CoMAGD: Annotation of Gene-Depression Relations

Rize Jin[1] **Jinseon You**[1] **Jin-Woo Chung**[1] **Hee-Jin Lee**[1]
Maria Wolters[2] **Jong C. Park**[1*]

[1]School of Computing
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Daejeon, Republic of Korea
{rizejin, jsyou, jwchung, heejin, park}@nlp.kaist.ac.kr

[2]School of Informatics
University of Edinburgh
Edinburgh, UK
maria.wolters@ed.ac.uk

Abstract

Clinical depression is a mental disorder involving genetics and environmental factors. Although much work studied its genetic causes and numerous candidate genes have consequently been looked into and reported in the biomedical literature, no gene expression changes or mutations regarding depression have yet been adequately collected and analyzed for its full pathophysiology. In this paper, we present a depression-specific annotated corpus for text mining systems that target at providing a concise review of depression-gene relations, as well as capturing complex biological events such as gene expression changes. We describe the annotation scheme and the conducted annotation procedure in detail. We discuss issues regarding proper recognition of depression terms and entity interactions for future approaches to the task. The corpus is available at http://www.biopathway.org/CoMAGD.

1 Introduction

Clinical depression, or major depressive disorder, is a mental disorder of the central nervous system with a pathophysiology involving the neocortex. Genetics and environmental factors are known to contribute to the development of mood disorders (Nestler et al., 2002). Many biomedical research efforts studied the causative factors of genetics in depression, with consequent rapid accumulation of candidate genes (Kao et al., 2011; Piñero et al., 2015). However, the accumulated information is not yet comprehensive enough to explain the role of genes involved in depression.

DisGeNET (Piñero et al., 2015) is a platform for discovering associations of genes and complex diseases including depression, defining gene-depression relations as simple binary relations that consist of *geneId*, *geneSymbol*, *geneName*, *diseaseId*, *diseaseName*, and *score*, where the score is a measure of relevancy based on the supporting evidence. DEPgenes (Kao et al., 2011) gives a prioritizing system that uses combined score to rank candidate genes for depression. Although DEPgenes is a nearly comprehensive candidate gene resource for depression in terms of its volume (5,055 candidate genes), its representation concepts are even simpler than DisGeNET and thus not quite adequate for the full understanding of depression-related phenomena.

In order to fully understand how a particular gene acts in depression, we need detailed information about gene expression changes or mutations and also how the depression level is changed along with the change in the gene. In this regard, we anticipate that text mining systems, which can identify and analyze both genes and depression changes comprehensively from text, would facilitate research on depression much further. Furthermore, if the mined information is annotated and then made available for reuse, key resources would be identified and constructed more effectively (McDonald and Kelly, 2012; Winnenburg

* Corresponding author

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 104–113,
Beijing, China, July 30, 2015.

et al., 2008). Such effort of making relevant corpora has already been made in the studies of genes (Kim et al., 2008; Poux et al., 2014) and of complex diseases such as cancers (Lee et al., 2013; Lee et al., 2014; Pyysalo et al., 2013), but has not yet been applied to depression.

In this paper, we present a depression-specific annotated corpus, CoMAGD, for future text mining systems that target specifically at providing comprehensive information of depression-gene relations as well as capturing complex information such as gene changes and biological events. For this purpose, we follow a multi-faceted annotation scheme for cancers (Lee et al., 2013) while tuning it extensively to depression. In this revised scheme, a piece of annotation is composed of four concepts that together express two events, *gene expression changes* and *depression level or antidepressant effect changes*, and the relationship between these two events. We anticipate that the present corpus and text mined results based on this corpus would contribute meaningfully to the successful exploration of the underlying functional correlation between genes and clinical depression.

The rest of the paper is organized as follows. Section 2 shows the corpus annotation. Section 3 gives details of inter-annotator agreement. Section 4 discusses issues about proper recognition of depression terms and entity interactions for future approaches to the task, before closing the paper in Section 5.

2 Corpus Annotation

2.1 Data collection and pre-processing

We collected PubMed IDs (PMIDs) that contain depression related terms in any of the three fields *title*, *abstract*, and *keyword*, using the query "*depress** OR *dysthymia* OR *cyclothymia*", and randomly selected 500 abstracts among them. The 500 abstracts were then segmented into sentences.

