



COHERENCE AND DISCOURSE STRUCTURE ANALYSIS FOR THE AMBIGUOUS CORPUS

¹Dr Shahebaz Ahmed Khan, ²Dr Mohammed Abdul Qadeer, ³Dr Abdul Ahad Afroz

¹Assistant Professor, ²Assistant Professor, ³Assistant Professor

^{1,2,3}Department of Computer Science Engineering

^{1,2,3}Avanathi Institute of Engineering and Technology, Hyderabad, India

Abstract : One of the major problems we come across in NLP is processing of discourse structure. To train the NLP models, we make use of discourse analysis. The discourse is full of exceptions, dualities and ambiguities which make the computers to understand remember and learn. Due to the corpus ambiguity it sometimes becomes very difficult to the computers to find the relationship among the various words and predicts the actual results. The coherent group of sentences in the corpus must make a sense and should produce finite quality meanings for the users. The imprecise expressions can be dealt with the concepts of coherence and discourse analysis. To reflect the original and actual meaning even though the proper context is not provided, the discourse analysis helps to train the NLP models to predict the actual meaning that can justify the use of Artificial Intelligence applications. The idea of coherence resolution finds all the 'mentions' to process natural languages in computer systems effectively with accurate contextual meaning predictions.

IndexTerms – Coherence, discourse structure, corpus, imprecise expression, resolution, context etc.

I. INTRODUCTION

The field of Artificial Intelligence faces challenges in using the concepts of Natural Language Processing. NLP refers to idea of communication between humans and machines which process the data. Discourse processing is a major problem in NLP, it has a role to be played in building the models and theories which talk about how the utterances stick together in a complete form to give coherent discourse. Any language we consider for processing in the world consists of some structures, collocations and coherences but it is not found in any isolated and unrelated text. The field of corpus linguistics is a point of deep research in Artificial Intelligence area for the last two decades. It has been flourishing mainly due to growing interest in linguistics in computers to train the robotic and artificial intelligence systems [1]. It can be both an easy and difficult task to define a corpus because we need to consider the purpose of creation, size of the text, type of the text and the way in which it has been understood and analyzed. The questions constitute the information and it is presented in questions using linear arrangements of syntactical elements in initial or final or any other positions.

Coherence in Natural Language Processing has the ability and characteristics to evaluate the quality of the results which are generated by the natural language systems. The corpus given should be always coherent and it should have some discourse. If there is no discourse, then the sentences do not exhibit the property of coherence. Coherence is an important step to perform and implement high level natural language processing tasks which include information retrieval, document summarization, document search, questioning and answering etc. In the vector representations of corpus and sentence embedding, there can be pronouns, adverbs in the sentences. When such pronouns or embeddings are not found with sufficient context the, these sentences do not reflect the original meaning and sentence understanding. In order to make the sentence embedding rich, we can apply the theme of coherence resolution to the text. Without the involvement of cohesion and coherence, the reader may find some gaps in the text and may feel that there is some choppiness. Any text without coherence and cohesion becomes difficult for the readers to understand [2]. There can be missing of paragraph unity and we can ensure the unity in paragraphs there comes a need to develop the coherence and cohesion of the corpus.

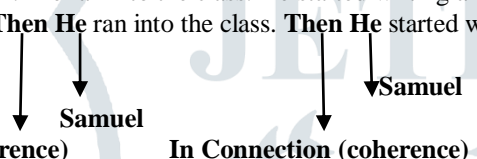
In the process of language translation, one major task is to identify the linguistic obstacles and triggers present in text source and sentences. The contextual and conceptual relations should have to be transferred into mental textual target. This can be done by using cohesion and discourse structure to build finite language conversions and transformations. One of the commonly attributed research point in corpus linguistics is discourse analysis and detection of discourse connectives. The different languages have various aspects of discourse connectivity and structures. Discourse analysis investigates the language functions and purpose, decodes how the meaning is constructed in the various contexts and discourse backgrounds. To examine the dynamics of communication, structures of conversation, social interactions the role of discourse analysis is effective in NLP. It determines the themes of how meaning is created through a language, how it is structured and how a particular context affects the content of the sentences and text.

