COREFERENCE RESOLUTION AND ENTITY'S ACTION IN LEGISLATIVE CORPUS

MR. SURAWAT POTHONG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR

THE DEGREE OF MASTER OF ENGINEERING

(COMPUTER ENGINEERING)

FACULTY OF ENGINEERING

KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI

2021

Coreference Resolution and Entity's Action in Legislative Corpus

Mr. Surawat Pothong B.Sc. (Computer Engineering)

A Thesis Submitted in Partial Fulfillment
of the Requirements for
The Degree of Master of Engineering
(Computer Engineering)
Faculty of Engineering
King Mongkut's University of Technology Thonburi
2021

Thesis Committee

....Chairman of Thesis Committee
(Assoc. Prof. Charnchai Pluempittiwiriyawej, Ph.D)

....Member and Thesis Advisor
(Asst. Prof. Nuttanart Facundes, Ph.D.)

....Member
(Asst. Prof. Santitham Prom-on, Ph.D.)

....Member
(Asst. Prof. Jumpol Polvichai, Ph.D.)

....Member
(Asst. Prof. Varin Chouvatut, Ph.D)

Thesis Title	Coreference Resolution and Entity's Action in Legislative Corpus
Thesis Credits	12
Candidate	Mr. Surawat Pothong
Thesis Advisor	Asst. Prof. Dr. Nuttanart Facundes
Program	Master of Engineering
Field of Study	Computer Engineering
Department	Computer Engineering
Faculty	Engineering
Academic Year	2021

Abstract

Understanding and interpretation of legislative corpus has problems related to unclear entity actions and coreference. Coreference is a linguistic expression in the corpus that refers to the same object, and entity's activities represent semantic meaning on what has been done to whom. This research addresses coreference resolution and meaning representation in legislative corpora using an integrated algorithm including Span Bidirectional Encoder Representation Transformer (SpanBERT) for coreference resolution and Abstract Meaning Representation (AMR) for meaning representation. Five-step frameworks are conducted: legal text preprocessing, coreference resolution, AMR, evaluation for meaning preservation, and complexity reduction. Smatch evaluation tool and Bilingual Evaluation Understudy (BLEU) scores were applied to evaluate overlapped meaning between resolved and unresolved coreference sentences. The AMR performance was evaluated by Smatch score with 12 experiments conducted on AMR nodes. Convention on the Right of the Child (CRC), Convention on the Rights of the Person with Disabilities (CRPD), and Convention and Protocol Relating to the Status of the Refugees (CPRSR) were the datasets. This experiment shows the SpanBERT algorithm's generalization for coreference resolution in the legal area with 72.08%, 78.03%, and 69.19% in CRC, CRPD, and CPRSR respectively. The ambiguity reduction shows the AMR semantic conversion, meaning preservation, and node fluctuation. The AMR nodes are fluctuated according to the complexity of coreference. When applying SpanBERT with AMR parsing, the legal text is simplified, and AMR is a potential tool to

perform legal text meaning representation. The AMR graphs after complexity being reduced can be applied for further legal text processing tasks with Neural Network such as legal inferencing.

หัวข้อวิทยานิพนธ์	การหานิพจน์ของคำนามและการกระทำในกฎหมาย
หน่วยกิต	12
ผู้เขียน	นาย สุรวัชร์ โพธิ์ทอง
อาจารย์ที่ปรึกษา	ผศ.ดร.ณัฐนาถ ฟาคุนเดซ
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2564

บทคัดย่อ

การเข้าใจและการตีความของกฎหมายมีความกำกวมและปัญหาเกี่ยวข้องกับการกระทำของคำนามและการหานิพจน์ของคำนามในตัวบท นิพจน์ของคำนามหมายเป็นความหมายเชิงภาษาศาสตร์ที่มีการอ้างอิงถึงคำนามที่มีความหมายเหมือนกันและการกระทำของนิพจน์เชิงความหมาย งานวิจัยนี้พูดถึงนิพจน์และความของการกระทำในกฎหมายโดยใช้อัลกอริทัมแบบบูรณาการของโครงข่ายประสาทเทียม SpanBERT เพื่อการหานิพจน์ และ วากยสัมพันธ์ของภาษาอังกฤษเพื่อหาการแสดงความหมายของภาษาได้มีการใช้ 5 กรอบวิธีในการทำวิจัยได้แก่ การเตรียมข้อมูลของภาษากฎหมาย หานิพจน์ วากยสัมพันธ์ การประเมินการคงความหมาย และการลดรูปและความกำกวม การประเมิน Smatch และการประเมินคุณภาพของข้อความโดยการทดลอง 12 ประเภทของโหนดวากยสัมพันธ์ ชุดข้อมูลประกอบไปด้วยอนุสัญญาว่าด้วยสิทธิของเด็ก อนุสัญญาว่าด้วยสิทธิคนพิการ และ อนุสัญญาปีค.ศ. 1951 ว่าด้วยสถานะผู้ลี้ภัย การทดลองได้แสดงให้เห็นว่าการใช้โครงข่ายประสาทเทียมได้ผล 72.08% 78.03% และ 69.19% ตามลำดับของชุดข้อมูล การลดความกำกวมแสดงถึงการเปลี่ยนแปลงวากยสัมพันธ์เชิงความหมาย ในด้านของโหนด การเปลี่ยนแปลงของจำนวนโหนดขึ้นอยู่กับความยากและกำกวมของกฎหมายนั้นๆ วากยสัมพันธ์ของภาษาที่ลดความกำกวมโดยการหานิพจน์สามารถถูกนำไปใช้ต่อได้ในการประยุกต์ต่างๆ เพื่อวิเคราะห์ข้อความทางกฎหมายในโครงข่ายประสาทเทียมได้แก่ การหาอนุมานของกฎหมายเป็นต้น

คำสำคัญ :   การประมวลผลภาษากฎหมาย/ การประมวลผลภาษาธรรมชาติ/ การหานิพจน์/ โครงข่าย
            ประสาทเทียมเบิร์ท/ รูปแบบทางความหมาย/ วายกสัมพันธ์ของภาษาอังกฤษ/
            สแปนเบิร์ท

ACKNOWLEDGEMENTS

CONTENTS

CONTENTS (Cont'd)

LIST OF TABLES

LIST OF FIGURES

LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS

AMR	=	Abstract Meaning Representation
BERT	=	Bidirectional Encoder Representation from Transformer
BLEU	=	Bilingual Evaluation Understudy
CNN	=	Convolutional Neural Network
Conj	=	Conjunction
CPRSR	=	Convention and Protocol Relating to the Status of Refugees
CRC	=	Convention on the Right of the Child
CRPD	=	Convention on the Rights of Persons with Disabilities
Dobj	=	Direct Object
LSTM	=	Long Short-Term Memory
ML	=	Machine Learning
NER	=	Name Entity Recognition
NLP	=	Natural Language Processing
No WSD	=	Word Sense Disambiguation
Nsubj	=	Nominal Subject
Nsubjpass	=	Passive Nominal Subject
Pobj	=	Object of a Preposition
POS	=	Part of Speech
RNN	=	Recurrent Neural Network
SpanBERT	=	Span Bidirectional Encoder Representation from Transformer
SRL	=	Semantic Role Labeling

CHAPTER 1 INTRODUCTION

Legislative corpora are commonly ambiguous and challenging to interpret and understand in coreferences and meaning. It is a potential Natural Language Processing (NLP) research area focusing on the semantic meaning of the corpus and its application in the legal domain. There are various NLP fields, including bringing scientific contribution and its applications for several domains. For instance, the medical NLP research area is widely known but NLP in legal processing is yet a potential field and application.

This thesis topic aims to expand the knowledge and application of the Natural Language Processing field into the area of legislative language. We capture the meaning of the law and know which entity refers to what circumstance and what precisely the entity is performing. It should be noted that legal text processing is a difficult task for both humans and machines. This thesis topic aims to study and develop the mechanism composed of Natural Language Processing to manipulate and capture coreferencing and actions in the legislative corpus. Coreference is the occurrence occasion of text or expression referring to the same object. Actions are activities that a noun is performing, which usually refers to the sentence's verb. The introduction chapter presents the evolution of NLP, the meaning of coreference, problem statements, brief methodology, and expected results.

1.1 Natural Language Processing

The Natural Language Processing field is the capability of a computer to process human or natural languages. The principle of Natural Language Processing field is to have the software automated while comprehension manipulated the language including text and speech. In other words, NLP is a field aiming to mimic human natural language capabilities.

Natural Language Processing has evolved over time and there have been various applications. It starts from the complex sets of hand-written rules base. The transformational-generative grammar approach is introduced by Chomsky. [1], who is called the "father of modern linguistics." Transformational-generative grammar is an attempt to simplify the language

structures to be processed by the system [2]. It mimics the possible faculties that focus on how humans try to generate the grammar rule for regulating our speech. Later in the years the 2000s, it is the start of applications of Machine learning (ML) to Natural Language Processing. The groundbreaking work with Deep Learning (DL) is introduced, such as developing language modeling, parsing and word embedding. The application of Natural Language Processing varies according to the objectives. Some of the applications are related to the information retrieval field for search engine development. Another application includes Name Entity Recognition (NER), which is a method of finding the entity in the corpus systematically such as person names, organizations and places [3].

1.2   Coreference and Entity's Action

Coreference resolution is an application area where Natural Language Processing is applied to resolve the coreference in the text corpus [4]. By the meaning of coreference, it is a condition in which two or more linguistic expressions in the corpus refer to the same object, which can be persons or things. Giving an example of this sentence, "Ben is coming home late because he has a late-night meeting," the proper noun of the sentence is Ben, and the pronoun of Ben is he. This represents the binding phenomenon, in which the hypothesis of this phenomenon is a pair that has the same semantic meaning. Coreference is often composed of antecedent and proform. The antecedent is the expression that fully presents the full form of the object while the proform is an abbreviated form. co-indexed numbering is applied for tagging the corresponding objects. For instance, Ben_i and he_i are indexed with the index i showing these two indexes should be interpreted as coreferential. The coreference action is the activity that coreference is performing. For instance, "Jan is thinking she will go to shopping mall," the coreference is "she" referring to "Jan" while the verbs of this sentence are "think and go." According to Abstract Meaning Representation (AMR), the core relationship is "think." This indicates that "think" has a higher priority than "go". The explanation of AMR is further explained on chapter 2.

1.3 Statement of Research Problems

To apply NLP into the legislation corpus, it is essential to solve the problems of coreference and actions of entities. There has been significant development from various researchers in the modern development of Natural Language Processing. Unlike most commonly used applications like sentiment analysis, text summarization, and search engines, semantic domain are relatively rare to be explored since the referent, entities, and actions are difficult to be understood. Automated Reference Resolution Legal texts paper shows that NLP is applied by using machine learning which takes four-step framework, including mention detection, contextual information extraction, antecedent candidate extraction, candidate extraction, and antecedent determination [5]. This thesis is conducted on the coreference resolution and entity's action in legislative corpora. Moreover, this research extends the area to capture the semantic meaning representation for the coreference actions for further processing.

1.4 Objectives

This thesis aims to develop a model that can resolve the coreference and the coreference's action in the legislation corpus. The objective is divided into two parts:

1. To develop and implement a machine learning model which can perform the coreference resolution in the legislation corpus, which are Conventional on the Rights of the Child (CRC), Convention on the Person with Disabilities (CRPD), and Convention and Protocol Relating to the Status of the Refugees (CPRSR) [6, 7, 8].
2. To develop and implement a machine learning model which can identify and link the coreference with the action after it has been resolved in the CRC, CRPD, and CPRSR corpus for further meaning processing.

1.5 Summary of Methodology

There have been various techniques for coreference resolution which have significantly improved in recent research such as pair mentions models [9] as the first-order model, Bidirectional Encoder Representation from Transformer (BERT) [10], SpanBERT model [11] and Abstract Meaning Representation (AMR) [12]. In this thesis, the SpanBERT model, a

pre-trained model, is chosen to resolve coreference resolution. While processing the coreference resolution, speech tagging is implemented to build the sentence structure. To obtain the purpose of AMR, the replacement of coreferences is transferred to the Abstract Meaning Representation [13]. AMR is a graph representation that seized the coreference actions. The legal corpora include Conventional on the Rights of the Child (CRC), Convention on the Person with Disabilities (CRPD), and Convention and Protocol Relating to the Status of the Refugees (CPRSR).

1.6   Deliverables

This thesis aims to develop and experiment on the machine learning model which can perform the coreference resolution and find coreference actions within a particular sentence. The machine learning model can perform coreference resolutions and capture the entity's action. Entity's action is the action that conferences are performing. The entity's action is a semantic meaning representation. The expected model is the model which can contextualize the meaning of the CRC, CRPD, and CPRSR datasets. The result of this thesis should facilitate the reader to resolve the coreference and process the semantic meaning representation. In corresponding, it will be the steppingstone for a new researcher in processing the legal corpus for further development.

