our research, we have created a tool to retrieve the IDs of the publications with a query and fetch the abstracts with the list of PubMed IDs. As first part of the application, we have used E-Utilies API's ESearch functionality with following parameters to retrieve the PubMed IDs of the related publications:

- Url : http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi
- Database : pubmed
- Return Mode : json
- Return Max : 1000
- UseHistory : y
- WebEnv : gkdnz
- Term : {PVT query explained above}

The response of this Http POST request is a json file with the IDs of the related publications, which can be seen in Figure 3.1.

```
{
  - header: {
        type: "esearch",
        version: "0.3"
    },
  - esearchresult: {
        count: "569",
        retmax: "569",
        retstart: "0",
        querykey: "1",
        webenv: "NCID_1_28622430_130.14.22.215_9001_1450260718_173766653_0MetA0_S_MegaStore_F_1",
      - idlist: [
            "26578902",
            "26538811",
            "26536818",
            "26481320",
            "26475506",
            "26472643",
            "26470810",
            "26455867",
            "26417679",
            "26398809",
            "26314785",
            "26262826",
            "26255593",
            "26228683",
            "26223289",
            "26194914",
            "26136671",
            "26096647",
            "26056031",
            "26042201",
            "26010947",
            "26008155",
            "25910577",
            "5068788",
            "18421830"
        ],
        translationset: [ ],
      + translationstack: […],
        querytranslation: "("paraventricular"[All Fields] AND ("thalamic"[All Fields] OR "thalamus"[All Fields]) AND ("nucleus"[All Fields]
        Fields]) OR ("Paraventricular Thalamic Nucleus"[All Fields] OR "paraventricular nucleus of thalamus"[All Fields] OR "paraventricular
        Fields])"
    }
}
```


Figure 3.1. Http response of the E-Search API call

After retrieving the list of ids of the publications, we downloaded the abstracts by using E-Fetch functionality of the E-Utilies API. This time we passed the PubMed IDs retrieved from the first step to the API call.

- Url : http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi
- Database : pubmed
- Version : 2.0
- Return Mode : abstract
- Return Type : text
- Return Max : 1000
- WebEnv : gkdnz
- Id : {each PubMed ID that is retrieved from ESearch}

The response of the EFetch Http request is a text which contains publishing details, abstract, author information and PubMed ID (Figure 3.2) .

```
1. Front Syst Neurosci. 2015 Oct 30;9:145. doi: 10.3389/fnsys.2015.00145.
eCollection 2015.

Prefrontal-amygdala fear networks come into focus.

Arruda-Carvalho M(1), Clem RL(1).

Author information:
(1)Fishberg Department of Neuroscience and The Friedman Brain Institute, Icahn
School of Medicine at Mount Sinai New York, NY, USA.

The ability to form associations between aversive threats and their predictors is
fundamental to survival. However, fear and anxiety in excess are detrimental and
are a hallmark of psychiatric diseases such as post-traumatic stress disorder
(PTSD). PTSD symptomatology includes persistent and intrusive thoughts of an
experienced trauma, suggesting an inability to downregulate fear when a
corresponding threat has subsided. Convergent evidence from human and rodent
studies supports a role for the medial prefrontal cortex (mPFC)-amygdala network
in both PTSD and the regulation of fear memory expression. In particular, current
models stipulate that the prelimbic (PL) and infralimbic (IL) subdivisions of the
rodent mPFC bidirectionally regulate fear expression via differential recruitment
of amygdala neuronal subpopulations. However, an array of recent studies that
employ new technical approaches has fundamentally challenged this interpretation.
Here we explore how a new emphasis on the contribution of inhibitory neuronal
populations, subcortical structures and the passage of time is reshaping our
understanding of mPFC-amygdala circuits and their control over fear.

PMCID: PMC4626554
PMID: 26578902  [PubMed]
```

Figure 3.2. Http response of the E-Fetch API call

As part of the PVT Corpus, 14 full text PVT related papers were manually annotated. This dataset was taken as gold standard for one of our PVT experiments and the results of evaluation can be found at Section 4.2.1. Several samples of the annotated brain regions are given in Table 3.1.

Table 3.1. Sample sentences from the annotated PVT corpus.

Sentence	**Brain Region 1**	**Brain Region 2**
These experiments confirm projections from Pa, Pt and other midline nuclei to the amygdala. [1]	Pa	amygdala
These experiments confirm projections from Pa, Pt and other midline nuclei to the amygdala. [1]	Pt	amygdala
In addition, we found that the aPVT was strongly innervated by the ventral subiculum but this projection largely did not involve the pPVT. [2]	aPVT	ventral subiculum
The paraventricular thalamus (PVT) , a midline thalamic nucleus , receives dense innervations from lateral hypothalamic orexin neurons (Peyron et al , 1998 ; Kirouac et al , 2005) and is involved in the regulation ofcognition , anxiety , emotionality and addiction behaviors (Huang et al , 2006 ; Li et al , 2009 , 2010a , 2010b and 2011) [31]	PVT	hypothalamic orexin neurons

3.1.1.2. Whitetext Corpus. Whitetext Corpus mainly consists of publications from Journal of Comparative Neurology which are retrieved from PubMed. French et al. provided 2 datasets as part of Whitetext corpus. The first data set was provided in 2009 with 1,377 abstracts that contain 3,097 connectivity relations, and the second dataset was released in 2015 containing 1,828 abstracts with 2,111 interactions. They have provided annotated data on The General Architecture for Text Engineering (GATE) tool and also in xml format. In our study, we used the airola xml file which contains the connectivity information. We parsed the airola xml using Java SAX (the Simple

API for XML) Parser which is an event-based parser for xml documents.

As can be seen in Figure 3.3, the relations are given in xml format for each document and sentence. It includes sentences only that have at least two brain regions. The brain region mentions are tagged as "entity" and the possible relations are tagged as "pair". If the interaction attribute of the pair is true, then the entities given in the pair are considered to be related.

```
<sentence id="WhiteTextNegFixFull.d1.s7" origId="1542" text="The medial subthalamic
nucleus(STh) sends strong projections to the medial part of the entopeduncular nucleus and the
adjacent lateral hypothalamic area.">
  <entity charOffset="4-29" id="WhiteTextNegFixFull.d1.s7.e0" origId="14656" text="medial
  subthalamic nucleus" type="Individual_protein"/>
  <entity charOffset="68-108" id="WhiteTextNegFixFull.d1.s7.e1" origId="14635" text="medial
  part of the entopeduncular nucleus" type="Individual_protein"/>
  <entity charOffset="127-151" id="WhiteTextNegFixFull.d1.s7.e2" origId="14697" text="lateral
  hypothalamic area" type="Individual_protein"/>
  <pair interaction="True" id="WhiteTextNegFixFull.d1.s7.p0" e1="WhiteTextNegFixFull.d1.s7.e0"
  e2="WhiteTextNegFixFull.d1.s7.e1"/>
  <pair interaction="True" id="WhiteTextNegFixFull.d1.s7.p1" e1="WhiteTextNegFixFull.d1.s7.e2"
  e2="WhiteTextNegFixFull.d1.s7.e0"/>
  <pair interaction="False" id="WhiteTextNegFixFull.d1.s7.p2" e1="WhiteTextNegFixFull.d1.s7.e2"
  e2="WhiteTextNegFixFull.d1.s7.e1"/>
```

Figure 3.3. Whitetext Corpus : Airola xml file that includes the brain region entities and connectivity information

In our study, first data set is used during our system development phase and the second data set is used as test set.

3.1.2. Creation of a Brain Region Dictionary

We used a dictionary-based approach to identify the brain region entities that participate in neuroanatomical relations and normalized their mentions to canonical (unique) names. We constructed a dictionary of brain regions including their acronyms and synonyms, where an acronym is the abbreviation of the brain region entity and a synonym is a similar word or phrase used for the same brain region entity in text. A portion of the created dictionary with sample entries is shown in Table 3.2. During the dictionary creation step, we initially gathered a dictionary of 892 brain regions and 562 acronyms from the NeuroNames ontology [32] and NeuroLex [33], which is a dynamic lexicon of neuroscience concepts. Additionally, we investigated a set of

neuroscience publications to identify and compile the different usages of brain region mentions in the neuroscience literature. We unified the brain region mentions that we extracted and chose the most common usages of the brain regions from the ontologies. The resulting enriched dictionary contains 3,044 brain region entities with their synonyms and acronyms. The created brain region dictionary is made publicly available as supplementary data for future text mining studies.[4]

Table 3.2. Brain region dictionary.

Brain Region	**Acronym**	**Synonym**
parietal lobe	PL	parietal cortex, parietal region, lobus parietalis
suprachiasmatic nucleus	SCN	suprachiasmatic nuclei
cingulate gyrus	CgG	cingular gyrus, cingulate area, cingulate region, gyri cinguli, gyrus cinguli
subthalamus	SbTh	subthalamic region, ventral thalamus, thalamus ventralis
superior frontal gyrus	SFG	marginal gyrus, superior frontal convolution, gyrus frontalis superior
parabrachial nucleus	-	parabrachial nuclei,parabrachial
paracentral nucleus	PC	paracentral thalamic nucleus, nucleus paracentralis, paracentral nucleus of the thalamus,paracentral
central medial nucleus	CM	central medial thalamic nucleus, nucleus centralis medialis, centralis medialis,central medial nucleus of the thalamus,central medial

3.1.3. Defining the Patterns

We manually designed a set of patterns, which are strings of keywords that mostly reveal a relation, when there are two or more brain region entities in a sentence. Especially in the neuroanatomical connections, we noticed that there are patterns like "projection to, innervate, receive input from" which are mostly used when there is

[4] *https://github.com/erincgokdeniz/relation_extraction*

a relation among brain regions. With this approach, we defined the list of patterns and individually assessed their existence when there is a relation in the sentences. For example, the following sentence contains a relation between "dorsal midline thalamus" and "accumbens nucleus" brain regions with the pattern of "projection to".

"An anterograde tracer injection into the dorsal midline thalamus revealed strong projections to the accumbens nucleus. " [1]

As shown in the second sentence, there are two relations connected with "innervate" keyword. The first relation is among pPVT and prelimbic cortex, and the second one is among pPVT and agranular portions of posterior insular cortex.

"For example, the pPVT was found to be distinctively innervated by the anterior most aspect of the prelimbic cortex and the agranular portions of the posterior insular cortex . " [2]

Neuroanatomical relations are in general signaled by pattern keywords. Since each keyword can have different prepositional suffixes (e.g. projection from, projection of, projection to) and different tenses (e.g. projects to, projecting to, projected to), regular expressions are used to cover the different usages of the patterns. As shown in the below regular expression for the pattern "project to", the patterns are considered to be case insensitive and are likely to contain additional words between their original keywords (i.e., between "project" and "to").

```
(?i)project(ing|s|ed){0,1} ((\w)* ){0,2}to
```

This sample regular expression matches sentences that contain the word "project", "projecting", "projects", or "projected", followed by zero, one, or two additional words, followed by the word "to". The following sentence is an example of a sentence that matches this regular expression. With this regular expression we match the sentences that have a structure of "Brain Region A projects to Brain Region B".

"Taken together, these cases confirm the anterograde data that Pa projects strongly to the accumbens nucleus and the rostromedial caudate nucleus. " [1]

For each pattern, while defining the regular expression, there has been an investigation phase to understand how they are used in the sentences and how they are related with the neuroanatomical relations. We started by creating the simple version of the regular expression, and then extended it according to the generic structure that patterns are used in the sentences of the publications. The steps below illustrates how the "receive input from" pattern was extended:

1. We created a basic regular expression for "receive input from".

Regex: `(?i)receiv(e|es){0,1} input(s){0,1} from`

Sentence: *"The spinal cerebellum (anterior lobe, paramedian lobule and pyramis) receives input from several separate regions in the dorsal accessory nucleus, the medial accessory nucleus and portions of the principal nucleus. "* [34]

2. We extended the regex since there were different tenses of the receive word.

Regex: `(?i)receiv(e|es|ing|ed){0,1} input(s){0,1} from`

Sentence: *"The results revealed that the pretectal nucleus of the optic tract received inputs from medial prestriate cortex, dorsomedial part of area 19, OAa, and PGa. "* [35]

3. Finally, we also noticed that in some sentences there were also additional words between the original keywords of the pattern. Considering the different usages, we limited the number of words between "receive" and "input" to 4 and between "input" and "from" to 3.

