UNIVERSITÉ DE LORRAINE

Loria
Laboratoire lorrain de recherche
en informatique et ses applications

Master Informatique

Traitement Automatique des Langues

Master Thesis

Mining Text at Discourse Level

Author:
Anh Duc VU

Supervisor:
Maxime Amblard
Yannick Toussaint

June 20, 2016

Contents

1 Introduction

According to the statistics of WorldBank in 2014, there are 40.7 Internet users per 100 people[1]. This number has grown continuously for several years. It means nearly a half world population are using Internet and as a result, there are more and more data generated on Internet each day under various formats such as text, video, spreadsheets, etc. It makes Internet a giant warehouse which host an enormous number of data sources with various types of information. Several research domains have been presenting due to various studies aiming to explore useful information from these sources. Text Mining is such a research domain that has the objective to study the process of extracting useful information and knowledge from text documents. It combines the technique from several domains such as Data Mining, Machine Learning, Natural Language Processing, etc. [Hotho et al., 2005].

This research concentrates on a specific type of text document, the biomedical text. The biomedical documents are texts, literatures, reports, etc. that come from the studies on medicine and biology applied for health-care purpose. There exists a branch of text mining called Biomedical Text Mining that aims to explore knowledge from biomedical texts. Its studies mostly about the Named Entity Recognition, Text Classification, Relation Detection, etc. [Cohen and Hersh, 2005]. For instance, Named Entity Recognition tasks focus on identifying drug names, gene names, while Relationship Extraction tries to find the relations between two given entities, etc.

These tasks mostly pay attention on delving knowledge from words or set of words by using the bag-of-words model in which a document can be represented by a vector of representative words it contains. This model has proven its utility since it has been employed for many applications. However, we believe that one can extract more precious knowledge that the model of vector of words could not provide. For a sake of illustration, let consider the example (1), if we pull apart (1a) into two separated sentences (1b) and (1c), we may not understand why *co-expressed paralogs is not an important source of false positives*. Otherwise, when these two sentences are put together into a context as in (1a), one can easily give the reason for the fact in the second sentence that *co-expressed paralogs is not important* since in first sentence stated that the *paralogous genes* has no influence on *the number of detected motifs* when they are removed *from each cluster*. In other words, we find that the (1a) reports information about *removal*, *paralogous genes*, *not affect* while (1b) mentions about *paralogous genes*, *not an important source*. Standing at the beginning of (1b) is the adverb *Consequently* that has grammatical function to link two clauses or sentences together and convey the meaning of cause-result relation. With this analysis, we can give the statement in (1b) a reason from (1a).

(1) a. Removal of paralogous genes from each cluster did not affect the number of detected motifs. Consequently, co-expressed paralogs is not an important source of false positives.

[1] http://data.worldbank.org/indicator/IT.NET.USER.P2/countries/1W?display=graph

b. Removal of paralogous genes from each cluster did not affect the number of detected motifs.

c. Consequently, co-expressed paralogs is not an important source of false positives.

The study of text at lexical (word), syntactic (sentence structure) or semantic (sentence meaning) level can not give us this kind of information. Therefore, to acquire more useful knowledge, we need to go beyond the sentence boundary to a higher level, the discourse level, where we will study the interactions and meaning between sentences.

Normally, a text transfers its author's ideas. These ideas will be express through the words to be selected and the sentences to be made from these selected words. Obviously, as we have seen in the previous example, that is not enough. The writer also have to use some linguistic techniques to link separated sentences together to express more information. Depending on the writing skill of the author, the text may have a high or low level of coherence, i.e. sentences are well connected or not. The means that help link sentences is called discourse relations.

Our objective is to experiment the contribution of discourse relation in mining the biomedical texts. In this literature, we present the first stage of this work concerning the process of extracting discourse relations from a collection of biomedical texts. Biomedical Discourse Relation Bank [Prasad et al., 2011], a corpus contains 24 full biomedical literatures, has been selected for our task because its discourse relation annotations will help us in evaluating the precision of our detected relations. For extraction task, we employ here in this work the pattern approach due to its simplicity of implementation the end-to-end product. The extracted relations then will be provided to the text mining step to delve knowledge from corpus, that is the subsequent stage and is not presented here, in this document.

For the rest of this text, we will first reserve section 2 to introduce the background knowledge. Section 3 talks about some copora with discourse relation annotation. We then describe our methodology to identify discourse relations in section 4. Section 5 is dedicated to present our results. Finally, we give some discussion on our method and our findings in section 6.

2 Background

In this section, we will consider some theories on discourse relations that provide us frameworks allowing carry out the discourse parsing task. However, before going deeper into each theory, since we are working on discourse, we should have a basic understanding about it. When talking about discourse, one usually think about a conversation or a discussion. But, in the study of discourse and related domains, things are more sophisticate. As stated in [Jurafsky and Martin, 2008], "language does not normally consist of isolated, unrelated sentences, but instead of collected, structured, coherent groups of sentences. We refer to such a coherent structured group of sentences as a discourse". In reality, the studies on discourse share the same point of view with this statement since they try to model the coherence of documents [Mann and Thompson, 1987], to explain the way sentences are connected [Asher and Lascarides, 2005], etc. Section 2.1 considers in detail two theories that are related to our research. Section 2.2 then presents the employment of these theories in the discourse parsing task. Section 2.3 finally introduces discourse inference theory that help deduce hidden discourse relations from evident ones.

2.1 Discourse Structure Theories

Despite the fact that different discourse theories share a common idea on connecting sentences in a text, they may differ from each other due to their approaches, their starting points, or their motivations, etc. Here in this section, we talk about two theories, Rhetorical Relation Theory with its perspective is considering the text as a whole comprise text spans, each of which is rhetorically connected to other. The other theory is Segmented Discourse Representation Theory with the approach from logical point of view. The definition of discourse relations of these two theories will help develop our discourse parser.

2.1.1 Rhetorical Structure Theory

Proposed by [Mann and Thompson, 1987], Rhetorical Structure Theory (hereafter RST) is a theory that studies the organization and the coherence of text documents. A document in this theory is viewed as a sequence of elements called *discourse units* which actually are a non-overlapped spans of text. A discourse unit should have an independent and functional integrity. In other words, it should have a complete grammatical structure and should convey a meaningful information. Two discourse units can be connected by a relation, named rhetorical relation. This relation has two representative structure, one express its semantic structure and other present its physical structure. Let delve into details of these properties.

The first property describes the role of units in a relation. A discourse unit that takes part in a relation can play the role as a nucleus or a satellite. The nucleus conveys the essential, important information of a relation while the satellite contains the supportive information for the nucleus. A discourse relation always expresses some information, and hence, it contains at least a nucleus. Therefore, the role structure of a rhetorical relation can be Nucleus-Satellite or Multi-Nucleus. The subsequent example will give an illustration for terms we have mentioned.

Look at example (2), one can find that (2b) adds more information to clarify the information conveyed by (2a). Therefore, an *elaboration* relation is established between these two sentences, (2a) plays the role of nucleus, contains the main information about John's good news while (2b) is a satellite that gives an explanation for this announcement.

(2) a. John have just found a good position.

b. He is going to work as system architecture at Google.

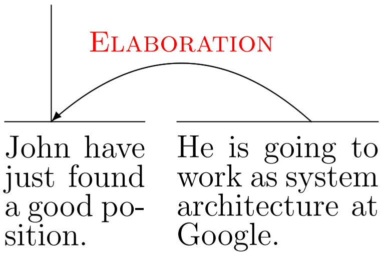

ELABORATION
John have just found a good position.
He is going to work as system architecture at Google.

Figure 1: Schema for Nucleus-Satellite relation in (2)

Example (3) demonstrates another type of role structure in which both discourse units express important information. The first unit presents an eventuality and the second unit negatives it. The two units are linked by the word *but*. In RST, this relation is called CONTRAST.

(3) a. John promised to take his son to the zoo

b. but he did not.

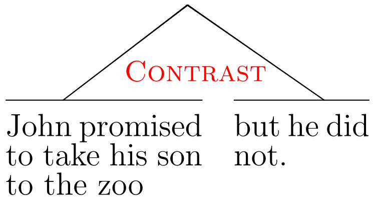


Figure 2: Schema for Multi-nucleus relation in (3)

The second structural type, named schema, determines the physical arrangement of units in a relation. RST have 5 sorts of schema as in fig. 3. To interpret a schema, we should know that the straight line marks the nucleus unit and the arc stands for the relation between units, the direction of arrow indicate the position of nucleus unit.

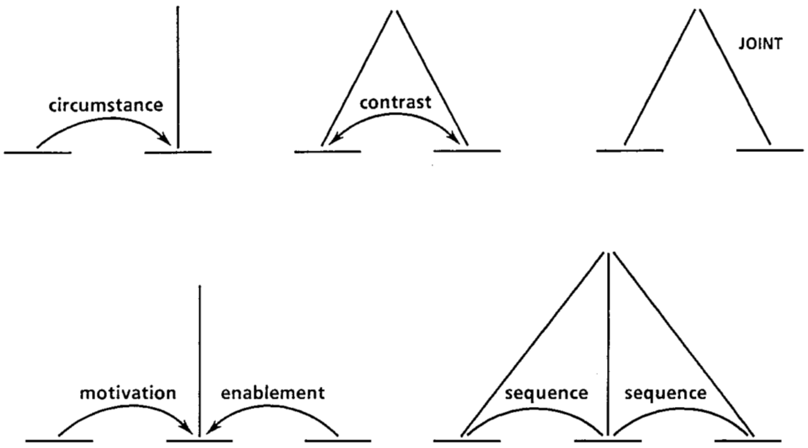


Figure 3: Schemas *[Mann and Thompson, 1987]*

We have discussed about the structure of a rhetorical relation. So, how many rhetorical relations does RST define? Table 1 presents the set of 12 main relations in RST, in which some relations can be divided into sub relation to get a higher level of detail. In total, it has 23 relations. According to [Mann et al., 1989], this set of relations is expandable, i.e., one can define more relations and add to the set. As far as we know, RST Discourse TreeBank [Carlson et al., 2002], a corpus annotating discourse relations based on RST theory, uses an extended set with 78 relations providing a high fine-grained granularity.