1.7   Requirements for this Research

This research utilizes Python programming language and libraries such as Spacy and AllenNLP. Also, it requires the knowledge of English's linguistic language, machine learning and deep learning, and the knowledge of English grammar.

1.8   Paper Organization

Firstly, chapter 1 introduces the background of Natural Language Processing with law, problems statements, objectives, methodology, and deliverables. Secondly, chapter 2 presents the literature review, which discusses the related work and research. Thirdly, chapter 3 defines the methodology, research method, and hypothesis. Fourthly, chapter 4 describes the results and discussion. Lastly, chapter 5 concludes the research work.

CHAPTER 2 LITRATURE REVIEW

In general, there are three main approaches when investigating NLP techniques applied to coreference resolution and meaning representation, which are linguistic-base, machine learning, and deep learning approaches. Linguistic approach discusses the linguistic based rules, dependency parsing, and text preprocessing techniques. Machine learning is used in addition to basic linguistic approach in Part of Speech tagging, and statistical methods including Hidden Markov Model, Pair Wised Model, and Abstract Meaning Representation. Regarding Abstract Meaning Representation, graph models are commonly used to trace the sentence's action. The last approach is Deep Learning models which include word vector representation, BERT model, SpanBERT model for seizing the coreference. Each of these has its advantages and disadvantages based on certain circumstances and types of datasets.

2.1 Linguistic Approach for Text Preprocessing

Linguistic approach is exploited in text preprocessing in 3 ways including, (1) Text Preprocessing, (2) Part of Speech Tagging, and (3) Dependency Parsing.
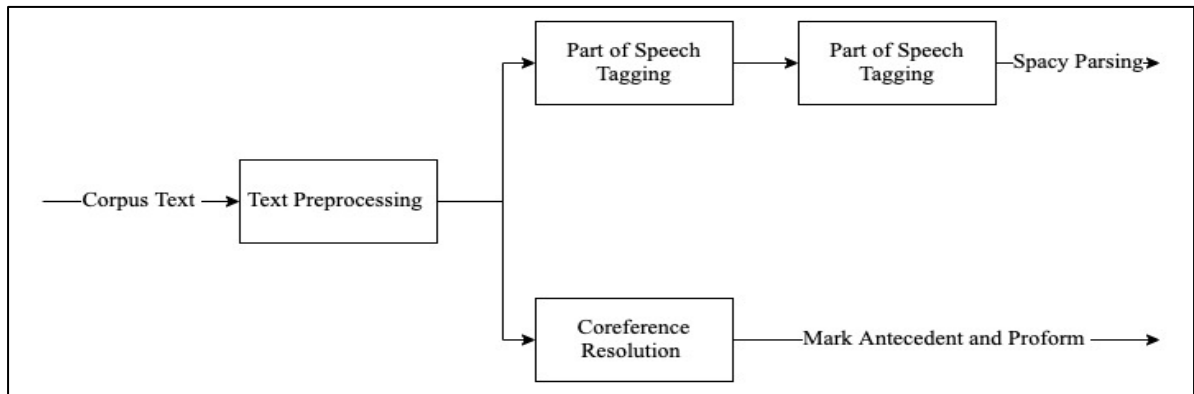
Figure 2.1 Natural Language Processing Pipeline

2.1.1 Text Preprocessing

Regarding the text preprocessing, there are several techniques such as tokenization, lemmatization, sentence segmentation, lower/upper case, removing the stop word, and normalization. This thesis focuses on tokenization, lemmatization, sentence segmentation, and lower case. Tokenization technique tokenizes a sentence into a single word while lemmatization technique changes the word into the root form. Sentence segmentation is a process of segmenting sentences into a single sentence. It is common to divide each sentence with a full stop, and all letters are lowered for easier compliance.

2.1.2 Part of Speech Tagging

Part of Speech (POS) is the category of each word in a language. POS includes nouns, pronouns, adjectives, prepositions, conjunctions, and interjection in English. This thesis focuses on nouns and pronouns in English for coreference. Christian Lehmann [14] mentions that Part of Speech has both semantic and structural aspects. The Part of Speech tagging also conveys the structure of the language. After tagging parts of speech of each word in sentences, the text is ready for further processing.

2.1.3 Dependency Parsing

According to [15], dependency parsing is discussed. Dependency parsing determines the structure of a sentence and relationship between each word in the sentence. For instance, Figure 2.2 shows the mapping of the dependency parsing when it is mapped into the sentence in the legislative corpus, which is "To diminish infant and child mortality." This shows the relationship between the tokens in the sentence. The grammatical structure and the dependency relationship between tokens in the sentence are the benefit of dependency parsing.
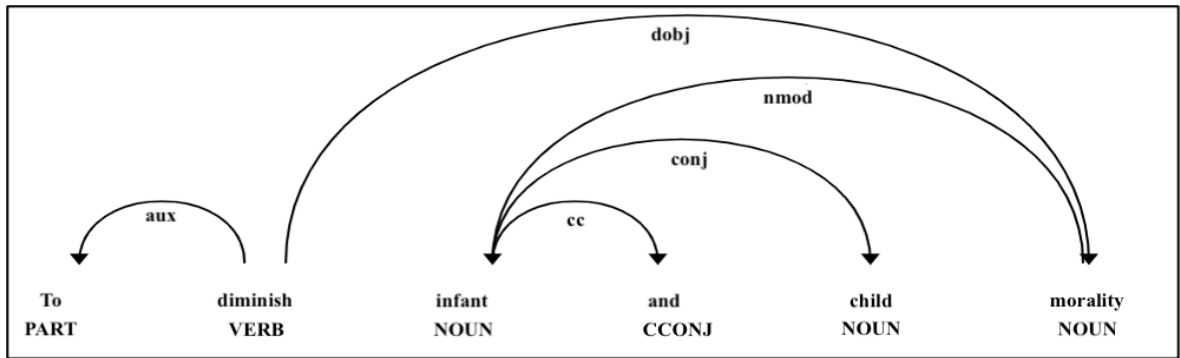
Figure 2.2 Example of Dependency Parsing of a Sentence from CRC Dataset

2.2 Machine Learning Approach

Machine learning approach focuses on exploiting the tools which lead to the understanding of relations and concepts in legal texts including (1) Spacy, (2) Part of Speech Tagging, (3) Pairwise Model, and (4) Abstract Meaning Representation (AMR). While Spacy is the text preprocessing tool, Part of Speech recognition can be performed using the Hidden Markov Model, which is the probability model that indicates the sequences of random stages. Pairwise Model is the classic coreference resolution machine learning model, which is the classifier to distinguish between two random noun phrases whether they have the similarity to be the antecedent and proform or not. Abstract Meaning Representation (AMR) is the graphical representation of the grammatical sentence structure, which is used to apply for feature extraction of relationships and concepts.

2.2.1 Spacy Library for Text Preprocessing

This Natural Language Processing technique can easily be achieved with the facilitation of the modern NLP library in Python known as Spacy [16]. Spacy is the preprocessing tool that facilitates the pipeline for tokenizing, lowering case, lemmatizing, Part of Speech tagging, and dependency parsing. Figure 2.3 shows the Spacy pipeline.
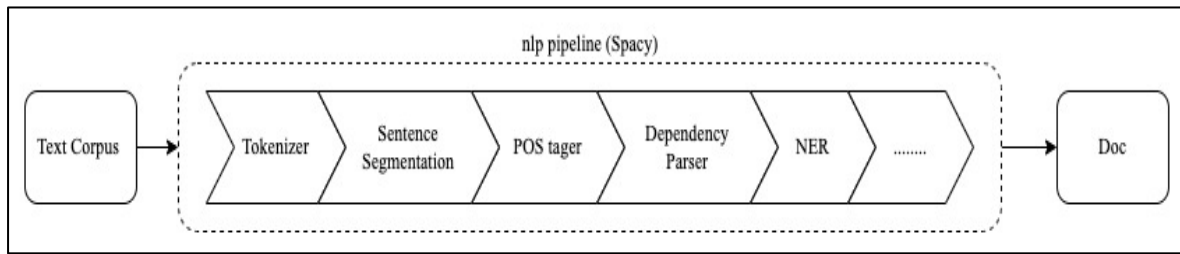
Figure 2.3 Spacy Pipeline

2.2.2 Markov Chain Model

The Markov Chain is a model that indicates the probabilities of the sequence from the random stages [17]. Figure 2.4 illustrates the process of the Hidden Markov Chain Model. In this model, the previous stages are taken in order to predict the current stages in sequence, however, there are constraints that previous stages from the pinpoint stage have no impact on the future stages unless the current stages are adjusted. As the ventricular language is in a sequential framework that suits the Markov Chain requirements, it is suitable to apply the Markov Chain Model with Part of Speech Tagging. While performing the model, the Hidden state is applied. Having the number of the possible tags or N stages, the matrix contains the tag transition, which has the probability of a tag based on the previous tag. After the probability of tagging associating with the words is calculated, the probability of the tagging falling into Part of Speech is given.

Although the Hidden Markov Model is the traditional way for the researcher to perform the Part of Speech Tagging, this thesis utilizes Spacy, which is more convenient to provide the speech tagging library.
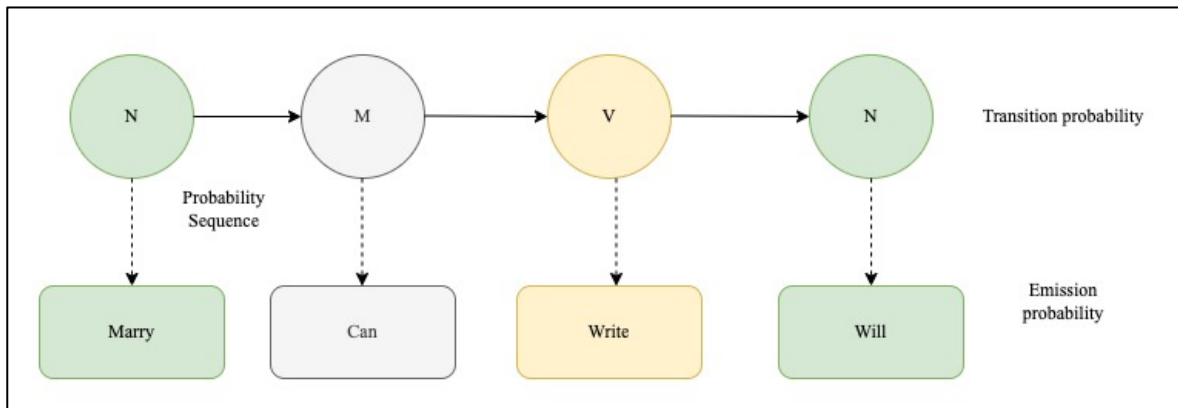
Figure 2.4 Sample of Markov Chain Model Capturing Part-Of-Speech (POS)