Regex:

```
(?i)receiv(e|es|ing|ed){0,1} ((\w)* ){0,4}input(s)
{0,1} ((\w)* ){0,3}(from)
```

Sentence: *"The PVT receives large and distinct inputs from several areas of the hypothalamus, including the suprachiasmatic , arcuate , dorsomedial and ventromedial nuclei and preoptic and lateral hypothalamic areas "* [36]

Sentence: *"The lateral ventral striatum receives input primarily from areas 24b, 24b' and 23b and medial portion of area 24c. "* [37]

The list of defined patterns and their corresponding regular expressions are shown in Table 3.3.

3.2. Neuroanatomical Relation Extraction

We developed a linguistically motivated knowledge-based approach for neuroanatomical relation extraction. The workflow of the proposed approach is shown in Figure 3.4. Automated relation extraction in general relies on finding the correct sentences that describe an interaction between brain regions. For this purpose, as a first step, the publications (abstracts or full text) are split into sentences. Next, these sentences are fed into our system and by using pre-defined patterns a list of candidate sentences, which might contain relations are selected. Then, NLP techniques are used to identify the brain regions that are described as being related in these sentences. We use the dependency and constituency parse trees of the sentences and apply linguistic rules over these parse trees to extract the portions of the sentences that are likely to contain brain region entities participating in a neuroanatomical relation, i.e., the candidate brain region entities. Based on predefined patterns, we also identify relation directionality by labeling the candidate brain region entities as "agents" or "targets". For example, from a sentence like "X receives input from Y", we obtain the information that Y is the agent and X is the target of the relation, i.e., the directionality of the relation is Y→X. In the relation decision step, the candidate brain region entities are searched in the

Table 3.3. List of the patterns and their corresponding regular expressions

Pattern	**Regular Expression**
innervate	`(?i)innervat(e|es|ing){1}`
innervation of	`(?i)innervation(s){0,1} of`
projection to	`(?i)projection(s){0,1} to`
projection to from	`(?i)projection(s){0,1} to ((\\w+)\\s){0,8} from`
projection of	`(?i)projection(s){0,1} of`
projection target of	`(?i)projection target(s){0,1} of`
projection from	`(?i)(the ){0,1}projection(s){0,1} from`
projection from to	`(?i)projection(s){0,1} from ((\\w+)\\s){0,8} to`
project to	`(?i)project(ing|s|ed){0,1} ((\w)* ){0,2}to`
project into	`(?i)project(ing|s|ed){0,1} ((\w)* ){0,2}into`
project from	`(?i)project(ing|s|ed){0,1} from`
project from to	`(?i)project(s|ed|ing){0,1} from ((\\w+)\\s){0,8} to`
receive input from	`(?i)receiv(e|es|ing|ed){0,1} ((\w)* ){0,4} input(s){0,1} ((\w)* ){0,3}(from)`
receive fiber from	`(?i)receiv(e|es|ing|ed){0,1} ((\w)* ){0,4} fiber(s){0,1} ((\w)* ){0,3}(from)`
receive innervation from	`(?i)receiv(e|es|ing|ed){0,1} ((\w)* ){0,4} innervation(s){0,1} ((\w)* ){0,3}(from)`
receive [ae]fferent from	`(?i)receiv(e|es|ing|ed){0,1} ((\w)* ){0,4} [ae]fferent(s){0,1} ((\w)* ){0,3}(from)`
send via to	`(?i)(((sen(d|ds|ding|t)) ((\w)* )*via ((\w)* )*to))`
send from	`(?i)(((sen(d|ds|ding|t)) from ((\w)* )*to))`
send to	`(?i)(sen(d|ds|ding|t)) ((\w)* ){0,2}to`
travelling from to	`(?i)travel(s|ling){0,1} ((\w)* ){0,2}from ((\w)* ){0,5}to`
travel through	`(?i)travel(s|ling){0,1} ((\w)* )*through`
exit through	`(?i)exit(s|ing){0,1} ((\w)* )*through`
exit from	`(?i)exit(s|ing){0,1} ((\w)* )*from`

brain region dictionary, and a neuroanatomical relation is identified if the candidate agent and target are matched in the dictionary. Finally, the agents and targets of the identified neuroanatomical relations are normalized to their canonical names using the brain region dictionary and a directional brain region connectivity graph is created. The graph can be further analyzed to generate new scientific hypothesis. The details of each step in our method are described in the following sub-sections.

3.2.1. Sentence Splitting

The Stanford Core NLP tool [38] is used for splitting the publications into their sentences. In order to have a unified structure in the sentences, some post processing is applied to the output of the Stanford Parser. For example, the Stanford Parser has a special syntax for the parenthesis (left bracket is represented as"–LRB-" and right bracket is represented as "–RRB-"). Therefore, we replaced these strings with the corresponding parenthesis signs in order to be able to match the sentences with the publication text. Additionally, specific to WhiteText corpus, Schwartz and Hearst Abbreviation Expansion algorithm [39] is applied for each sentence. This algorithm requires the replacement of short forms of the abbreviations with long forms and the addition of short form after the long form. Therefore, for each abstract, we also applied abbreviation expansion algorithm and mainly we needed this enhancement for the accurate evaluation of our findings with the Whitetext corpus.

3.2.2. Pattern-based Sentence Selection

After preparing the data, the first phase of the relation extraction is to scan the publications and extract the sentences which contain the predefined patterns. The extracted sentences at this step are the first candidates that might include the brain regions and the relations. We use the regular expressions to match the patterns in the sentences as described in Section 3.1.3. A sample sentence with a pattern can be found below:

"The NAc receives a strong dopaminergic projection from the ventral tegmental

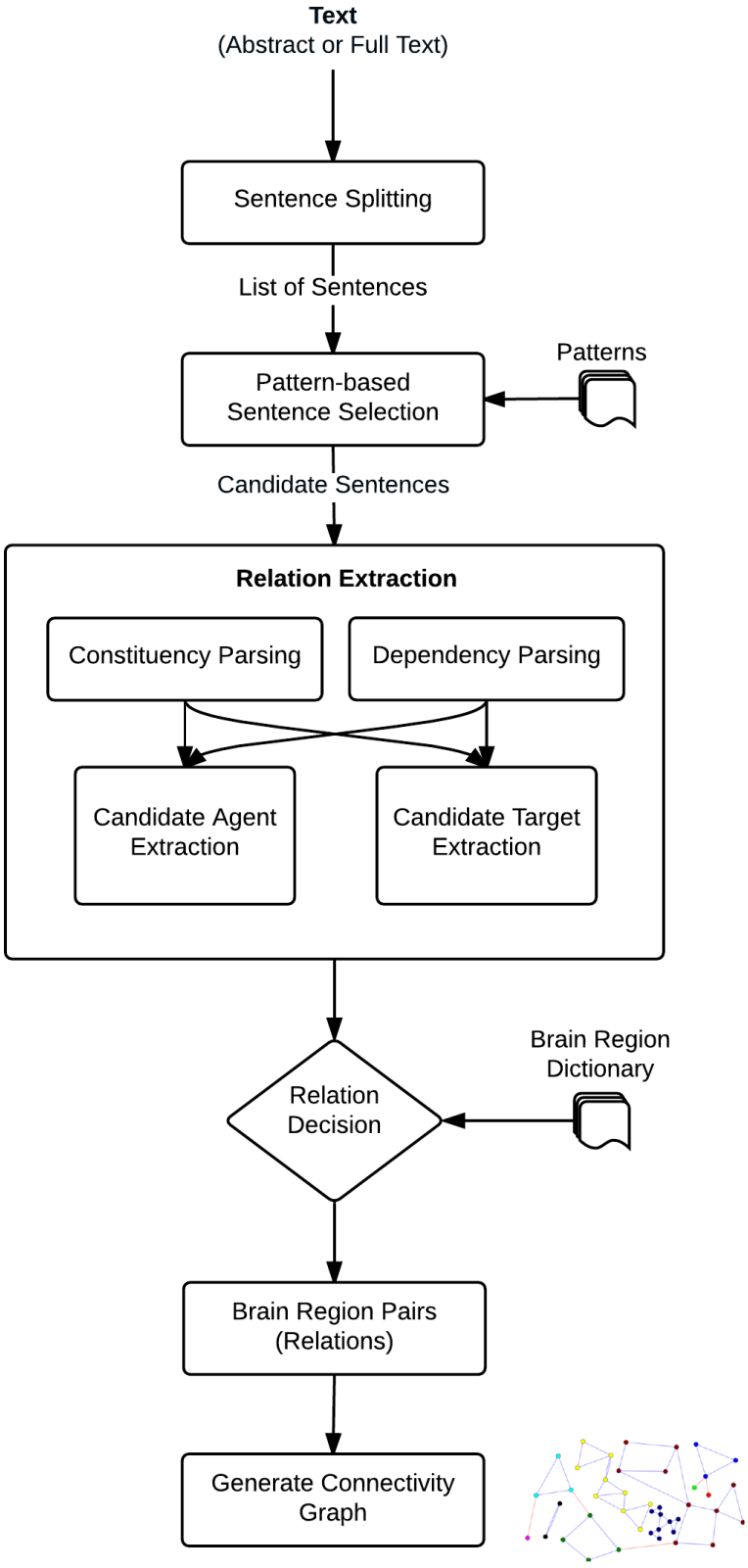

Figure 3.4. Steps to extract the neuroanatomical relations

area , and dopamine released from these projections acts upon dopamine receptors of the D1 and D2 families located on postsynaptic targets." [40]

The sentences in the publications that match these patterns (regular expressions) are selected as candidate sentences and provided as input to the relation extraction component described in the next sub-section.

3.2.3. Candidate Generation Using NLP Techniques

After generating the list of sentences which are candidates for hosting brain region relations, a detailed syntactic analysis of each sentence is done. There are two dependents of the patterns: agents and targets. If both of these dependents include brain region entities, then we consider that there is a relation between these entities. There can be more than one relation within a given sentence if dependents include more than one brain region. To be able to identify whether a dependent is an agent or target, we need the directionality of the relation and this information is gathered directly from the patterns. For example, for the patterns like "receive input from, projection from, efferent from", it is likely that the text string that follows the pattern is agent. On the other hand, for the "project into, innervate, terminate in" patterns, the same text reveals the target.

Table 3.4. Directionality of the relation is decided by the pattern and this information helps to identify the agent and the target in the sentences.

Sentence	**Agent**	**Target**
The suprachiasmatic nucleus is well known to project densely to Pa in rats [1]	suprachiasmatic nucleus	Pa
These experiments confirm projections from Pa, Pt and other midline nuclei to the amygdala [1]	Pa, Pt	amygdala

The Stanford Parser is used to syntactically parse the sentences and obtain their constituent elements [41]. One of the dependents (agent or target) in general occurs right after the pattern keyword. The constituency (phrase structure) parse tree is

traced until we reach the pattern and then we select the first Noun Phrase (NP) following the pattern in the bracketed notation of the parse tree. After finding the NP, all the leaves under this NP are used to generate the candidate dependent. In Figure 3.5, a bracketed notation of the parse tree for the "The suprachiasmatic nucleus is well known to project densely to Pa in rats" [1] sentence is presented and in Figure 3.6 the tree representation of the same sentence is shown. The identified NP is enclosed in a box in these figures.

```
(ROOT
  (S
    (NP (DT The) (JJ suprachiasmatic) (NN nucleus))
    (VP (VBZ is)
      (ADVP (RB well))
      (VP (VBN known)
        (S
          (VP (TO to)
            (VP (VB project)
              (ADVP (RB densely))
              (PP (TO to)
                (NP
                  (NP (NNP Pa))
                  (PP (IN in)
                    (NP (NNS rats))))))))))
    (. .)))
```

Figure 3.5. Bracketed notation of parse tree for the sentence: *"The suprachiasmatic nucleus is well known to project densely to Pa in rats "* [1]. First noun phrase after the pattern(project to) is selected.

In some sentences, the prepositional phrase (PP) following the detected NP modifies the NP and may contain candidate dependents for the relation. Therefore, if a detected NP is followed by a PP, which contains the keyword "including", then it is also added as part of the candidate brain region text (dependent). An example sentence is provided below.