- Circumstance
- Solutionhood
- Elaboration
- Background
- Enablement and Motivation
  - Enablement
  - Motivation
- Evidence and Justify
  - Evidence
  - Justify
- Relation of Cause
  - Volitional Cause
  - Non-Volitional Cause
  - Volitional Result
  - Non-Volitional Result
  - Purpose
- Antithesis and Concession
  - Antithesis
  - Concession
- Condition and Otherwise
  - Condition
  - Otherwise
- Interpretation and Evaluation
  - Interpretation
  - Evaluation
- Restatement and Summary
  - Restatement
  - Summary
- Other Relations
  - Sequence
  - Contrast

Table 1: RST relations *[Mann and Thompson, 1987]*

RST has acquired many attentions in the domain of discourse study recent years. [Marcu, 1997], one of the first work on discourse parsing that using RST as a representation framework for discourse relation, has inspired many other works in the same domain. One can make use of RST to represent a text as a discourse parse tree in which, the leaves of this tree are elementary discourse units while its nodes are rhetorical relation that are generated by relating elementary discourse units or nodes. A RST discourse parse tree is illustrated in fig. 4.

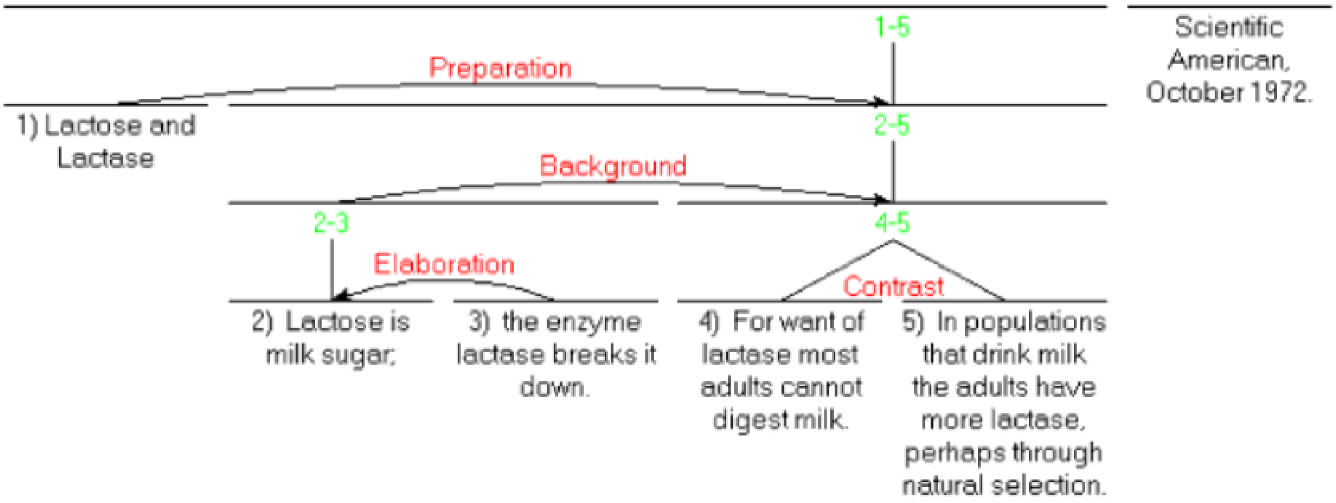

1-5
Preparation
Scientific American, October 1972.
1) Lactose and Lactase
2-5
Background
2-3
4-5
Elaboration
Contrast
2) Lactose is milk sugar;
3) the enzyme lactase breaks it down.
4) For want of lactase most adults cannot digest milk.
5) In populations that drink milk the adults have more lactase, perhaps through natural selection.

Figure 4: RST discourse parse tree
Taken from http://www.sfu.ca/rst/01intro/intro.html

2.1.2 Discourse Representation Theory

Discourse Representation Theory (hereafter DRT) is introduced by [Kamp and Reyle, 1993] in order to overcome the problem of Montague grammar [Montague, 1973]. Let use an example to illustrate the problem. Based on Montague grammar, one interprets (4a) as (4b). But this representation is not correct because the variable x in the last predicate is not bounded – x is beyond the scope of the quantifier ∀x. Therefore, (4b) does not express exactly the meaning of (4a). The correct interpretation of (4a) would be (4c), in which, variable x from all predicates refers to the same donkey beaten by Pedro.

(4) a. If Pedro owns a donkey, he beats it.

b. ∀x[donkey(x) ∧ own(Pedro,x)] → beat(Pedro,x)

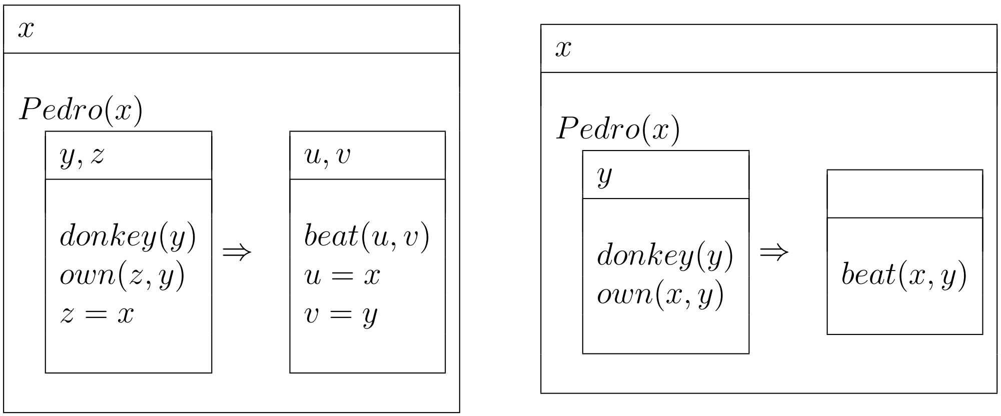
c. ∀x[[donkey(x) ∧ own(Pedro,x)] → beat(Pedro,x)]

The elementary structure of DRT is Discourse Representation Structure (hereafter DRS), which consists of a universe of referents and a set of conditions that are predicates of referents. Let consider the example (5). It has one discourse referent x and two predicates donkey(x) and bray(x). This can be represented with DRS under the form of a box with two part, the top part consists of referents and the bottom part contains predicates, as illustrated with the figure on the right of the example.

(5) A donkey brays. $\begin{array}{|l|}\hline x \\ \hline donkey(x) \\ bray(x) \\ \hline \end{array}$

A DRS can hold other DRSs inside it, they are called sub-DRSs. A sub-DRS is always a part of a condition of DRS. In previous example, we also have talked about the conditions a DRS as predicates of referents, but they are the simple ones. Here we will present a more complex type of condition that is the combination of sub-DRS(s) with one of two operators implication (⇒) or negation (¬).

Now the problem in example (4) can be solved with DRS as fig. 5. In fig. 5a, referents like z, u, and v are anaphoric. They can be resolved by referring back to their previous contexts. DRT proposes a principle called accessibility to perform this resolution. Accessibility can summarize in two words: *left* and *up*. In other words, to resolve a referent of a DRS, with *left* principle, one can look into the universe of referents of DRS standing on the left of the operator that current DRS is a part, with *up* principle, one should inspect the set of referents of the DRS that the current DRS belongs to. For instance, in fig. 5a, the referents in DRS contains condition beat(u, v) can access to referents of its sibling and its parents. Applying the accessibility principle, one can even achieve a more simple interpretation for the donkey sentence problem as in fig. 5b.



(a)

(b)

Figure 5: DRT interpretations for donkey sentence

2.1.3 Segmented Discourse Representation Theory

Segmented Discourse Representation Theory (hereafter SDRT) [Asher and Lascarides, 2005], presented by Nicholas Asher and Alex Lascarides, is an extension of DRT that takes into account the relations that hold between discourse segments. For sake of illustration, let consider example (6). We find that π_b and π_e add more information to π_a, this type of relation in SDRT is defined as Elaboration relation that holds between 2 discourse units. Between π_b and π_e, there is a temporal relation in which, an event (π_e) happens after the another event (π_b). This sort of relation is called Narration. Figure 6 illustrates the discourse structure for sentences in example (6), the top sentence express main information, the other sentences that stand below top sentence will have the responsibility to expend main information by adding more knowledge to top sentence and hence, making the main information clearer. We can find that, this structure has the graph shape as in figure 7.

(6) a. Max had a great evening last night. (π_a)

b. He had a great meal. (π_b)

c. He ate salmon. (π_c)

d. He devoured lots of cheese. (π_d)

e. He then won a dancing competition. (π_e)
(*source:[Asher and Lascarides, 2005]*)

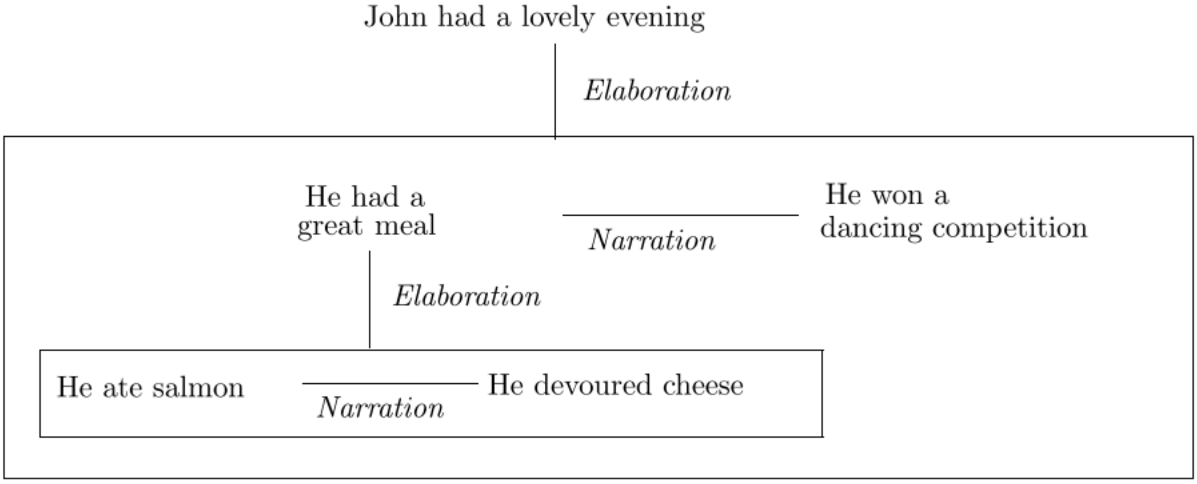


Figure 6: SDRT Representation of (6) *[Lascarides and Asher, 2007]*

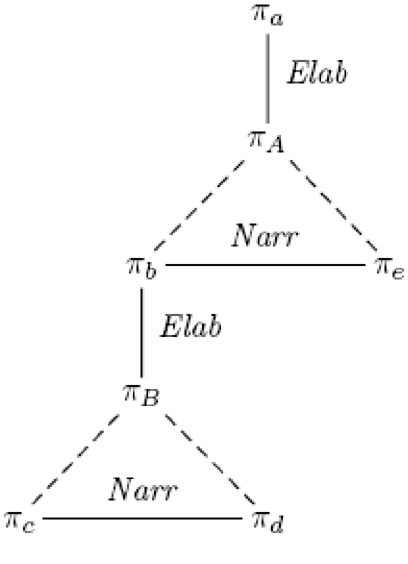


Figure 7: SDRT graph representation of (6)*[Asher et al., 2007]*

DRT is a theory that help deal with linguistic phenomena such as anaphora, but it lacks of the expression of structure between discourse relations. RST in turn can provide the discourse structure but when we want to study the effects between theses relations, we do not have a means to do it. SDRT seems to mix the idea of logical representation for discourse of DRT and structure representation of RST. It makes SDRT more powerful since it provide the ability to infer through discourse relations (as the property of logic) and give a structure of text as property of RST.