We extracted only the sentences that contain at least one pair of gene and depression/antidepressant related terms. BANNER (Leaman and Gonzalez, 2008) and Moara (Neves et al., 2010) were used to identify and normalize gene names. For depression and antidepressant terms, the system used dictionary-based longest matching. The dictionary consists of 303 entries of depression and antidepressant related terms collected from NCI Thesaurus and other relevant articles. The entries were then edited by a domain expert in mental health.

For the sentences that contain more than one pair, we made their copies, matching the number of depression-gene pairs. We call each of these copies a *co-occurrence*. For example, if there are three gene names and two depression related terms in a sentence, the system makes six co-occurrences for this sentence.

We then tokenized, part-of-speech tagged, and parsed the co-occurrences, using the Charniak-Johnson parser (Charniak and Johnson, 2005) with a biomedical parsing model (McClosky, 2010). The resulting phrase structures were then converted into dependency structures with the Stanford conversion tool (Marneffe et al., 2006). We identified mentions of gene expression changes, using the Turku event extraction system (Björne et al., 2009). Most of the processes above are included in a preprocessed dataset, or EVEX (Landeghem et al., 2012); however, we modified the system and utilized some part of the system separately where necessary.

Finally, we performed manual work to validate automatically identified co-occurrences in order to produce confirmed annotation units, such as manually constructing predicates (i.e., 'depression of [non-human subjects]') to filter out false positives from the dictionary matching outputs of depression-related terms and manually eliminating false relations (hypothesis sentences).

2.2 A multi-faceted annotation scheme

We modify a multi-faceted annotation scheme of (Lee et al., 2013), originally designed to represent ternary relations among genes, cancers and gene changes, in order to address relations not only between depression and genes, but also between antidepressants and genes, so as to provide more details and enable further insights for follow-up studies such as prioritizing depression candidate genes and designing effective treatments and therapy. For example, one may assign a lower weight to a gene if the gene shows expression changes only in antidepressant studies. We also introduce directed causal relations between genes and depression/antidepressants. Identification of the cause and effect not only reflects the methodologies of individual studies, but also provides the facts. While the undirected causality claim usually is interpreted as a necessary and sufficient clause, we find that it could result in false conclusions,

Concept	Value	Definition
Change in Gene Expression (CGE)	increased	Expression level of the gene is increased
	decreased	Expression level of the gene is decreased
Change in Depression Level (CDL) or Change in Antidepressant Effect (CAE)	increased	The depression level/antidepressant effect is increased as CGE
	decreased	The depression level/antidepressant effect is decreased as CGE
	unidentifiable	The information about whether or not CGE accompanies the depression level/antidepressant effect change is not provided
Causality Claim (CC)	none	CGE accompanied by CDL/CAE is reported but the causality between the two is not claimed
	g2x	The causality is claimed as CGE causes CDL/CAE
	x2g	The causality is claimed as CDL/CAE causes CGE

Table 1: Annotation concept values and their definitions

Sentence	CGE	CDL	CC
Example 1. In particular, we found decreased NF-L, PSD95, and SAP102 transcripts in bipolar disorder, and [*decreased*]$_e$ [*SAP102*]$_g$ levels in [*major depression*]$_d$. [PMID: 15054476]	dec.	uni.	non.
Example 2. In conclusion, chronic forced swim stress was a good animal model of [*depression*]$_d$, and it induced depressive-like behavior and [*decreased*]$_e$ [*P-Erk2*]$_g$ in the hippocampus and prefrontal cortex in rats. [PMID: 17050000]	dec.	inc.	x2g
Sentence	**CGE**	**CAE**	**CC**
Example 3. [*Fluoxetine*]$_a$ substantially [*inhibits*]$_e$ [*CYP2D6*]$_g$ and probably CYP2C9/10, moderately inhibits CYP2C19 and mildly inhibits CYP3A3/4. [PMID: 9068931]	dec.	uni.	x2g
Example 4. [*Inhibition*]$_e$ of [*neuronal nitric oxide synthase*]$_g$ in the rat hippocampus induces [*antidepressant-like*]$_a$ effects. [PMID: 9068931]	dec.	inc.	g2x

Gene names, depression related terms, antidepressant related terms, and the keywords for gene expression change are noted in matching square brackets and marked with subscripts 'g', 'd', 'a', and 'e', respectively.