"Studies in rats show that the caudal DR projects strongly to limbic structures including the amygdala and hippocampus." [1]

In this sentence, "limbic structures" is the NP following the pattern. Applying the above rule, we select "amygdala and hippocampus" as candidate dependents since

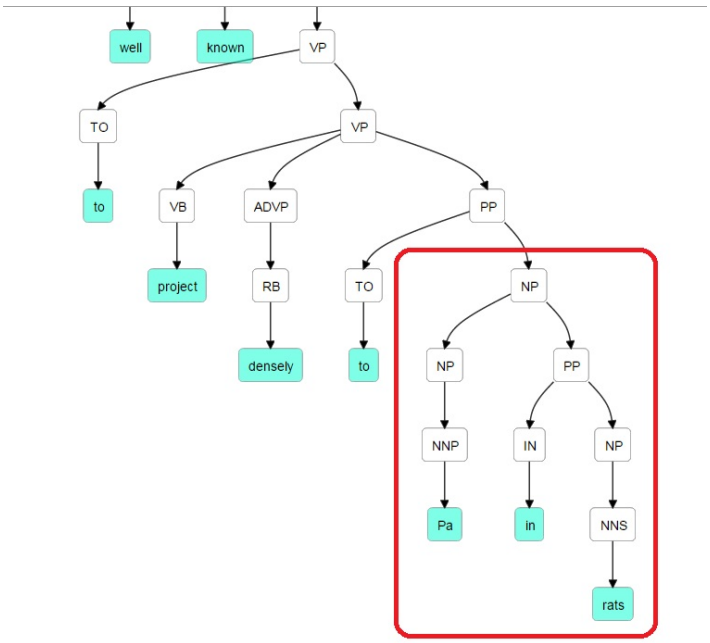

Figure 3.6. Parse Tree for the sentence: *"The suprachiasmatic nucleus is well known to project densely to Pa in rats"* [1]. Selected dependent is "Pa in rats"

they are also part of the affected brain regions, but knowing that this decision may also bring false-positives.

To find the first dependent (brain region candidate) that follows the pattern keyword, we used the constituency parser. On the other hand, for the second dependent, the text extraction phase was more complex. The second dependent can be found in different locations of the sentence. It can be at the beginning, right before the pattern, or close to the end of the sentence after the pattern. The dependency tree of a sentence can capture the long-distance relations among its words. We used the Stanford Dependency Parser [42] to analyze the dependency structures of the sentences and obtain the second candidate dependent, which does not necessarily occur close to the pattern.

The output of the Stanford Dependency Parser is the Stanford Dependencies representation , which is a description of the grammatical relationships among the words in a sentence [42]. There was also a work in progress for Universal Dependencies representation when we started to work with Stanford Parser. The version of the Stanford Parser libraries that we used in our study only supports Stanford Dependencies. And

as the representation of these dependencies we used Propagated Collapsed Dependency Tree representation in which the dependencies are collapsed into one relation in case there are prepositions, conjunctions in the relations.

A dependency is given as relation(governor-pos1, dependent-pos2) where the governor and the dependent are words in the sentence and pos1 and pos2 indicate the positions of the two words in the sentence. "relation" is one of the 50 grammatical relations defined in the Stanford Parser [42]. A sample dependency tree with SD representation can be found below:

Stanford Dependencies Representation:

```
det(AP-2, The-1)
nsubj(projected-3, AP-2)
root(ROOT-0, projected-3)
advmod(projected-3, heavily-4)
det(complex-9, the-6)
amod(complex-9, dorsal-7)
amod(complex-9, vagal-8)
prep_to(projected-3, complex-9)
advmod(projected-3, especially-11)
det(subnuclei-17, the-13)
amod(subnuclei-17, commissural-14)
conj_and(commissural-14, medial-16)
amod(subnuclei-17, medial-16)
prep_in(projected-3, subnuclei-17)
det(NTS-20, the-19)
prep_of(subnuclei-17, NTS-20)
det(nucleus-26, the-23)
amod(nucleus-26, dorsal-24)
nn(nucleus-26, motor-25)
prep_in(projected-3, nucleus-26)
conj_and(subnuclei-17, nucleus-26)
det(vagus-29, the-28)
prep_of(nucleus-26, vagus-29)
```

Figure 3.7. Stanford dependencies representation for the sentence : *"The AP projected heavily to the dorsal vagal complex, especially in the commissural and medial subnuclei of the NTS , and the dorsal motor nucleus of the vagus"* [43]

As the starting point of identifying the second dependent, when a pattern is found in a sentence, one of the dependency types below is searched in the dependency tree. The pattern keyword in these types can be either governor or dependent. The

descriptions of all the dependency types, including the ones briefly described below, can be found in the Stanford Parser dependencies manual with sample sentences and dependency trees.

(1) Direct Object (dobj): A noun phrase which is the object of the verb

"Orx/Hcrt neurons receive projections from the medial prefrontal cortex. . . " [36]

dobj(receive, projections)

(2) Nominal Subject (nsubj): A noun phrase which is the syntactic subject of a clause

"Studies in rats show that the caudal DR projects strongly to limbic structures including the amygdala and hippocampus. . . " [1]

nsubj(show, studies)

(3) Passive Nominal Subject (nsubjpass): A noun phrase which is the syntactic subject of a passive clause

"The pPVT was found to be distinctively innervated by the prelimbic cortex." [2]

nsubjpass(found, pPVT)

(4) Controlling Subject (xsubj): A controlling subject is the relation between the head of an open clausal complement (xcomp) and the external subject of that clause

"The PBN is reported to project directly to the NAcc." [44]

xsubj(project, PBN)

(5) Noun compound modifier (nn): Any noun that serves to modify the head noun in an NP

"Also, many of the projection targets of PVT neurons, including PFC and amygdala, show strong stress responses" [45]

nn(targets, projection)

(6) Reduced non-finite verbal modifier (vmod): These are used to modify the meaning of an NP or another verb

"The fifth major region projecting to the BSTvl was the brainstem." [46]

vmod(region, projecting)

3.2.3.1. Relations where the pattern keyword is in nsubj/ nsubjpass/ xsubj/nn relations.

This rule set is applied for the pattern keywords that contain nsubj, nsubjpass, xsubj, or nn type of relations. In these cases, the governor/dependent that is found in this relation is directly considered as a candidate brain region. Additionally, two different rules are applied when the pattern keyword is found in these relations.

1. If the pattern keyword is found as a dependent, then the Prepositional Modifier (prep) of the governor is retrieved. The dependent of the prep relation is selected as a candidate brain region. Then the Adjectival Modifier (amod) or Noun Compound Modifier (nn) relations are also gathered as parts of the candidate brain region. A sample sentence specific to nsubj for the extended version of this scenario will be given in the next subsection.

2. If the pattern keyword is found as a governor, all the relations that contain the dependent as a governor are selected. The dependents of these relations are retrieved as candidate brain regions. A portion of the dependency tree for a sample sentence,

for which this rule applies, is presented in Figure 3.8. The extraction steps of the candidate brain regions from this sentence are presented below.

Sentence: *"This topography is consistent with findings in rats , in which the external lateral parabrachial subnucleus projects strongly to the anterior Pa, and less so to the middle and posterior Pa (Krout and Loewy, 2000a)."* [1]

Step 1: The pattern keyword "projects" is found in a nsubj relation.

nsubj(projects-17, subnucleus-16)

Step 2: The dependent of the first step "subnucleus" is searched as a governor in all dependencies. The dependents of the identified relations are retrieved as candidate brain regions.

det(subnucleus-16, the-12)

amod(subnucleus-16, external-13)

amod(subnucleus-16, lateral-14)

nn(subnucleus-16, parabrachial-15)

Step 3: The candidate brain regions retrieved in the previous steps are returned in sorted order by their positions in the sentence.

the-12, external-13, lateral-14, parabrachial-15, subnucleus-16

3.2.3.2. Special case for nsubj where the pattern keyword is in dobj. This specific case is an extension of the rule set described in the previous subsection for nsubj relations that have a pattern keyword in a Direct Object (dobj) relation. Our candidate brain region detection algorithm starts by finding the pattern keyword in a dobj relation type.

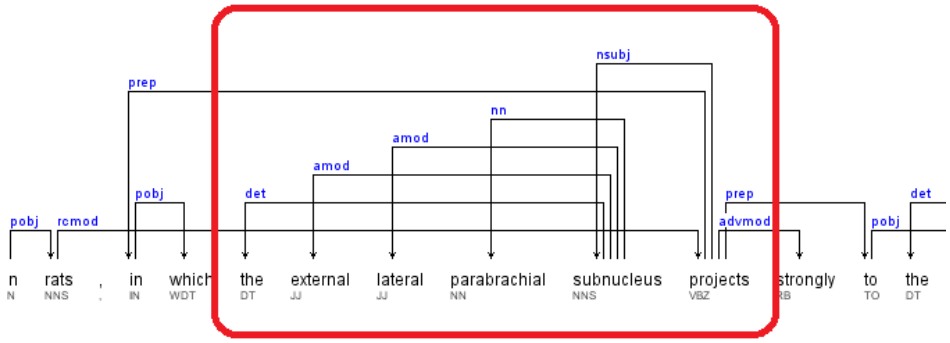

Figure 3.8. Dependency Tree for the sentence : *"This topography is consistent with findings in rats , in which the external lateral parabrachial subnucleus projects strongly to the anterior Pa, and less so to the middle and posterior Pa (Krout and Loewy, 2000a)."* [1]

Then, the governor of the dobj relation is searched as the governor of a Nominal Subject (nsubj) relation. The dependent of the nsubj relation is taken as a candidate brain region. Differently from the other nsubj cases (described in the previous subsection), in this case, the nsubj relation does not need to contain the pattern keyword.

Additionally, this rule is extended to consider the modifiers of the nominal subject. Each dependent retrieved from a nsubj relation, is searched in the Adjectival Modifier (amod), Noun Compound Modifier (nn), and Prepositional Modifier (prep) relations as a governor. If such a relation is identified, the dependent of the relation is selected as a candidate brain region. Lastly, all identified candidate brain region words are returned in sorted order based on their sentence position information.

A portion of the dependency tree of the following sample sentence is shown in Figure 3.9.

Sentence: *"An anterograde tracer injection into the dorsal midline thalamus revealed strong projections to the accumbens nucleus, basal amygdala, lateral septum, and hypothalamus."* [1]

The extraction steps of the candidate brain regions from this sentence are presented below.

Step 1: The pattern keyword “projections” is found in a dobj relation.

dobj(revealed-10, projections-12)

Step 2: The governor of the first step “revealed” is searched in the dependencies as the governor of a nsubj relation. The dependent “injection” of the identified nsubj relation is retrieved as a candidate brain region.

nsubj(revealed-10, injection-4)

Step 3: The dependent of the second step “injection” is searched in the amod, nn and prep relations as a governor. The dependents (i.e., “anterograde, tracer, and thalamus”) of the identified relations are retrieved as candidate brain regions.

amod(injection-4, anterograde-2)

nn(injection-4, tracer-3)

prep_into(injection-4, thalamus-9)

Step 4: Additionally for the prepositional modifier of the third step “thalamus”, the amod and nn relations are gathered. The dependents of the identified relations are selected as candidate brain regions.

amod(thalamus-9, dorsal-7)

amod(thalamus-9, midline-8)

Step 5: The candidate brain regions retrieved in the previous steps are returned

in sorted order by their positions in the sentence.

anterograde-2, tracer-3, injection-4, dorsal-7, midline-8, thalamus-9

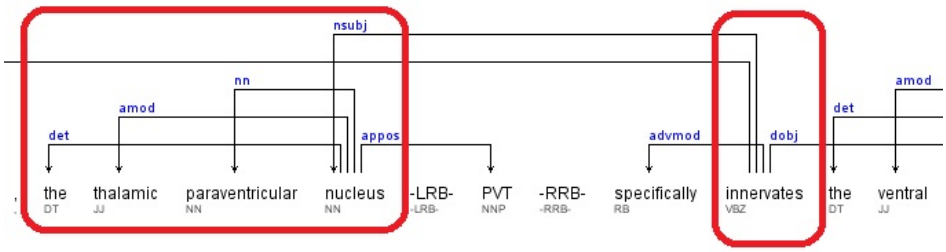

Figure 3.9. Dependency tree for the sentence: "*An anterograde tracer injection into the dorsal midline thalamus revealed strong projections to the accumbens nucleus, basal amygdala, lateral septum, and hypothalamus.*" [1]

3.2.3.3. Relations where the pattern keyword is a vmod. This group of rules first finds the Reduced non-finite verbal modifier (vmod) relations where the pattern keyword is a dependent. In the next step, the complementary relations Adjectival Modifier (amod) or Noun Compound Modifier (nn) involving the governor of the identified vmod relation are retrieved. If any relation is found in this step, the dependent of the relation is retrieved as a candidate brain region. A sample sentence with the steps for retrieving the candidate brain regions from this sentence are outlined below.

