ISSN 2007-9737

Cause and Effect Extraction from Biomedical Corpus

Sindhuja Gopalan[1], Sobha Lalitha Devi [2]

[1,2] Anna University, AU-KBC Research Centre, MIT, Chennai, India

{sindhujagopalan, sobha}@au-kbc.org

Abstract. The objective of the present work is to automatically extract the cause and effect from discourse analyzed biomedical corpus. Cause-effect is defined as a relation established between two events, where first event acts as the cause of second event and the second event is the effect of first event. Any causative constructions need three components, a causal marker, cause and effect. In this study, we consider the automatic extraction of cause and effect realized by explicit discourse connective markers. We evaluated our system using BIONLP/NLPBA 2004 shared task test data and obtained encouraging results.

Keywords. Discourse relation, cause-effect, discourse connective, causal entity, discourse parser, named entity recognition.

1 Introduction

The cause and effect is defined as a relation established between two events, where the first event acts as the cause of the second event and the second event is the effect of the first event. One cause can have several effects. A cause is why an event happens. The effect is an event that happens because of cause. The cause and effect occurs based on the following criteria, where cause has to occur before effect, and whenever the cause occurs the effect has to occur. Causality is temporally bounded and hence cause always precedes their effect. Any causative constructions need three components- causal marker, cause and effect. The cause-effect can be explicit or implicit. The explicit cause-effect can be figured out by exploring some signal words. In case of implicit cause-effect, the relation will be established between the events, but they are not explicitly realized using a causal marker. The cause-effect at discourse level is classified into two, reason and result. The signal words for causation can be expressed in many ways as causal verbs (cause, make, kill, etc.), discourse connectives (because, thus), causal words (as a result of) and etc.

Example 1:
Emerging evidences indicate that Snail causes a metabolic reprogramming, bestows tumor cells with cancer stem cell-like traits, and additionally, promotes drug resistance, tumor recurrence and metastasis.

In Example 1, the causal verb “causes” establishes a simple causality between two noun phrases “Snail” and “a metabolic reprogramming”, where the first noun phrase is the cause and the second noun phrase is the effect. It is relatively a simple task to establish the cause-effect and extract them in this example.

Consider the Example 2, where the cause and effect is established between two clauses. Here, the automatic extraction of this relation is relatively a difficult task.

Example 2:
M2 leukemic blast cells behave differently because they undergo monocytic differentiation with both the differentiation inducers.

In the above Example 2, the event “they undergo monocytic differentiation with both the differentiation inducers” is the cause and “M2 leukemic blast cells behave differently” is the effect. Here, the explicit discourse connective “because” is the causal trigger that indicates the presence of a cause-effect. This example shows the type of cause-effect known as reason causal relation. The causal connective like because, since, as, due to, etc. shows reason cause relation.

Consider another example, Example 3, where the cause-effect is realized between sentences.

Example 3:
Spinal delivery of 8-Br-cGMP, a PKG-activating cGMP analog, without subsequent formalin treatment also caused an increase of PKG-I protein expression. ***Hence****, the up regulation of PKG-I might possibly be mediated by cGMP itself*.

In the above example, the cause-effect occurs across sentence, where the first sentence is the cause and the second sentence bounded to the causal trigger "Hence" is the effect. Here, the connective shows the type of cause-effect known as result causal relation. The causal connectives like consequently, thus, therefore, hence, as a result, etc. shows result causal relation. It is not easy for a machine to identify such cause-effect and requires knowledge of the structure of the language. Hence, the task of extracting the cause and the effect is difficult. Consider the example below.

Example 4:
E2F is controlled by the Retinoblastoma Tumor Suppressor Protein Rb. Implicit (Because) Rb recruits chromatin remodeling factors, DNMT1 and a histone methyl transferase.

Example 4 shows an implicit cause-effect between two discourse units. The second unit justifies the claim in first unit. Here the connective "Because" is inserted implicitly. An Implicit cause-effect is inferred in two adjacent sentences or clauses without any explicit lexical or grammatical marker. In this work the implicit cause-effect are not considered. This study pertains to the identification of explicit cause-effect from biomedical text.