2.2.3 Mention Pair Model for Classic Coreference Resolution Task

Mention Pair Model is the classic machine learning algorithm for coreference identification. The Pairwise Model has one local classifier to decide whether the two-noun phrase (mentions) has the coreference or not. The sample mapping among mentioned pairs for coreference resolution is shown in Figure 2.5. To begin with, the Mention Pair Model groups related words by using single classifier. Then, all the mentioned pairs are checked to identify their distribution in the corpus. Emanuel Lassallle [9] states that space hierarchy learning improves the coreference resolution model. Using different types of pairs to develop the model for each of the mention pairs is the fundamental concept of this model. According to [9], the randomness of the mention pairs is taken into consideration. The data sparsity has a major downfall which is an overload of mention pairs, so the model scopes down the sparsity by choosing the space among the vast possibilities although the a priori remains unknown. Dividing the mention pairs in the model has the same logic as building the sample feature space. To build the hierarchy, the same decision tree technique is applied. The classification can be stopped before obtaining the last individual feature, however, the behavior must be at the same level. Up to this point, the identical copies of indicators stay in the subtree hierarchy in the complete decision tree. The linguistic divination is added to develop the relevant sequence of the indicator. Then, the classifier with the feature, which is Online Passive-Aggressive algorithm, is trained by the unique path from the root to the tree's leaf. In addition, the dynamic programming is used to optimize the computation of the best hierarchy by

removing the tree and maintaining the classifier on the left. The removal of the subtree begins when the subtree has a lower score than the node [18].
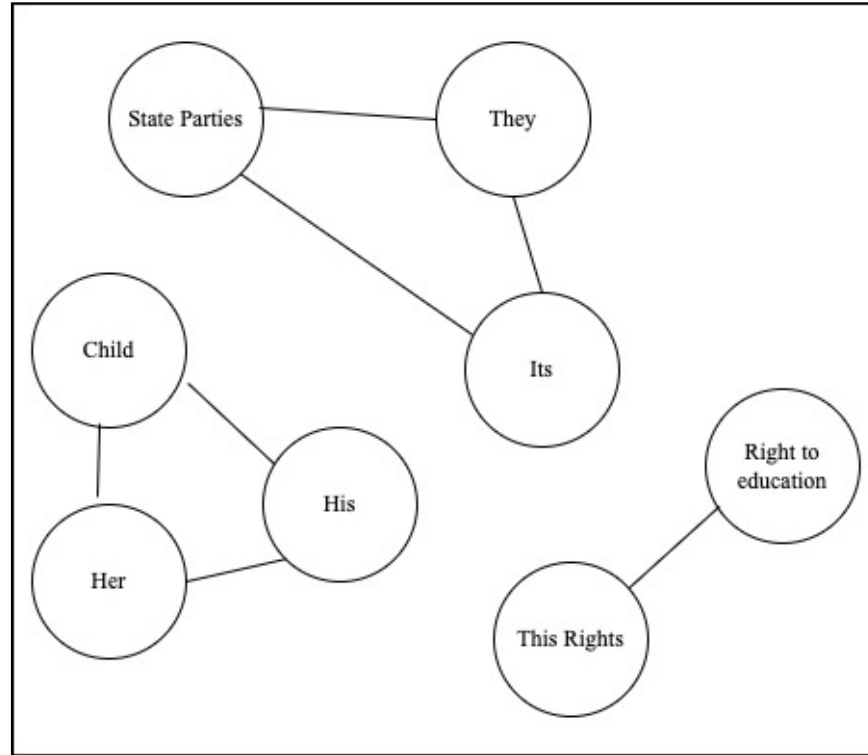
Figure 2.5 Sample of Mention Pairs (Coreferences)

2.2.4 Abstract Meaning Representation and Smatch Evaluation

Abstract Meaning Representation (AMR) is the semantic meaning of a sentence. This thesis focuses on the AMR of English which is designed to capture relationships between who is doing what and to whom [19, 20]. The result is shown in graph structure. The node of graphs is called "concept," and the relation is called "relation." The concepts (node) in AMR can be classified into four main categories: events concept, open class concept, constant, and special concept. Event class concept is the verb of the sentence and is followed by a number like go-01. The number ranges from 01 to 90. Open class concept is anything Part of Speech that does not contain numbers and be lemmatized as boy, women, or red. Constant is a number, string, and symbol such as date and year. Special concept is the concept that contains ARGN arguments which resolves the backlog of open class concept by extending the number beyond 90. Figure 2.6 shows the AMR of the sentence, "The girl wants to go to the store." There are

four concepts which are "girl", "want-01", "go-02", and "store". In this sentence, the verbs, "want" and "go", are followed by numbers according to a sequence of actions. The format of the graph varies according to the research paper; however, it is compulsory to remain the message of the relationship of who is doing what [21]. The structure of the AMR is represented in two main forms which are the PENMAN structure and the graph structure. PENMAN is the standard format to encode the rooted and directed acyclic graph. After bracketing the sentence to capture the spanning tree on the graph, the tree bracket is converted to label relations and nodes for AMR graph conversion. Both PENMAN and graphical structure are implemented in most research [22]. For the evaluation, the Smatch proposed by [19] is used to evaluate the output of the parser to check the similarity between AMR graphs. In the Smatch task, the level of precision, recall and F1 are calculated to evaluate the concepts and relations between two corpora of AMR graphs.

Smatch evaluation is exploited to test the overlapped meaning in a sentence pair [23]. [24] extends the Smatch evaluation to evaluate 12 experiments including the Smatch, Unlabeled, No Word Sense Disambiguation (WSD), Non_Sense_Frames, Name Entity Recognition, Wikification, Negations, IgnoreVars, Concepts, Frames, Reentrancies, and SRL. For Unlabeled, Smatch score ignores the relationship between nodes of the sentences. No WSD is computed based on the ignorance of the propbank tree. Name Entity Recognition is the F-score after computing on the name entity recognition (:name roles). Wikification is computed based on Wikification roles. The Negations is the F-score after computing negation or the polarity of the sentence. Concepts is the F-score that is computed on the task of the concept. To illustrate, Figure 2.7 compares the simple Smatch nodes between "The girl wants to go to the store" and "The girl wants to eat meat pie." Even though both sentences have similar structure, after comparing the nodes, it is found that the two sentences have completely different objectives.
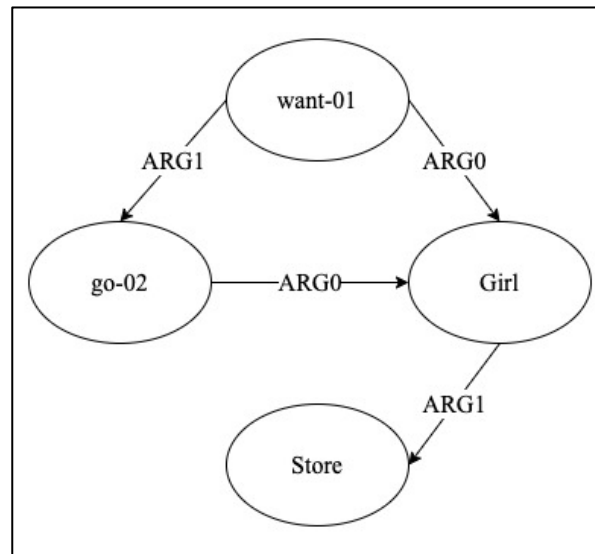
Figure 2.6 AMR for the Sentence "The girl wants to go to the store"
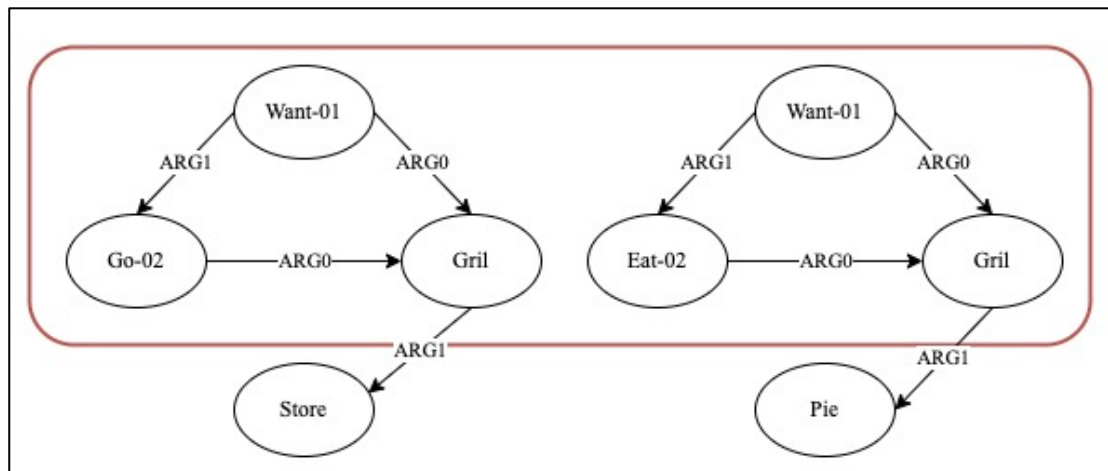
Figure 2.7 Smatch Evaluation in Sentence Pair

2.3 Deep Learning Approach

Deep Learning is a subfield of machine learning that mimics human brain working process. It comprises of neural network nodes inspired by the structure and function of human brain. Since the neural network has evolved, the Natural Language Processing gains significant benefits from this development. This approach consists of 6 techniques including (1) AllenNLP, (2) Word Embedding Technique, (3) Attention, (4) Transformer, (5) BERT, and (6) SpanBERT. To give more details on each technique, starting from the AllenNLP, it is a

platform that is built on the Pythorch library, which benefits researchers in several aspects such as text data processing, NLP abstraction and experimental framework. Secondly, Word Embedding is used in AllenNLP to facilitate the coreference resolution. The remaining techniques, which are Attention, Transformer, BERT, and SpanBERT, share the same concept whereas the SpanBERT is the latest version.

2.3.1 AllenNLP

AllenNLP library converts the representation of text into a vector sequence which is then transformed into a set of new vector sequences before merging them into a single vector [25]. The token embedder, one of the functions provided, takes text arrays as the input before generating the new outputs which are sequences of embedded vector that can be adjusted. The configuration can be made by choosing between the pre-trained word embedding, and word embedding concatenated with a character level. Convolutional Neural Network (CNN) encoding or the pre-trained model with context embedding, and Seq2SeqEncoder are also provided in AllenNLP by taking in the sequence of word vector and Recurrent Neural Network (RNN) and returning the output as the new vector sequence. In addition, Seq2VecEncoder is the model which merges the sequence of a vector into a single vector by using RNN or feature extraction by using the pooling, or Convolutional Neural Network. Lastly, the Span Extractor is a module provided in AllenNLP, which is beneficial for the coreference resolution. The span of text is determined by computing the sequence of token vectors. AllenNLP is designed for good science practice where the researcher can control the attributes for experimenting. Due to its benefit, the researcher can configure the model outside the code.
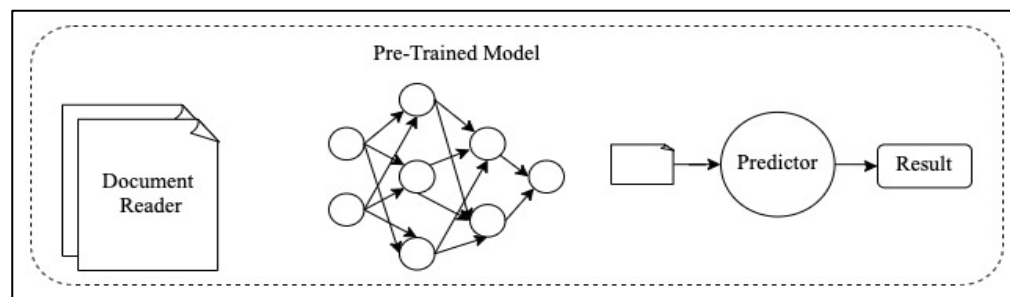
Figure 2.8 Component of AllenNLP

2.3.2 Embedding Technique

The word embedding technique is used to vectorize the single word token or the single sentence to find the closest meaning of the word to others. To illustrate, in Figure 2.9, the word "queen" and "king" are closed to each other in vectorized format. With the meaning measurement, word embedding is a potential tool for comparing coreference resolution. However, after the synthesis of [26], the word embedding does not significantly improve the coreference resolution task. The overall performance improves only by 0.5% in all test cases. The experiment is conducted on the hypothesis that word embedding can facilitate the coreference resolution. Henlein and Melher. [26] have carried out the experiment to test six different embedding methods. Focusing on the Word2vec, the datasets are trained with a fixed vocabulary. Figure 2.9 depicts the simple Word2Vec concept. As previously mention, a preprocessing method is applied, such as the lemmatization and tokenization to the 400,000 most used words. The Word2vec is a method that creates the word embedding. It is mostly used in neural network for the model to seize the relationship and association among the text corpus. It represents a word token in a form of a vector which indicates the semantic similarity between the word vectors. The word embedding is trained with the 300 standard paraments. After the embedded technique is completed, the token word embedding represented the relationship between tokens. This technique is utilized in the Attention model to enhance overall performance. The coreference resolution uses SpanBERT, along with the help of AllenNLP. The comparison of the unmodified datasets and the version of the pronoun with the completed antecedent replaces the mention head. As a result of the investigation, there is no correlation between coreference resolution and word embedding.
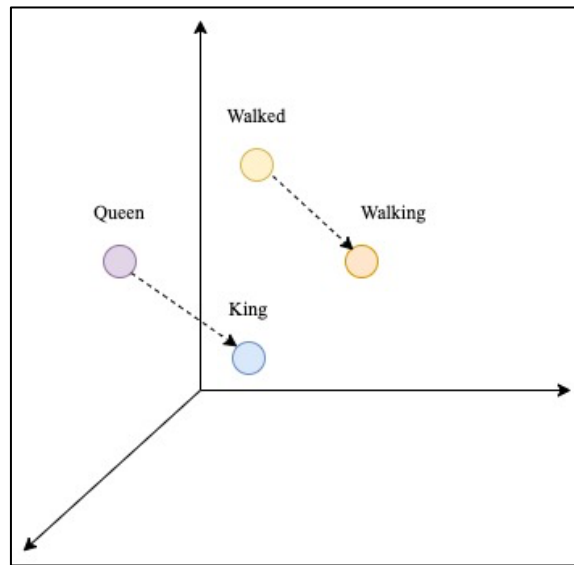
Figure 2.9 Sample of Word2Vec Embedding