Sentence: *"Here, we combined neuronal tract-tracing using the retrograde tracer cholera toxin b (CTb) with Fos expression to examine the effect of acute nicotine administration on orexin neurons projecting to the basal forebrain or PVT."* [47]

Step 1: The pattern keyword (i.e., projecting) is found in a vmod relation as a dependent, the governor (i.e., neurons) is selected as a candidate brain region.

vmod(neurons-30, projecting-31)

Step 2: The governor of Step 1 (i.e., neurons) is searched as a governor in amod and nn dependencies and the corresponding dependents (i.e., orexin) are collected.

nn(neurons-30, orexin-29)

Step 3: The candidate brain regions retrieved in the previous steps are returned in sorted order by their positions in the sentence.

orexin-29, neurons-30

3.2.4. Relation Decision

After the candidate generation phase (Section 3.2.3), the identified candidates are searched in the Brain Region Dictionary, in which a brain region is represented with its name, acronyms, and synonyms. A neuroanatomical relation is extracted, if at least two different brain region entities are matched in the dictionary, and one of them has the role of agent, whereas the other has the role of target. For the success of the dictionary matching process, we applied several steps as described below.

First we checked whether there was a full match between the agent/target and the dictionary entity (Step 1 in Figure 3.10). If there is no match, this might mean that the agent/target consists of more than one brain region. Therefore, we split the text into strings from the conjunctions "and" and "or", and the punctuation marks "comma" and "semicolon" (Step 2 and Step 3.a in Figure 3.10) (i.e., "the NAS, PFC, and amygdala" text is split as "NAS", "PFC", and "amygdala"). In addition, for each text string, there is a post-processing step, which removes some commonly used words like "of", "the", "area", "part", and "pole". If there is still no match for that text string, two more steps are applied. First, a substring search for this text string is done in all dictionary entities and the candidates are retrieved and secondly this text string is split into tokens from the spaces and then, each token is searched in the brain region dictionary separately (Step 3.b in Figure 3.10).

After finding the brain regions from the dictionary, only the longest version of the overlapping brain regions are selected. For example, if we retrieve "thalamus", "midline thalamus", and "amygdala" as candidates for one of the dependents (i.e.,

Step	Candidate Agent/Target Found by the Application	Annotated BRs in publication	Available Entities in Dictionary	Matching Type
1	Text: "thalamus" Candidate Brain Regions: thalamus	Thalamus	thalamus	Full Match
2	Text: "the NAS, PFC and amygdala" Candidate Brain Regions: NAS PFC amygdala	NAS PFC Amygdala	NAS PFC Amygdala	Full Match
3	Text: "dorsal thalamus and SCN"	dorsal midline thalamus SCN	dorsal midline thalamus thalamus midline thalamus SCN	Full Match for SCN
3.a	Candidate Brain Regions: (tokenization by "and") dorsal thalamus SCN			No match for dorsal thalamus
3.b	Candidate Brain Regions: (tokenization by space) dorsal thalamus			Partial match for thalamus

Figure 3.10. Extraction of agent/target and dictionary matching

target or agent), then we select “midline thalamus” and “amygdala” as the extracted brain regions. “thalamus” is not selected, since it overlaps with “midline thalamus”, which is a longer match.

As the last step of relation extraction we define whether the extracted brain regions are “full match” or “partial match” when compared with the annotated data set. If an extracted brain region matches only a part of the brain region in the annotated sentence, this is considered as a partial match. For example, assume that the application retrieves “thalamus” as a brain region and the manually annotated brain region text in the sentence is “dorsal midline thalamus. In this case, the extracted brain region is considered as a partial match and the evaluation results are shown as ‘Lenient’ in Section 4, which means that the extracted brain region might be equal to or part of the annotated brain region.

The roles of agent and target are determined based on the pattern of the sentence. For each pattern in Table 3.3, we define a rule that determines the direction of the relation. For example, in the sentence “The present study found a projection from the lateral portion of parabrachial nucleus to the anterior PVT” [1], a relation between “parabrachial nucleus” and “anterior PVT” is identified and the directionality is from “parabrachial nucleus” to “anterior PVT”, i.e., “parabrachial nucleus” is the agent and “anterior PVT” is the target.

Let us summarize the steps of our relation extraction approach with a sample sentence.

First, sentences matching at least one of the patterns in Table 3.3 are retrieved as candidate neuroanatomical relation describing sentences. For example, the “projection to” pattern is matched with the sample sentence below.

“An anterograde tracer injection into the dorsal midline thalamus revealed strong projections to the accumbens nucleus, basal amygdala, lateral septum, and hypothalamus .” [1]

The defined pattern also provides the directionality of the relation. Therefore, the candidate relation is in the form of "[agent] projection to [target]".

After matching the pattern in the sentence, by using the constituency parser, the closest NP following the pattern is retrieved as the first candidate dependent (i.e., agent or target).

Candidate target: "the accumbens nucleus, basal amygdala, lateral septum, and hypothalamus"

The second candidate dependent is retrieved by using the dependency parser. The pattern keyword is found in a dobj relation (dobj(revealed-10, projections-12)) and the second dependent is extracted as described in Section 3.2.3.2.

Candidate agent: "anterograde tracer injection dorsal midline thalamus".

The brain region dictionary is used to extract the relations among the candidate agents and targets. For the candidate target, there is no exact match for "anterograde tracer injection dorsal midline thalamus". "thalamus", "midline thalamus" and "dorsal midline thalamus" is matched from the entities of the dictionary. Since dorsal midline thalamus is more specific, we retrieve the longest version possible as target brain region: "dorsal midline thalamus". For the candidate agent, we tokenize the "the accumbens nucleus, basal amygdala, lateral septum, and hypothalamus" by using "comma", "semicolon", "and", and "or" as separators. As a result of this step, "lateral septum", "basal amygdala", "hypothalamus", "accumbens nucleus" are retrieved as brain regions. In the last step, we pair the brain regions and return the relations as follows.

- dorsal midline thalamus-lateral septum
- dorsal midline thalamus-basal amygdala
- dorsal midline thalamus-hypothalamus
- dorsal midline thalamus-accumbens nucleus

3.3. System Development and Evaluation

We use the precision [48], recall [48], and F-Measure [48] metrics to evaluate our relation extraction approach.The automatically extracted neuroanatomical relations (i.e., pairs of brain region entities) are compared with the manually annotated (gold standard) pairwise neuroanatomical relations. Precision is defined as the proportion of correctly retrieved neuroanatomical relations (i.e., related pairs of brain regions) to all the relations that the application retrieves, whereas recall is defined as the proportion of correctly retrieved neuroanatomical relations to all the neuroanatomical relations in the gold standard annotation. F-Measure is the harmonic mean of the precision and recall values. The harmonic mean is always less than either the arithmetic or geometric mean, and often quite close to the minimum of the two numbers. This makes sense especially in the case that one of the values are very high and the other is low.

$$\textbf{Precision} = \frac{CorrectlyRetrievedRelations}{AllRetrievedRelations} \tag{3.1}$$

$$\textbf{Recall} = \frac{CorrectlyRetrievedRelations}{AllAnnotatedRelations} \tag{3.2}$$

$$\textbf{F} - \textbf{Measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.3}$$

For system development, we used the 1,377 abstracts of the WhiteText corpus as the gold standard with their manually annotated brain region entities and interactions [5]. The refinements that we performed during system development and the results obtained are summarized in Table 3.7. Initially, our application found 543 interactions correctly and retrieved 221 false-positive interactions, whereas the total number of true interactions in the corpus is given as 3,097. The precision of our pattern-based

approach was 71.07% and the recall value was 17.53%, which led to an F-Measure of 28.13% (Step 1 in Table 3.7).

We improved these initial values by removing some of the conjunctions and articles (i.e., "or", "the", "and") from the candidate agents and targets prior to the brain region dictionary matching phase. We also removed some words such as "part" and "area", which are sometimes used in brain regions names in the publications, but are not always included in the names of the brain regions in the dictionary. As can be seen in sample sentence below, annotated brain region mention is "anterior part of the basal nucleus", and the candidate agent extracted by the application is "anterior basal nucleus". Here, we remove the "part, of, the" words from the mention and make the comparison of these two strings to decide whether they match or not. This increased the recall value by %1.5, which also led to an improved F-score (Step 2 in Table 3.7).

Sentence : *"All of area 13 also sends efferents to the anterior part of the basal nucleus."* [49]

Entity in corpus: *"anterior part of the basal nucleus"*

Removed terms: *part, of, the*

Candidate agent found by application: *anterior basal nucleus*

After this step, the success of each pattern is evaluated separately and patterns that achieve high precision are selected. By creating a knowledge-based system, our aim was to obtain higher precision on relation extraction than the machine learning based methodologies that have been applied in the neuroscience domain. We experimented with different precision thresholds for pattern selection and observed that when we preferred only very high precision patterns (i.e., patterns with precision above 90%), only a few patterns were selected to extract relations. This resulted in very high precision, but in very low recall. On the other hand, when we used infrequent patterns with lower precision values, this increased the total number of relations that we extracted,

but the precision of the predictions was lower. Therefore, we targeted at least 60% precision for each pattern. It's important to have a higher value than 50% since it can be even defined as coincidence if it's 50%. We also tried higher success levels (i.e.,

Table 3.5. Success level of each pattern during the system development. This table is a snaphot of the patterns at Step 3 of Table 3.7 and it does not include the patterns that are matching with the acronyms. Acronyms are added later with Abbreviation Expansion Algorithm.

Patterns	**True**	**False**	**Total**	**Precision**
innervate	26	10	36	72.22%
project from		1	1	
project to	155	50	205	75.61%
projection from	120	65	185	64.86%
projection of	2	1	3	66.67%
projection to	63	27	90	70.00%
receive [ae]fferent from	5		5	100.00%
receive innervation from	3	1	4	75.00%
receive input from	32	9	41	78.05%
terminate in	13	10	23	56.52%

One of the main observations during the system development phase was that a pattern could have different semantic meanings depending on the sentence structure. For example, both "amygdala's projections to PVT" and "Efferent projection to forebrain from lateral septum" match the pattern "projection to". However, the directionality of the relations is different. The first phrase can be semantically represented as "agent's projections to target", whereas the second phrase can be represented as "projection to target from agent". Therefore, we split the existing pattern "projection to" into two patterns "projection to" and "projection to from". (Step 3 in Table 3.7).

Following new pattern types were generated during this phase. The relation extraction method was also different for these new patterns. For regular patterns, we

obtained the closest noun phrase coming after the pattern and the other dependent is obtained by dependency tree. In these new patterns, the text string between two prepositions (from, to) was picked as candidate brain region text instead of using dependency tree for extraction.

Table 3.6. New patterns that are extended from the existing patterns during the system development phase.

Existing pattern	**New pattern**
Projection from	Projection from X to Y
Projection to	Projection to X from Y
Project to	Project from X to Y
Travel to	Travel from X to Y

The following sentence shows that the pattern "projection from" directly reveals the relation with a "from X to Y" notation.

"There is a substantial projection from the dorsal midline thalamus to the hypothalamus , which appears from the retrograde tracer labeling to originate primarily in Pa." [1]

With this improvement of the patterns, we retrieved the related brain region pairs with a recall of 18.44% and an increased precision of 76.44%, which is 5% higher than the initial value.