2.2 Discourse Parsing

A discourse relation links two text spans together. Discourse relation (also called rhetorical relation [Mann and Thompson, 1987] or coherent relation) always consists of a discourse connective (also called discourse marker) and discourse arguments (also called discourse units). Discourse relation has several types depending on the discourse theory that we adopt. A discourse connective can be visible, in this case, the discourse relation is called explicit relation. When discourse connective is invisible, if there exists a relation between two discourse arguments, the relation is said *implicit*. Each discourse relation conveys a meaning, that links its arguments. This meaning is called discourse types.

2.2.1 Discourse parsing tasks

The task of identifying discourse relations (explicit and/or implicit) in a text is called discourse parsing. Due to the structure of a discourse relation, a typical discourse parsing process includes the following steps:

1. **Discourse connective identification** has the objective to detect all possible discourse markers from the text. There are many words or phrases can be used as discourse marker. As information recorded by [Pitler and Nenkova, 2009], there are only 11 over 100 words or phrases annotated in PDTB have been used as a discourse connective more than 90%. In other cases, they do not express a discourse relation. Let consider the following example, the word *but* in (7a) is an adverb to express the exclusion of the love from things that John possesses, in (7b), *but* is a conjunction that link two clauses and express the contrast relation between these clauses.

(7) a. John have everything but love.

b. John loves Mary but she does not love him.

In [Pitler and Nenkova, 2009], the authors claim that the syntactic knowledge will help improve the performance of the discourse connective disambiguation task. They proposed a set of features to be used for this task including: Self-Category, Parent-Category, Left-Sibling-Category, Right-Sibling-Category, Right-Sibling-Contain-Verb, Right-Sibling-Contain-Trace.

[Lin et al., 2014] develops the idea of [Pitler and Nenkova, 2009] by considering the information about the connective's context and the part-of-speech. The authors showed that by adding the feature related to the POS tag of connective, its right and left word and the parsed path from connective to tree root, the performance of discourse connective identification task will be increased.

2. **Discourse argument identification** step aims to extract text spans for discourse arguments. According to [Prasad et al., 2008], a discourse relation has two discourse arguments arg1 and arg2. Arg2 always link to discourse connective syntactically. Arg1 can appear in the same sentence as connective, stand in precedent sentence or locate in subsequent sentence of connective. Therefore, it is always a difficult problem when we want to extract discourse arguments precisely.

   [Wellner and Pustejovsky, 2007] proposed to restatement this problem in the way that instead of finding the full text span of arguments, we only have to find the head words that stick the argument by using syntactic parse tree. However, this proposition can not help solve the problem completely, it can be useful in certain cases such as when we want to compare the matching of two arguments, we can compare 2 head words rather than comparing two whole spans of texts that is error-prone.

   [Lin et al., 2014] also use the syntactical approach but in a different point of view. Based on the analysis of [Prasad et al., 2008] on position of arguments, the authors proposed two sets of feature that will be used for classifiers training process, one for locating the relative position of arguments and other for extracting the text span of arguments.

3. **Discourse relation type classification** is the task of identifying the type of discourse relations. Based on the discourse theory we use, we have different sets of discourse relation types. The problem is there are several connectives that take different type in variant contexts.

   For the approach of [Marcu and Echihabi, 2002] and [Sporleder and Lascarides, 2008], they use a set of patterns to detect discourse relations. Each pattern corresponds to a discourse relation and belongs to a discourse relation type. Therefore, we do not have to struggle with this task.

   For the statistic approach, we have to train a discourse type classifier from data. With the emergence of some discourse corpora such as PennDTB, RST-DT, etc., and the development of the machine learning techniques, most of the recent research are on this approach.

.

2.2.2 Discourse parsing classification

Based on the method one adopts, the corpus one uses, we can have two types of discourse parsing.

- **Structural parsing**: This type of parsing apply a discourse structure theory such as RST, and then, the result of the parsing process is the discourse structure of the input text. This output help us have an overview of the coherent relations between segments in text. This will be useful for subsequent task such as text summarization of sentiment analysis since it provides us the important and less important parts through the semantic of discourse relation.

For example, once applying RST theory as [Feng, 2015] or [Joty et al., 2015], we can obtain a tree structural representation of the text. To carry out this type of parsing, first the text should be divided into segments that are not overlapped, we called them elementary discourse units (EDUs) and the process the segmentation. Second, a machine learning method will be apply on a set of features to learn a classifier to help determine whether there exist a relation between two discourse units, and what type is it.

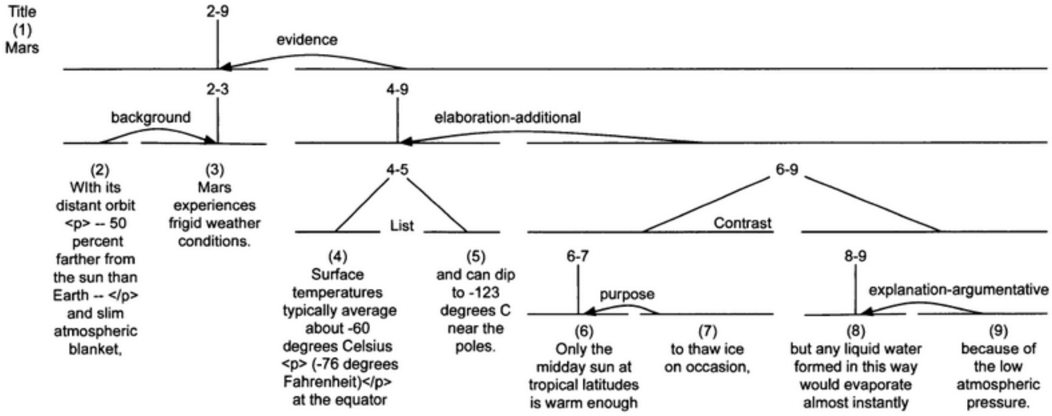

Title
(1)
Mars
2-9
evidence
2-3
background
4-9
elaboration-additional
(2)
WIth its distant orbit <p> -- 50 percent farther from the sun than Earth -- </p> and slim atmospheric blanket,
(3)
Mars experiences frigid weather conditions.
4-5
List
(4)
Surface temperatures typically average about -60 degrees Celsius <p> (-76 degrees Fahrenheit)</p> at the equator
(5)
and can dip to -123 degrees C near the poles.
6-9
Contrast
6-7
purpose
(6)
Only the midday sun at tropical latitudes is warm enough
(7)
to thaw ice on occasion,
8-9
explanation-argumentative
(8)
but any liquid water formed in this way would evaporate almost instantly
(9)
because of the low atmospheric pressure.

Figure 8: Structural representation of text

source:http://www.inf.ed.ac.uk/teaching/courses/anlp/lectures/29/

- **Nonstructural parsing** Nonstructural parsing or shallow discourse parsing is the process of analyze the text to extract the relations between discourse units. It may not give us a shaped connected structure but several disjointed structures that represent the discourse dependency between text segments. Normally, the PennDTB-based discourse parsers such as [Lin et al., 2014] belong to this type because the PennDTB annotation give us only the relation between discourse argument. So the annotation data looks like a dependency structures with a predicate and two arguments.

2.3 Discourse Relation Inference

A discourse is coherent if ideas expressed in the discourse are linked to each other through discourse relations. Usually, ideas are provided by discourse units, such as clauses, sentences, or larger textual units. Discourse relations are realized by overt linguistic markers, called cue words and cue phrases (e.g. *because*, *so*, *in addition*, *nevertheless*, *therefore*). One may employ certain kind of syntactic constructions to signal a discourse relation. Some choice of verbs or phrases can also trigger a discourse relation. In other words, the discourse relations that discourse theories study are realized by explicit linguistic means. However, in certain cases, a text may give rise to relations that are *implicit*, that is, they are not expressed by linguistic means, but can be *inferred*. For sake of illustration, let us consider a piece of text consisting of three discourse units π_1, π_2, and π_3. Assume that the linguistic information in the text indicates that π_1 and π_2 are related by a discourse relation Γ_1, whereas π_2 and π_3 are related by a discourse relation Γ_2. The problem

I am tackling is to identify whether there is some kind of relation Γ_3 between π_1 and π_3. While we call Γ_1 and Γ_2 explicit discourse relations, we refer to Γ_3 as an implicit discourse relation. Hence, given $\Gamma_1(\pi_1, \pi_2)$ and $\Gamma_2(\pi_2, \pi_3)$, the question of study is whether one can infer $\Gamma_3(\pi_1, \pi_3)$ for some implicit discourse relation Γ_3.

[Roze, 2011b] presented a theory which can be used to deduce between discourse relation. This help to calculate the discourse closure, a set that contains all possible discourse relations of a group of sentences. Consider example (8), (8a) has Result relation with (8b) and (8b) has Elaboration with (8c). To our knowledge, we can recognize that there exist a relation Result between (8a) and (8c). So we have an inference rule here: Result(π_1, π_2) $\wedge$ Elaboration(π_2, π_3) $\rightarrow$ Result (π_1, π_3)

(8) a. It has rained a lot today.
    b. So John cooked.
    c. He made a pie.