2.3.3 Attention

As the development of the neural network, the Sequence-to-Sequence learning weight of the Attention algorithm is introduced as the baseline of the Transformer [27]. The paper named "Attention is All You need" describes the Transformer and the Sequence-to-Sequence model. The neural network model changes the sequence of text into other sequences, which is commonly known as encoding technique. Seq2Seq model is applied in several aspects of NLP, such as the translation task where the sequence of the text in the current language is encoded and transformed into another language. The LSTM or the Long Short-Term Memory-based model is developed from the RNN, which takes the sequence of text as an input. This deep learning method is highly effective when dealing with the data in sequence. The modification of the RNN is the LSTM where the model complies with the electronic concept in which it tries to imitate human capability of learning. Humans and LSTM have one common ability which is the forgetfulness. However, forgetfulness eventually leads to an ability to retain some information back to our memory. This technique seems to be very effective. LSTM module not only provides the meaning of the sequence, but it also has the ability to remember the sequence of data and sentence. The architecture of Seq2Seq is explained by having the encoder and decoder. Encoder acts like the feature reduction where

the data is taken in high dimensional space (dimensional vector) while the decoder transforms the pair encoding to the output sequence, such as language. The Attention model comes to play in the pipeline at the stage of Weight of Attention. At this stage, the decision is made to identify which Attention Weight is important. The Attention model enhances the Seq2Seq capability by remembering the important part of the sentence with the semantic meaning while processing the decoding process with its standard output. To elaborate, Figure 2.10 shows that the LSTM takes input into the encoding process, and the Attention mechanism continuously takes several inputs into account and makes a decision on which weight is not essential. The decoding process takes encoding sentence as the input and weight from the Attention Mechanism is further processed by Transformer.
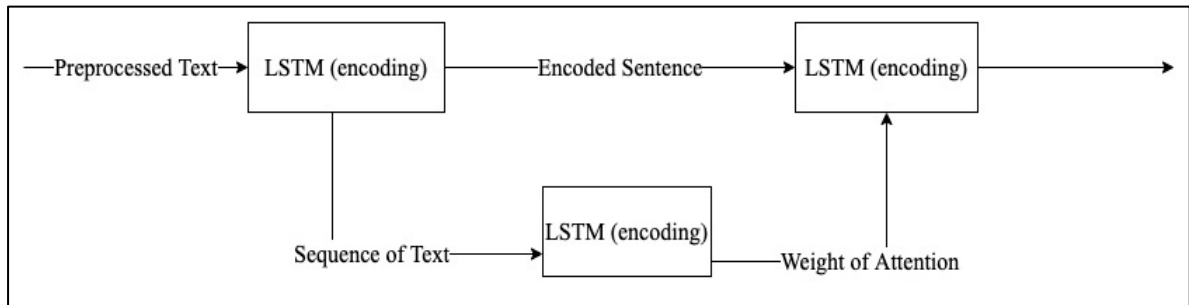
Figure 2.10 Attention Mechanism

2.3.4 Transformer

The Transformer is composed of the encoder and decoder which consist of several parts that can be stacked on top of each other several times [27]. The Transformer architecture is depicted in Figure 2.11. The model has a variety of multi-head Attention and feed-forward neural network layers. The input token is embedded into an n-dimensional space. The embedded representation of the multi-dimensional space acts like RNN to remember the sequence's position. The model takes the value of the vector representing all the words in the sequence. Then, the weight is represented by the word of sequence. After that, the activation function called Softmax is applied. The Softmax function returns the value as the probabilities ranging from 0 to 1. The weight of SoftMax is applied to all the words in the sequence. While the processing is running, the Attention Mechanism is parallelized, and the

several procedures are looping. The matrix shows the variation of the Attention mechanism depending on the encoder and decoder. This is highly beneficial because the system can walk through the encoder input sentence or the part of the decoder input sentence. The help of the multi-head Attention that links the encoder and decoder ensures that both inputs are paralleled up to the certain position.
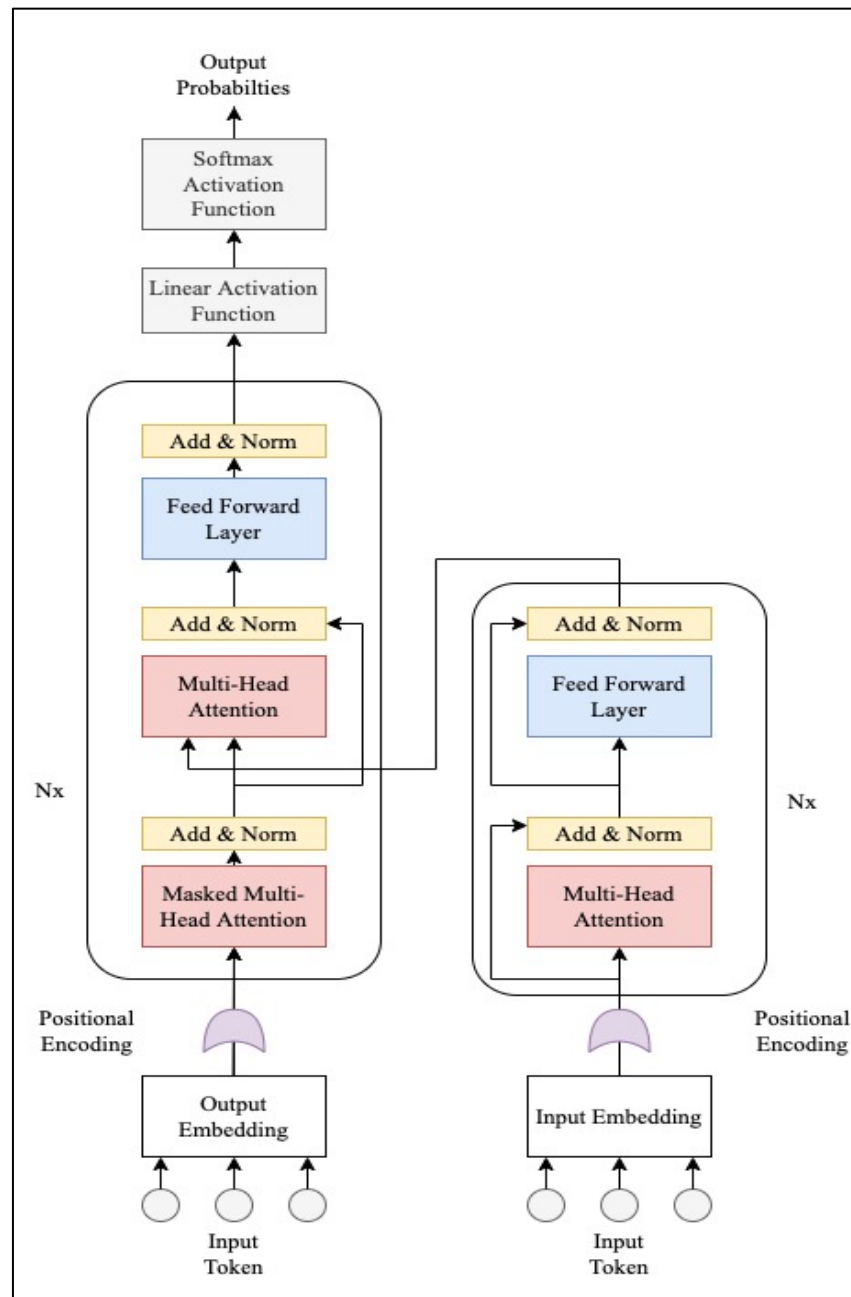
Figure 2.11 Transformer Architecture

Attention and the Transformer model are the baselines of modern deep learning technique. Attention takes the actual weight into account to resolve the gradient vanishing when it reaches its maximum. In addition, Transformer provides the architecture for processing the text classification, baseline for the Bidirectional Encoder Representation Transformer (BERT), and Span Bidirectional Encoder Representation Transformer (SpanBERT). BERT and SpanBERT models develop based on the state-of-the-art Transformer model, which uses multi-Attention, feed-forward neural network, and Softmax activation function for classifying output.

2.3.5 Bidirectional Encoder Representation Transformer (BERT)

The Bidirectional Encoder Representation from Transformer is a pre-trained neural network with bidirectional representation [10]. The bidirectional neural network takes the input from left to right and right to left at the same time. As the name implies, the BERT model uses the Transformer neural network as part of the mode. There are two steps in BERT, which are pre-training and fine-tuning process. Figure 2.12 illustrates the BERT architecture. The architecture of the model is the multi-layer bidirectional Transformer encoder. It is developed based on the idea of the Transformer, which includes the encoder and decoder technique. The input is taken as the single sentence and a pair of sentences and represented it as one token sequence. Then, the sentence pairs are combined as a single sequence. Recalling the prior knowledge, the Transformer takes input tokens in the sequence. The model distinguishes the sentence by using two different methods: separating with the special token ([SE]), and adding a learned embedding to every single token to pinpoint if it belongs to sentence A or sentence B. In summary, the inputs are taken as the summation of the corresponding tokens, segments, and the position's embedding. The first part of the BERT algorithm is the Pre-training BERT. The bidirectional neural network allows the model to read input from left to right and right to left, giving the model to predict the target word in multiple contexts trivially. The model masks token of the complete input sequence. The "masked LM" or MLM is introduced. The masked token is usually covered from the entire corpus when it is being processed. The model usually masks 15 percent of the total sequence input. The mask token, which corresponds to the final hidden vector, is passed to the output layer from SoftMax to vocabulary. The second task of the Pre-training BERT is the Next Sentence Prediction (NSP). The NSP allows the

model to understand the relationship between two sentences which are extremely helpful in coreference resolution. The NSP can be trained in any monolingual corpus. There are sentence A and sentence B for the pre-training sample. Fifty percent of sentence A is followed by sentence B while the remaining 50 percent is the random sentence. Another task of the BERT is Fine-tuning BERT. Transformer allows the mode to formulate the downstream tasks to check whether the mode is the single text of the pair or not. This can be performed by exchanging the input and output. The encoding is done to the next pair, which is applied to the bidirectional cross Attention. The self-Attention mechanism combined two stages which are encoding a concatenated test pair and feeding self-Attention between two sentences.
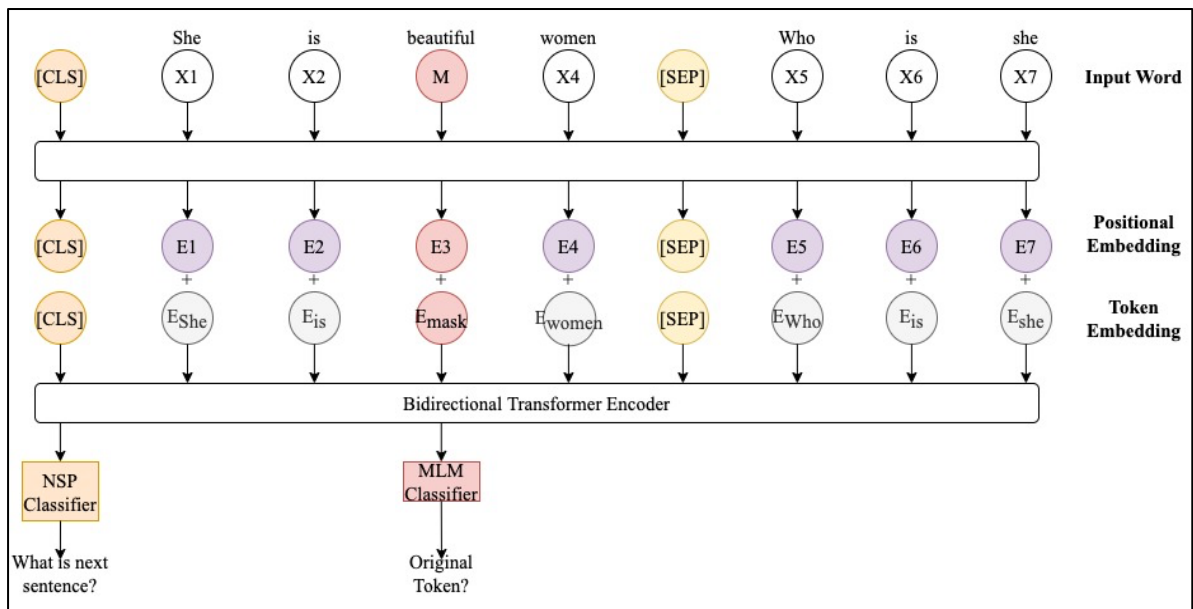
Figure 2.12 BERT Architecture

2.3.6 Span Bidirectional Encoder Representation Transformer (SpanBERT)

SpanBERT is an extension of the BERT model. It is claimed that the model is designed to improve the representation and prediction text span. Unlike BERT model where individual token is masked, SpanBERT masks the random contiguous spans. The Span Boundary Objective (SBO) is introduced to capture the prediction of the entire masked span. The SBO predicts the entire span rather than focusing on the overall context. It also stores span-level information. The geometric distribution is used to sample the span length, which skews the

distribution to the shorter span. Moreover, the Span Boundary Objective facilitates the model in representing the end of t-span for summarizing the internal span content. The SBO involves the prediction of the masked span token by representing the observed token. The summation of the loss of both span boundary and regular Masked Language Model are calculated. Meanwhile, SpanBERT reuses the input embedded in MLM and SBO. The Single Boundary Objective characteristic is depicted on Table 2.1, and Figure 2.13. The Next Sentence Prediction (NSP) is trained with the single segment sentences, not the two half-length segments. As the actual input, the document is divided into the segment of predefined length. The segment is encoded individually by the Transformer. The feedforward neural network processes the scoring function. It is confirmed that the single sequence training and the SBO improve performance, which is different from NSP that has the adversary effect on the performance.