As the last phase of the pattern evaluation we introduced one new pattern (receive fiber from) and some infrequent and unlikely patterns ('terminate in', 'innervate in') were removed from the list of patterns (Step 4 in Table 3.7).

received fiber from : *"Both the amygdaloid nucleus (AC) and the lateral amygdaloid nucleus (AL) receive fibers from the prelimbic and infralimbic areas."* [50] (There is a relation)

terminate in: *"Fibers from the ventral half of the dentate nucleus terminate in the lateral bend and ventral lamina of the principal olive.."* [51] (There is a relation)

"The DAB reaction was terminated by rinsing in PBS before sections were mounted onto gelatincoated slides and coverslipped" [52] (No relation)

Until this stage, all the acronyms were manually evaluated in the sentences. Finally, by applying abbreviation expansion algorithm on the corpus, we obtained 536 true-positive and 173 false-positive results as highest precision from the training set with 75.60% precision and 17.31% recall and 28.17% F-Measure value (Step 4 in Table 3.7)

Table 3.7. Progress of the evaluation during the system development phase for WhiteText Corpus.

	Precision	**Recall**	**F-Measure**
Step 1.Pattern-based Approach (Initial)	71.07%	17.53%	28.13%
Step 2.Removal of Conjunctions and Articles	71.00%	18.89%	29.84%
Step 3.Extraction of New Patterns	76.44%	18.44%	29.71%
Step 4.Automated Acronym Addition	75.60%	17.31%	28.17%

In the system development phase, after observing the results, we focused on two major questions: how to improve the accuracy of our findings, and how to catch the missing relations. We intensified our research on the patterns and dependency parsing rules.

By using the patterns, we faced several limitations of a rule based system. During the relation candidate generation phase, the very first step is to find the matching sentences that include one of the predefined patterns, which means that only if we find the pattern then we look for the relation in that sentence. In the light of this information, our maximum recall is limited with the maximum number of sentences that the patterns can match. As a consequence, even though we find all the interactions

correctly that the list of patterns provide (1,787 out of 3,097), the maximum recall value that we can reach is 57.7%. To be able to catch missing relations, we have analyzed some new patterns. Some new candidates like "input to", "arise from" were added as patterns and their individual precision and recall values were evaluated. Both of the patterns supported less than 55% precision and did not provide consistent output. Therefore we eliminated these patterns with the approval of our domain experts.

To improve the accuracy, we investigated parse trees of the different sentences and came up with new rules/structures. For example, in some rare cases, we noticed that controlling subject (xsubj) is used as part of the relation. In addition to that, usage of complementary relations helped us to return partial matches to the full matches.

4. RESULTS

With our linguistically motivated approach, we present three different set of results. Firstly, we provide a comparison with the present studies which have applied kernel-based machine learning methods to a manually annotated corpus, i.e., the WhiteText corpus (French et al., 2012; 2015). We show how the knowledge-based approach can introduce a different point of view on the topic by using NLP techniques. Secondly, the case study on paraventricular nucleus of the thalamus (PVT) and its interactions with other brain regions are presented. For the PVT case study, we have two different evaluation sets. First, evaluation results are given for 14 full text publications, which are manually annotated by domain experts. Second, to provide automated extraction results on the PVT corpus, which consists of 558 publications, we executed our application on the abstracts of the 451 publications (the full text of which are not publicly available) and 107 full text publications (which are publicly available). We further used the output of this evaluation on connectivity graph generation. Lastly we evaluate the success of our approach on finding the directionality of the relations.

4.1. Comparison with Previous Related Work

After making improvements on the system during the training phase, we use the second set of abstracts from the WhiteText corpus to execute the testing phase. The evaluation results are gathered from 1,828 abstracts which included a total number of 2,111 true interactions (relations).

To be able to compare our results with the existing studies, we adapted our approach during the corpus creation and dictionary usage phases as follows. Firstly, the abbreviation expansion algorithm of Schwartz and Hearst [39] is applied to each sentence to obtain the abbreviations of the brain regions . This algorithm requires the replacement of short forms of the abbreviations with long forms and the addition of short form after the long form. First it gathers the candidate abbreviations and their full forms and then applies it to the abstract wherever a match of an abbreviation

is obtained. A sample sentence is given both in original structure as written in the abstract and with the structure the abbreviation expansion algorithm applied below:

Sentence in the abstract : *"In contrast, the projections of PAG neurons to the A5 cell group and the locus coeruleus may mediate the cardiovascular and motor effects produced by stimulation of sites in the ventrolateral PAG."* [53]

Sentence with algorithm applied : *"In contrast, the projections of periaqueductal gray(PAG) neurons to the A5 cell group and the locus coeruleus may mediate the cardiovascular and motor effects produced by stimulation of sites in the ventrolateral periaqueductal gray(PAG)."* [53]

Secondly, French et al. [4–6] flag the brain region mentions from the publications as entities. In addition, most previous studies evaluate the relation extraction step by assuming that the brain region mentions are given. As can be seen in Figure 4.1, one of the brain region mentions is "entopeduncular nucleus and/or subthalamus". This does not match with our brain region dictionary structure since we only have one brain region for each entity. In our evaluation on the WhiteText corpus, instead of using our brain region dictionary, we evaluated our findings with the brain region mentions that were defined in the interactions for each sentence in the WhiteText corpus.

```
<sentence id="WhiteTextUnseenEval.d2732.s3" origId="1896" text="Multiple HRP injections were
then made bilaterally in the substantia nigra and the entopeduncular nucleus and/or
subthalamus in order to label the entire population of pedunculopontinus(PPN) neurons
projecting to the basal ganglia.">
  <entity charOffset="58-73" id="WhiteTextUnseenEval.d2732.s3.e0" origId="2958" text=
  "substantia nigra" type="Individual_protein" />
  <entity charOffset="83-123" id="WhiteTextUnseenEval.d2732.s3.e1" origId="2959" text=
  "entopeduncular nucleus and/or subthalamus" type="Individual_protein" />
  <entity charOffset="217-229" id="WhiteTextUnseenEval.d2732.s3.e2" origId="2960" text="basal
  ganglia" type="Individual_protein" />
  <pair interaction="False" id="WhiteTextUnseenEval.d2732.s3.p0" e1=
  "WhiteTextUnseenEval.d2732.s3.e1" e2="WhiteTextUnseenEval.d2732.s3.e0" />
  <pair interaction="True" id="WhiteTextUnseenEval.d2732.s3.p1" e1=
  "WhiteTextUnseenEval.d2732.s3.e2" e2="WhiteTextUnseenEval.d2732.s3.e0" />
  <pair interaction="True" id="WhiteTextUnseenEval.d2732.s3.p2" e1=
  "WhiteTextUnseenEval.d2732.s3.e2" e2="WhiteTextUnseenEval.d2732.s3.e1" />
</sentence>
```

Figure 4.1. Brain region mentions in the Whitetext corpus

In these relations, there are also some interactions that include the same entities

more than once in a sentence. Since we provide one pair for each sentence with the same entities in our application, we removed the redundant records from the evaluation. Additionally, when the entities of a pair in the given interaction were the same, we discarded that interaction as well. In the final evaluation, total number of true-interactions that were used as gold standard was 1,898 and we achieved to retrieve 277 relations correctly whereas we misinterpreted 83 of these relations.

The WhiteText corpus has been provided as two different datasets in time. In French et al. [4, 5], the first dataset with 1,377 annotated abstracts were shared, and then 1,828 more abstracts were provided as the second dataset of the WhiteText corpus (French et al. [6]). Richardet et al. [27] also used the first dataset during their research. In our approach (Linguistically Motivated Approach) we used the first dataset while developing our system to improve the patterns and the NLP techniques that we applied (Table 3.7) and the evaluation is mainly done on the second dataset with 1,828 abstracts. Table 4.1 summarizes the results of our linguistically motivated approach and the results of the previous studies obtained on the WhiteText corpus. The corresponding data set information used by each study for evaluation is also shown. Table 4.1 shows the best results obtained by French et al. [5, 6] by using the Shallow Linguistic Kernel and the results of Richardet et al. [27], who extend the study by French et al. [5] with filters and Ruta rules [28]. These filters include discarding the sentences longer than 500 characters or containing more than 7 brain regions; discarding the sentences that do not contain specific trigger words such as project; and keeping only the nearest neighbors co-occurrences. In Table 4.1, Kernel represents the machine learning model (i.e., the Shallow Linguistic Kernel) and the Ruta rules are the ones that are manually crafted on the Apache UIMA Ruta workbench [28], according to the structures of the sentences, which is an approach similar to our approach for defining the patterns.

Using a knowledge-based approach comes with more accurate results with the cost of missed relations when it is compared with the semi-automated or fully automated machine learning techniques. Therefore, if we compare our approach with the Kernel results of French et al. [5, 6] and Richardet et al. [27], the precision that we obtain is higher, whereas the recall is lower. On the other hand, comparing with Richardet et al.'s

rule based approach [27]; we achieve higher recall since we have more fine-grained rules at the linguistic level. Finding an optimum level for combining these three different approaches could be the next challenge to improve the automated neuroanatomical relation extraction task.

Table 4.1. Evaluation results for WhiteText corpus.

	Precision	**Recall**	**F-Measure**
2nd Dataset (1828 abstracts)			
Linguistically Motivated Approach	76.94%	14.59%	24.53%
French et al.,2015, Shallow Linguistic Kernel	51.00%	67.00%	57.92%
1st Dataset (1377 abstracts)			
Linguistically Motivated Approach	75.60%	17.31%	28.17%
French et al.,2012, Shallow Linguistic Kernel	50.30%	70.10%	58.30%
Richardet et al.,2015, Kernel	60.00%	68.00%	64.00%
Richardet et al.,2015, Ruta Rules	72.00%	12.00%	21.00%
Richardet et al.,2015, Filter-Kernel	66.00%	19.00%	29.00%
Richardet et al.,2015, Filter-Kernel-Rules	82.00%	7.00%	12.00%

4.2. PVT Case Study

A particular point of interest and a motivating factor in our undertaking the present study is due to a bottom-up view of depression proposed by one of us [54, 55]. Briefly, it is proposed that mood and depressive symptoms can be modulated by varying the intensity, duration and quality of stimulation by means of sensory input via visual, auditory, taste and olfactory systems, among others, as well as physical exercise. This bottom-up approach, in contradistinction to the more established account of depression and its therapies by top-down processes, is able to integrate a large body of evidence from studies that have manipulated depression by sensorimotor modulation in animal models of mood and depression and offers a new avenue of potential treatments for depression in humans. Canbeyli [55] proposed a circuitry for the integration of bottom-up sensorimotor peripheral input to the neurocircuitry underlying depression

in humans and animals with 'top-down' - potentially more cognitive influences - from the neocortex. The amygdala in particular was proposed as a key element in the nexus of the top-down and bottom-up processes. While the amygdaloid complex is a critical component of the neurocircuitry of depression, it is remarkable that the PVT, particularly with its connections to lower brainstem structures involved in visceromotor input and its connections to the amygdala, the infra- and prelimbic cortices as well as the subgenual cingulated gyrus area, is also in a position to integrate the bottom-up sensorimotor influences. As the PVT connectivity graph and the following discussion will show, the PVT may be a new target of research in mood assessment.

4.2.1. Evaluation on the Annotated PVT Corpus

For the evaluation of the PVT Case Study, we have used the 14 manually annotated full texts which are PVT specific publications.

As the output of the Relation Extraction phase (Section 3.2), we generate the candidate relation pairs which are constructed of the agents and the targets. The brain region dictionary that we created is used to validate the existence of brain region entities in the texts of the agents and targets. Therefore, the impact of a comprehensive dictionary is very high on the accuracy of the evaluation results.

The manually annotated data set of PVT from the 14 publications used in the present study contains 322 relations; 97 of them do not have one of our pre-defined patterns in their corresponding sentences. Therefore, they are already missed since the corresponding sentences are not selected as candidates for further processing. In the light of this information, the maximum level of recall that our approach can reach is 69.88%. Using NLP techniques, our application extracted 161 relation candidates out of 225 "pattern-including" relations. When we compare each relation candidate with the annotated dataset, the number of full matches is 107 and the number of partial matches is 15, whereas the number of incorrect predictions is 20. For the remaining 19 relation candidates, we have evaluated the results in two different ways. These 19 candidates included the agents and the targets and were matching with the brain region entities in

the brain region dictionary. This meant that we hit a relation with correct brain regions; therefore we evaluated these values as full or partial matches. We shared these results as NLP-based results in Table 4.2. On the other hand, during the annotation process, these relations are found to be too generic or ambiguous and eliminated depending on the sentence structure. In this second approach they are considered as incorrect predictions and are given as part of Strict and Lenient evaluations.