Table 2: Examples of annotated co-occurrences

especially in the studies of depression. For example, depression may decrease the expression level of a particular gene; however, increasing the expression level of that gene may not necessarily reduce the symptom. One reason is that the genetic factor is not the only cause of depression. It is also believed that, compared to oncogenesis, much more genes act together and render a person to become vulnerable to depression (Belmaker and Agam, 2008). As such, a more fine-grained annotation of causal directions will be essential for more complex diseases such as depression. In an answer to these needs, we use a flexible schema for annotating concepts and ever-changing metrics and facts in genetic studies of depression. The flexibility would allow the schema to exploit the location information as well, as studies show that genes may respond differently to the same antidepressant if they are in different parts of a body. More details will be discussed in Section 4.

2.3 Annotation concept

The proposed corpus contains four core annotation concepts: *Change in Gene Expression* (CGE), *Change in Depression Level* (CDL), *Change in Antidepressant Effect* (CAE), and *Causality Claim* (CC). CGE captures whether the expression level of a gene is 'increased' or 'decreased'. CDL/CAE captures the way how the depression level/antidepressant effect changes together with a gene expression level change. If information about such changes is not provided in the sentence,

```
<?xml version="1.0" ?>
<!DOCTYPE gene_depression_corpus [
        <!ELEMENT     gene_depression_corpus (annotation_unit+)>
        <!ELEMENT     annotation_unit (sentence, annotation+)>
        <!ATTLIST     annotation_unit type (depression | antidepressant) #REQUIRED >
        <!ELEMENT     sentence (#PCDATA)>
        <!ATTLIST     sentence pmid CDATA #REQUIRED >
        <!ELEMENT     annotation (gene, expression_change_keyword_1,
                              expression_change_keyword_2, depression_term+, CGE, CDL, CC)>
        <!ATTLIST     annotation id CDATA #REQUIRED>
        <!ELEMENT     gene (#PCDATA)>
        <!ATTLIST     gene offset CDATA #REQUIRED >
        <!ELEMENT     expression_change_keyword_1 (#PCDATA)>
        <!ATTLIST     expression_change_keyword_1 offset CDATA #REQUIRED
                              type (Negative_regulation | Positive_regulation) #REQUIRED>
        <!ELEMENT     expression_change_keyword_2 (#PCDATA)>
        <!ATTLIST     expression_change_keyword_2 offset CDATA #REQUIRED
                              type (None | Gene_expression) #REQUIRED>
        <!ELEMENT     depression_term (#PCDATA)>
        <!ATTLIST     depression_term offset CDATA #REQUIRED>
        <!ELEMENT     CGE EMPTY>
        <!ATTLIST     CGE value (increased | decreased) #REQUIRED>
        <!ELEMENT     CDL EMPTY>
        <!ATTLIST     CDL value (increased | decreased | unidentifiable) #REQUIRED>
        <!ELEMENT     CC EMPTY>
        <!ATTLIST     CC value (x2g | g2x | none) #REQUIRED>
]>
```

Table 3: The XML DTD of the corpus

we assign ‘unidentifiable’. CC captures whether the causality between the gene expression change and the CDL/CAE is claimed in the sentence or not, with values ‘none’, ‘x2g’, and ‘g2x’. Each concept is assigned with one of the pre-specified values to complete a *facet* of annotation. Table 1 shows the pre-specified values and the definitions of the respective values. Three of the four concepts together complete a piece of annotation that express information about a gene’s expression level change with a change in depression level or antidepressant effect.

Table 2 shows examples of the annotated sentences and Table 3 shows the DTD schema of the corpus. As mentioned earlier, we collected sentences from PubMed that describe gene expression changes in depression/antidepressants. Each sentence was presented to the annotators as one or more copies with markings for a gene term, keywords for gene expression change, and a depression/antidepressant-related term. The annotators read the sentence with such markings and selected proper values for the annotation concepts. Note that the four annotation concepts are semantically orthogonal, in that the value of a concept can be identified without knowing the values of the other concepts.

2.4 Corpus statistics

The corpus consists of 210 annotation units, where an annotation unit is simply a mention of gene expression change that co-occurs with at least one depression or antidepressant related term in a sentence. These annotation units are derived from 106 different sentences, which in turn are extracted from 73 PubMed abstracts. The corpus contains 82 gene types, 5 depression terms, and 20 antidepressant terms (cf. Table 4).

Tables 5 and 6 show the distribution of annotation concept values and the distribution of the annotated genes, respectively. The values of CGE show a uniform distribution, whereas the others show skewed distributions. In particular, for values of CDL/CAE, ‘unidentifiable’ is frequently chosen (89% for CDL, 87% for CAE). The value distribution of the concept CC associated with

CAE also exhibits dominance of a single value, or 'x2g'. We compared the genes in our corpus with previous studies: 58% (48) and 95% (79) of our annotated genes (83) are included in DisGeNET and DEPgenes, respectively. Note that DEPgenes only published 169 core genes that exhibit a high chance to be associated with depression from 5,055 candidate genes.