II. COHERENCE

The word coherence is used to refer the relationship between sentences of a text which produce real discourses simple arrangements of words. Coherence is the making if the sense for the utterances in terms of discourse. Coherence is to make meaningful correlations and connections between the various tokens of a sentence. The clarity of expression is determined by coherence and this coherence is created when the scope of proper grammar and correct vocabulary is used in a context. The structured relationships found between text units which indicate some reason are called coherence relations. The coherence discourses are given a structure using such forms of coherence relations. The coherence relations are investigated from different points of view like computational linguistics, theoretical linguistics and psycholinguistics [3]. The theoretical linguistics deal with those factors which contribute to the discourse coherence and these categorize the various relations of coherence by connecting the clauses, phrases and sentences. The computational linguistics tends to decode the language corpus in terms of context so that the reader can evaluate the correct form of text to take basic internal meaning. This is also found with psycholinguistics, where it deals with the mental state receiving of actual meaning from the discourse by forming logical connections. Coherence is preceded by cohesion. The difference between cohesion and coherence is that, cohesion is achieved when sentences are connected and form relation at textual level, whereas as coherence is achieved when ideas are connected and forms the logical, consistent and understandable theme of a context.

If we read a part of the text or paragraph from a language source like news papers, articles etc, all that part or paragraph is interrelated and has interconnectivity from one point to another. Here, in this context we can say that the discourse of the text is coherence and at the same time if we take the headings of the articles or newspapers then we miss the discourse and might not be able to determine the relations among these headings. This scenario is considered as non coherence and it is not to be treated as discourse [4]. When there are structured group of sentences with coherence, then we refer it as discourse. The discourse coherence characterizes the relationships and connections among the textual units in the given structure. The example of coherence can be understood as given below.

Eg: Samuel took his book. He ran into the class. He started writing an essay.
 Samuel took his book. **Then He** ran into the class. **Then He** started writing an essay.



In Connection (coherence) In Connection (coherence)

2.1 Local and Global Coherence

The discourses exhibit both local coherence and global coherence. In local coherence, the clauses or sentences are related to the nearby clauses or sentences in semantic manner. For example, let us take the clauses given below.

Eg: Michael took a book from the library. He likes mango.

In the above sentence, the given sequence is incoherent. The reason is that, it is unclear to the reader why the second sentence has appeared and what it has to do with first sentence in its following. This means, what connection Michael's going to library has with mangoes? In this context, the reader may take efforts and tries to find out how the discourse can be coherent in this context. By contrast to this, let us take another example which is as follows.

Eg: Tom booked a ticket to London. He needs to attend a conference.

In the second example, there we can find the structural relationship and reason between the two text units. These coherence relationships are considered as locally coherence relations.

In addition to the local coherence which is found between adjacent and nearby sentences, the discourses are also found in the form of global coherence. Many classes and categories of text are associated with particular conventional discourse structures. Let us consider the academic or story articles. These have different sections which describe the methodology or results. Stories may follow some traditional plotline. In global coherence, the overall structure of a discourse is coherent in terms of dependency on the classes of the discourse. Here the examples can be the comparison and analysis of the structure of stories, research articles, scientific papers, critical arguments etc.

2.2 Coherence Properties

There are two properties of coherence namely Coherence relation between utterances and Coherence relation between entities [5]. The coherent relation between the utterances tells that there will be always some kind of interconnection between the utterances of a text or sentence. It has some kind of explanation which justifies the connections between the textual units. Whenever there exists some relationship among the entities or elements of the corpus, then the discourse is said to be coherent for entities. This coherent between the entities is called entity based coherence. The entity based coherence acts as a central point to centering theory which says that "the way we refer to entities will have an impact and will influence the way how a discourse coherent should be.

III. DISCOURSE ANALYSIS AND STRUCTURE

In NLP context, discourse is a sequence of clauses or sentences which occur one after the other in speech or writing. In any speech or linguistic execution, there will be some entities which are involved and discussed with possible references in that context. These references are sometimes called 'mentions'. Discourse analysis is used to extract and determine the meaning out of the corpus or the given text. Using discourse analysis, the MLP models are trained to produce efficient results. In simple words, the analysis done to find the meaning of a context in NLP is treated as discourse analysis. It is a method of analyzing a language used in various contexts to find how people use words to create meaning [6]. The discourse analysis examines the ways to understand

how a language influences decision making, how it creates and maintains relationships, how it reflects the power dynamics and how a language can be used to construct actual theme of context or reality. Discourse analysis explores the cultural views and public opinion for a given context.