The following sentence contains three of these 19 relations. Our application retrieved the relation candidates "PVT"-"PFC", "PVT"-"NAS", and "PVT"-"AMG" and they are likely to refer to a relation. However, these relations were considered either too generic or ambiguous, and therefore, have not been manually annotated in the data set.

"..., it appears likely that there are no substantial differences in the degree to which stress activates PVT neurons that innervate the PFC, NAS and AMG." [56]

Actually, this is one of the core points that we would like to highlight with automated relation extraction. Using different techniques, we can automatically extract brain region relations, but this is still an input for further evaluation and the domain knowledge is crucial to turn this input to valuable information. We consider this NLP-based evaluation as valuable and share it in addition to the Strict and Lenient results. Table 4.2 shows these evaluation results by classifying them as Strict Comparison, which is the full-match of brain regions from the dictionary, Lenient Comparison, which is the full matches and partial matches of the brain regions, and lastly NLP-based comparison, which additionally includes the true-positive relations that the application finds but not annotated by domain experts.

When we compare and evaluate the WhiteText and PVT corpora, we can reach two conclusions. Firstly, recall value is higher with the PVT corpus, and the main reason for that is the percentage of the sentences that we can match with the patterns. For the WhiteText corpus, the maximum recall that we can reach is 57.7%, whereas for PVT annotated corpus, it is 69.88%. Thus, the PVT corpus contains more relations

Table 4.2. Evaluation results of the PVT case study.

	Precision	**Recall**	**F-Measure**
Strict (Full Match)	66.43%	33.23%	44.30%
Lenient (Full Match + Partial Match)	75.78%	37.89%	50.52%
NLP-based	87.58%	43.79%	58.39%

aligned with the patterns. Secondly, the precision values of the patterns are similar across the two data sets. Although the patterns were tuned based on the WhiteText corpus, they can effectively be applied to other data sets in this domain with precision levels of at least 70-75%.

Lastly, out of 322 relations in the 14 publications, only 7 of the annotated relations were in the abstract part of the publications which means that only 2% of the relations are available in the abstracts within this corpus. Using full text publications instead of abstracts mostly assures to obtain more relations to be extracted. A strength of our system is that it obtained the same success level on full text documents as well as on abstracts.

The PubMed IDs of the 14 annotated PVT papers and the annotated sentences are shared as supplementary data. Considering that some of the publications are not publicly available, the publications are not fully provided.

4.2.2. Full PVT Corpus and Connectivity Graph

We have run our application for the dataset which consists of 558 publications (451 abstracts and 107 full text publications) and 811 relations have been extracted from this corpus including 343 different brain regions. Further analysis on the relations shows that PVT is the target of 75 relations, and the source of 92 relations. Table 4.3 shows the top five brain regions with the highest number of total relations and Table 4.4 shows the ten most frequent relations that are extracted from the PVT dataset.

Table 4.3. Top 5 brain regions as agent or target in a relation.

Brain Region	**Agent**	**Target**	**Total**
pvt	92	75	167
locus coeruleus	39	23	62
nucleus accumbens	8	47	55
suprachiasmatic nucleus	30	18	48
amygdala	10	29	39

Table 4.4. Top 10 Relations that are automatically extracted from PVT Corpus.

Agent	**Target**	**Number of Relations**
pvt	nucleus accumbens	23
pvt	prefrontal cortex	13
suprachiasmatic nucleus	pvt	10
pvt	amygdala	8
pvt	medial prefrontal cortex	6
pedunculopontine nucleus	thalamus	6
paratenial nuclei	nucleus accumbens	4
lateral hypothalamus	pvt	4
brainstem	pvt	4
ventral tegmental area	nucleus accumbens	3

In Figure 4.2, we apply these 811 relations to a connectivity network graph. The brain regions are defined as nodes and the edges between them represent the relations. We use a color map where the green and yellow represent low edge counts for a node whereas the orange and red are used for higher edge counts for a node. The nodes of the graph get larger according to the edge count. Similarly, for the edge color mapping we used edge betweenness. The edge betweenness of an edge is defined as the number of the shortest paths between pairs of vertices that run along it [57]. High edge betweenness score means that if this edge is removed it will have a high impact on the connections between the nodes.

While creating the graph, the agents and the targets are matched with the unique entities in the dictionary. Directional connectivity graph (with the arrows showing the direction) can be found in Appendix A.

4.3. Directionality of The Relations

One of the contributions of our research to existing works is to define the direction of the relations.

During the evaluation phase, the accuracy of the directions of the extracted directions, in the WhiteText corpus is calculated as 100.00%, which is also validated by one of the authors (RC). The accuracy of the directions that we extracted for the relations in the PVT corpus is 97.54%. These results are shown at Table 4.5.

Table 4.5. Accuracy of the direction prediction for each corpus.

Corpus	**Accuracy**
WhiteText corpus	100.00%
PVT Corpus (14 annotated publications)	97.54%

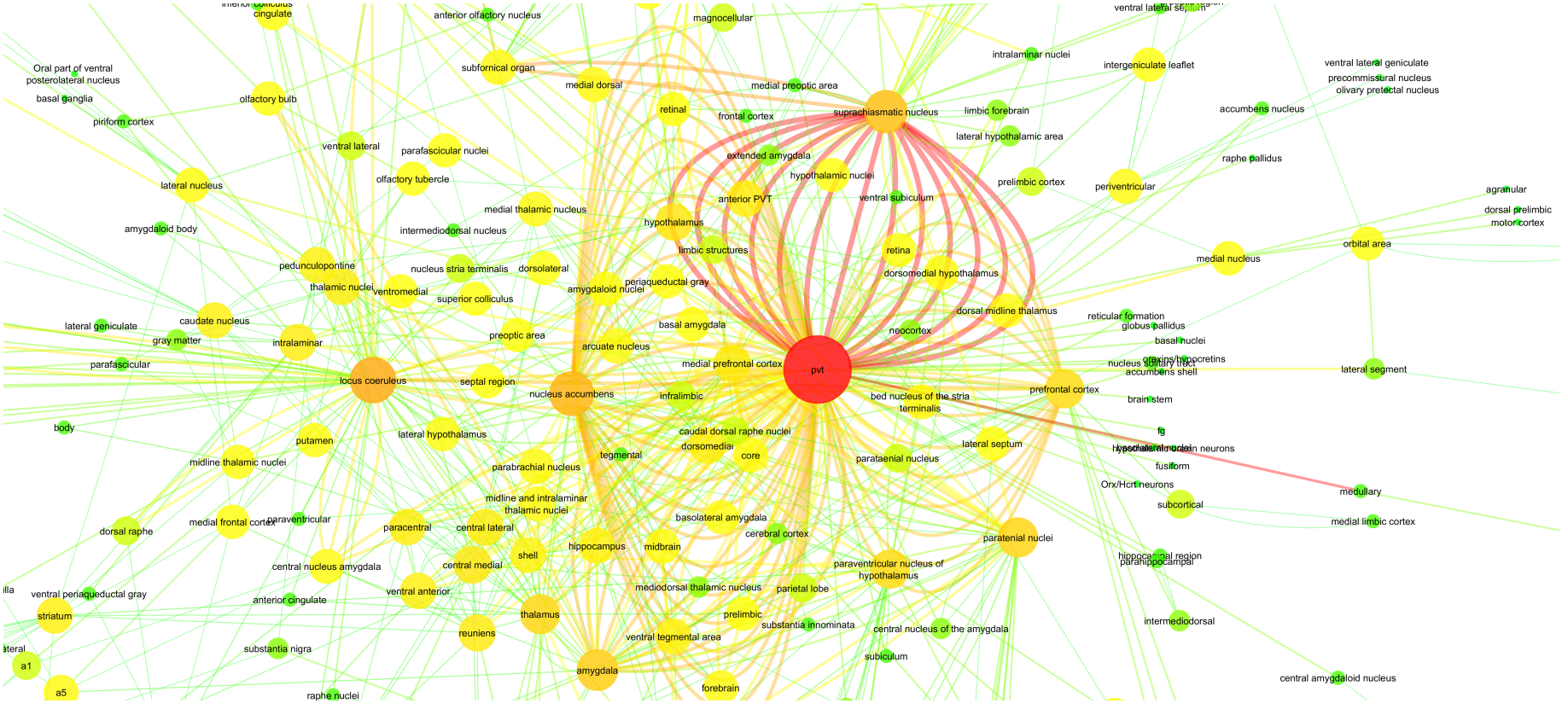

Figure 4.2. PVT connectivity graph

5. CONCLUSION

A major aim of the present study was to provide a new approach in text mining to chart out neuroanatomical connections of a specific brain structure. We have presented a linguistically motivated approach to extract neuroanatomical relations from the scientific publications by using NLP techniques. Compared to the previously reported semi-automated and automated machine learning based approaches, our approach leverages the constituency and dependency parse trees of the sentences and defines the agents and the targets by also providing the directionality of the relation.

Our motivation was to create a linguistically empowered approach and we targeted to build a high precision system which provides more accuracy than the semi-automated or fully-automated relation extraction systems. We managed to extract the neuroanatomical relations from the annotated Whitetext Corpus with higher precision compared to kernel-based methods (French et al., 2015), and with higher recall compared with the studies that have both kernel based methods and some linguistic rules and filters on top of it (Richardet et al., 2015). Particularly for PVT case study, we obtained higher recall in comparison with the results that we obtained from Whitetext corpus and the F-Measure value was more than 50% both for the Lenient and the NLP-based evaluations on PVT case study. On NLP-based approach we also obtained the brain region connectivities with 87.58% precision.

The strength of our approach comes from the patterns and rules that are defined over the parse trees of the sentences. The selection criteria for the patterns heavily depend on the individual success of each pattern to lead to a relation. We use the patterns to identify the candidate sentences for further processing and relation extraction. A limitation of our approach is that only relations from sentences that match one of our pre-defined patterns can be extracted. On the other hand, whenever a pattern is found in a sentence, it is very likely that a relation extracted after further processing is correct. Therefore, our expectation from the present study was to obtain high precision and low recall values.We preferred to have a target of at least 60% precision level for

each pattern, and as a consequence, the maximum recall value that our application could reach was approximately 70% (on the PVT data set). It is up to the researchers to define the optimum level for their evaluations. In this study, our goal was to design a high precision system so that many false positive relations are not included in the brain region connectivity graph, which could lead to incorrect interpretations.

Additionally, by using the predefined patterns to find the agent and the target, we were able to contribute on a missing feature of prior work on relation extraction: directionality of the relation. According to the grammatical structure of the sentences and the pattern usages, we identified the relation directionality between the brain regions and overall accuracy of extracted directions was more than 97%.

An additional aim of the present study was to provide by means of a connectivity graph an overview of the neuroanatomical relations of PVT that may suggest potentially new functions for the midline thalamic structure. As demonstrated in Figure 4.2, PVT has a far reaching direct connectivity with a large number of brainstem, subcortical and cortical structures. These neuroanatomical connections have yet to be adequately interpreted in terms of potential functions that may be served by subcircuits involving a more restricted number of PVT connections. Nevertheless, there is a growing realization that the PVT is not merely a component of a general behavioral arousal mechanism or a stress circuitry [3, 56], but is likely to be critically involved in more specific functions. For example, presence of connections with hypothalamic structures such as the SCN, dorsomedial hypothalamus and the fact that it is the recipient of strong orexinergic hypothalamic projections have suggested to researchers that the PVT may be an important factor in sleep/waking cycle [58]. Furthermore, due to its prominent connection with the nucleus accumbens, PVT has been investigated in connection with reward mechanisms and drug addiction [36].