As mentioned above, the coreference resolution section is handled by a deep learning model. This section shows the development of deep learning according to the Natural Language Processing and coreference resolution lens. It starts from the libraries commonly used in dealing with Coreference. AllenNLP is feasible, and configurable. Also, it contains the pre-trained model for the application, which is the strength of this library. In addition, it concludes all the models which are required to complete the coreference resolution tasks. Transformer and Attention have the strength to process the text sequentially. The Transformer model takes Attention models to another step by taking encoding and decoding in the loop. BERT elaborates the idea by processing the text in a bidirectional format which strengthens the text processing sequentially in both directions. To conclude, SpanBERT is commonly known for the coreference resolution because it is a pre-trained model that is trained by many datasets. It is assumed that the deep learning model becomes more accurate and precise. The coreference tasks are more predictable when having Span text and masked tokens. In this thesis, AllenNLP and SpanBERT are utilized in complex sentence in the system. The system should return the set of references and replace the proform with the antecedent.

Table 2.1 Sample of Mask Language Model in BERT and SpanBERT

Original Sentence: "Ben is coming home late because he as a late-night meeting"	
BERT with Masked Token	Ben is coming home [masked] because he has a late-night meeting.
BERT prediction with incorrect word	Ben is coming home [behind time] because he has a late-night meeting.
BERT prediction with correct word	Ben is coming home [late] because he has a late-night meeting.
SpanBERT masks using Span Boundary Objective	Ben is coming home [masked1] [masked2] [masked3] [masked4] a late-night meeting.

The	state	Parti es	unde rtak e	To	ensur e	the	chil d	such	protecti on	and	care	as	is	nece ssar y	for	his	or	her	Well - bein g
									Transformer										
x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20

Figure 2.13 SpanBERT on the Span Boundary Objective

2.4. Related Work in Legislative corpus Processing

In processing legal texts, understanding of language is required. NLP has been used for the analysis and exploitation of legal information in various areas including drafting, managing, retrieving, linking, reasoning and argumentation, for example. One of the works which is closely related to our work is [5]. [5] shows the reference resolution in a legal corpus. This research takes two major steps for coreference resolution task including the text processing, and confusion matrix. [5] also shows the coreference resolution tasks in the Japanese National Pension Law corpus. The paper shows the coreference resolution in the legal engineering domain. The primary investigation is to work on the mention pair detection, coreference extraction, and mention resolution. This work is completed on the Japanese National Pension Law corpus written in the Japanese language. The work takes four-step framework: mention detection, contextual information extraction, antecedent candidate extraction, and antecedent determination. The mention detection process achieves an 80.06% F1 score, and the antecedent determination step achieves 87.03% accuracy. In overall, it achieves a 67.02% F1 score on End-to-End performance.

Figure 2.14 shows overall four-step framework from [2]. Firstly, mention detection which is the first step for identifying the occurrences of mentions is further applied to context information extraction and antecedent candidate extraction. Secondly, context information extraction identifies the context information. The task is divided into three subparts: mention splitting, mention classification, and position recognition. Mention splitting splits the mention into two parts: position and content. The position part locates the referent's location while the content part extracts the mention that is referred. Mention classification decides whether the located mention is in the existing document or in other documents. Position recognition determines the scope of the antecedent on article level, and paragraphs. Thirdly, antecedent candidate extraction receives output from step 2 to identify the scope of antecedent and to extract the antecedent candidate in a sentence by dependency trees. Lastly, the antecedent determination receives the list of candidates extracted from step 2 to determine the fragment of the text. Antecedent determination chooses the best antecedent candidates to be the output of coreference.

To summarize, three main approaches are applied in our research. The preprocessing of texts is done by Spacy while Coreference Resolution is completed by SpanBERT. Lastly, the Abstract Meaning Representation including concept and relation is used for the logical semantic meaning representation. The PENMAN logic format creates the logical meaning of the sentence before AMR creation. After that, AMR is converted to a graph to visualize semantic meaning representation, which is evaluated by the Smatch evaluation to check the overlap meaning between sentence pair.
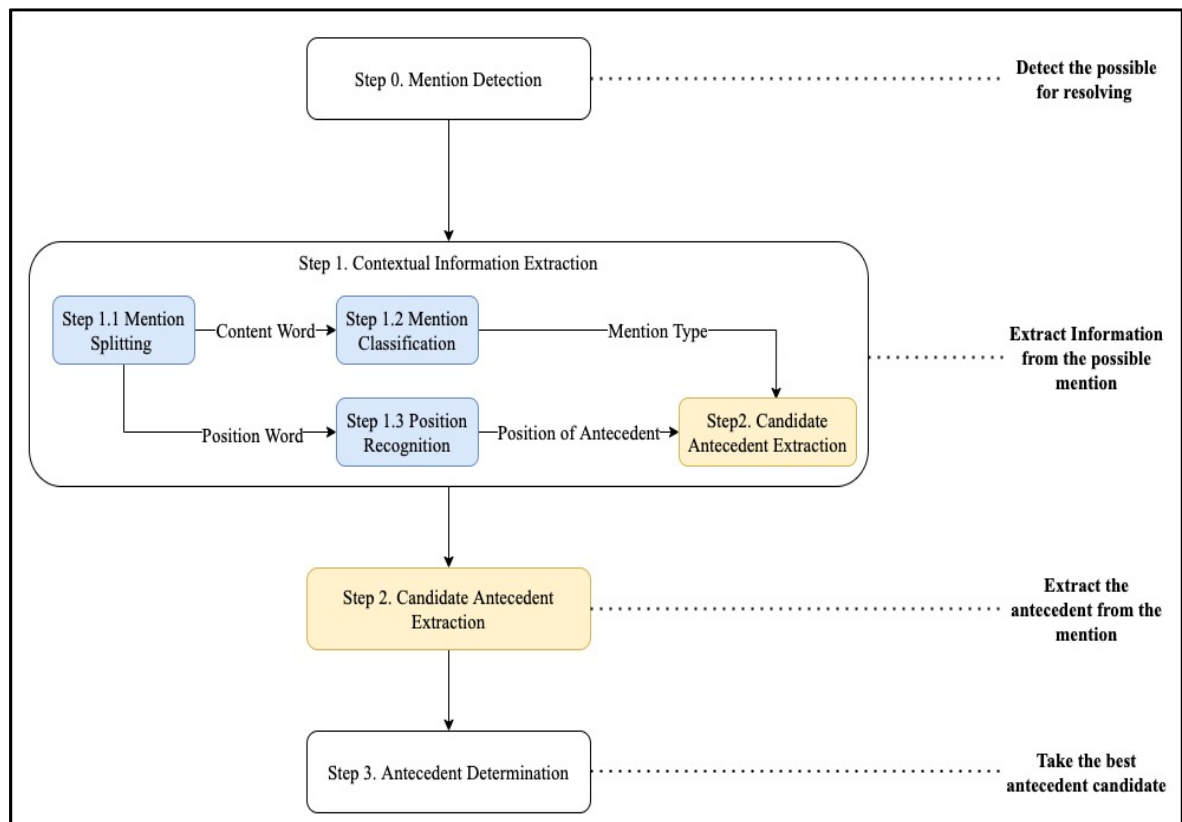
Figure 2.14 Four-Step Framework from Related Work in Legal Text.

CHAPTER 3 METHODOLOGY

This chapter starts with the summarization of methodology before elaborating each section in more details. This thesis focuses on coreference resolution and determines the entity's action in the legislative datasets which are Conventional on the Rights of the Child, Conventional on the Rights of Person with Disabilities, and Convention and Protocol Relating to the Status of Refugees.

Figure 3.1 shows the overall process of the model, which concluded the input-process-output in one high-level diagram. It should be noted that dependency parsing was not used in this experiment because AMR structure and parsing yielded in different characteristics. The datasets were processed in a variety of methods according to the objectives as shown in the diagram. The input was transferred to each main process starting from POS tagging, coreference resolution, and Abstract Meaning Representation. Figure 3.1 also shows the overall high-level diagram of this research methodology. The CRC, CRPD, and CPRSR datasets were passed through the text preprocessing process. The output of text preprocessing was the sentences organized in the structure. Firstly, the organization method was described in text-preprocessing section. Secondly, the Part of Speech tagging was conducted to explore the datasets. Thirdly, the coreference resolution was performed by SpanBERT before the output was fed into AMR. Lastly, Smatch and BLEU were used to evaluate the final output of AMR [28].
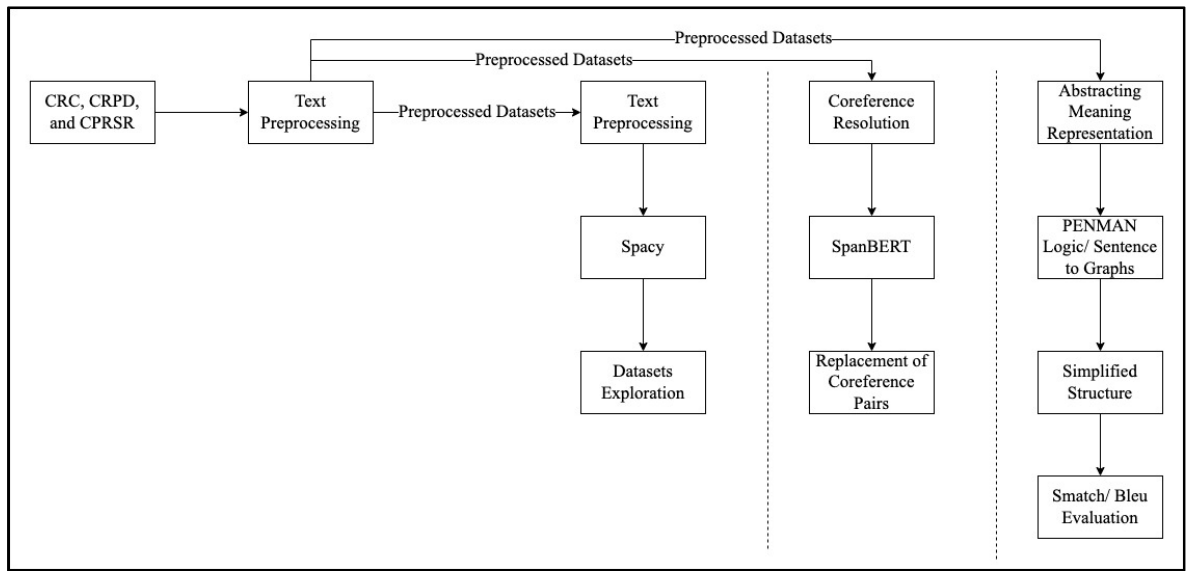
Figure 3.1 High-level Diagram of Methodology

3.1 Overview of the Methodology

The methodology was divided into the following procedures:

1. The datasets were received as an input which was known as text corpus. The text preprocessing step was processed according to the following stage's objective, as shown in Figure 3.1.

2. The preprocessed datasets were transferred to the POS tagging stage where the Spacy tagged the Part of Speech. The output of this model was the tagged sentence.

3. The preprocessed datasets were fed into the coreference resolution where the SpanBERT model resolved the coreference and replaced the antecedent back to the sentence.

4. The preprocessed datasets were fed into the Abstract Meaning Representation together with the mentioned input from the above section.

5. The simplified structures from step 4 were fed into the Smatch and BLEU evaluation process. The output of this stage was used to compare meaning preservation between resolved and non-resolved coreference pairs and to check the logical semantic structure of AMR.