There are two forms of inference as illustrated in figure fig. 9

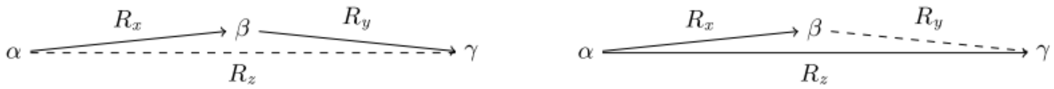


Figure 9: Forms of discourse inference rules *[Roze, 2011a]*

3 Data collection

This section presents some familiar corpora in the discourse study domain like Penn Discourse TreeBank, RST Discourse Tree Bank and Biomedical Discourse Relation Bank. We will study their properties, structures to determine which one is the most suitable corpus for our work.

3.1 Penn Discourse TreeBank

Penn Discourse TreeBank (hereafter PDTB) [Prasad et al., 2008] is one of the most popular corpora for discourse relation research. It provides the discourse-level relation annotation for Penn TreeBank on Wall Street Journal set. Its emergence has stimulated the development of the data-driven approach on discourse relations study by providing pre-annotated data on rhetorical relations. This sort of data gives researchers a means to evaluate their result on the precision of their extracted data. PDTB annotation system is based on D-LTAG [Forbes et al., 2002], the annotation data is represented as dependency graphs and the discourse relations are lexical based. A discourse relation in PDTB compromises of a discourse connective, two discourse arguments, and discourse relation types (or *senses*). Discourse connective can be coordinating conjunctions, subordinating conjunctions or discourse adverbials. Discourse relations can be explicit and be signaled by discourse connective. In the cases that the discourse connective is not explicitly present, the discourse relation is called implicit. The second discourse argument is always linked to discourse connective syntactically. The first discourse argument can appear right after the second argument, right before the discourse connective or before discourse connective but is separated from its connective by some discourse units. Discourse relation types in PDTB are organized in classes, types and subtypes as a hierarchy structure. A new type of relation added to type system will inherit the meaning of its parent. Figure 10 presents the type system of PDTB with 4 big classes: TEMPORAL, COMPARISON, CONTINGENCY, EXPANSION and its types and subtypes.

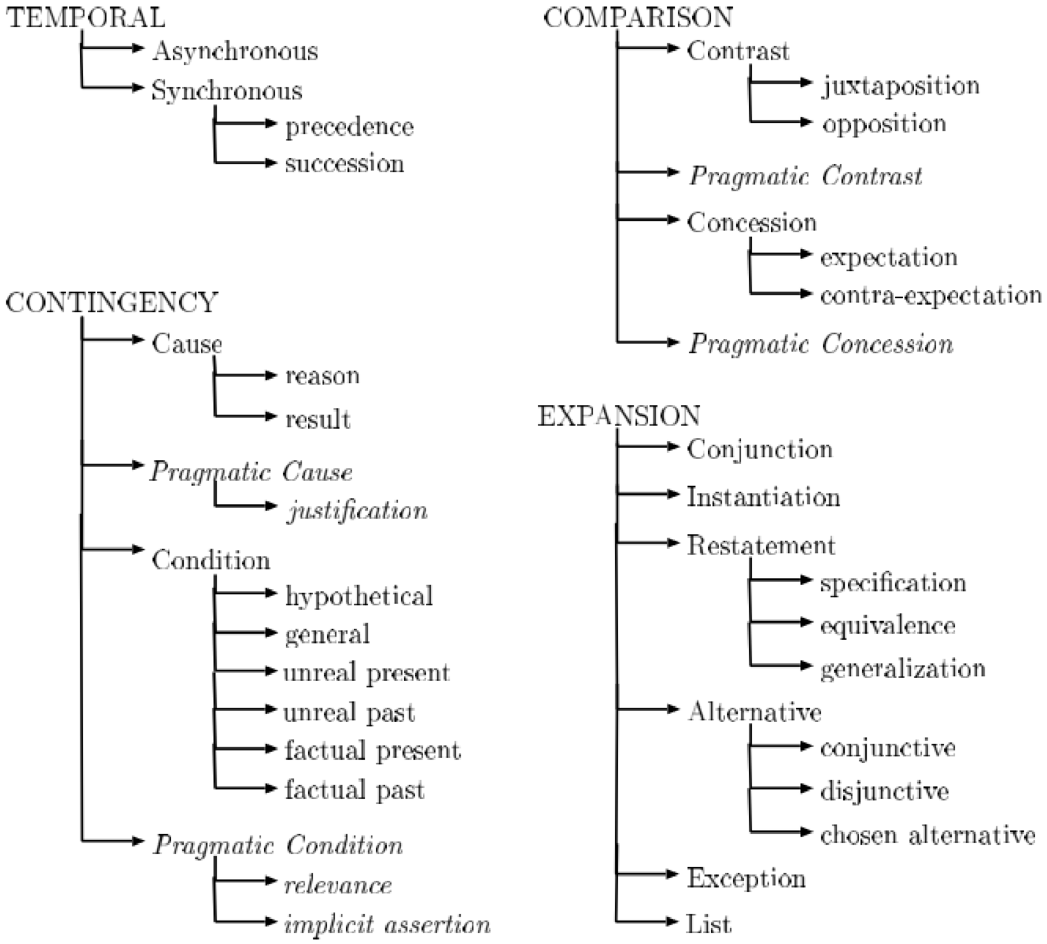
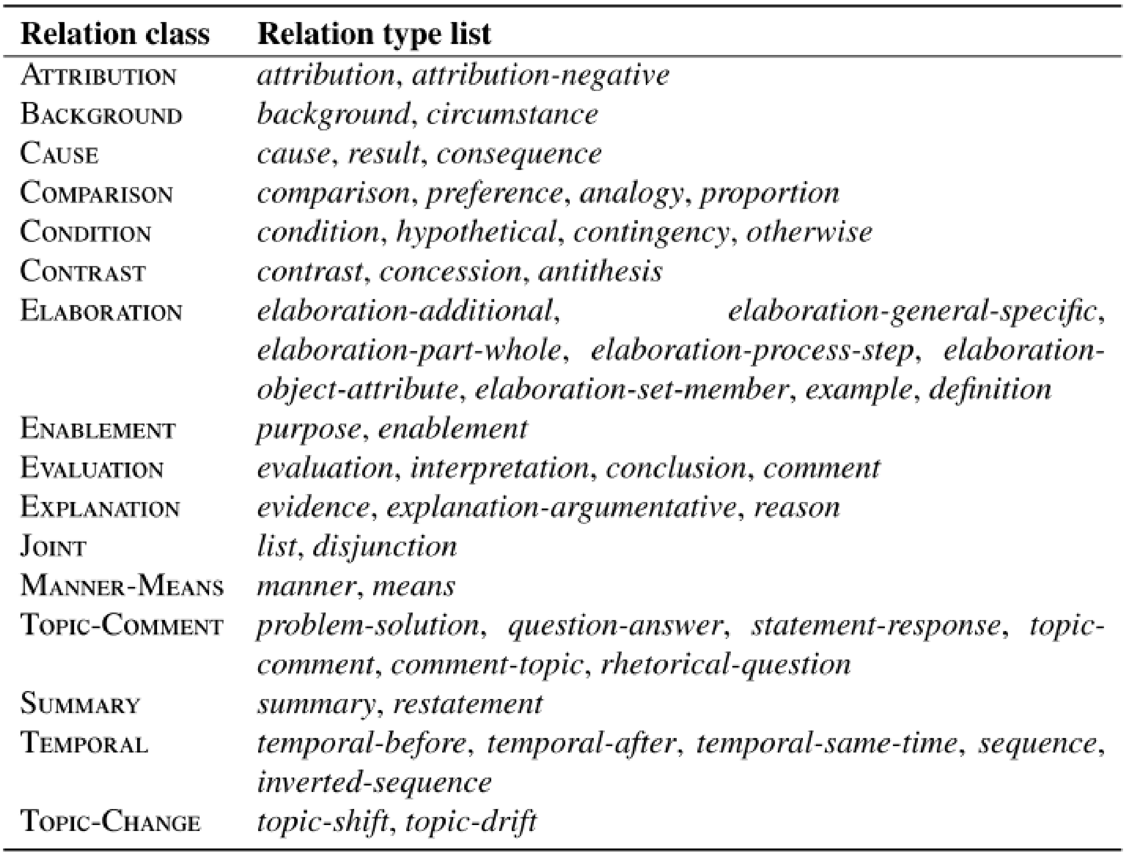


Figure 10: Penn Discourse TreeBank relations *[Prasad et al., 2008]*

3.2 RST Discourse TreeBank

RST Discourse TreeBank (hereafter RST-DT) [Carlson et al., 2002] is another well-known corpus that is widely used in discourse relation research. It consists of discourse relation annotations follow RST theory for 385 Wall Street Journal articles. Its relation set are extended from the RST framework [Mann and Thompson, 1987] with 53 mononuclear relations and 25 multi-nuclear relations. A mononuclear relation link two spans of text in which, a span, called nucleus, holds the main information of the relation while the other span, called satellite, contains the auxiliary information to support that help and divided into 16 classes (refer to table 2 for more detail)

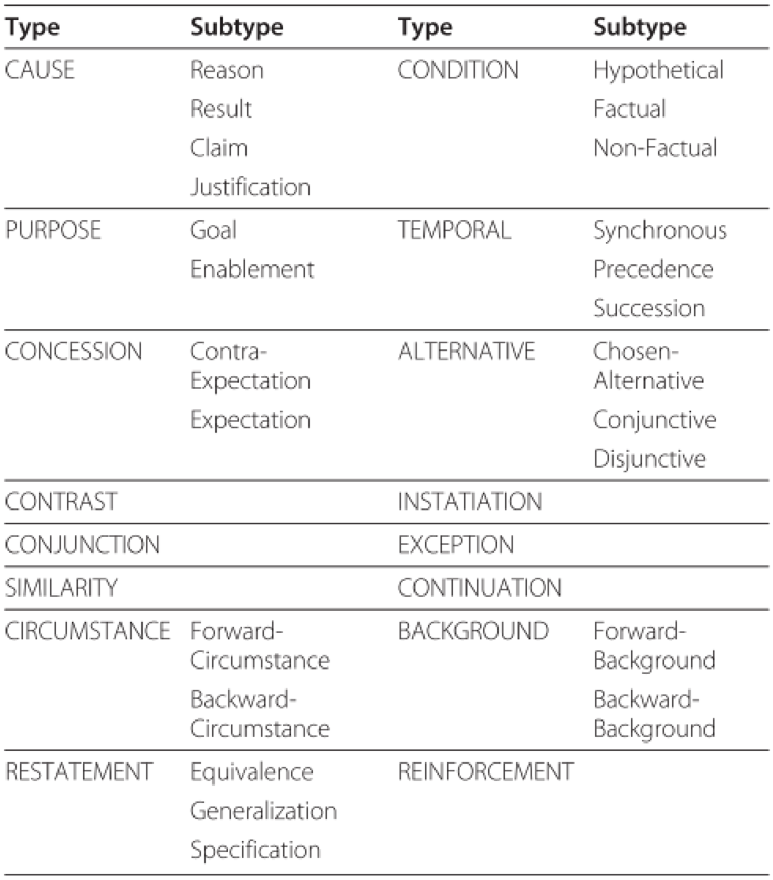


Relation class	Relation type list
Attribution	*attribution, attribution-negative*
Background	*background, circumstance*
Cause	*cause, result, consequence*
Comparison	*comparison, preference, analogy, proportion*
Condition	*condition, hypothetical, contingency, otherwise*
Contrast	*contrast, concession, antithesis*
Elaboration	*elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition*
Enablement	*purpose, enablement*
Evaluation	*evaluation, interpretation, conclusion, comment*
Explanation	*evidence, explanation-argumentative, reason*
Joint	*list, disjunction*
Manner-Means	*manner, means*
Topic-Comment	*problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question*
Summary	*summary, restatement*
Temporal	*temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence*
Topic-Change	*topic-shift, topic-drift*