In the light of the vast connectivity uncovered by our present study, we hope that there may be more interest in delineating neuroanatomical subcircuits involving the PVT as potential substrates for various functions. Towards that goal, we hereby propose in outline form a PVT circuitry that we hope to elucidate in a future article

that may be underlying a mood modulatory mechanism. Briefly, our analysis of PVT connections have uncovered a strong connectivity between the PVT and several structures known to be involved in mood and depression in both humans and animals. As demonstrated in Tables 4.3 and 4.4 and Figure 4.2, PVT has its strongest connection with the SCN. It is also connected with the nucleus accumbens, the amygdaloid complex and the extended amygdala that includes the bed nucleus of the stria terminalis (BNST) and the ventromedial prefrontal cortex. Along with other functions that they may share, these structures are also involved in mood and depression especially as indicated by studies on animal models of depression. Thus, depression as measured by forced swimming in rats is reduced with SCN [59] or amygdala lesions (Avlar and Canbeyli, manuscript in preparation), aggravated by BNST lesions [60,61], while stimulation of the ventromedial prefrontal cortex reduces depression in both humans [62] and rats [63]. Animal studies also indicate that disruption of the nucleus accumbens results in anhedonia which is a major symptom of depression in both humans and animals [64,65]. Despite such evidence, there is a paucity of studies that have directly addressed the issue of PVT involvement in depression. In the only relevant study so far, Zhu et al. (2011) [66] have shown that co-increase in c-fos positive neurons in the PVT and the central nucleus of the amygdala (CE) in rats subsequent to forced swimming rats may indicate that PVT neurons are engaged in acute depressive events.

5.1. Discussions

In the PVT case study, we preferred to have a dictionary-based approach while extracting the brain regions from publications. It is known that in the neuroscience literature brain region entities are not used in a unique and standardized way. There are several different names of each brain region and the corresponding abbreviations may vary. By using a dictionary, we accepted the possible loss on finding all the brain regions from the texts, but on the other hand we leveraged the dictionary usage on the connectivity graph by providing unique identifiers for each brain region. Most of the present text mining studies on neuroscience domain use the brain region mentions directly without normalizing them to canonical brain region names. We did not prefer

to use these mentions since it would cause to have redundant entities (nodes) that refer to the same brain region in the connectivity graph.

Another decision point for us was whether to use the existing ontologies or to create our own dictionary. Before constructing the dictionary, we investigated the existing brain ontologies. Brain Architecture Management System (BAMS) [67] was one of ontologies that includes brain regions and their relations for rats. Neuroscience Information Framework Standard Ontology [68] and Textpresso [69] were also comprehensive resources on neuroscience domain. These ontologies are very helpful to have a standard consistent terminology of the brain regions with their acronyms and synonyms. The missing part of these ontologies is that they are not defined for text mining purposes. The authors of the publications do not commonly use the brain regions as they are referred to in these ontologies. For example, most of the brain regions are given with "nucleus" in the ontologies, on the other hand, in the publications the authors can omit "nucleus" (i.e. dorsomedial is used instead of dorsomedial nucleus). Secondly, authors may prefer to use different acronyms instead of the known acronyms of the brain regions. For example, BrainInfo portal[5] , which contains NeuroNames knowledgebase, uses "PV" and "PVT" as acronyms of paraventricular nucleus of the thalamus, whereas Hsu et al. (2009) [1] used "Pa" acronym for the same brain region. Lastly, we needed to obtain the anatomical directions during the text mining process. For this purpose, we created the brain region entities in the dictionary with the direction information like anterior, posterior, ventral, dorsal, rostral, caudal, etc. Therefore we had both anterior PVT and PVT as separate brain region entities in our dictionary. As a result of our investigation, we decided to create our own brain region dictionary and use it for relation extraction.

During the relation extraction phase, we faced several difficulties. One was related to the WhiteText corpus. This manually annotated corpus is considered as gold standard for the first evaluation phase of our research. Since this corpus is enhanced with the abbreviation expansion algorithm, we also needed to use the same approach. Schwartz and Hearst Abbreviation Expansion Algorithm [39] is used for this purpose

[5] $http://braininfo.org$

and it requires the replacement of the short forms of the abbreviations with their long forms. The short form is also added right after the long form. We skipped this step on the PVT Case study since the abbreviations are already included as part of the brain region dictionary under the name of acronyms. The second challenge was the ambiguity on the brain regions which are given with conjunctions (i.e. "dorsal and ventral cortex" or "basolateral and basomedial nuclei of the amygdala"). We initially decided to evaluate these phrases as one brain region entity since the WhiteText corpus considered such phrases as one brain region mention. For the PVT corpus, we needed to remove the conjunction and create two different brain region entities from these mentions. After the implementation of this phase, we noticed that the overall precision was reduced due to false-positives, hence, we kept ambiguity resolution as a project for future work.

5.2. Future Work

In our study, we have focused on automated relation extraction of brain regions on neuroscience domain. Hence, our defined patterns and rules might not be generic enough to be used in Protein-Protein and Gene-Disease interactions. This is considered as a possible future work. Additionally, the current research identifies only the neuroanatomical relations of the brain regions (circuitry). As future work, the chemical connections between brain regions (neurotransmitters) and the functional connections (by the attributed cognitive function of the relation) will be our focus of interest.

APPENDIX A: Connectivity Graph - with directions

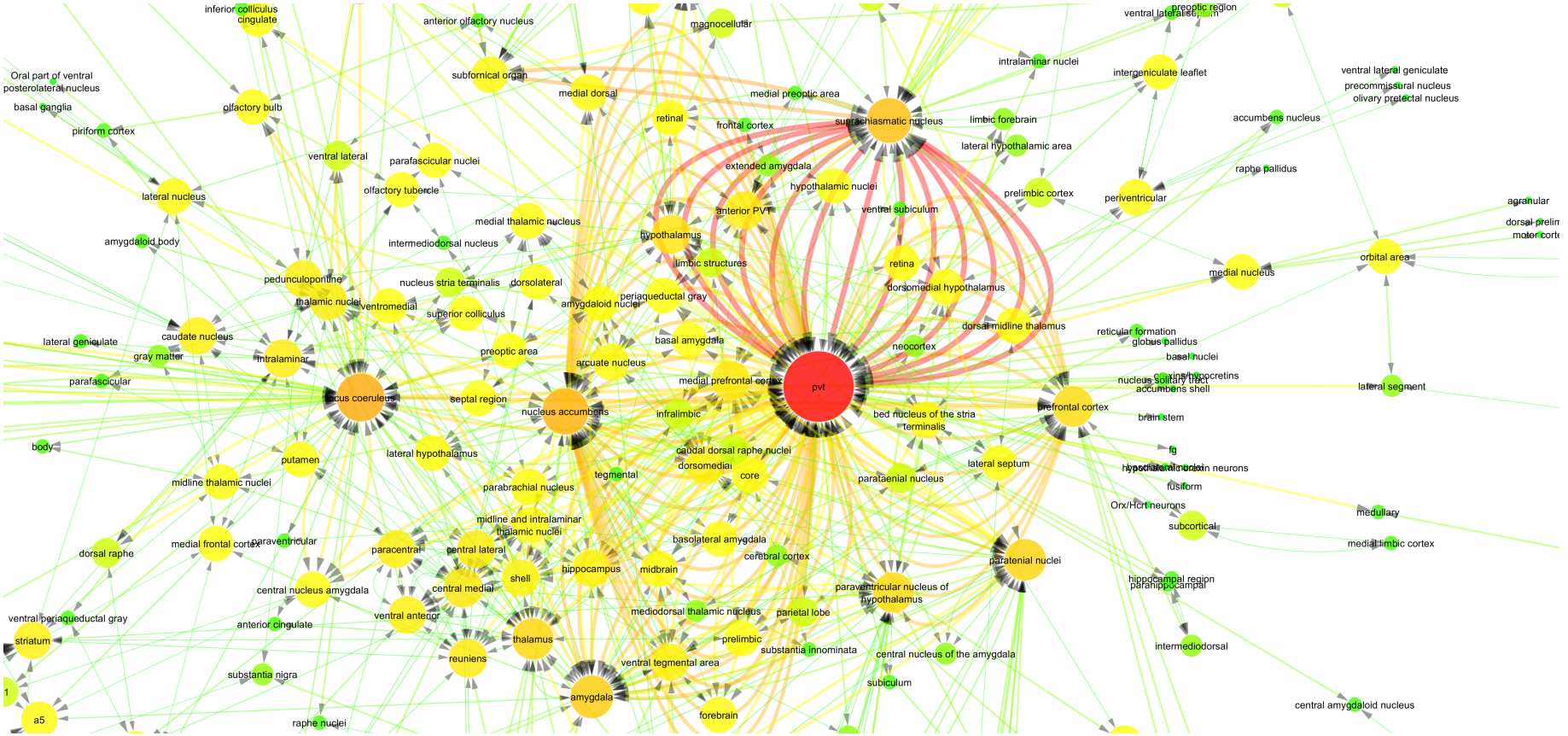

Figure A.1. PVT connectivity graph with directions

REFERENCES

1. Hsu, D. T. and J. L. Price, "The Paraventricular Thalamic Nucleus: Subcortical Connections and Innervation by Serotonin, Orexin, and Corticotropin-Releasing Hormone in Macaque Monkeys", *J Comp Neurol 2009 February 20; 512(6): 825–848. doi:10.1002/cne.21934.*, 2009.

2. Li, S. and G. J. Kirouac, "Sources of inputs to anterior and posterior aspects of the paraventricular nucleus of the thalamus", *Brain Struct Funct.217(2):257-273*, 2012.

3. Vertes, P. R., S. B. Linley and W. B. Hoover, "Limbic circuitry of the midline thalamus", *Neuroscience and Biobehavioral reviews. 54, 89-107*, 2015.

4. French, L., S. Lane, L. Xu and P. Pavlidis, "Automated recognition of brain region mentions in neuroscience literature", *Front. Neuroinform.*, Vol. 3, p. 29, 2009.

5. French, L., S. Lane, L. Xu, C. Siu, C. Kwok, Y. Chen, C. Krebs and P. Pavlidis, "Application and evaluation of automated methods to extract neuroanatomical connectivity statements from free text", *Bioinformatics, 28(22), 29632970*, 2012.

6. French, L., P. Liu, O. Marais, T. Koreman, L. Tseng, A. Lai and P. Pavlidis, "Text mining for neuroanatomy using WhiteText with an updated corpus and a new web application", *Front. Neuroinform. 9:13. doi: 10.3389/fninf.2015.00013*, 2015.

7. Krallinger, M., F. Leitner, C. Rodriguez-Penagos and A. Valencia, "Overview ofthe protein-protein interaction annotation extraction task of BioCreative II", *Genome Biol. 9(Suppl. 2), S4. doi: 10.1186/gb-2008-9-s2-s4*, 2008.

8. Arighi, C. N., Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman and C. H. Wu, "Overview of the biocreative III workshop", *BMC Bioinform. 12(Suppl. 8):S1. doi: 10.1186/1471-2105-12-S8-S1*, 2011.

9. Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction", *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Tas, Stroudsburg, PA, 1–9.*, 2009.

10. Kim, J.-D., S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen and J. Tsujii, "Overview of BioNLP shared task 2011", *Proceedings of the BioNLP Shared Task 2011 Workshop, Portland, OR, 1–6.*, 2011.

11. Nédellec, C., R. Bossy, J. D. Kim, J. Kim, T. Ohta, S. Pyysalo and et al., "Overview of bionlp shared task 2013", *Proceedings of the BioNLP Shared Task 2013 Workshop, Sofia, 1–7*, 2013.

12. Fukuda, K., A. Tamura, T. Tsunoda and T. Takagi, "Toward information extraction: identifying protein names from biological papers", *Proceedings of the Pacific Symposium on Biocomputing, Hawaii*, p. 707–718, 1998.

13. Hur, J., A. D. Schuyler, D. J. States and E. L. Feldman, "SciMiner: web-based literature mining tool for target identification and functional enrichment analysis", *Bioinformatics 25, 838–840. doi: 10.1093/bioinformatics/btp049*, 2009.

14. McDonald, R. and F. Pereira, "Identifying gene and protein mentions in text using conditional random fields", *BMC Bioinformatics 6(Suppl 1):S6. doi: 10.1186/1471-2105-6-S1-S6*, 2005.

15. Hsu, C.-N., Y.-M. Chang, C.-J. Kuo, Y. S. Lin, H.-S. Huang and I.-F. Chung, "Integrating high dimensional bi-directional parsing models for gene mention tagging", *Bioinformatics 24, i286-i294. doi: 10.1093/bioinformatics/btn183*, 2008.