In this study, we consider the automatic extraction of cause-effect realized by explicit discourse connective. The cause-effect expressions that function at discourse level in biomedical corpus were studied to efficiently identify the cause and effect. It is difficult to extract semantic relation from English texts. There are many applications that benefit from deep semantic relations. Cause-effect is one such important semantic relation for many applications [2]. Identification of biomedical named entities (BNEs) with information of explicit causal discourse relation would benefit the development of more sophisticated information extraction systems. Hence we have identified the BNEs from discourse parsed output. This will further influence the development of multiple tasks, such as question answering system, discovering new facts and information retrieval systems.

The following section details the related works to the identification of cause-effect. In section 3 the present work is discussed. In section 4 we have presented our approach in developing the cause-effect extraction system. The results are discussed in Section 5. The paper is concluding in section 6.

2 Related Works

The literature survey for cause-effect extraction shows that different perspectives were used to achieve the goal of extraction of cause-effect.

i) The first perspective is the automatic acquisition of causal knowledge using causal verbs, causal words, compound nouns and etc. as shown in Example 1. Although they are not directly expressed as a cause-effect in the text, they show the presence of causality in the text based on the contextual knowledge.

ii) The second perspective is that of works which aimed at identification and extraction of discourse relations marked by discourse connectives. The cause-effects are only a part of this bigger set of relations. This work concentrates on a grammatical sub-class, namely connectives.

A method for automatic detection and extraction of causal relations was devised by [2]. The lexico-syntactic pattern noun phrase 1 – causal verb – noun Phrase 2 was used to identify the causality relation on TREC9 text collection corpus. C4.5 decision tree learning algorithm was used and they obtained precision of 73.91% and recall of 88.69%. [5] focused on a sub-task of causal relations detection. An open-class set of linguistic markers, AltLex recognized in PDTB were analyzed for causality relation. They expanded the definition of AltLex to include these markers when they are present within the sentences and obtained an accuracy of 79.58% and true F-score of 80.90%.

A minimally supervised approach, based on focused distributional similarity methods and

discourse connectives, for identification of causality relations between events in context was presented by [1]. They obtained F-score of 43.9% for cause-effect associations and 46.9% for cause-effect associations between two events and discourse connectives. [12] introduced the theory of granularity and described different approaches to identify granularity in natural language. The causality in granular theory is defined as where an event is caused by a set of sub-events. A part of LDC - New York Times Annotated corpus was used for their experiments.

A knowledge extraction and knowledge discovery system that extracts causal knowledge from textual databases was developed by [7]. Causal links and causing verbs were used as causal identifiers. They obtained F-score of .763 for causality identifier, .508 for cause identification and .578 for effect identification.

An annotation scheme for enriching biomedical domain corpora with causality relations was defined by [11]. This schema has subsequently been used to annotate 851 causal relations to form Biocause, a collection of 19 open access full text biomedical journal articles belonging to the subdomain of infectious diseases. The semantic relation of causality or cause-effect, how it is marked in Tamil, how the causal markers in Tamil manifest in texts, their syntactic and semantic properties and how this information can be represented so as to handle causal information and reasoning was studied by [6].

The cause-effect expression in Tamil text was identified by [10]. They have classified the causal links and causal verbs for cause-effect extraction. The system performs with an overall precision of 73.89 % and Recall of 72.09 %. From the literature survey, we observed that the works on cause and effect in biomedical domain are limited. Causality lies at the heart of biomedical knowledge and plays a key role in diagnosis, pathology, systems biology, etc. The automatic identification of cause-effect can greatly reduce the human workload and help in curation of disease models. Hence we have tried to apply the discourse parser and BNE recognition (BNER) models to extract the cause-effect expressions from the biomedical text. In the next section the aim of the present study is briefed.

3 Present Work

Our work focuses on the analysis of the causality in biomedical corpus based on various discourse connectives that denote cause-effect. In this work we consider the identification causal relations realized across clauses and sentences using discourse connectives. We have applied the discourse parser and BNER system to identify the causal relations and causal entities from the text. Identifying the causal relation across sentence or clause is a difficult task as the cause and effect does not involve the complete sentence connecting the causal marker. Consider the below Example 5.