Table 2: RST Discourse TreeBank relations *[Feng, 2015]*

3.3 Biomedical Discourse Relation Bank

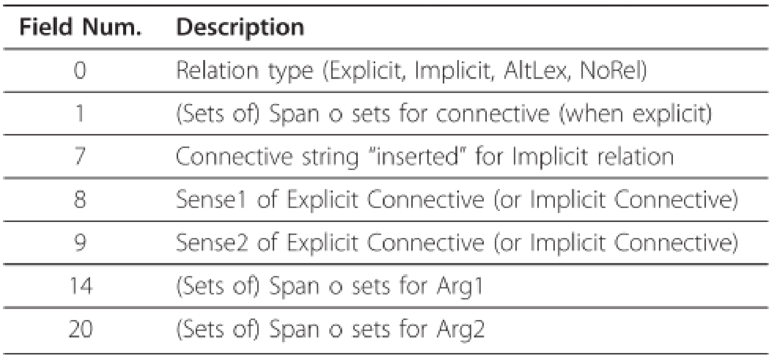
The Biomedical Discourse Relation Bank (hereafter BioDRB) [Prasad et al., 2011] is a corpus that annotated 24 full text biomedical articles. BioDRB is a project tends to help the study in the biomedical domain by supplying some domain oriented discourse relations. The corpus contains around 112000 words and approximately 5000 sentences. The relation types of BioDRB are distributed into 16 classes and 31 subclasses as in table 3



Type	Subtype	Type	Subtype
CAUSE	Reason Result Claim Justification	CONDITION	Hypothetical Factual Non-Factual
PURPOSE	Goal Enablement	TEMPORAL	Synchronous Precedence Succession
CONCESSION	Contra-Expectation Expectation	ALTERNATIVE	Chosen-Alternative Conjunctive Disjunctive
CONTRAST		INSTATIATION	
CONJUNCTION		EXCEPTION	
SIMILARITY		CONTINUATION	
CIRCUMSTANCE	Forward-Circumstance Backward-Circumstance	BACKGROUND	Forward-Background Backward-Background
RESTATEMENT	Equivalence Generalization Specification	REINFORCEMENT	

Table 3: BioDRB relations *[Prasad et al., 2011]*

It follows the annotation system of PDTB but with some modifications. It uses the same stand-off format of annotation as PDTB but with less number of fields (see table 5).



Field Num.	Description
0	Relation type (Explicit, Implicit, AltLex, NoRel)
1	(Sets of) Span offsets for connective (when explicit)
7	Connective string "inserted" for Implicit relation
8	Sense1 of Explicit Connective (or Implicit Connective)
9	Sense2 of Explicit Connective (or Implicit Connective)
14	(Sets of) Span offsets for Arg1
20	(Sets of) Span offsets for Arg2

Table 4: BioDRB annotation format *[Prasad et al., 2011]*

<table>
<tr><td>Implicit||Wr|Comm|Null|Null||and|Conjunction||||||1021..1132|Inh|Null|Null|Null||1134..1358|Inh|Null|Null|Null||</td></tr>
<tr><td>AltLex|1360..1387|Wr|Comm|Null|Null|||Cause.Claim||||||300..1358|Inh|Null|Null|Null||1360..1678|Inh|Null|Null|Null||</td></tr>
<tr><td>Explicit|1462..1468|Wr|Comm|Null|Null|||Temporal.Precedence||||||1388..1461|Inh|Null|Null|Null||1469..1499|Inh|Null|Null|Null||</td></tr>
</table>

Table 5: BioDRB annotation sample

Our study aims to the biomedical domain, therefore BioDRB is the most suitable corpus. However, we also have studied the other corpora that support discourse annotation data to see whether they can provide some useful knowledge. According to [Ramesh et al., 2012], in BioDRB, 56% of its explicit discourse connectives appear in PDTB, 33% are not annotated in PDTB, 11% do not appear in PDTB texts. Therefore, the authors claimed that there exists some domain specific discourse connectives. Because PDTB and RST-DT all use article from Wall Street Journal, we speculate that the same phenomena happens between BioDRB and RST-DT (As we do not have RST-DT corpus, we could not verify this claim). While PDTB and RST-DT propose a very fine granularity set of relations, BioDRB gives us a more coarse one. It helps reduce the complexity. Hence, we decide to use the BioDRB corpus for our work because the following reasons:

- Since we planned to work on biomedical text, this corpus fits exactly what we are interested in because it gives us the annotations for discourse relations in biomedical literatures.
- Among the corpora presented in this work, the set of relations of BioDRB is rather similar to our selected set presented in section 4.1. Therefore, this might help reduce the complexity of the comparison between two set of relation.

4 Discourse relation extraction

This section presents the main content of our works concerning the identification of discourse relations in text corpus. To carry out this task, we need to answer the following questions: what discourse relations will be employed, what discourse markers will be used for each chosen discourse relation, and finally what is the method for the extraction. These problems will be considered in detail in the following sub-sections.

4.1 Selecting discourse relations

We have considered some discourse theories, each of which proposes its own set of discourse relations. In [Marcu and Echihabi, 2002], the author studied different discourse theories like [Asher and Lascarides, 2005] or [Mann and Thompson, 1987]. They pointed out that despite discourse relations defined in each theory are varied, these relations share some common points. For demonstration, let consider example (9). According to [Mann and Thompson, 1987], two sentences (9a) and

(9b) are in a EVIDENCE relation since sentence (9b) gives the evidence to support the statement given in sentence (9a). Considering these two sentences with relations in[Asher and Lascarides, 2005], one can find that there exists EXPLANATION relation between them.

(9) a. John is a really good typewriter.
    b. He can type 200 word per minute without a typo.

Based on this statement, [Marcu and Echihabi, 2002] proposed a small set of relations that covers relations defined in various discourse theories and hence, it has a high coarse-grained granularity. This set contains four relations including CONTRAST, CAUSE-EXPLANATION-EVIDENCE, CONDITION, and ELABORATION. Each of these four relations might be the union of several relations taken from some discourse theories. For instance, the relation CAUSE-EXPLANATION-EVIDENCE is composed by the EVIDENCE, VOLITIONAL-CAUSE from [Mann and Thompson, 1987], RESULT and EXPLANATION from [Lascarides and Asher, 1993], etc.

In section 3, we decided to use the BioDRB corpus. This data collection presents annotations for a different set of relation in comparison with the one that has been chosen here. Therefore, after performing the discourse extraction method, we will obtain different relations annotations to the existing one in BioDRB. To make the results comparable, we should mapping BioDRB relations onto our relations. Let go deeper in definition of each relation of BioDRB to find the similarities that help the mapping between two set of relations.

Refer back to table 3, with the relations in that table that have same names with the ones in our set such, we simply map them together. For the other relations, look into their detail definitions, we find that, some relations have the same meaning with our relations. For example, PURPOSE and REINFORCEMENT can be mapped to CAUSE-EXPLANATION-EVIDENCE or CONCESSION might correspond to CONTRAST. For the remaining relations, we cannot match them to any of our relations. Though this does not mean that these relations will not be used, but, in contrast, we will consider their number of occurrence before making the decision of their employment.

BioDRB relation	**Our chosen relations**
Contrast, Concession	Contrast
Cause, Purpose, Reinforcement	Cause-Explanation-Evidence
Continuation	Elaboration
Condition	Condition

Table 6: BioDRB-Our chosen relations mapping

4.2 Choosing discourse markers

After determining the set of relations to be used, we now need to study how to identify these relations. In [Marcu and Echihabi, 2002], the authors proposed the pattern matching approach to help detect discourse relations. We decide to employ this approach for our discourse parser because it allows identifying discourse relations including their discourse markers and their discourse units easily. This approach bases on patterns which are built through two stage as follows: first, considering a discourse relations, one will collect as many discourse markers as possible such that they signal this relation. Then, for each marker, the generalized syntactic structures of clauses, sentence, or multi-sentences in which this marker appears will be extracted and be used to construct a pattern. So, a pattern is a mapping from an abstract structure to many concrete clauses or sentences in a text.

Let take a look into example (10) that illustrates a particular pattern. The word *If* here is a discourse marker, it signals the condition, hence, this pattern express the Condition relation. *BOS* and *EOS* stand for *Begin Of Sentence* and *End Of Sentence*, respectively. The ... represents clause in this case because this pattern has only one BOS and one EOS, i.e., it demonstrates a single sentence. To be noticed that in certain cases, when the structure of a sentence is complex with several segments separated by commas, while applying the pattern (10), the first comma will be used to separate sentence into two parts correspond to two discourse units of this conditional relation.