3 Inter-annotator agreement

We annotated the sentence units through two main annotation phases (cf. Table 7) and revised annotation guidelines after each annotation phase. Table 8 shows the IAA values obtained from each annotation phase as well as from the whole corpus. We measured IAAs in three different ways, using simple IAA (the proportion of annotations in common between two annotators over the total number of annotations provided by either annotator), Cohen's kappa, and G-index. IAA values from the final phase show that adequate agreement among the annotators is achieved. The overall IAA values, obtained from the whole corpus, also suggest internal consistency. We resolved all disagreements in the published corpus.

3.1 Disagreements

We identify the following as the major sources for conflicts between the annotators: simple mistakes, subjective readings, the use of reasoning, and the judgements by using prior knowledge. Disagreement rate is greatly reduced in the second annotation phase, as we revised the guidelines after the completion of the first phase.

Simple mistakes are inevitable in manual annotations, contributing a small number of conflicts to all the four annotation concepts. In detail, simple mistakes take up 1% (1 out of 142), 8% (11 out of 142), and 24% (34 out of 142) of the disagreements on CGE, CDL/CAE, and CC values, respectively, in Phase 1, and 9% (6 out of 67), 0% (0 out of 67), and 3% (2 out of 67) in Phase 2.

Disagreements also arise from subjective readings, contributing to most of the disagreements on CC values.

Example 5. [*CRF*]$_{g}$ is [*increased*]$_{e}$ during anxiety, [*depression*]$_{d}$ and pain as well as functional disorders of the pelvic viscera. [PMID: 15538210]

For the annotation unit above, one annotator subjectively interpreted the preposition 'during' as implying a causal relation and assigned 'x2g' to CC, but the other interpreted the word as having its literal meaning and assigned 'none' to CC. After annotator meeting, the annotators agreed to include instructions on such subjectivity issues in the annotation guidelines, and the IAA values on CC show significant improvement in the second annotation phase. Subjective readings induce disagreements on CAE values as well.

	Type	Count
Depress.	Depression	48
	Major depression	17
	Bipolar disorder	14
	Dysthymia	14
	Mood disorder	4
Antidep.	Antidepressant	47
	Fluoxetine	31
	Electroconvulsive therapy	4
	Imipramine	4
	Mirtazapine	4
	Citalopram	3
	Escitalopram	3
	Trazodone	3
	Lithium	2
	SSRI	2
	Carbamazepine	1
	Chlorpromazine	1
	Fluvoxamine	1
	Haloperidol	1
	Papaverine	1
	Perphenazine	1
	Quetiapine	1
	Reboxetine	1
	Sertraline	1
	Venlafaxine	1

Table 4: Statistics of depression/antidepressant related terms

Example 6. BACKGROUND: Indirect evidence suggests that loss of brain-derived neurotrophic factor (BDNF) from forebrain regions contributes to an individual's vulnerability for depression, whereas [*upregulation*]$_{e}$ of [*BDNF*]$_{g}$ in these regions is suggested to mediate the therapeutic effect of [*antidepressants*]$_{a}$. [PMID: 16697351]

For the annotation unit in Example 6, one annotator interpreted the verb 'mediate' as conveying the meaning of 'positive regulation' and as-

	CGE		CDL/CAE			CC		
	Inc.	**Dec.**	**Inc.**	**Dec.**	**Uni.**	**Non.**	**g2x**	**x2g**
Depress.	54(56%)	43(44%)	4(4%)	7(7%)	86(89%)	56(58%)	8(8%)	33(34%)
Antidep.	61(54%)	52(46%)	15(13%)	1(1%)	97(86%)	1(1%)	9(8%)	103(91%)
Total	115(55%)	95(45%)	19(9%)	8(4%)	183(87%)	57(27%)	17(8%)	138(65%)