Example 5:
*CNP may produce its effect directly on dopaminergic neurons **because** we found that its receptor, guanylyl cyclase GC-B, was expressed in the mesencephalon where dopaminergic neurons originate, as well as in their projection fields*.

At discourse level, a connective connects two discourse units, argument 1 (arg1) and argument 2 (arg2) [13]. Here, “because” is a subordinator connective that connects two clauses. The clause that follows the connective is arg2 and the other clause is arg1. The arg2 acts as the reason for the event occurred in arg1. “CNP may produce its effect directly on dopaminergic neurons” is the arg1 and “its receptor guanylyl cyclase GC-B was expressed in the mesencephalon” is the arg2. The minimal unit required to realize the cause-effect is considered. Here, arg2 that acts as the reason is the “cause” and arg1 is the effect. Guanylyl cyclase GC-B, CNP, mesencephalon, dopaminergic neurons BNEs and thus the example shows that the identification of BNEs in cause and effect aids in explaining the relation further.

This system makes a significant contribution towards the question answering system particularly the answering of Why-Questions and text summarization. Causality in the medical domain is mainly concerned with health and diseases. It mainly involves the causation or origination of a disease. Hence extraction of causal information will play an important role in the identification of cause and effect of abnormal conditions, which may be caused by various reasons. Identification of BNEs will be helpful in

identifying the cause involved in the abnormal condition. This system can also be used in text summarization to extract the text causal relations that produce a coherent text.

4 Our Approach

In this study we have applied the task of explicit discourse relation identification and BNER to the task of extraction of cause-effect and causal entities from biomedical text. First, the input text was preprocessed and the preprocessed output was given as input to the discourse parser module for identification of connective, sense and its arguments. The required columns from discourse analyzed output was given as input to BNER module.

After identification of the BNEs, the discourse output and BNER output were merged. The final output obtained after merging discourse and BNER output contained connective, sense, arguments and BNEs. The causal relations were extracted from this output based on the sense of the connectives.

If the sense of the connective belongs to class "cause", then they were extracted. The cause-effect and the causal entities were extracted from this output. The system architecture is detailed in further sections.

4.1 Corpus Used for Evaluation

For evaluating the system, we used BioNLP/NLPBA 2004 test data. The data includes the annotated collection of MEDLINE abstracts from the GENIA project. 404 abstracts were used. The data contains 4260 sentences and 96780 tokens.

4.2 Preprocessing

In the preprocessing step the input text was preprocessed to convert it to the format required for explicit discourse relation extraction and BNE identification. The sentences were split, tokenized; PoS and chunk tags were added using Genia Tagger [14].

This tagger is a useful pre-processing tool. After preprocessing, the text was in column format with tokens, PoS and Chunk added in consecutive columns required for further processing. This output was passed as input to the discourse parser module.

4.3 Discourse Parser Module

We extracted the explicit discourse relations using supervised machine learning technique, Support Vector Machines (SVMs). We used Yamcha tool [8], an open source implementation of SVMs algorithm.

a. Corpus Used

We developed the biomedical corpus annotated with explicit discourse relations, by following the guidelines of PDTB, a large-scale resource of annotated discourse relations and their arguments [13]. We concentrated on the extraction of explicit discourse relations. The corpus was built on abstracts collected from PubMed Central. The corpus was annotated with explicit discourse relations.

There was 2957 explicit discourse relation. The Cohen's kappa measure was calculated for explicit connectives and its argument boundaries. The Kappa score for explicit connectives is .94, for arg1 start is .86, for arg1 end is .0872, for arg2 start is .863 and for arg2 end is .832. In the case of annotation of arguments there is a substantial agreement between the annotators for all the argument boundaries. The language models for connective, sense and argument identification were built using this corpus.

b. Feature Extraction

After analyzing the preprocessed text the features were extracted. Feature selection is important for statistical machine learning, as they play an important role in improving the system's performance. We used a set of linguistic features for identification of connectives and its arguments. For our work we have used simple and minimal number of features. Lexical features like word, PoS, chunk, clause and their combinations were used for connective identification. For sense identification, we have used lexical features and connective itself is as a feature. Connective is an exceptional feature for sense identification.