(10) $[BOS\ If\ \ldots,\ \ldots\ EOS]$

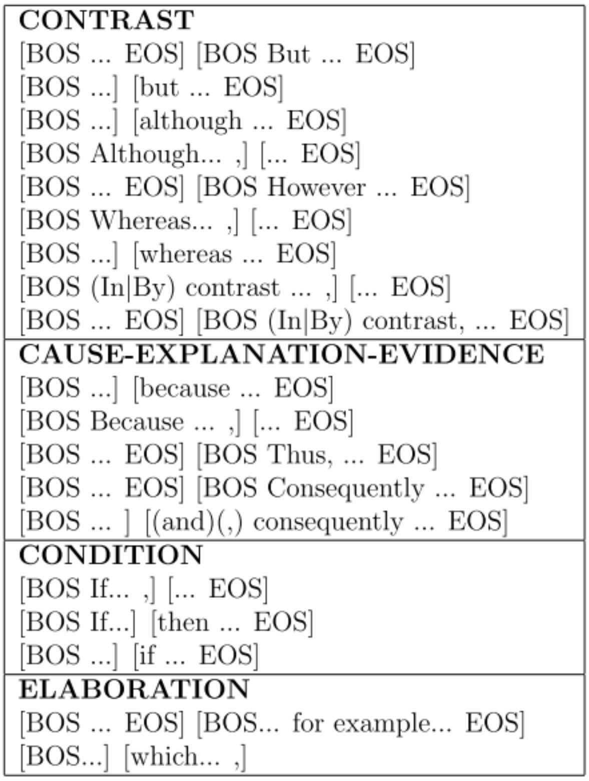
[Marcu and Echihabi, 2002] presented a list of patterns for their work. This list is rather simple with 12 patterns for 4 relations. In [de Moosdijk, 2014], the author also followed the approach of discourse relations identification by patterns with a small improvement in the set of pattern of [Marcu and Echihabi, 2002] such that it is more suitable for discourse relations in medical literatures. In our work, at the very beginning stage, we are going to reuse the pattern collection described in [de Moosdijk, 2014]. This collection contains 19 patterns. To identify CONTRAST relation, [de Moosdijk, 2014] built 9 patterns based on the following markers *but, although, however, whereas, in/by contrast.* For CAUSE-EXPLANATION-EVIDENCE relation, the author constructed 5 patterns with 3 markers *because, thus, consequently* and the similar constructions for remaining relations (refer to table 7 for more detail).

We find that there are some patterns express relations inside a sentence and the others demonstrate relations between two adjacent sentences. We call these types of relation are *intra-relation* and *inter-relation*, respectively.

4.3 Extracting discourse relations

In [de Moosdijk, 2014], the author proposed to create a corpus by collecting a set of biomedical articles from PubMed, an online service that consists of about 26 millions citations for biomedical literature from MEDLINE, a bibliographic database compromise of about 22 million references concentrating on biomedicine. Then, they make use of patterns that describe the *surface structure* of discourse relations in order to identify these relations from collected papers.

In this work, we also follow this approach. In our first experiment, we decide to reuse the pattern list in table 7 in order to gather discourse relations from BioDRB corpus. The process of detecting discourse relations from corpus comprise two stages described in fig. 12.



CONTRAST
[BOS ... EOS] [BOS But ... EOS]
[BOS ...] [but ... EOS]
[BOS ...] [although ... EOS]
[BOS Although... ,] [... EOS]
[BOS ... EOS] [BOS However ... EOS]
[BOS Whereas... ,] [... EOS]
[BOS ...] [whereas ... EOS]
[BOS (In|By) contrast ... ,] [... EOS]
[BOS ... EOS] [BOS (In|By) contrast, ... EOS]
CAUSE-EXPLANATION-EVIDENCE
[BOS ...] [because ... EOS]
[BOS Because ... ,] [... EOS]
[BOS ... EOS] [BOS Thus, ... EOS]
[BOS ... EOS] [BOS Consequently ... EOS]
[BOS ...] [(and)(,) consequently ... EOS]
CONDITION
[BOS If... ,] [... EOS]
[BOS If...] [then ... EOS]
[BOS ...] [if ... EOS]
ELABORATION
[BOS ... EOS] [BOS... for example... EOS]
[BOS...] [which... ,]

.

Table 7: Discourse relation patterns *[de Moosdijk, 2014]*

First, our parser will detect all intra-relation by browsing through each sentence of a document, and check whether it matches a pattern in our list. For each matched result, one can acquire a discourse relation. The matching test in this stage is performed as follows: look into a sentence, if it contains a discourse marker as in a pattern then it will be divided into 2 segments based on the position of the first occurrence of character comma. The text spans that stand on the left and the right of this comma will be the first and second discourse unit, respectively.

Second stage aims to discover all inter-relation. Our parser now will consider each pair of sentences taking from a text and test if there is any pattern in our list is satisfied. A pair is formed by 2 adjacent sentences. Similar to the previous stage, whenever a matched pair is found, one gathers one more discourse relation. For this stage, when a pair is matched if it contains a discourse marker that appear in any pattern of our list. We see that, in inter-relation pattern, the discourse marker always stand in the second sentence. Therefore, in this case, the parser will take the first sentence of the pair as the first discourse unit. The second discourse unit will be the remained text span after removing the discourse marker from the second sentence.

These two stages are performed one after another. It means, whenever the intra-relation parsing process is finished, the inter-relation parser will be started.

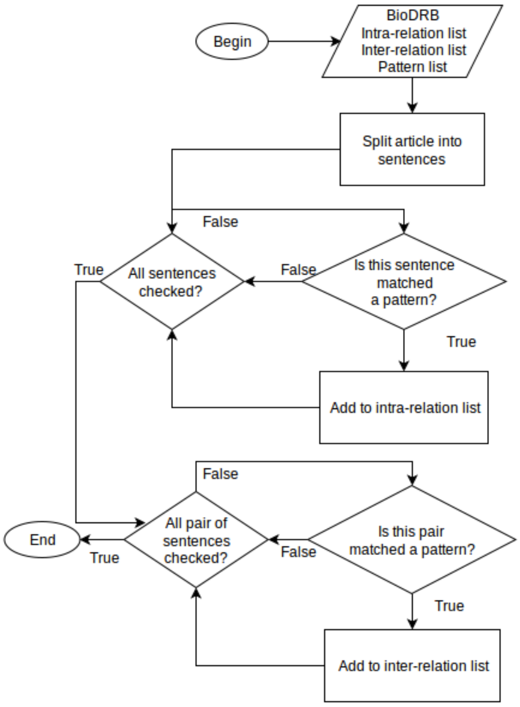

Begin
BioDRB
Intra-relation list
Inter-relation list
Pattern list
Split article into sentences
False
True
All sentences checked?
False
Is this sentence matched a pattern?
True
Add to intra-relation list
False
End
True
All pair of sentences checked?
False
Is this pair matched a pattern?
True
Add to inter-relation list

Figure 11: Discourse Relation extraction process

5 Experiments

To have a basic intuition about the result, we carry out a simple comparison as in figure 12, in which, given a detected relation maker offset and a list of pre-annotated relation marker offsets, we consider this relation is correctly extracted if its marker offset belong to the list of offsets of pre-annotated markers. This approach is not sufficient for determining the precision of a relation extraction method since it does not take into account the discourse units. But it gives us a sketch of our method.

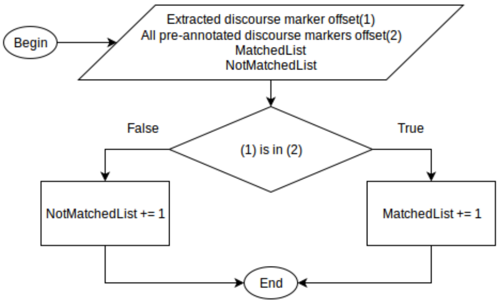


Figure 12: Simple algorithm for comparing detected relation

After having the evaluation on the relative precision of each extracted discourse relation, we tend to calculate some values such as precision and recall to help understand the performance of our method. Since our approach only finds the set of correct discourse relations but does not look for incorrect relation, we cannot calculate the precision and recall value by using typical formula that is used widely in machine learning. Therefore, we decided to use another formula that is adopted in Information Retrieval domain [Rijsbergen, 1979].

$$Precision = \frac{Number\ of\ correct\ extracted\ relations}{Number\ of\ extracted\ relations}$$

$$Recall = \frac{Number\ of\ correct\ extracted\ relations}{Number\ of\ preannotated\ relations}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Performing the parsing process with pattern in table 7 and making the comparison, we first get the following result in table 8. We find that with there is not any correct result for the extraction of Elaboration relations. This does not mean our algorithm failed in this case but there is a problem with the mapping between two set of relations, particularly is the mapping between Elaboration relation and Continuation relation. It means when we perform the evaluation, we only compare the detected relations with the ones annotated as Continuation an annotation data of BioDRB. And obviously, we can see that is no matched relation. But, beside comparing by group of relations, we also do the evaluation on the whole annotation data set and we obtain 7/209 relation matched (all of them are *for example* markers, the other not matched marker is *which*). In other words, Elaboration in our relation set is another relation that differs from Continuation of BioDRB corpus. Due to this result, for subsequent experiments, we will consider Elaboration as a new independent relation to the set of BioDRB.

Relation	# pre-annotated	# detected	# positively detected
Contrast	466	455	283
Cause	989	94	77
Condition	25	45	18
Elaboration	25	209	0(7)

# positive	395
# negative	408
# total	2636

Relation	Precision	Recall	F-1
Contrast	0.621978	0.607296	6.508834
Cause	0.819149	0.077856	28.129870
Condition	0.400000	0.720000	7.777778
Elaboration	0.000000	0.000000	-
Total	0.491905	0.149848	17.412658

Table 8: Relation extraction result with patterns in table 7

For the case of the Cause relation, there is a big difference between the pre-annotated data and our detected results. This is because in the pre-annotated set, two discourse markers *by* and *to* that are not included in our patterns, signal the purpose relations, which is a component of causal relations and these two markers have the highest number of occurrences in the total number of Cause relations (462/989).

Consider the Condition relation, we find that the parser has detected more than our expectation. There are cases in which it found the if word but it is not annotated as condition in BioDRB. Let consider a concrete example in (11), the marker *if* in this case, according to me, expresses a conditional meaning or maybe a Circumstance relation in BioDRB but in fact, it is not annotated.