Table 5: Distribution of the annotation concept values

		Gene	
		inc.	**dec.**
Depress.	**inc.**	PRKCA^{d}, MAPK3^{d}, MAPK1^{d}	ALB, TNF^{d,p}, IL2^{d}, IL1B^{d,p}, MAPK1^{d}
	dec.	MAPK1^{d}, BDNF^{d,p}, LEP^{d}, SLC6A4^{d,p}	
	uni.	DLG4, NEFL^{d}, DLG3, GFAP^{d,p}, AVP^{d}, ESR1^{d,p}, NR3C1^{d,p}, TRP, CRHR1^{d}, S100A10^{d,p}, INS^{d}, BDNF^{d,p}, GRM2^{d}, GRIA3^{d}, SV2A, IGFBP2^{d}, PENK, HTR1A^{d,p}, CD19, CD8^{d}, GRIN2A^{p}, GRIN1^{p}	PDLIM5^{d,p}, CRH^{d,p}, IL6^{d,p}, CAMK2A^{p}, CAMK2B, IL1B^{d,p}, TNF^{d,p}, IFNA1^{d}, IL2^{d}, AVP^{d}, PDYN^{p}, FCGR3A, CD4^{d}, CD8^{d}, DRD4^{d,p}, PCNT^{d}
Antidep.	**inc.**		TNF^{d,p}
	dec.	CHRM1, NOS1^{d,p}, CYP2D6^{dp}	HTR1A^{d,p}, NR3C1^{d,p}, BDNF^{d,p}, PLCG1^{d}
	uni.	HTR3A^{p}, IL1B^{d,p}, HTR2A^{d}, TNF^{d,p}, HTR1A^{d,p}, FOS^{d}, FZD3^{d}, ABCB1^{d,p}, PLA2G4A^{p}, IL6^{d,p}, CACNA1G, CACNA1I, CACNA1H, GSK3A^{d}, SLC6A3^{d,p}, SLC6A4^{d,p}, KCNK2^{d}, Defa5, VIM, TRA, BRCA1^{d}, CKB, ACTB, GFAP^{d,p}, PDE4A^{d}, CREB1^{d,p}, CCNA2, CKS1B, BAX	FOS^{d}, IL6^{d,p}, HTR2A^{d}, ALB^{d}, ADRA2A^{d,p}, HTR1A^{d,p}, BDNF^{d,p}, PDE4A^{d}, ABCB1^{d,p}, IGF1^{d}, S100A10^{d,p}, HTR1B^{d,p}, CREB1^{d,p}, PRL^{d}, PLA2G4A^{p}, SYP^{d}, NCAM1^{d}, NTRK2^{d,p}, PLCG1^{d}, SPR^{d}, Hspa9, RASEF, PDIA3, SLC6A4^{d,p}, CDKN1A, CDKN1B, BCL2^{d}, MAPK1^{d}

Genes marked with superscripts d and p are validated with DisGeNET (Piñero et al., 2015) and DEPgenes (Kao et al., 2011), respectively. The reader is referred to the published corpus for more details.

Table 6: Distribution of the annotated genes

signed 'increase' to CAE. However, the other annotator interpreted the word as conveying only the meaning of 'regulation' with no directionality and assigned 'unidentifiable' to CAE. After annotator meeting, the CAE value of the annotation unit above was set to 'increase'.

Example 7. Repeated treatment with antidepressant drugs, [*imipramine*]$_{a}$ (Imi) and fluoxetine (Flu), significantly reduced the plasma corticosterone concentration and [*enhanced*]$_{e}$ the [*BDNF*]$_{g}$ and CREB levels. [PMID: 16519925]

For the annotation unit above, one annotator interpreted the phrase '*repeated treatment*' as conveying the meaning of 'enhance' and assigned 'increase' to CAE. However, the other annotator argued that the nature of the antidepressant drugs did not change and assigned 'unchanged' to CAE.

Another cause of disagreements was the use of reasoning and prior knowledge during annotation.

Example 8. In the current paper, we propose that the rapid [*decrease*]$_{e}$ in [*insulin*]$_{g}$ level during the postpartum period may be one of the causes of [*postpartum mood disorders*]$_{d}$. [PMID: 16321476]

For the annotation unit in Example 8, one annotator claimed that there is no association between the gene *insulin* and the depression *mood disorders*, as he did not find any explicitly stated piece of information. The other annotator, however, assigned 'decreased' to CGE, as he inferred that the *mood disorders* co-occurs with *insulin* in *postpartum period*. After annotator meeting, the annotators agreed on 'decreased', and added an instruction that allows the inference using logical reasoning to the annotation guidelines.