The argument boundary in most of the cases will be the start or end of the sentence or will succeed or precede the connectives. Since the arguments are also clauses, the clause tagging helps in the identification of arguments. The positions of the connectives were also used as feature for argument boundary identification. The sentence positions with respect to the connective and sentence boundaries were used as features to identify the argument boundaries. We developed separate models for argument boundaries and hence the previous boundaries identified were used as feature for subsequent boundary identification.

c. Discourse Relation Identification Task

The whole task of connective and argument identification was divided into three sub tasks. Motivated by the work [9] and [4], we designed our system as a pipeline, where the relations are identified in sequential order. First, the system identified and predicted the discourse connectives and their sense. Then, using the identified connectives, arg1 and arg2 spans were identified and extracted.

Connective *classification.* The system was trained using lexico-syntactic features like word, Parts-of-Speech (PoS), chunk, combination of word, PoS and chunk and clause in a window size of 5 for the task of connective identification. The lexicon itself acts as a good feature to identify the discourse connectives. The average of 10-fold cross validation is presented in Table 3.

Sense *identification.* After classifying the connectives, we identified the sense of the connectives. The main class, type and sub type of the connective sense was identified. Lexical features and connectives were used as features for sense identification. We developed gold standard and automatic parser for sense identification. For gold standard parser the gold standard connectives were used as feature and for automatic parser the connectives identified in first sub-task were used as feature. After identifying the connectives and its sense, the third sub task, argument identification was performed.

Argument *identification.* The overlapping sequences shown is Example 6, has made the task of argument identification using ML techniques difficult. Example 6(a) shows that the two relation share same arguments and hence it is difficult to identify the boundary of the arguments.

Example 6:

(a) <argj1><argi1>Lmx1b promotes Wnt1/Wnt signaling</argi1></argj1>, <CONi>and</CONi> <argi2><CONj>thereby</CONj> <argj2>increases midbrain size and dopamine progenitor allocation</argi2></argj2>.

(b) <argi1>Lmx1b promotes Wnt1/Wnt signaling</argi1>, <CONi>and</CONi> <argi2>thereby increases midbrain size and dopamine progenitor allocation</argi2>.

(c) <argi1>Lmx1b promotes Wnt1/Wnt signaling</argi1>, and <CONi>thereby</CONi> <argi2>increases midbrain size and dopamine progenitor allocation</argi2>.

To overcome the problem of overlapping sequences, we processed one connective at a time. Since the connectives were identified in the first sub task, we first processed the connective “and” and then “thereby” as shown in 6(b) and 6(c).

Hence connective is an important constraint for identification of arguments. We developed two types of parsers gold parser and automatic parser. For gold standard parser the gold standard connectives were used to partition the data and were used as feature for argument identification. For automatic parser the output from the connective identification task was used to partition the data.

The development of the models include four phases i.e. identification of arg1 start, arg1 end, arg2 start and arg2 end. The argument boundaries were identified in the following series, arg2 start, arg1 end, arg1 start and arg2 end. The output from one phase was fed as input to next phase. The choice of order of identification of boundaries was made based on the idea that it will be easier to identify the boundaries that are close to the connective. The boundaries of the arguments were identified, instead of identifying the whole argument.

After identifying the boundaries, the outputs were merged. In cases of overlapping boundaries the boundary with higher probability score was selected. The results are detailed in next section. All four types of discourse relations (comparison, contingency, expansion and temporal) were identified. The output from discourse parser

module had connectives, sense and arguments identified. This output was given as input to BNER module.

d. Biomedical Named Entity Recognition

The required column from discourse parser output was obtained and was given as input to the BNER module. We used the system developed by [3] for BNER. The BNER module identifies Protein, Gene, DNA, RNA, chemical, cell type, cell line, disease, cellular component, and gene-protein complex. The output in column format was converted to row format as shown in Example 6(b). This output was given as input to cause-effect extraction module.

e. Cause-Effect and Entity Extraction

The output from discourse parser module and BNER module is merged in such a way that the final output contains token and connective, sense, argument and NEs identified. The explicit causal relation from this output was extracted based on the sense of the connectives. The discourse relations that had sense class contingency, type cause were extracted. Consider the below Example 7, where 7(a) is the input text and 7(b) is the output from discourse parser and NER module.