3.2 Text Preprocessing

After exploration, it was found that the terminology exercised under this convention concentrated on children who were defined by the convention as any human being under the age of eighteen. Convention on the Rights of Person with Disabilities was the treaty of the United Nation where the purpose of this convention was to protect the rights and dignity of the person with disabilities. The CRPD was adopted on 30th March 2007 and was drafted on 13th December 2016. The terminology of this convention focused on the rights of person with disabilities. Convention and Protocol Relating to the Status Refugees was the multi-lateral that took the refugee context into account. CPRSR, which was adopted on 28th July 1951, discussed the definition and protocol of the refugee, including the war criminal. The terminology in this convention focused on the humanitarian context.

In a classic NLP pipeline, the text corpus was preprocessed according to the objective of the next stage, such as coreference resolution, Part of Speech tagging, dependency parsing, and Abstract Meaning Representation. Each stage had unique and similar requirements from the preprocessed text stages. The datasets description, text-preprocessing requirements according to the stage, and tools were provided accordingly. Figure 3.2 presents the part of the data in the CRC.

Figure 3.2 demonstrates the text sample in the CRC; the text was formed in academic writing and the legislative corpus. The datasets were divided into sub-articles, and each article had its unique consideration. In addition, the datasets showed some coreference samples in the corpus, such as "his" and "her" representing the child.

Article 1

For the purposes of the present Convention, a child means every human being below the age of eighteen years unless under the law applicable to the child, majority is attained earlier.

Article 2

1. States Parties shall respect and ensure the rights set forth in the present Convention to each child within their jurisdiction without discrimination of any kind, irrespective of the child's or his or her parent's or legal guardian's race, colour, sex, language, religion, political or other opinion, national, ethnic or social origin, property, disability, birth or other status.
2. States Parties shall take all appropriate measures to ensure that the child is protected against all forms of discrimination or punishment on the basis of the status, activities, expressed opinions, or beliefs of the child's parents, legal guardians, or family members.

Article 3

1. In all actions concerning children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities or legislative bodies, the best interests of the child shall be a primary consideration.
2. States Parties undertake to ensure the child such protection and care as is necessary for his or her well-being, taking into account the rights and duties of his or her parents, legal guardians, or other individuals legally responsible for him or her, and, to this end, shall take all appropriate legislative and administrative measures.

Figure 3.2 Sample of the Convention of the Rights of the Child

Article3

General Principles

The principles of the present Convention shall be:

(a) Respect for inherent dignity, individual autonomy including the freedom to make one's own choices, and independence of persons;

(b) Non-discrimination;

(c) Full and effective participation and inclusion in society;

(d) Respect for difference and acceptance of persons with disabilities human diversity and humanity;

(e) Equality of opportunity; Accessibility;

(f) Accessibility;

(g) Equality between men and women;

(h) Respect for the evolving capacities of children with disabilities and respect for the right of children with disabilities to preserve their identities.

Figure 3.3 Sample of Convention of the Rights of the Person with Disabilities

---

Article 46

Notifications by the Secretary-General of the United Nations

The Secretary-General of the United Nations shall inform all Members of the United Nations and non-member States referred to in article 39:

   (a) Of declarations and notifications in accordance with section B of article 1;
   (b) Of signatures, ratifications and accessions in accordance with article 39;
   (c) Of declarations and notifications in accordance with article 40;
   (d) Of reservations and withdrawals in accordance with article 42;
   (e) Of the date on which this Convention will come into force in accordance with article 43;
   (f) Of denunciations and notifications in accordance with article 44;
   (g) Of requests for revision in accordance with article 45.

Figure 3.4 Sample of Convention and Protocol Relating to the Status of the Refugee

Several preprocessing techniques were applied in this experiment, including sentence segmentation, text tokenization, text lemmatization, and legal text preprocessing. Legal text preprocessing was conducted to facilitate the machine to process the legal corpus. Text tokenization was a technique that divided the sentence into a single word. In addition, the text lemmatization was completed to return the word into an infinitive form.

Common characteristics were shared among the three datasets. Each dataset had the introduction part, each part had its articles, and each sentence had its sub-sentences. The introduction part was excluded from this experiment because it was composed of a noun phrase which had no implication toward the meaning of the datasets. The preprocessing work began with transforming the corpus complex structure into a more simplified version for machine processing. There were sub-sentences in the sentence, which was mandatory in this experiment to rearrange and conclude structure. Figure 3.3 and Figure 3.4 illustrate the sample of the sentences and sub-sentences. The datasets were being rearranged in the text according to the following priorities: Sentence level, Sub-Sentence number, Sub-Sentence alphabet, and Sub–Sentence Roman numerals.

In Figure 3.5, the datasets were preprocessed and divided into article levels. Then, the sentences were preprocessed into appropriate structure. The Regular Expression was used to preprocess the datasets by checking Roman numerical and Arabic numerals [29]. The white space was checked to ensure that sentences had been arranged correctly while Spacy were

applied for sentence segmentation. After the segmentation, the datasets were compiled according to the article levels and saved as individual files. Then, all the datasets were changed to lower character. From this stage, the datasets were ready to be processed and counted for the benchmark. To ensure Spacy performance, the "." was counted to identify the total number of sentences. The sample outcome of the text preprocessing process is shown on Figure 3.5.
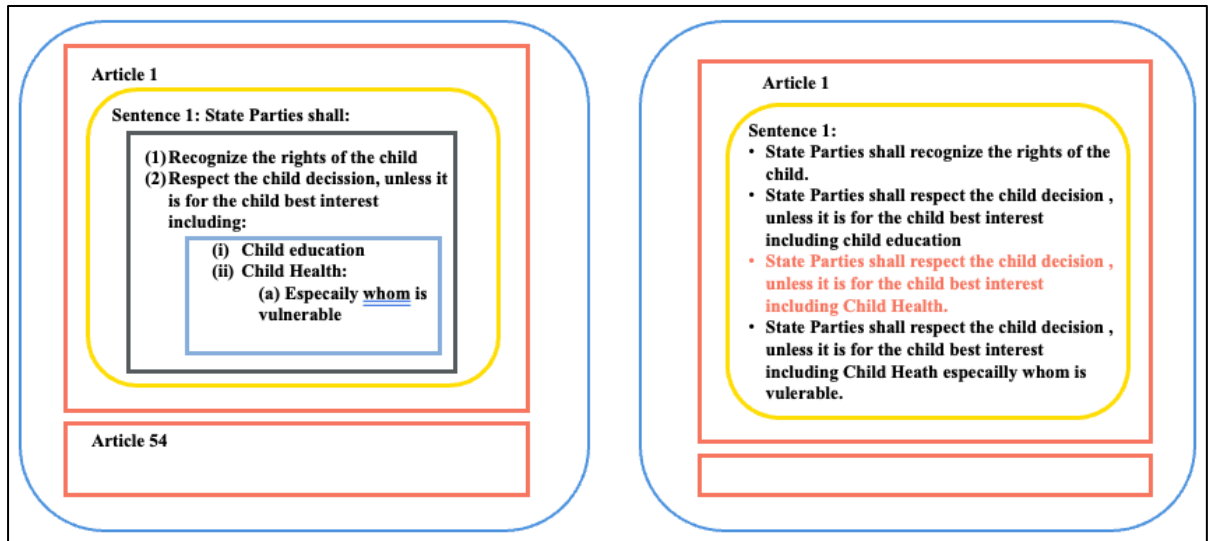
Figure 3.5 Corpus Structure Before and After Preprocessing

3.3   Spacy for Part of Speech Tagging and Dependency Parsing

After receiving the input from text preprocessing, Part of Speech and dependency parsing were performed by Spacy. The Part of Speech tagging demonstrated the dataset exploration and characteristics of nouns, pronouns, and verbs. Figure 3.6 shows the Part of Speech tagging which was completed by Spacy. The preprocessed datasets were fed into Spacy to identify the Part of Speech. The dependency parsing was completed by the identical method from the Part of Speech tagging. However, the result of the dependency was not used in the Abstract Meaning Representation and the reason was discussed in chapter 4.
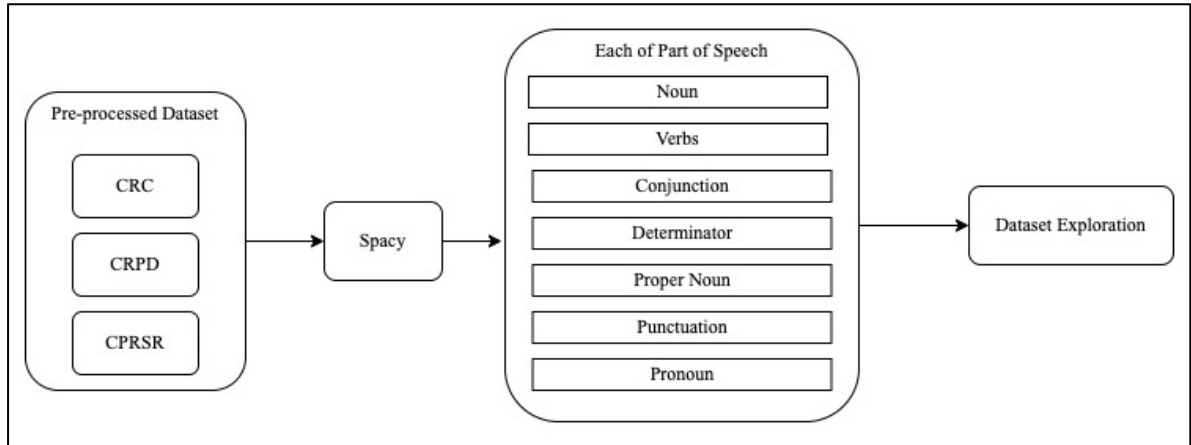
Figure 3.6 Part of Speech Tagging Process

3.4 Coreference Resolution

The coreference resolution was performed in several ways. The pre-training method was the favored method to obtain the coreference resolution. The preprocessed texts were added to the SpanBERT which was provided by AllenNLP.

Figure 3.7 shows that the model received sentences as input. The preprocessed sentences were passed to the coreference resolution process. The model called "coref-spandbert-large-2020.0" was loaded to use the pre-trained model to achieve the thesis objective. The model was divided into clusters and document parts, cluster predictor, and document predictor. The loop began with the sentence input process to find and update the number of mentions found in the sentences. After finding the coreference, the coreference object was appended into one single list. The coreference clusters were displayed corresponding to antecedent and proform of each sentence. In addition, the Predictor library provided functions that resolved the coreference back into the sentences while antecedent was being replaced into sentences. Then, the output of this stage was fed into the Abstract Meaning Representation stage to resolve the entity's action. The expected result of this stage was to achieve the coreference resolution cluster, which included a cluster of antecedent and proform.

The experiments utilized the pre-trained model, as shown on the Figure 3.7. The algorithm, in Figure 3.8, processed the embedding Span Boundary Objectives on the bidirectional

Transformer model. The max text length was 512, and the feature size was 20. The maximum span width was 30, with 1,024 Transformer dimensions [30]. The mention was detected with the Span head and Span Representation level. The feed-forward neural network and the antecedent score were utilized. The mentioned feed-forward neural network received 61,440 dimensions as input, with two layers of the neural network, and 1,500 hidden layers. The activation function was ReLu. The dropout rate was 20%. The antecedent feed-forward neural network received 3,686,400 dimensions as input. The number of hidden layers, activation function, and dropout rate remained the same from the previous neural network. The outputs of feed-forward neural networks were fed to the Softmax Activation function to return the probabilistic classification output. The span per word rate was 40%. The model was trained with 40 epochs, patience was 10, and optimizer was "hugging face_adamw".