# Phase	# Units	#Depression	#Antidepressant	#Genes	Data source
Phase 1	142	75	67	47	PubMed abstracts
Phase 2	68	22	46	42	PubMed abstracts
Total/Unique	210/106	97/5	113/20	89/82	PubMed abstracts

Table 7: The annotation phases

	CGE			CDL/CAE			CC		
	Simple	**Kappa**	**G**	**Simple**	**Kappa**	**G**	**Simple**	**Kappa**	**G**
Phase 1	1	1	1	0.92	0.69	0.88	0.76	0.47	0.64
Phase 2	0.91	0.81	0.82	1	1	1	0.97	0.93	0.96
Total	0.95	0.91	0.91	0.96	0.85	0.94	0.87	0.7	0.8

Table 8: IAA values

Example 9. All [*antidepressants*]$_a$ [*increased*]$_e$ [*c-fos mRNA*]$_g$ in the central amygdala, as previously shown, while c-fos was also increased in the anterior insular cortex and significantly decreased within the septum. [PMID: 15812568]

One annotator considered the phrase "All antidepressants increased c-fos mRNA" a universal affirmative, and just modified the antidepressant term as the universal quantifier, "All antidepressants". However, the other annotator anchored on the pre-annotated keyword "antidepressants". After annotator meeting, the annotators agreed to specify the quantification type of a term and check the scope of that quantifier.

As we refined annotation guidelines after Phase 1, the disagreements among the annotators were greatly reduced. In Phase 2, almost all the disagreements were found due to simple errors. Compared to the values from Phase 1, IAA values on CDL/CAE and CC from Phase 2 show 13.6% and 50.0% increases in terms of *G index*, respectively.

3.2 Annotation guidelines

The initial annotation guidelines were taken from Lee et al. (2013). After each annotation phase in this work, the annotators held meetings to resolve the disagreements and to revise the guidelines. Table 9 shows the final version of guidelines.

4 Discussion

In this section, we show suggestions to further automating some of the processes described in the

#	Instruction
1	Annotators should annotate the sentences only if the gene exhibits changes in its expression level and this has relations with the depression or anti-depressant related term
2	Annotators can annotate the relations between CGE and CDL/CAE utilizing linguistic clues and textual evidence
3	Annotators can infer omitted fact utilizing reasoning
4	Annotators should interpret the sentences from an 'objective point of view'
5	Annotators need not consider gene expression level changes in healthy people and people with a past history of clinical depression
6	Annotators should not infer information using their prior experience or knowledge about properties of various kinds of depression
7	Annotators should not infer information (i.e., the effects of antidepressants) using their prior knowledge about the functions of genes
8	Annotators should not infer information by using inductive reasoning
9	Annotators need not consider the certainty level of propositions.
10	Annotators need consider universal propositions and particular propositions
11	Annotators should not annotate relations between genes and mania in bipolar disorder

Table 9: Annotation guidelines

previous section, especially those of extracting depression-gene relations.

- ***ML-based event relation recognition***

Example 10. OBJECTIVE: To examine whether the pathogenesis of [*depression*]$_d$ is associated with altered [*activation*]$_e$ and expression of [*Rap-1*]$_g$, as well as expression of Epac, in depressed suicide victims. [PMID: 16754837]

Example 10 shows that there are co-occurrences whose depression and gene name pairs were identified as correct but whose relation was nonetheless incorrect. The present co-occurrence has a relation of study description rather than that of gene expression change event. Besides training to come up with the event relation classifier, we can also build a system that automatically filters out false relations (i.e., hypothesis sentences) based on the previous work such as topic-classified relation recognition (Chun et al., 2006; Kilicoglu and Bergler, 2008) and deep-syntactic parser (Ballesteros et al., 2014; Hara et al., 2005; Masseroli et al., 2006; Skounakis et al., 2008).

- ***Location and contrasting information***

Example 11. Animal studies demonstrate that some antipsychotics and [antidepressants]$_a$ [increase]$_e$ neurogenesis and [*BDNF*]$_g$ expression in the hippocampus, which is reduced in volume in patients with depression or schizophrenia. [PMID: 16652337]

Example 11, and Example 9 too, show that location information turn out to be important in studies of depression and genes may respond differently to the same antidepressant in different parts of a body. Many annotation units do not explicitly provide such location information. However, missing such information will lead to conflicts and even paradoxes among annotated or mined results.

Although the annotation concepts of the presented corpus are originally designed to represent relations between gene changes and depression/antidepressant changes, they must be made to accept other concepts and constantly changing metrics in genetic studies of depression. In this regard, we should extend the annotation scheme to include parts of a body as the location and their hierarchical relationship information.