Example 7(a)

As mercury is nephrotoxic and neurotoxic, it is interesting to note that post-treatment of vitamin E showed more protection in the kidney compared to pre-treatment.

Example 7(b)

<CON>As</CON> (SENSE: Contingency: Cause: reason) <ARG2><CHEMICAL> mercury </CHEMICAL> is nephrotoxic and neurotoxic</ARG2>, it is interesting to note that <ARG1>post-treatment of <CHEMICAL>vitamin E</CHEMICAL> showed more protection in the kidney compared to pre-treatment</ARG1>.

From the Example 7(b), we observe that the connective “As” has two arguments arg1 and arg2. While extracting cause-effect based on the explicit discourse connectives, the arg2 of reason connectives like because, since, as and etc. is the cause and arg1 is the effect. Whereas, for result connectives like hence, therefore, thus, as a result and etc. arg1 is the cause and arg2 is the effect as shown in Example 8.

Hence, while extracting the explicit cause-effect from discourse analyzed corpus, it is important to note whether the connective is reason connective or result (this information can be obtained from sense of the connective). Based on this, cause and effect were extracted. Since one discourse connective is processed at a time, the problem of multiple cause-effects present in a sentence is handled efficiently.

Example 8:

The results of the present study clearly indicated that <ARG1><CHEMICAL> quercetin </CHEMICAL> has a protective role against reserpine-induced <DISEASE>orofacial dyskinesia</DISEASE> and <DISEASE>memory impairment</DISEASE></ARG1>. <CON> Consequently </CON> (Sense: Contingency: Cause: Result), <ARG2>the use of <CHEMICAL>quercetin</CHEMICAL> as a therapeutic agent for the treatment of <DISEASE>TD</DISEASE> should be considered</ARG2>.

Example 9:

<CON>As</CON> (SENSE: Contingency: Cause: reason) <ARG2><CHEMICAL> mercury </CHEMICAL> is nephrotoxic and neurotoxic </ARG2>, it is interesting to note that <ARG1> post-treatment of <CHEMICAL> vitamin E </CHEMICAL> showed more protection in the <ORGAN>kidney</ORGAN> compared to pre-treatment</ARG1>.

Connective: *As*.

Cause: *mercury is nephrotoxic and neurotoxic*.

Effect: *post-treatment of vitamin E showed more protection in the kidney compared to pre-treatment*.

Causal Entity: *mercury – CHEMICAL*.

Effect Entity: *vitamin E – CHEMICAL*.

Kidney – *ORGAN*.

The final output from cause-effect extraction module is shown in Example 9. The results are discussed in next section.

Table 1. Result for Connective Identification (in %)

Task	Precision	Recall	F-score
Connective Classification	89.69	82.3	86

Table 2. Result for Sense Identification (in %)

Task	Gold Parser			Automatic Parser		
	P	R	F	P	R	F
Sense Identification	97.9	91.7	94.8	90.1	82.4	86.3

Table 3. Results for Argument Boundary Identification (in %)

Task	Gold Parser			Automatic Parser		
	P	R	F	P	R	F
Arg1 start	84.2	80.3	82.2	82.9	79.4	81.2
Arg1 end	93.5	87.5	90.5	92.9	85.9	89.4
Arg2 start	93.1	90.4	91.8	90.8	88.9	89.9
Arg2 end	85.8	79.4	82.6	82.9	72.8	77.9
Average	89.2	84.4	86.8	87.4	81.8	84.6

Table 4. Results for Cause-Effect System in %

Task	Precision	Recall	F-score
Causal Marker	88.24	78.36	83.3
Cause	89.02	54.48	71.75
Effect	94.94	55.97	75.46

5 Results and Discussion

The test data of BIONLP/NLPBA 2004 task data was used to evaluate the system. We evaluated the performance of our system using precision, recall and F-score measure. Precision is the number of labels correctly perceived by the system from the total number of labels identified, Recall is the number of labels correctly detected by the system by the total number of labels contained in the stimulus text and F-score is merely the mean of precision and recall.

5.1 Discourse Parser

In this section we present the results for identification of explicit connectives, sense and its arguments. For connective identification we obtained a precision of 89.69%, Recall of 82.3% and F-score of 86%. The results are tabulated in below Table 1. The analysis of the output for connective identification showed that the errors generated were mainly due to data sparsity and propagation of errors from previous modules. Another major error was due to the fact that all conjunctions are not connectives and hence in some cases identification of conjunctions that are not connectives results in fall in performance measures.