The expected output was the list of antecedents and the proform of the coreference. Then, the output of this stage was replaced by the antecedent of the coreference. For example, if the input was "States Parties undertake to ensure the *child* such protection and care as is necessary for *his* or *her* well-being, taking into account the rights and duties of *his* or *her* parents, legal guardians, or other individuals legally responsible for *him* or her*, and, to this end, shall take all appropriate legislative and administrative measures.", the outputs were ['child', 'his', 'her', 'his', 'her', 'him', 'her'] and ['ensure', 'this', 'end']. The replacement was completed, and the result was "States Parties undertake to *ensure the child* such protection and care as is necessary for the child's well-being, taking into account the rights and duties of the child's parents, legal guardians, or other individuals legally responsible for the child, and, to *ensure*, shall take all appropriate legislative and administrative measures."
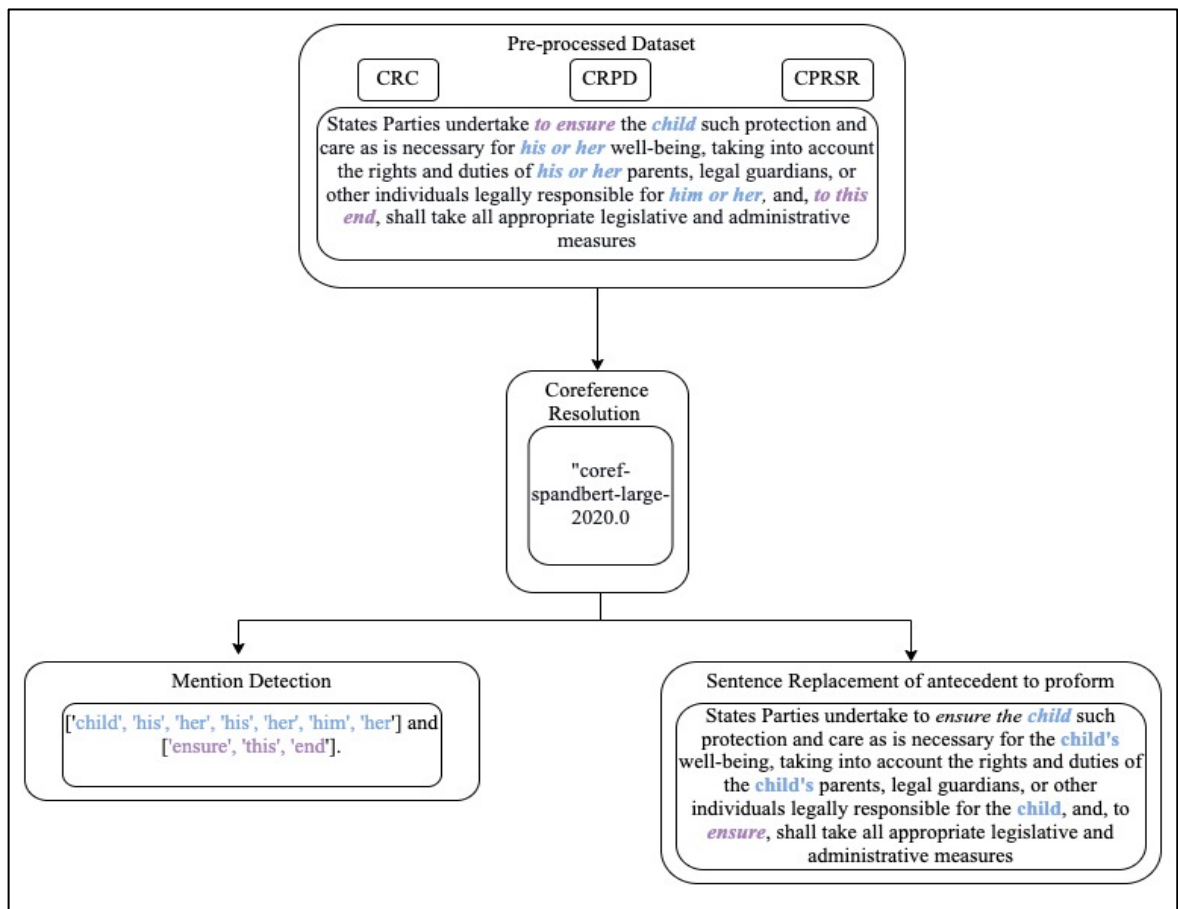
Figure 3.7 Input-Process-Output of Coreference Resolution Stage
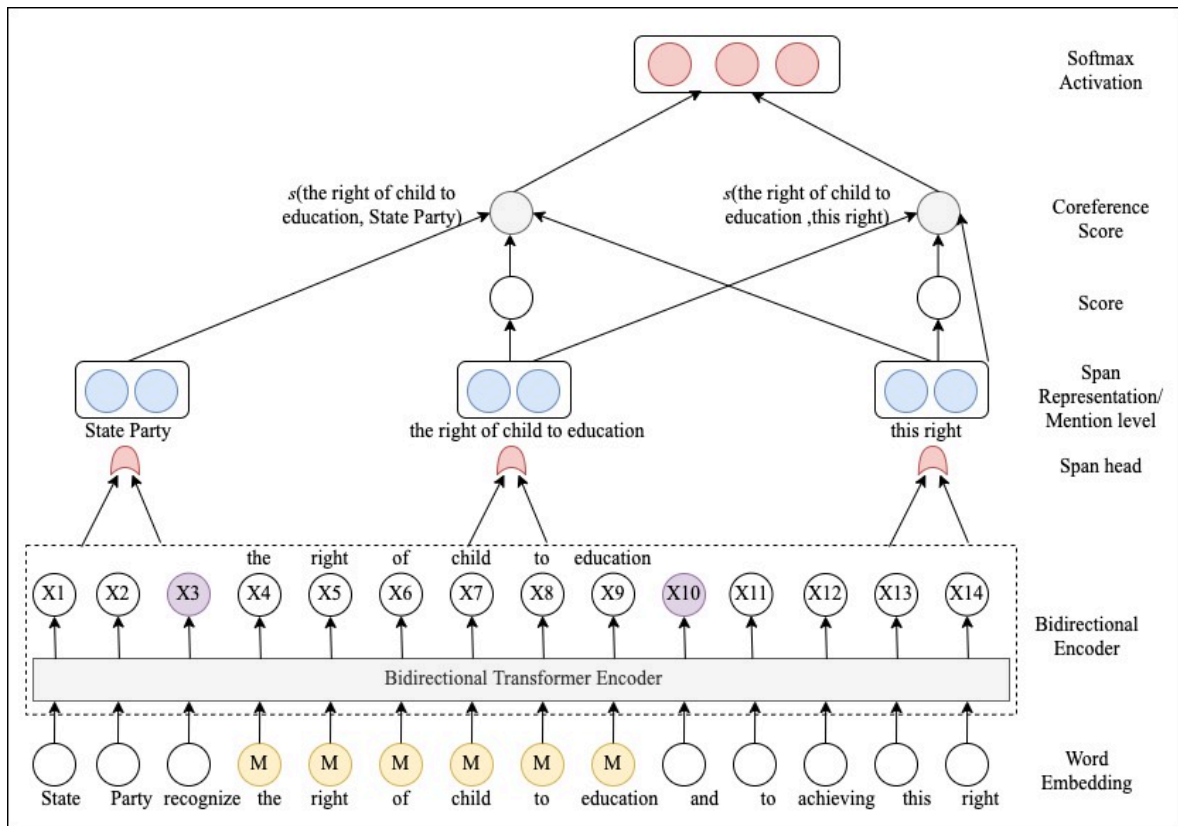
Figure 3.8 Process in the Coref-SpanBERT-Large Model

3.5 Abstract Meaning Representation

Abstract Meaning Representation was completed to achieve the relationship between coreference. The preprocessed texts and SpanBERT outputs were fed into the AMR. As mentioned above, output from the coreference resolution was fed into Abstract Meaning Representation (AMR). Figure 3.9 shows the process of receiving the sentence from coreference resolution to "model_parse_t5_model_stog" for AMR conversion.
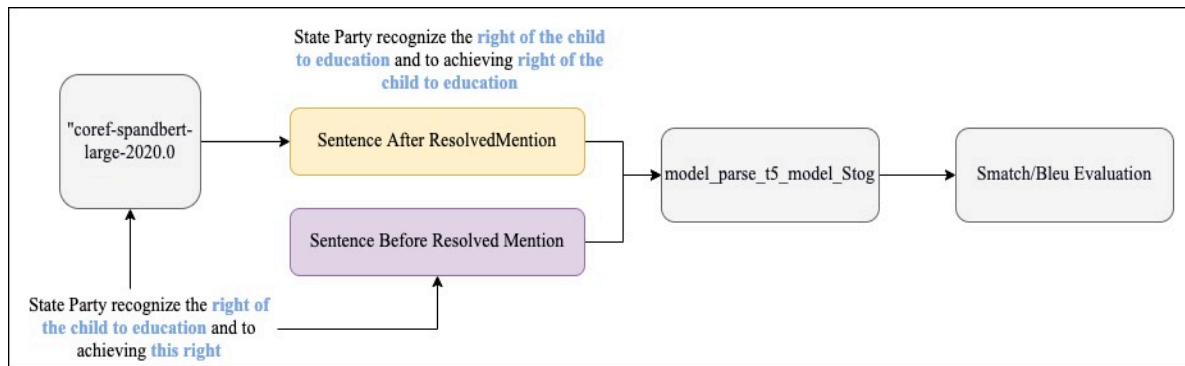
Figure 3.9 Flow Input into the AMR Model

Abstract Meaning Representation was the semantic representation that represented the relations and concepts between sentences. The AMR was created in the sentence level to see the relationship between concepts. Figure 3.11 shows the example of text in datasets that were mapped into the AMR structure. For example, to resolve the coreference, the sentence "a child means every human being below the age of eighteen years" was put into coreference resolution. Then, the AMR converted the outputs to the graph. After the sentence was mapped, the core relation represented the main actions of the statement. In addition, the AMR outputs were kept in the link list structure where the nodes' head was pointed to *human*, which was the main subject of this statement. The sentence structure was represented in a graph. Figure 3.11 represents the structure of the AMR.

The experiment was set up based on the pre-trained model, and Seq2Seq model called "model_parse_t5-v0_1__0". The Transformer concepts were used for building and training the Sentence to Graph algorithm known as Stog format. The Transformer received input in the form of "input_ids", "attention_mask", "target_ids", and "target_attention_mask". The "input_ids" was a raw text to be processed by the Seq2Seq model. The Text-to-Text Transfer Transformer model was used by the model_t5 to generate the AMR graphs. The essential component of the Transformer was called self-Attention. The Transformer, which consisted of encoder and decoder, was initiated for the sequence-to-sequence processing task.

The T5 model took the Transformer implementation since it was closest to the original model. Firstly, the input of the sequence was mapped into the embedding layers which comprised of

the self-Attention layer and the feed-forward neural network [31]. The normalization layer was embedded inside the input of self-Attention and feed-forward neural network. The model took a simplified normalization layer by activating only rescaled. Secondly, the data was transferred to the residual skip connection layers, where component of encoder layer input was added to its output. The dropout was applied to feed-forward neural networks, Attention weights, skip connection, and input-output stack. At this stage, the decoder layer was identical to the encoder layer. The only difference was that the decoder layer contained the standard Attention mechanism. Unlike encoder layer, decoder layer used the self-Attention mechanism which used the Auto Autoregressive function to present the last output. In the last stage, the final decoder was passed to the dense layers together with the Softmax Activation function. The weight of dense layers was shared between the inputs and the embedding which was a scalar scale for computing the Attention weight. The position embedding parameter was shared to all layers of the model. Figure 3.10 demonstrates the T5 model for creating AMR graphs.

This model had 32 embedding layers with an increase in logarithmic size to 128 to assign the relative position to the same embedding. The insensitive layers were related to a relative position beyond 128 tokens. The "text-to-text" format was used by the model to gain input for contextualizing and conditioning purposes. The model was forced to produce some output texts. This enhanced the model to provide consistency in the pre-training and fine-tuning process.