- ***Pronouns, acronyms, and appositions***

Other difficulties we faced during recognition were in dealing with grammatical constructions such as pronouns, acronyms, and appositions. They may have coped better by using the full resolved forms of pronouns and acronyms for annotation, which in turn require the access of preceding sentences or the whole abstract in the worst case. We also found that text mining tools we used extract both the appositive phrase and the phrase in apposition, but it would be better to utilize only appositives. For example, for the following phrase, we should not annotate the word "*Tricyclic antidepressants*" an antidepressant related term, or annotate "*serotonin reuptake*" a gene.

"*Tricyclic antidepressants, selective serotonin reuptake inhibitors, and serotonin-noradrenaline reuptake inhibitors, as well as the immediate precursor of serotonin*"

Instead, we should identify the three appositives as antidepressant related terms, even if they were not included in the dictionary.

- ***Sense ambiguity of 'depression'***

We also see that using simple dictionary-based matching for detecting depression-related terms produces many ambiguous terms, some of which are not related to the mental disorder at all. In particular, the term 'depression' could also be used in a situation where a certain amount, value, or function is lowered or decreased, among others. We notice that such cases are frequently observed in biomedical texts as exemplified below:

Example 12. Lack of enteral stimulation with PN impairs mucosal immunity and [*reduces*]$_e$ [*IgA*]$_g$ levels through [*depression*]$_d$ of GALT cytokines (IL-4 and IL-10) and GALT specific adhesion molecules. [PMID: 16926565]

Example 13. LTA causes cardiac [*depression*]$_d$ by [*activating*]$_e$ myocardial TNF-alpha synthesis via [*CD14*]$_g$ and induces coronary vascular disturbances by activating Cox-2-dependent TXA2 synthesis. [PMID: 16043646]

In our initial dataset that has 1,251 occurrences of depression-related terms obtained via the simple dictionary-based matching, the term 'depression' is found 730 times, which amounts to more than half of the entire occurrences. Our corpus statistics in Table 4 also show that 'depression' is the most frequent depression-related term. This means that not a few of such terms still have potential sense ambiguities. Although we manually filtered out false positive examples in our corpus, this issue is still important since it could hinder the performance of extracting depression-related

terms in a fully pipelined system. Although a few named entity recognizers for biomedical text have been developed (Leaman and Gonzalez, 2008; Campos et al., 2013), none of these tools are capable of recognizing terms referring to depression, especially identifying 'depression' as the mental disorder, to the best of our knowledge.

It is anticipated that the disambiguation of the term 'depression' can be addressed with the conventional methods of word sense disambiguation with various features such as context information or external knowledge resources. Our data analysis suggests that local semantic features would be effective in many cases, among others. In particular, the following three types of syntactic construction could act as strong indicators for false positives: (1) prepositional phrases, (2) prenominal modifiers, and (3) coordinate constructions. First, prenominal modifiers often signal the context where some activity or amount is decreased, such as the physical malfunction ("cardiac depression"), the object or cause of inhibition ("Orx-B-induced depression", "AMPAR depression"), and the degree of decrease ("significant depression", "moderate depression"). Second, prepositional phrases provide information about the location or inhibition of a biological process ("depression in synapses", "depression of synaptic transmission", "depression of gamma interferon"). Last, coordinate constructions allow for exploiting the semantic similarity ("depression and anxiety" vs. "long-term potentiation and depression"). All of these features are highly local; syntactic dependencies do not cross the boundary of noun phrases.

Another possible approach would be to employ the document topic features by assuming that if the abstract of a document discusses the mental disorder, the term 'depression' in the abstract is also likely to refer to the mental disorder. In order to figure out what kind of terms are best indicative of documents that discuss the depressive disorder, we collected a set of 5,000 Medline abstracts that contain unambiguous domain-specific terms in our depression term dictionary such as 'depressive disorder', 'bipolar disorder', and 'antidepressant', and also collected another set of 10,000 abstracts that do not contain any of those terms including 'depression'. The chi-square statistics are employed to measure the discriminative power of terms found in each set of abstracts. Table 10 shows the 10 top-ranked terms for each of two types of term: terms that partially match one of the terms in our depression term dictionary (on the left column) and terms that are not found in the dictionary (on the right column). It is shown that many of the terms in the latter set are used in the context of diagnosis or treatment of depression. One of the possible methods is to use terms of this kind as features for training a binary classifier that determines whether a given document containing 'depression' discusses the mental disorder or not.