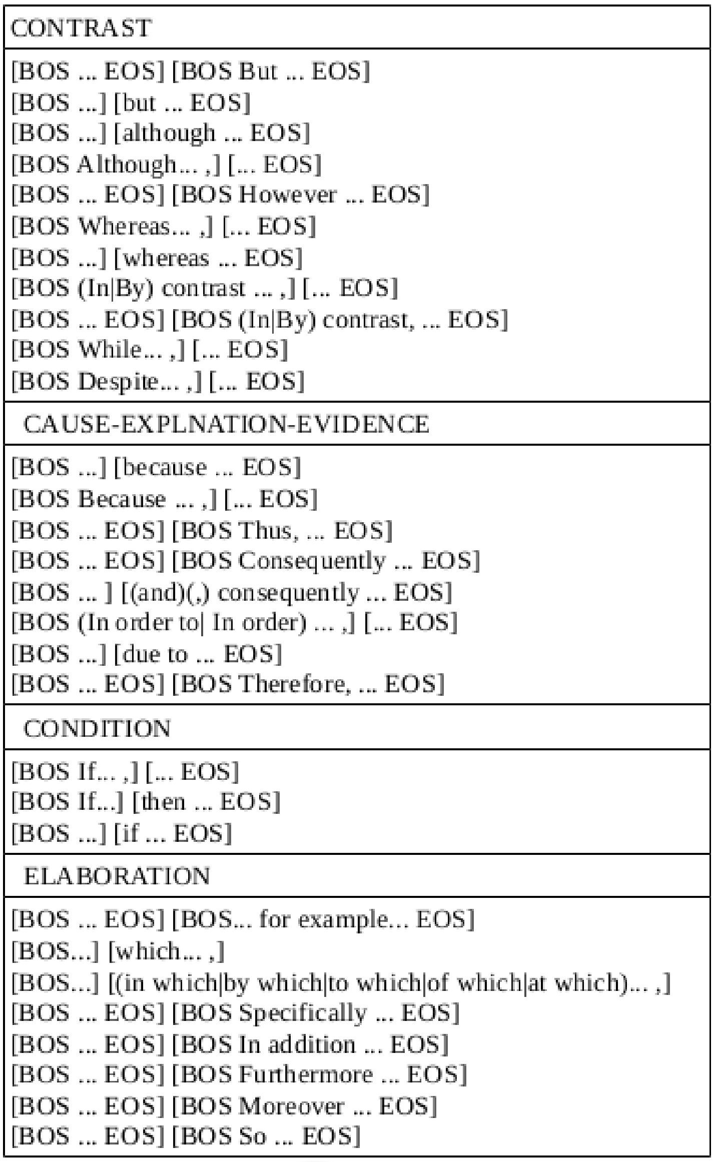
(11) However limitations may arise from deficient expression in circulating tumor cells or low level illegitimate expression in haematopoietic cells, particularly **if** a nested PCR approach is used.
[source: article 115840]

Observing the evaluation results, we recognize that there are discourse markers that are not presented in our patterns. Therefore, in order to increase the quantity of extracted relations, we analyze manually the annotation data and decide to add more patterns to describe some popular lacking discourse markers that have high frequency of occurrence such as *despite, therefore, furthermore* as demonstrated in example (12). In particular, we include *despite, while* patterns for Contrast relation set, *due to, therefore, in order to* patterns for Cause-Explanation-Evidence relation set, *specifically, in addition, furthermore, moreover, so* for Elaboration relation set. For Condition relation, we find that, the marker *if* appear most frequently in the total number of conditional relations (20/29), hence, we did not add more pattern in this case. Table 9 presents the new set of discourse relation patterns.

(12) a. In short, despite the complex picture of Treg in autoimmunity, it can be envisioned that it will become feasible to manipulate regulatory T cells for therapeutic benefit. [*Source:article 1065338*]

b. It will therefore be important to further elucidate the role of Id-1 in human B cells by selective over expression or inhibition of Id-1 gene expression. [*Source:article 1134658*]

c. Furthermore, we report that BMP-6 has an antiproliferative effect in B cells stimulated with anti-IgM alone or the combined action of anti-IgM and CD40L. [*Source:article 1134658*]



CONTRAST

[BOS ... EOS] [BOS But ... EOS]
[BOS ...] [but ... EOS]
[BOS ...] [although ... EOS]
[BOS Although... ,] [... EOS]
[BOS ... EOS] [BOS However ... EOS]
[BOS Whereas... ,] [... EOS]
[BOS ...] [whereas ... EOS]
[BOS (In|By) contrast ... ,] [... EOS]
[BOS ... EOS] [BOS (In|By) contrast, ... EOS]
[BOS While... ,] [... EOS]
[BOS Despite... ,] [... EOS]

CAUSE-EXPLNATION-EVIDENCE

[BOS ...] [because ... EOS]
[BOS Because ... ,] [... EOS]
[BOS ... EOS] [BOS Thus, ... EOS]
[BOS ... EOS] [BOS Consequently ... EOS]
[BOS ...] [(and)(,) consequently ... EOS]
[BOS (In order to| In order) ... ,] [... EOS]
[BOS ...] [due to ... EOS]
[BOS ... EOS] [BOS Therefore, ... EOS]

CONDITION

[BOS If... ,] [... EOS]
[BOS If...] [then ... EOS]
[BOS ...] [if ... EOS]

ELABORATION

[BOS ... EOS] [BOS... for example... EOS]
[BOS...] [which... ,]
[BOS...] [(in which|by which|to which|of which|at which)... ,]
[BOS ... EOS] [BOS Specifically ... EOS]
[BOS ... EOS] [BOS In addition ... EOS]
[BOS ... EOS] [BOS Furthermore ... EOS]
[BOS ... EOS] [BOS Moreover ... EOS]
[BOS ... EOS] [BOS So ... EOS]

Table 9: Our set of patterns for relation extraction

Applying this new set to our algorithm, we obtain the results in table 10. We state that some small improvements are achieved. For example, in the ensemble of contrast relations, the number of true-positive results increased significantly due to the added patterns. The same situation happened to the universe of Elaboration relation. The precision of causal relation detection process is also slightly augmented. For the case of Condition, we added no pattern, hence, there is no change.

Relation	# pre-annotated	# detected	# positively detected
Contrast	466	439	321
Cause	989	177	139
Condition	25	46	18
Elaboration	25(0)	285	0 (85)

# positive	593
# negative	376
# total	2636

Relation	Precision	Recall	F-1
Contrast	0.731207	0.688841	5.638629
Cause	0.785311	0.140546	16.776978
Condition	0.391304	0.720000	7.888889
Elaboration	0.000000	0.000000	-
Total	0.611971	0.224962	12.158516

Table 10: Relation extraction result with new patterns

Although the performance of our method achieved some small advancements, it is not enough efficient since there are many things we can do to improve it. In the subsequent section, we will consider in detail the limitation of our method and propose some ideas to deal with this.

6 Discussion

Analysing the results and, we recognize several limitations of our method. We now delve into each limits, discover their reasons, propose some idea to improve or even to solve them.

The first and most obvious problem relating to the quantity of extracted relation set since it is rather small in comparison with the annotation in BioDRB corpus. The statistic number show that, we have only obtained one over five relation as expected. The big F1 value means there is a long distance between the precision and recall value. In other words, the performance of this method is not good enough because the number of correctly extracted relations holds a small part in the total number of discourse relation annotated. There might be two ways we could do to solve this problem. The manual way, we have to study each discourse relation annotated in BioDRB, try to collect all possible discourse markers and build patterns correspond to these markers. This process requires a lot of human efforts. Another way is reapplying a discourse parser developed for another domain and adapting it to biomedical literature. We believe this task is also painful. Therefore, depending on real problem, we should evaluate and make a suitable decision.

Let look deeper into another problem relating to the constructing of a set of patterns for discourse relations. This job demands the analysis on several samples and generalize them to patterns that express the surface structure of a rhetorical relation. Apparently, it is difficult to define a good pattern since we are working on natural language, and trying to formalize unstructured relation. In this work, we base on the discourse marker, an explicit means signaling the discourse relation, to build a pattern. In a text, there may exist several positions in a sentence or group of sentences where a discourse marker can appear such as at the beginning, in the end, in the middle. For instance, in (13a), the discourse marker *nonetheless* appears at the beginning of the sentence while in (13b), it stands in the middle. Thus, in most cases, we need good linguistic knowledge to assure our patterns cover all possible structures that a discourse marker may have. Deciding to follow the pattern approach, we need to be well prepared to face to this problem because it is inevitable. We believe that one can still use this approach in the case that they want to deal with a small and explicit set of discourse relation where the knowledge of experts in linguistics will be more effective than the stochastic processing of computer. But in general, the statistic approach will give us a higher number of extracted discourse relations. One big disadvantage of the automatic method is it requires a big training data, i.e., the pre-annotated data, to work effectively. In our case, it is a problem because the BioDRB is the biggest manual annotated corpus available in biomedical domain to our knowledge, but BioDRB has a rather small size with only approximately 5000 sentences.

(13) a. So the Wilson campaign did not use it. Nonetheless, the slogan appeared all over the country, thanks to independent pro-Wilson groups. [2]

b. The officer later told reporters Carthan had been set up. A white judge nonetheless sentenced Carthan to three years in prison, and Carthan was forced to leave office in 1981. [3]

In the experiments, to verify the exactitude of a detected discourse relation, we use an algorithm that simplifies the real problem. A discourse relation consists of a discourse marker and discourse units. Our algorithm compares only the discourse marker and ignores the discourse units. Identifying discourse unit actually is a big and difficult problem. Our method solve it by using a *comma.* For a sake of illustration, let have a look on example (14). When dealing with a complex sentence that comprise several segments separated by commas, our method will take the first comma it encountered and divide the sentence into two part, before and after this comma. Applying this procedure to (14a), we will obtain (14c) but the good and accepted separation should be (14b) in which the division should happen at the second comma. To get over this problem, we are considering to employ the knowledge gather from a syntactic parse tree. It means, every time we need to separate a sentence into to discourse unit, we will build a syntactic tree. Each discourse unit to be extracted should have a structure of a clause or a sentence on this tree. In this concrete example, they should have the POS tag: SBAR and S_1 as illustrated in fig. 13.

(14) a. If I have a dog, a cat and a monkey, I will take them to the park every weekends.

b. If *I have a dog, a cat and a monkey,* **I will take them to the park every weekends.**

c. *If *I have a dog,* **a cat and a monkey, I will take them to the park every weekends.**

Another approach for dealing with this problem is that instead of detecting full text span, [Wellner and Pustejovsky, 2007] proposed to find a *head word* that is a representative of a discourse unit. Example (15) gives us an overview about head

[2] https://www.washingtonpost.com/opinions/in-2016-were-going-to-campaign-like-its-1916/2015/01/02/7c2fab58-8a08-11e4-a085-34e9b9f09a58_story.html

[3] https://www.washingtonpost.com/national/a-man-elected-before-his-time-in-mississippi-gives-politics-one-more-go/2015/08/03/67ca61bc-356b-11e5-8e66-07b4603ec92a_story.html

words which are underline, the discourse marker is inside a box, the first discourse unit is in italic, the second one is bold.