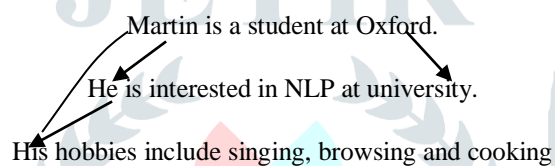
Discourse has a structure and the structure it should have is dependent on discourse segmentation. Discourse segmentations will determine the structure types for possible the discourses. The discourse segmentation has a lot to do in information retrieval, text summarization, language processing etc. The discourse segmentation is done using some algorithmic tools. There can be supervised and unsupervised discourse segmentations.

3.1 Algorithmic models for Discourse Segmentation

Unsupervised segmentation means classification of the textual units with the help of lexical cohesion or else coherent discourse. It is indirectly grouping of similar textual units using coherent discourse. We use lexical cohesion for unsupervised discourse segmentation to indicate the connections, identities and relationships among similar textual units like synonyms. In this, an algorithm will classify the similar textual units and will group these together with the help of some linguistic devices. For example, when we have a text of a story, then we can segment this text into several units of multiple paragraphs and in further, these paragraphs, a single unit will represent a passage of the text. The unsupervised discourse segmentation is also called linear segmentation [7].

The supervised segmentation of discourse means, classification of the text based on some labels and boundaries. In this, there is training data text which is found with certain label boundaries. Then according to these labels, the textual units are grouped together. We use some cue words called discourse makers to separate the discourse segments. These cue words will identify the discourse structure category based on similarity signals. The cue words are domain specific and so it becomes easy to separate the discourses as per various segments [8]. An example of discourse is as follows.

Eg: Martin is a student at Oxford. He is interested in NLP at university. His hobbies include singing, browsing and cooking.



Here, "Martin", "NLP" and "Oxford" are possible entities.

"He" and "His" are references to the entity "Martin" and "university" is a reference to the entity "Oxford".

3.2 Discourse Coherence

The idea of discourse coherence is to examine and find the coherence relationship among the textual units of the context. We use lexical repetition in order to find the discourse structure. But, if we use the lexical repetition, we cannot meet the conditions of discourse coherence like substitutions, ellipses, lexis etc. So, in such cases, Hebb has given a better solution to find the relationship among the text.

Let us consider two sentences as S0 and S1

Let S0 = Sam got the first rank in the class.

S1 = Sam will be given prize.

From the above two sentences, we can say that S1 is a cause of the sentence S0.

Similarly, S0 can also be the cause of sentence S1. For example,

S0 = Sam did not write the examination.

S1 = He was ill.

The text can be in parallel contexts. For example, S0 as John went to college and S1 as Jane went to market. Both these scenarios and contexts are parallel with context of 'went'. The assertions tell that both are parallel. Here parallel means that, the assertion from S0 i.e, $P(a_1, a_2, a_3, \dots)$ and the another assertion from statement S1 i.e, $P(b_1, b_2, b_3, \dots)$ that 'ai' and 'bi' will be similar for all the values of I where $i = 1, 2, \dots, n$

There is also elaboration in this contextual analysis. This can be a kind of proposition where P is inferred from both the given statements or assertions S0 and S1.

For example S0 and S1 be John is at college and Jane is in house respectively. So here, both these assertions are inferred from some proposition P.

The occasion also takes place here. Occasion is the change in the state inferred from the first assertion. The last or final state is transferred from the statement S1 and from S0 to S1. The example of this kind of relationship can be John took the bread. He gave it to Jane.

3.3 Hierarchical Discourse Structure

We build the hierarchical discourse structure for a large discourse context with the help of statement classes. This is done to form context segments. The hierarchical structure is created to get the complete discourse among the coherence relations. This can be constructed as follows in figure 3.1.

Let us consider a few sentences or textual units

S1 = Martin went to the college to pay the fee

S2 = He then went to Jane's office.

S3 = He wanted some files.

S4 = He had some work with the files.

S5 = He also wanted a print out from the office of Jean.

The above discourse can be framed into hierarchal structure as given below.