Terms in our dictionary		**Terms not in our dictionary**	
Term	**Score**	**Term**	**Score**
major	3414	treatment	807
antidepressant	2533	reuptake	504
disorder	1957	serotonin	475
depressive	1615	MDD	464
bipolar	986	psychiatric	450
mood	874	rating	356
disorders	695	diagnostic	340
unipolar	523	DSM-IV	312
tricyclic	441	criteria	301
depressed	409	patients	296

Table 10: Discriminative terms for documents related to the depressive disorder

5 Conclusion

In this paper, we presented a depression-specific corpus in support of the development of advanced text mining systems that target specifically at providing a comprehensive information of depression-gene relations. The annotation scheme of current version can express two events, *gene expression changes* and *depression level or antidepressant effect changes*, and the relationship between these two events. The presented corpus shows a high inter-annotator agreement. We also discussed several issues in the domain of depression and made suggestions to extend the annotation scheme further to resolve conflicts and sometimes paradoxes in the acquired knowledge for depression.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2014R1A2A1A11052310).

References

M. Ballesteros, B. Bohnet, S. Mille, L. Wanner. 2014. Deep-Syntactic Parsing. In *Proceedings of the 24th International Conference on Computational Linguistics*. 1402-1413

R. H. Belmaker, G. Agam. 2008. Major depressive disorder. *New England Journal of Medicine*, 358:55-68.

J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, T. Salakoski. 2009. Extracting complex biological events with rich graph-based features sets. In *Proceedings of the BioNLP'09 Shared Task on Event Extraction Association for Computational Linguistics*, 10-18.

D. Campos, S. Matos, J. L. Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics*, 14:54.

E. Charniak, M. Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd ACL*, 173-180.

H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, J. Tsujii. 2006. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. *BMC Bioinformatics*, 7(Suppl 3):S4.

Hee-Jin Lee, Sang-Hyung Shim, Mi-Ryoung Song, Hyunju Lee, Jong C. Park. 2013. CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations. *BMC Bioinformatics*, 14:323.

Hee-Jin Lee, Tien Cuong Dang, Hyunju Lee, Jong C. Park. 2014. OncoSearch: cancer gene search engine with literature evidence. *Nucleic Acids Research*, 42(W1):W416-W421.

T. Hara, Y. Miyao, J. Tsujii. 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In *Proceedings of IJCNLP*, 199-210.

C. F. Kao, Y. S. Fang, Z. Zhao, P. H. Kuo. 2011. Prioritization and evaluation of depression candidate genes by combining multidimensional data resources. *PLoS ONE*, 6(4):1-9.

H. Kilicoglu, S. Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 2008, 9(Suppl 11):S10.

J. Kim, T. Ohta, J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

S. V. Landeghem, K. Hakala, S. Rnnqvist, T. Salakoski, Y. Peer, F. Ginter. 2012. Exploring biomolecular literature with EVEX: connecting genes through events, homology and indirect associations. *Advances in Bioinformatics*, 2012:582765.

R. Leaman, G. Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Proceedings of the Pacific Symposium on Biocomputing*, 652-663.

M. C. D. Marneffe, B. MacCartney, C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the LREC*, 449-454.

M. Masseroli, H. Kilicoglu, F. Lang, T. Rindflesch. 2006. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics*, 7:291.

D. McClosky. 2010. *Any domain parsing: automatic domain adaptation for natural language parsing*. *PhD Thesis*, Brown University, Department of Computer Science.

D. McDonald, U Kelly. 2012. The value and benefits of text mining. *UK JISC*, [Online. Available: http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining].

E. J. Nestler, M. Barrot, R. J. DiLeone, A. J. Eisch, S. J. Gold, L. M. Monteggia. 2002. Neurobiology of depression. *Neuron*, 34:13-25.

M. Neves, J. M. Carazo, A. Pascual-Montano. 2005. Moara: a Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics*, 11: 157-169.

J. Piñero, N. Queralt-Rosinach, À . Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, L. I. Furlong. 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028.

S. Poux, M. Magrane, C. N. Arighi, A. Bridge, C. O'Donovan, K. Laiho, The UniProt Consortium. 2014. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database*, 2014:bau016.

S. Pyysalo, T. Ohta, S. Ananiadou. 2013. Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, 58-66.

M. Skounakis, M. Craven, S. Ray. 2003. Hierarchical hidden Markov models for information extraction. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 427-433.

R. Winnenburg, T. Wachter, C. Plake, A. Doms, M. Schroeder. 2008. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Briefings in Bioinformatics*, 9(6):466-78.