16. Jelier, R., G. Jenster, L. C. Dorssers, C. C. Van Der Eijk, E. M. Van Mulligen, B. Mons and A. J. Kors, "Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes", *Bioinformatics 21, 2049–2058. doi: 10.1093/bioinformatics/bti268*, 2005.

17. He, M., Y. Wang and W. Li, "PPI finder: a mining tool for human protein-protein interactions", *PLoS One. 2009; 4(2):e4554*, 2009.

18. Blaschke, C. and A. Valencia, "The frame based module of the SUISEKI information extraction system", *IEEE Intell. Syst. 17, 14–20. doi: 10.1109/MIS.2002.999215*, 2002.

19. Giuliano, C., A. Lavelli and L. Romano, "Exploiting shallow linguistic information for relation extraction from biomedical literature", *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento*, p. 401–408, 2006.

20. Fundel, K., R. Küffner and R. Zimmer, "RelEx—Relation extraction using dependency parse trees", *Bioinformatics*, Vol. 23, p. 365–371, 2007.

21. Erkan, G., A. Ozgur and D. R. Radev, "Semi-supervised classification for extracting protein interaction sentences using dependency parsing", *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 228–237, 2007.

22. Airola, A., S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter and T. Salakoski, "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning", *BMC Bioinformatics 9:S2. doi: 10.1186/1471-2105-9-S11-S2*, 2008.

23. Tikk, D., P. Thomas, P. Palaga, J. Hakenberg and U. Leser, "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature", *PLoS Comput.Biol. 6:e1000837. doi:10. 1371/journal.pcbi.1000837*, 2010.

24. Quan, C., M. Wang and F. Ren, "An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature", *Raghava GPS, ed. PLoS ONE. 2014;9(7):e102039. doi:10.1371/journal.pone.0102039*, 2014.

25. Burns, G., D. Feng and E. Hovy, "Intelligent approaches to mining the primary research literature: Techniques, systems, and examples", *In Computational Intelligence in Medical Informatics*, Vol. 85, pp. 17–50, 2008.

26. Segura-Bedmar, I., P. Martinez and C. de Pablo-Sánchez, "Using a shallow linguistic kernel for drug–drug interaction extraction", *Journal of biomedical informatics, 44(5), 789-804*, 2011.

27. Richardet, R., J.-C. Chappelier, M. Telefont and S. Hill, "Large-scale extraction of brain connectivity from the neuroscientific literature", *Bioinformatics. 2015;31 (10):1640-1647. doi:10.1093/bioinformatics/btv025*, 2015.

28. Kluegl, P., M. Toepfer, P.-D. Beck, G. Fette and F. Puppe, "UIMA ruta: Rapid development of rule-based information extraction applications", *Natural Language Engineering (2014): 1-40*, 2014.

29. Vasques, X., R. Richardet, S. L. Hill, D. Slater, J.-C. Chappelier, E. Pralong, J. Bloch, B. Draganski and L. Cif, "Automatic target validation based on neuroscientific literature mining for tractography", *Front. Neuroanat. 9:66. doi: 10.3389/fnana.2015.00066*, 2015.

30. Sayers, E., "A General Introduction to the E-utilities. In: Entrez Programming Utilities Help [Internet]", *Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25497/,accessed at September 2015*.

31. Choi, D. L., J. F. Davis, I. J. Magrisso, M. E. Fitzgerald, J. W. Lipton and S. C. Benoit, "Orexin signaling in the paraventricular thalamic nucleus modulates mesolimbic dopamine and hedonic feeding in the rat", *Neuroscience 210, 243–248 (2012)*, 2012.

32. Bowden, D. M., E. Song, J. Kosheleva and M. F. Dubach, "NeuroNames: An Ontology for the BrainInfo Portal to Neuroscience on the Web", *Neuroinformatics*

2012;10(1):97-114. doi:10.1007/s12021-011-9128-8., 2012.

33. Larson, S. D. and M. M. E., "NeuroLex.org:an online framework for neuroscience knowledge", *Front. Neuroinform. 7:18. doi:10.3389/conf.neuro.11. 2009.08.140*, 2013.

34. Linauts, M. and G. F. Martin, "The organization of olivo-cerebellar projections in the opossum, Didelphis virginiana, as revealed by the retrograde transport of horseradish peroxidase.", *FJ Comp Neurol. 1978 May 15;179(2):355-81*, 1978.

35. Lui, F., K. M. Gregory, R. H. Blanks and R. A. Giolli, "Projections from visual areas of the cerebral cortex to pretectal nuclear complex, terminal accessory optic nuclei, and superior colliculus in macaque monkey.", *J Comp Neurol 363(3): 439–460*, 1995.

36. Matzeu, A., E. R. Zamora-Martinez and R. Martin-Fardon, "The paraventricular nucleus of the thalamus is recruited by both natural rewards and drugs of abuse: recent evidence of a pivotal role for orexin/hypocretin signaling in this thalamic nucleus in drug-seeking behavior", *Front Behav Neurosci. 3;8:117. doi: 10.3389/fnbeh.2014.00117*, 2014.

37. Kunishio, K. and S. N. Haber, "Primate cingulostriatal projection: limbic striatal versus sensorimotor striatal input", *J Comp Neurol. 1994;350(3):337–56. Epub 1994/12/15. doi: 10.1002/cne.903500302*, 1994.

38. Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60*, 2014.

39. Schwartz, A. S. and M. A. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text", *Pac. Symp. Biocomput. 8, 451–462*, 2003.

40. Janak, P. H. and N. Chaudhri, "The Potent Effect of Environmental Context on Relapse to Alcohol-Seeking After Extinction", *The Open Addiction Journal, 3, 76–87. http://doi.org/10.2174/1874941001003010076*, 2010.

41. Klein, D. and C. D. Manning, "Stanford Parser, Accurate Unlexicalized Parsing", *Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430., 2003*, 2003.

42. De Marneffe, M.-C., B. MacCartney and C. D. Manning, "Generating typed dependency parses from phrase structure parses", *In Proceedings of LREC, vol. 6, no. 2006, pp. 449-454. 2006*, 2006.

43. Ito, H. and M. Seki, "Ascending projections from the area postrema and the nucleus of the solitary tract of Suncus murinus: anterograde tracing study using Phaseolus vulgaris leucoagglutinin.", *Okajimas Folia Anatomica Japonica, 75 (1), 9±31*, 1998.

44. DenBleyker, M., D. Nicklous, P. Wagner, H. Ward and K. Simansky, "Activating mu-opioid receptors (MOPRs) in the lateral parabrachial nucleus increases c-Fos expression in forebrain areas associated with caloric regulation, reward and cognition", *Neuroscience 162: 224–233*, 2009.

45. Lin, Y., G. J. Ter Horst, R. Wichmann and et al., "Sex Differences in the Effects of Acute and Chronic Stress and Recovery after Long-Term Stress on Stress-Related Brain Regions of Rats.", *Cerebral Cortex (New York, NY). 2009;19(9):1978-1989. doi:10.1093/cercor/bhn225*, 2009.

46. Shin, J.-W., J. C. Geerling and A. D. Loewy, "Inputs to the ventrolateral bed nucleus of the stria terminalis", *Journal of Comparative Neurology 511, no. 5 (2008): 628-657*, 2009.

47. Pasumarthi, R. K. and J. Fadel, "Activation of orexinhypocretin projections to basal forebrain and paraventricular thalamus by acute nicotine.", *Brain Res Bull*

77:367–373. 10.1016/j.brainresbull.2008.09.014, 2008.

48. Manning, C., P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press., 2008.

49. Turner, B. J. and J. Zimmer, "The architecture and some of the interconnections of the rat's amygdala and lateral periallocortex", *J Comp Neurol. 1984;227:540–557*, 1984.

50. Ottersen, O. P., "Connections of the amygdala of the rat. IV. Corticoamygdaloid and intraamygdaloid connections as studied with axonal transport of horseradish peroxidase.", *J Comp Neurol 205: 30–48*, 1982.

51. Tolbert, D. L., L. C. Massopust, M. G. Murphy and P. A. Young, "The anatomical organization of the cerebello-olivary projection in the cat", *J Comp Neurol. 1976;170:525–544*, 1976.

52. Li, S., Y. Shi and G. J. Kirouac, "The Hypothalamus and Periaqueductal Gray Are the Sources of Dopamine Fibers in the Paraventricular Nucleus of the Thalamus in the Rat.", *Frontiers in Neuroanatomy 8 (2014): 136.*, 2014.

53. Bajic, D. and H. K. Proudfit, "Projections of neurons in the periaqueductal gray to pontine and medullary catecholamine cell groups involved in the modulation of nociception", *J Comp Neurol. 1999;405:359–379*, 1999.

54. Canbeyli, R., "Sensorimotor modulation of mood and depression: an integrative review", *Behavioural Brain Research*, Vol. 207.2, pp. 249–264, 2010.

55. Canbeyli, R., "Sensorimotor modulation of mood and depression: In search of an optimal mode of stimulation", *Frontiers in Human Neuroscience 2013;7:428. doi:10.3389/fnhum.2013.00428.*, 2013.

56. Bubser, M. and A. Y. Deutch, "Stress induces Fos expression in neurons of the tha-

lamic paraventricular nucleus that innervate limbic forebrain sites", *Synapse.32:13-22.*, 1999.

57. Girvan, M. and M. E. Newman, "Community structure in social and biological networks", *Proceedings of the national academy of sciences*, Vol. 99(12), pp. 7821–7826, 2002.

58. Colavito, V., C. Tesoriero, A. Wirtu, G. Grassi-Zucconi and M. Bentivoglio, "Limbic thalamus and state-dependent behavior: The paraventricular nucleus of the thalamic midline as a node in circadian timing and sleep/wake-regulatory networks", *Neurosci Biobehav Rev. 2015 Jul;54:3-17. doi: 10.1016/j.neubiorev.2014.11.021*, 2015.

59. Tataroğlu, O., A. Aksoy, A. Yılmaz and R. Canbeyli, "Effect of lesioning the suprachiasmatic nuclei on behavioral despair in rats", *Brain Research,1001(1), 118-124*, 2004.

60. Schulz, D. and R. Canbeyli, "Lesion of the bed nucleus of the stria terminalis enhances learned despair", *Brain Res Bull. 52:83-87*, 2000.

61. Pezuk, P., E. Aydin, A. Aksoy and R. Canbeyli, "Effects of BNST lesions in female rats on forced swimming and navigational learning", *Brain Research. 1228:199-207*, 2008.

62. Koenigs, M. and J. Grafman, "The functional neuroanatomy of depression: Distinct roles for ventromedial and dorsolateral prefrontal cortex", *Behav Brain Res. 2009 Aug 12; 201(2): 239–243. doi: 10.1016/j.bbr.2009.03.004*, 2009.

63. Hamani, C., M. Diwan, S. Isabella, A. M. Lozano and J. N. Nobrega, "Effects of different stimulation parameters on the antidepressant-like response of medial prefrontal cortex deep brain stimulation in rats", *Journal of Psychiatric Research*, Vol. 44(11), pp. 683–687, 2010.

64. Russo, S. J. and E. J. Nestler, "The brain reward circuitry in mood disorders", *Nat Rev Neurosci. 2013 Sep;14(9):609-25. doi: 10.1038/nrn3381*, 2013.

65. Willner, P., "Animal models of depression: an overview", *Pharmacol Ther. 45:425-455*, 1990.

66. Zhu, L., L. Wu, B. Yu and X. Liu, "The participation of a neurocircuit from the paraventricular thalamus to amygdala in the depressive like behavior", *Neuroscience letters.488(1):81-86*, 2011.

67. Bota, M. and L. W. Swanson, "BAMS neuroanatomical ontology: design and implementation", *Frontiers in neuroinformatics, 2*, 2008.

68. Bug, W. J., G. A. Ascoli, J. S. Grethe, A. Gupta, C. Fennema-Notestine, A. R. Laird and et al., "The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience.", *Neuroinformatics 6, 175–194. doi: 10.1007/s12021- 008-9032-z*, 2008.

69. Muller, H. M., A. Rangarajan, T. K. Teal and P. W. Sternberg, "Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers", *Neuroinformatics 6, 195–204*, 2008.