For sense classification, using gold standard parser we obtained F-score of 94.8% and by using automatic parser we obtained F-score of 86.3%. The sense identification mainly depends on the connective feature and hence the performance of sense identification also depends on performance of connective identification. The same connective has multiple senses and hence in some cases the sense was wrongly predicted by the system. The results for sense identification are presented in Table 2. The letters “P”, “R” and “F” in Table 2 and Table 3 refers to Precision, Recall and F-score respectively.

The average F-score of argument identification using gold parser is 86.8% for automatic parser is 84.6%. The error analysis showed that the errors were mainly due to paired connectives (e.g.: neither-nor, not only -but also), the argument containing multiple sentences and position of the connectives. The position of the connectives varies, where it may occur at sentence initial, medial or final position. Since the arguments remain syntactically bound to connective, the position of the connective generates error in argument boundary identification. The results for argument boundary identification are presented in Table 3. The output from discourse parser is given as input to BNER module. The BNEs in the text were identified

5.2 Cause-Effect Extraction

The output obtained from discourse parser module contained 1332 explicit discourse relations. There

were totally 134 explicit causal relations in the evaluation data. Out of 134 causal connectives, the system retrieved 119 causal connectives and out of which 105 connectives were correctly identified. We obtained a precision of 88.24%, recall of 78.36% and F-score of 83.3% for causal trigger.

After identifying the connectives, the arguments were identified. Out of 134 causes, 82 causes were retrieved by the system and 73 were correct. The precision for cause identification is 89.02%, recall is 54.48%, and F-score is 75.46%. Out of 134 effects, 79 effects are identified by the system and 75 were correct. The precision for effect is 94.94%, recall is 55.97% and F-score is 75.46%. Out of 134 cause-effects, for 68.65% cause-effect expression at least one boundary was identified properly and for 73.13% cause-effect expression at least one boundary was identified properly for effect. Totally 63 cause-effect (causal trigger, cause and effect) were correctly identified by the system. The results for cause-effect extraction are tabulated in Table 4. The results obtained are encouraging and are "comparable" with the results of state-of-art systems described in section 2. [7] have developed a system for biomedical domain and when compared with this system, we obtained better results.

Then we analyzed the errors generated by the system. Consider the Example 10 below, where "as" is not a connective in this particular instance, still is identified by the parser as connective resulting in false positives.

Example 10:
<ARG1> Detailed and exact illustration of the process of hematopoiesis will provide an opportunity to revive hematopoiesis </ARG1> <CON> as </CON> (sense:contingency: cause:reason) <ARG2>one of the most fascinating targets of research in developmental biology </ARG2>.

Example 11:
<CON>Since</CON> (sense: contingency: cause: reason) <ARG2>the effect of {CHEMICAL}cortisol{/CHEMICAL} was additive to that of {CHEMICAL}PgE2{/CHEMICAL} and was not changed by phosphodiesterase inhibitors</ARG2>, it is conceivable that <ARG1>the hormone acts at a level different from the {PROTEIN}adenylate cyclase{/PROTEIN} – {CHEMICAL}phosphodiesterase{/CHEMICAL} system</ARG1>.

Example 11 shows a complex causal relation. The causal trigger in this example is "since" and occurs at the sentence initial position. It is a reason connective. The arg2 usually follows the connective and in this instance, the discourse unit following the connective is arg2 and the other unit is arg1 as shown in Example 11. Here, arg2 is the cause and arg1 is the effect. Due to this complex sentence structure the system was not able to identify the cause-effect boundaries properly.

Example 12:
<ARG1>No alterations of {CELL_TYPE}thymocyte subpopulations{/CELL_TYPE} were seen, suggesting that changes in the percentage of {CELL_TYPE}CD4+ CD8+ thymocytes {/CELL_TYPE} after administration of {CHEMICAL}androgens{/CHEMICAL} depend on the presence of functional {PROTEIN}androgen receptors{/PROTEIN}</ARG1>.
<CON>Thus</CON> (sense:contingency: cause:result), <ARG2>it is concluded that {CHEMICAL}androgens{/CHEMICAL} indirectly accelerate {CELL_TYPE} thymocyte {/CELL_TYPE} apoptosis in vivo</ARG2>.