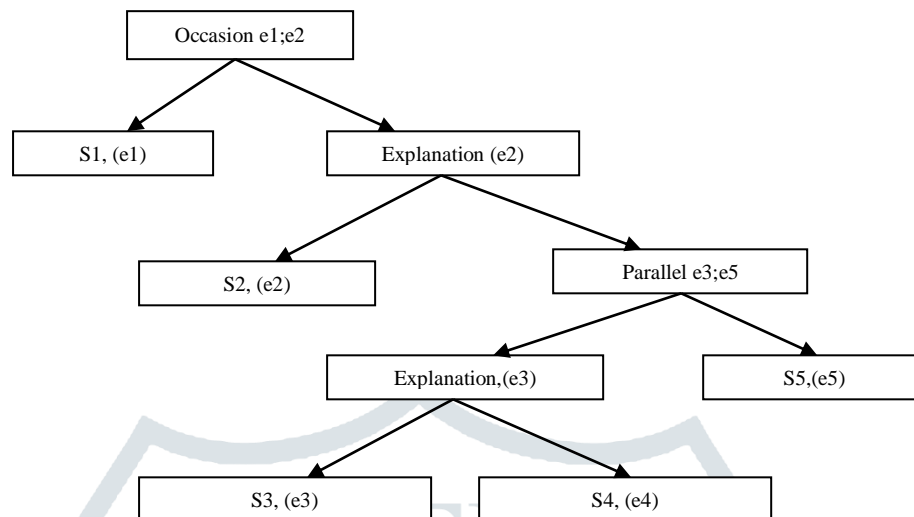


Figure: 3.1

IV. REFERENCE RESOLUTION

Reference is a linguistic expression which is used to indicate an entity or individual in a context [9] [10]. It is a form of pointed device to an entity of a textual part. Interpretation is an important task from a discourse to know what entity is talked about and referred in that particular context. The discourse is examined to extract the actual meaning of the context for decision making. Reference, in NLP, is a linguistic process where one word in a sentence or discourse may refer to another word or entity. The task of resolving such references is known as Reference Resolution. Reference resolution is a task which determines and estimates the entities that are referred to by the linguistic expressions in the given context. The reference resolution makes us to understand the type of the entity that is in the discussion or textual reference.

For example, let us consider the below sentences.

S1 = Tom went to the office.

S2 = He took his job.

S3 = His office is located in New York.

S4 = It is really a big one in that city.

In the above sentences, Tom, He, and His are the references made to entity of the context Tom. Similarly, 'it' is used as a reference for office and 'that' is used for New York. So, we can simply say that, the reference resolution as the task of determination of the entities that are being referred to by the linguistic expressions.

There are various referential devices used in linguistic discourse analysis like referent, referring expression, co-refer, discourse model, antecedent, anaphora and antecedent, which all will form a designed and elaborated meaning for the context. Reference creates cohesion by using possessive pronouns [11] and Substitution is the use of a different word in place of a previously mentioned word. Here, we make use of co-reference resolution and entity extraction to compute the entity grids so as to cluster these into the discourse mentions and to parse these sentences to get the syntactic role and semantic base.

V. CONCLUSION

The paper discusses the NLP technique which is used to compute and resolve the issues of ambiguity and confusion in discourse structure. The area of artificial intelligence primarily uses the NLP mechanisms to process the information signals and inducement of task. The concept of discourse analysis and coherence plays an important role in decoding the textual units so that a given context is understood with accurate outcomes. Indirect role of coherence and discourse analysis manages the decision making when used in the artificial intelligence field. Machine translation, information summarization and extraction achieves more efficiency and easiness when reference resolution and discourse analysis is applied. By doing so, sentences become self contained and no additional context is needed for the computer to understand their meaning.

REFERENCES

- [1] Speech and Language Processing, 3rd Edition by Dan Jurafsky and James H. Martin.
- [2] Rhea Sukthanker, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu (July 2020) Anaphora and coreference resolution: A review.
- [3] Auger, Alain & Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. Terminology 14(1). 1–19.
- [4] Tanskanen, Sanna-Kaisa. 2006. Collaborating Towards Coherence: Lexical Cohesion in English Discourse. Amsterdam/Philadelphia: John Benjamins Publishing Company.

- [5] Lapata, Mirella. 2005. Automatic evaluation of text coherence: Models and representations. In Proceedings of IJCAI, 1085–1090.
- [6] Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In Proceedings of ACM SIGIR, pages 137–144.
- [7] Barzilay, R. and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- [8] Grosz, B. J. 1977. The representation and use of focus in a system for understanding dialogs. IJCAI-77. Morgan Kaufmann.
- [9] Coreference Resolution and Discourse Coherence by Dr Mithun Balakrishnan.
- [10] Webber, B. L., M. Egg, and V. Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.
- [11] Sidner, C. L. 1979. Towards a computational theory of definite anaphora comprehension in English discourse. Technical Report 537, MIT Artificial Intelligence Laboratory, Cambridge, MA.