(15) *Choose 203 business executives, including, perhaps, someone from your own staff,* and **put them on the street**, to be deprived for one month of their homes, families and income. [*source: [Wellner and Pustejovsky, 2007]*]

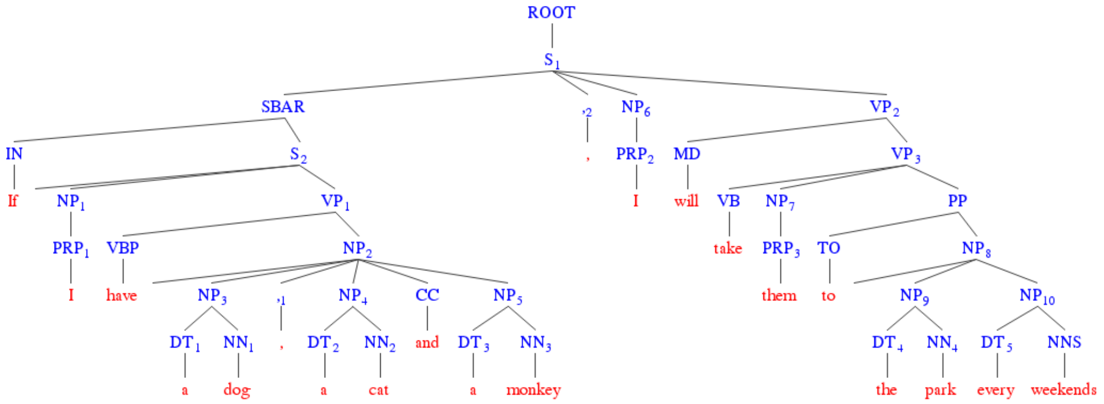


Figure 13: Syntax tree for example (15)

The pattern approach allows us to detect the explicit discourse relations based on their discourse markers. But, in fact, while the discourse markers are not present, there may be a discourse relation between two sentences or two text segments. In these cases, this approach can not help us to detect this sort of relations. In example (16), there exist a CAUSE relation between two discourse units. We can put *Because* at the beginning of the sentence to make it clearer: Because *SNCF has organized a strike*, **I have to walk to work**. So the now problem is how could we determine whether there exist a discourse relation between two text spans (here we do not use the term discourse unit because it is not assured that there exists a discourse relation between these two text spans). [Marcu and Echihabi, 2002] presents the idea about using pairs of words extract from 2 discourse unit of an explicit discourse relation as training data for implicit discourse classifier. [Wellner et al., 2006] proposed to use Multi Knowledge Source in order to produce feature classes that are in turn used to train a discourse relation classifier.

(16) SNCF has organized a strike, I have to walk to work.

Beside the suggests for our method itself, we also propose here some methods to improve the precision of evaluation step. Now we are using a simple comparison to verify the exactitude of a discourse relation by matching its offset with the pre-annotated ones. This method does not take into account the discourse units, so the result may not be exact as we expected. Therefore, we propose the following methods to compare 2 discourse units.

We can calculate the cosine distance between 2 units that are considered as a vector of words now. The result is a number between 0 and 1. The nearer 1, the more similar 2 discourse units. This method is simple since it is just a word comparison.

Another tool can be used here is Cohen's kappa coefficient, in which we will calculate the similarity between 2 discourse unit through a matching matrix. This method has the same output as the first one and can be interpreted in the same way.

For the first two methods, we do not care about the linguistic knowledge of discourse units. In this method, we propose to use syntactic tree to calculate the distance between 2 units. First we build syntactic parse tree for each unit. The distance between these two parse trees is the number of necessary steps to transform a tree to another.

By applying these methods, we have to give some threshold value to determine the good result. It depends closely on the data we process and our experiment results.

7 Conclusion

During this research, we have study mainly about the discourse parsing step of the subject Mining Text at Discourse Level. We find that there exist many ways in discourse parsing nowadays, each one has its own advantages and limitations. Our work apply the pattern approach to identifying the discourse relations inside biomedical literature.

This method is simple by mean of defining and developing end-to-end product. But it has a big limit when it does not use the deep linguistic knowledge of the discourse relation but its surface structure. It means we need to take care of many cases in which a discourse marker appears and formalize these cases into patterns. Hence, it leads to the inexactitude of the task of recognition of the discourse arguments. It also limits the ability of the relation detector when whenever one want to detect a new sort of discourse relation, a new pattern representing to this relation must be defined.

This approach has advantages such as we do not have to deal with the problem of discourse relation type disambiguation since every relations that are detected must be a type that its pattern belongs to at the moment we define our set of patterns. Since the set of patterns consists of 4 classes, the task of classifying a relation is rather simple.

As the results show us, the results is sparse. In order to mine more relations from texts, we are considering to employ the idea of [Roze, 2011b]. Considering the case where we have 3 sentences π_1, π_2, π_3. There exists a relation R_1 between π_1 and π_2, and a relation R_2 between π_2 and π_3. So a question rises as whether there is a relation between π_1 and π_3. [Asher and Lascarides, 2005] presents in their work a set of inference rules that allow us to deduce this sort of relation. [Roze, 2011b] developed this idea in her PhD thesis and gives us a more complete picture about the problem as well as her solution. For instance, in example (17), the relation Result exists between (17a) and (17b), while the relation between (17b) and (17c) is Contrast. [Roze, 2011b] deduced that there exists relation Result between (17a) and (17c)

Relation	# of occurrence
Cause	898
Contrast	466
Conjunction	421
Temporal	394

Table 11: Most popular relations in BioDRB

(17) a. Et dans le dernier acte, le BBCD souffrait encore terriblement devant une défense remarquablement agressive. Il fallait donc un final de guerriers aux Bisontins pour tenir jusqu'au bout leur avantage.

b. Kirksay, Vuleta et Mélicie enfilaient donc joyeusement le bleu de chauffe

c. alors que Mantcha Traore était prié de regagner le banc prématurément. [source:[Roze, 2011b]]

Another job we can to do our method to help find out more relations is adding more relation to our set of relations. As we recorded, among the relations annotated in BioDRB, the relations that have the highest frequency of occurrence are Cause (including Purpose and Cause), Conjunction, Temporal, Contrast (including Contrast and Concession). They take 82% (2179/2636) of the explicit relation annotated (see table 11). For the Cause relations, we should add more patterns for the markers *by* and *to* that appear much more frequently than the other markers. But in fact, it is also very hard to detect these to marker since, their structures are varied.

To conclude, we conducted the experiments on discourse parsing and found that the result of pattern-based approach is rather sparse. However, this approach can be suitable for the problem that want to delve into a small set of relations where the human effort is sufficient. As far as we know, there are few researches on discourse parsing for biomedical text. Therefore, it is worthy to pay more attention to this domain since the result will help other tasks such as text mining, text summarization, or text classification which in turn will be really useful for end user such as doctors and patients.

References

[Asher and Lascarides, 2005] Asher, N. and Lascarides, A. (2005). *Logics of Conversation.*

[Asher et al., 2007] Asher, N., Prévot, L., and Vieu, L. (2007). Setting the background in discourse. *Discours. Revue de linguistique, psycholinguistique et informatique.*, (1).

[Carlson et al., 2002] Carlson, L., Marcu, D., and Okurowski, M. E. (2002). Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current Directions in Discourse and Dialogue.*

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

[de Moosdijk, 2014] de Moosdijk, S. F. V. (2014). Mining texts at the discourse level.

[Feng, 2015] Feng, V. W. (2015). *RST-Style Discourse Parsing and Its Applications in Discourse Analysis.* PhD thesis.

[Forbes et al., 2002] Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A., Webber, B., Joshi, A., and Webber, B. (2002). D-ltag system: Discourse parsing with a lexicalized tree adjoining grammar. *Journal of Logic, Language and Information*, 12:261–279.

[Hotho et al., 2005] Hotho, A., Nrnberger, A., and Paa, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology.*

[Joty et al., 2015] Joty, S., Carenini, G., and Ng, R. T. (2015). Codra: A novel discriminative framework for rhetorical analysis. *Comput. Linguist.*, 41(3):385–435.

[Jurafsky and Martin, 2008] Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*

[Kamp and Reyle, 1993] Kamp, H. and Reyle, U. (1993). *From Discourse to Logic.* Springer.

[Lascarides and Asher, 1993] Lascarides, A. and Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy*, 16(5):437–493.

[Lascarides and Asher, 2007] Lascarides, A. and Asher, N. (2007). *Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure*, pages 87–124. Springer Netherlands.

[Lin et al., 2014] Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184.

[Mann et al., 1989] Mann, W. C., Matthiessen, C. M., and Thompson, S. A. (1989). Rhetorical structure theory and text analysis. Technical report, DTIC Document.

[Mann and Thompson, 1987] Mann, W. C. and Thompson, S. A. (1987). *Rhetorical Structure Theory: A Framework for the Analysis of Texts*, pages 79–105. International Pragmatics Association.

[Marcu, 1997] Marcu, D. (1997). The rhetorical parsing of natural language texts. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 96–103.

[Marcu and Echihabi, 2002] Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375. Association for Computational Linguistics.

[Montague, 1973] Montague, R. (1973). The proper treatment of quantification in ordinary English. In Thomason, R., editor, *Formal Philosophy: Selected Papers of Richard Montague*, pages 247–270. Yale University Press, New Haven, CT.

[Pitler and Nenkova, 2009] Pitler, E. and Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16. Association for Computational Linguistics.

[Prasad et al., 2008] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, pages 2961–2968.

[Prasad et al., 2011] Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The biomedical discourse relation bank. *BMC bioinformatics*, 12:188.

[Ramesh et al., 2012] Ramesh, B. P., Prasad, R., Miller, T., Harrington, B., and Yu, H. (2012). Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association : JAMIA*, (5):800–8.

[Rijsbergen, 1979] Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, 2nd edition.

[Roze, 2011a] Roze, C. (2011a). Towards a discourse relation algebra for comparing discourse structures. In *CID 2011 - Constraints In Discourse*, pages 1–7.

[Roze, 2011b] Roze, C. (2011b). Vers une algèbre des relations de discours pour la comparaison de structures discursives. In *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*.

[Sporleder and Lascarides, 2008] Sporleder, C. and Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: An assessment. *Nat. Lang. Eng.*, 14(3):369–416.

[Wellner and Pustejovsky, 2007] Wellner, B. and Pustejovsky, J. (2007). Automatically Identifying the Arguments of Discourse Connectives. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 92–101.

[Wellner et al., 2006] Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., and Saurí, R. (2006). Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 117–125. Association for Computational Linguistics.

List of Tables

List of Figures