The example 12 shows that "thus" is the causal trigger and result connective that connects two sentences (cause and effect). In this example we observe that both cause and effect has complementizer clause. Here, the cause and effect boundaries were wrongly identified. Since, "thus" is a result connective, arg1 is the cause and arg2 is the effect.

Error analysis have paved way for further improvement in the performance of the system and hence as future work we will try to improve the system's performance based on these observations.

6 Conclusion

In this paper the application of explicit discourse relation extraction and BNER recognition to the task of cause-effect extraction is discussed in detail. In this study the case-effect is studied at discourse level. The explicit causal discourse connective is used as causal marker to identify the

cause-effect from the text. We used BIONLP/NLPBA test corpus to evaluate the system. We obtained F-score of 83.3% for causal marker, 71.75% for cause, 75.46% for effect. The results obtained are encouraging and shows state-of-art performance. The error analysis has paved way for the improvement in system's performance.

References

1. **Do, Q.Z., Chan, Y.S., & Roth, D. (2011).** Minimally Supervised Event Causality Identification. *Conference on Empirical Methods in Natural Language Processing,* Edinburgh, UK, pp. 27–31.
2. **Girju, R. (2003).** Automatic Detection of Causal Relations for Question Answering. *ACL 2003 workshop on multilingual summarization and question answering*, Sapporo, 12, pp. 76–83.
3. **Gopalan, S. & Lalitha Devi, S. (2016).** BNEMiner: mining biomedical literature for extraction of biological target, disease and chemical entities. *International Journal of Business Intelligence and Data Mining,* Vol. 11, No. 2, pp. 190–204.
4. **Gopalan, S. & Lalitha Devi, S. (2016).** BioDCA Identifier: A System for Automatic Identification of Discourse Connective and Arguments from Biomedical Text. *Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Osaka, Japan, pp. 89–98.
5. **Hidey, C. & McKeown, K. (2016).** Identifying Causal Relations Using Parallel Wikipedia Articles. *54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp.1424–1433.
6. **Lalitha Devi, S. & Menaka, S. (2011).** Semantic Representation of Causality. *Language in India, Special Volume: Problems of Parsing in Indian Languages,* Vol. 11, No. 5, pp. 60–63.
7. **Khoo, C.S.G., Chan, S., & Niu, Y. (2000).** Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. *8th Annual Meeting on Association for Computational Linguistics*, Hong Kong, China, pp. 336–343.
8. **Kudo, T. & Matsumoto, Y. (2000).** Use of Support Vector Learning for Chunk Identification. *Fourth Conference on Computational Language Learning and the Second Learning Language in Logic Workshop*, Lisbon, Portugal, pp.142–144.
9. **Lin, Z., Ng H.T., & Kan, M.Y. (2012).** A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering,* Vol. 1, No. 1, pp. 1–35.
10. **Menaka, S., Malarkodi, C.S., & Lalitha Devi, S. (2014).** A Study on Causal Relations and its Automatic Identification in Tamil*. 2nd Workshop on Indian Language Data: Resources and Evaluation*, Reykjavik, Iceland, pp. 34–40.
11. **Mihaila, C., Ohta, T., Pyysalo, S., & Ananiadou, S. (2013).** BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics,* Vol. 14, No. 2.
12. **Mulkar-Mehta, R., Welty, C., Hobbs, J.R., &Hovy, E. (2011).** Using Part-Of Relations for Discovering Causal Relations, *Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, Florida,* US, pp. 57–62.
13. **Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., & Webber B. (2008).** The Penn Discourse Treebank 2.0. *In: Sixth International Conference on Language Resources and Evaluation (LREC'08),* Marrakech, Morocco, pp. 2961–2968.
14. **Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. (2005).** Developing a Robust Part-of-Speech Tagger for Biomedical Text. *Lecture Notes in Computer Science,* 3746, pp. 382–392.

Article received on 22/12/2016; accepted on 26/02/2017.
Corresponding author is Sindhuja Gopalan.