The standard encoder-decoder structure was used since it generated a significant result in forming classification tasks. The encoder and decoder of Attention mechanism had 12 clocks of self-Attention and a feed-forward neural network. The dense feed-forward layer contained 3,072 dimensionality with the ReLU Nonlinearity. The Attention mechanism had 64 dimensionality. The model had 12 head Attention mechanisms. The summation of dimensionality between sub-layers and embedding layers was 768. To summarize, there were 220 million parameters while the dropout rate was 10%. The model was trained with 524,288 steps. The sequences of 65,536 tokens were packed in the batch. The number of batch sizes and steps in the pre-training process were 34 billion tokens.
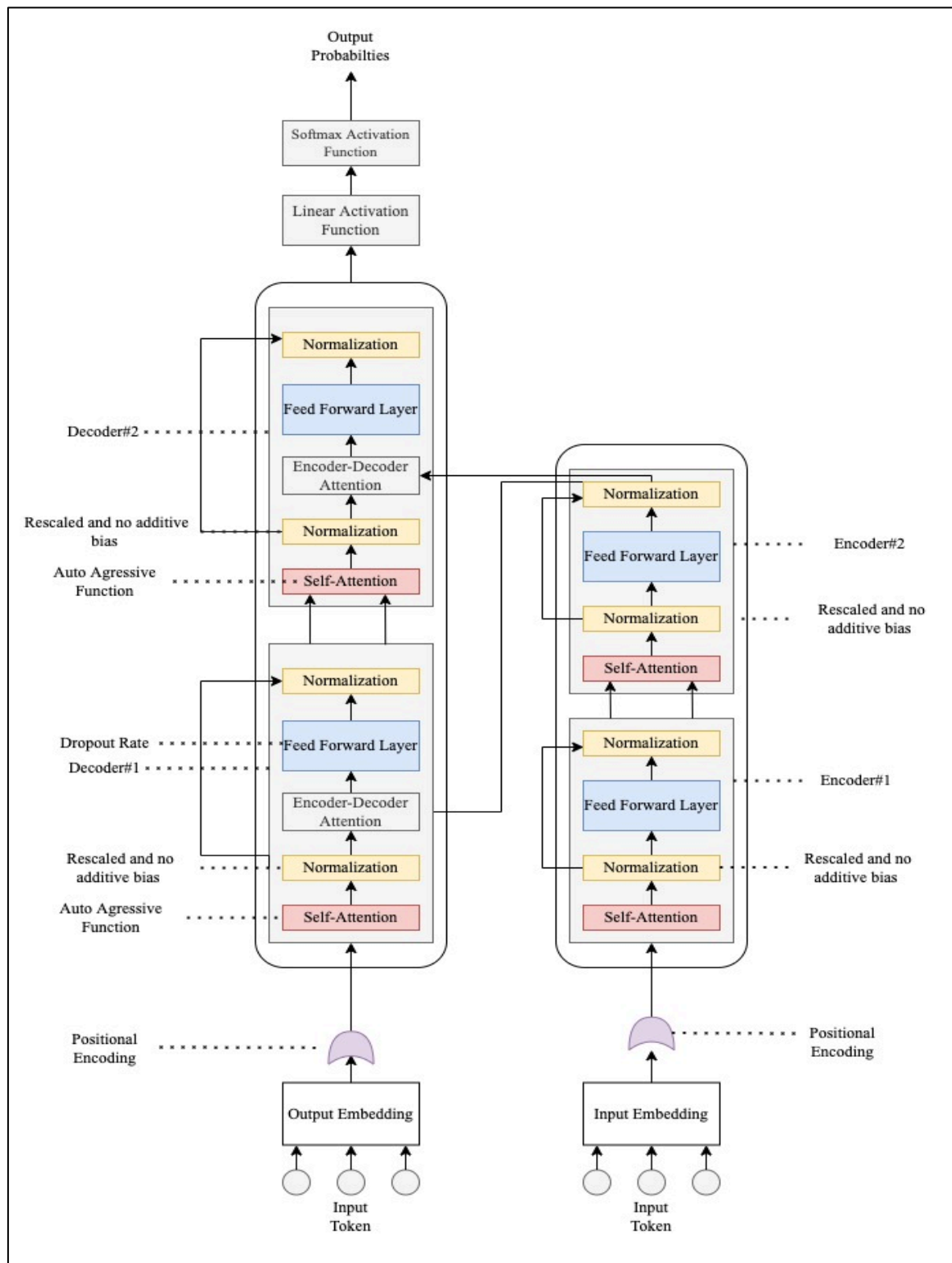
Figure 3.10 T5 Model Processing AMR Graphs
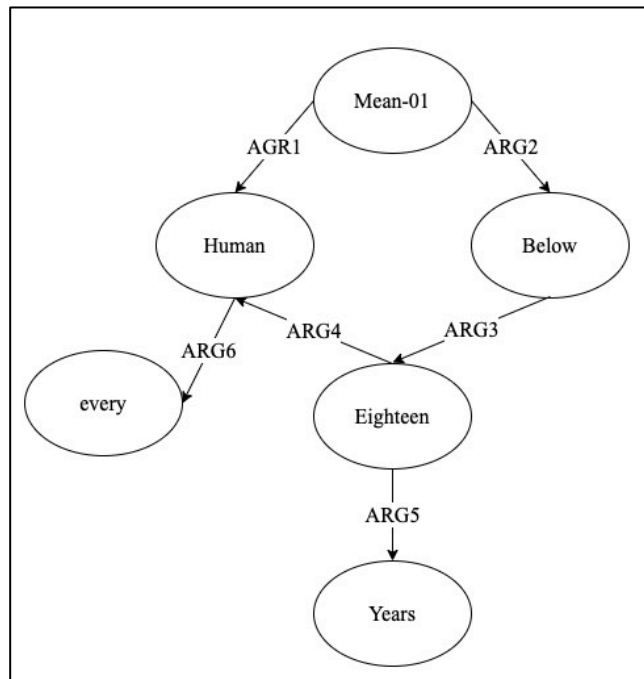
Figure 3.11 Sample AMR sentence

The core methods of test preprocessing, coreference resolution, Part of Speech tagging, and Abstract Meaning Representation were applied in the thesis. The test case for experimentation was created for the best result in these datasets. There were two main test cases which were test for SpanBERT, and Abstract Meaning Representation. The first hypothesis was to test whether SpanBERT was an effective tool to solve the coreference resolution. It also aimed to test the large pre-trained model of SpanBERT by having "coref-soandbert-large-2020.0" as the trained corpus. The large pre-trained corpus often delivered better results. However, this technique was rarely trained in legislative vocabulary. In the second hypothesis, the Abstract Meaning Representation was tested to check whether the T5 model could convert and capture semantic meaning of the legislative corpus.

The output of AMR structure is displayed on Figure 3.11 and Figure 3.12. The ARG showed the relation between nodes. As shown on the Figures, "undertake" was the core relation of this sentence while "child" was an antecedent of the coreference. Both Figures showed that "child" should be ensured with protection and well-being. Figure 3.13 is the expected result when the entity's action was resolved.

```
(Un /  undertake-01
      :ARG1(St/ state party-
      :ARG2(En/ ensure-02
      :ARG3(Ca/ care-03
              :ARG2(Ch/ child
              :ARG2(Pr/ protection
              :AGR3(We/ well-being
              :AGR1(Ta/  take-04
                      :ARG1(Ri/ right
                      :ARG4(Du/ duty
                                  :ARG4(Pa/ parent
                                  :ARG4(Gu/ guardian
                                  :ARG4(In/ individual
              :ARG1(Ap/ appropriate
              :AGR1(Le / legislative
              :ARG1(Ad/ administrative
              :ARG1(Me/ Measure)))))))))))))))))
```

Figure 3.12. AMR PENMAN Logic for the Output of Coreference
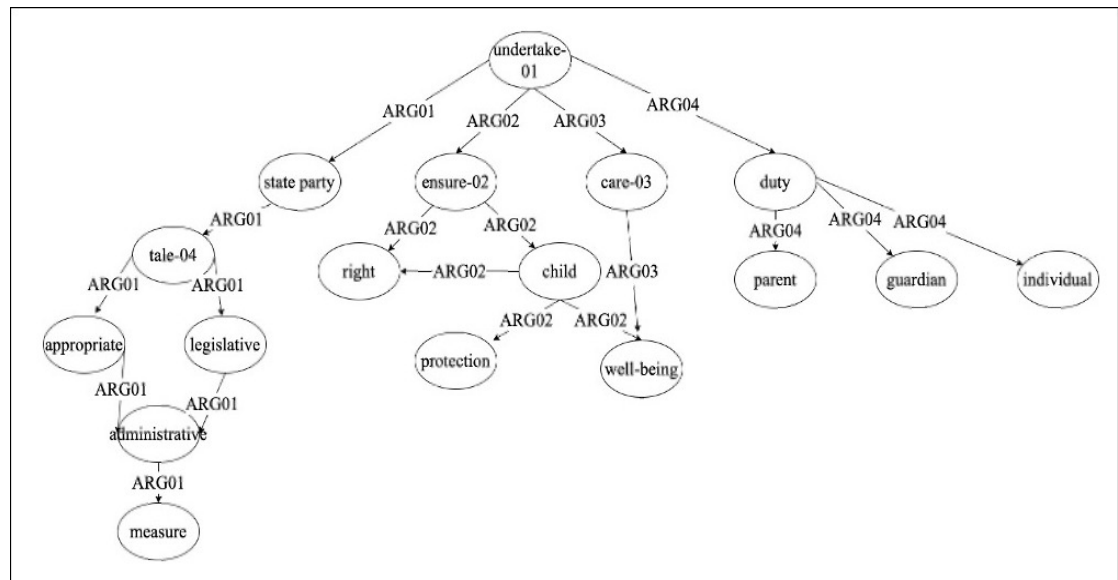
Figure 3.13. AMR Graph for the Output of Coreference

To summarize, the methodology and experiments setup were composed of five components including text-preprocessing, Part of Speech exploration, coreference resolution, AMR, and Smatch evaluation. Firstly, the gold datasets, which were created manually by the author, were used to evaluate the coreference resolution process. Secondly, the AMR evaluation outputs were evaluated by the Smatch, and BLEU technique. Both experiments were set up according to the requirements of each stage. Text-preprocessing required the raw datasets and Regular Expression to modify the set of strings. Then, the gold datasets for SpanBERT

evaluation were created on the preprocessed output. The SpanBERT algorithm completed the coreference resolution with the facilitation of the Pytorch and AllenNLP library. The evaluation of SpanBERT was completed on the generalization purpose by using gold datasets. Coreference resolution generated sentences with resolved coreference as outputs which were fed into Abstract Meaning Representation (AMR) for semantic representation. PENMAN logic and AmrLib were used for creating the AMR graphs. The Sentence to Graphs (StoG) function from the library was provided to generate the graph visualization. The AMR graphs was created by the Spacy extension, and AmrLib. In addition, AMR parsing was applied to unresolved coreference sentence for Smatch evaluation purpose. The output of each step affected the other algorithm performance, excluding POS tagging. For example, the SpanBERT and AMR were affected by the text preprocessing while AMR graphs and coreference resolution performance affected the Smatch evaluation. To check the meaning preservation and ambiguity reduction, Smatch and BLEU evaluation were conducted on the before and after resolved sentences. The meaning overlapping was compared by Smatch score on precision, recall, and F1. Moreover, BLEU scores were computed on resolved coreference AMR graphs to evaluate the human understanding level.

However, the Abstract Meaning Representation had some limitations. First, it could be used only with the grammatically correct sentence. To explain more, it was hard to apply the AMR with the spoken language. For example, "I will go work it out" was spoken language that bypassed the "and" conjunction. The full form of this sentence was "I will go and work it out." If the sentence in spoken language was being processed in the AMR algorithm and propbank tree, it was unable for the algorithm to distinguish the significant level of verbs. Second, the AMR could only be used in a sentence level, not a paragraph level. For instance, the semantic representation showed "who is doing what and to whom" in the single sentence while the coreference resolution simplified "who and whom" with antecedent in the legal corpus to reduce ambiguity. The simplified legal sentences were fed into AMR model individually to extract and represent the semantic meaning of "doing what" in the legal context. To conclude, the coreference resolution emphasized on the nouns and noun phrases by simplifying the "who, what, and whom" while the AMR converted legal sentences to graphs to represent the meaning of the sentences.

CHAPTER 4 RESULT AND DISCUSSION

4.1 Preprocessing Result

Convention on the Rights of the child (CRC) was the first dataset explored in this experiment. The dataset comprised of four main parts: introduction, part1, part2, and part3. The introduction part was excluded in this experiment because it contained the noun phrase without a full stop and had no implication towards the rest of the articles. The sample of CRC is depicted in Figure 3.2. Table 4.1 shows the benchmark of the datasets. The preprocessed work revealed the benchmark of the CRC. To explain, there were a total of 217 sentences after being preprocessed by the algorithm. The gold datasets (hand-tagged datasets) were developed. The CRC gold dataset, which was tagged from the output of preprocessed sentence, revealed a total of 108 sentences with mention (coreferences) and 109 sentences without mention. The number of mentioned pairs in the CRC gold dataset were 316.

The gold mentions, which were used for the evaluation of SpanBERT performance, were the datasets that was manually or automatically tagged before corrected by humans. In this experiment, the gold datasets and the preprocessed sentences were manually annotated to resolve mention pairs. The gold mentions were used to evaluate the SpanBERT output.

The Convention on the Rights of the Person with Disabilities (CRPD) was transformed and counted for benchmarks. The preamble part was excluded because this experiment concentrated on the article level of the text. Figure 3.3 depicts a sample of the CRPD. The CRPD dataset was unique from the other datasets because there was no division of the "Part level." In general, the CRPD was composed of the preamble (introduction), article level, sentence level, and sub-sentence level. The sub-sentence level was in the same pattern as mentioned in chapter 3. There was a total of 50 articles. Table 4.1 shows the benchmark of the CRPD. The total number of CRPD preprocessing outputs were 233 sentences. The gold CRPD dataset was tagged in the preprocessed stage: 99 sentences with mentions, 134 sentences without mention, and 115 mentioned pairs.

Convention and Protocol Relating to the Status of the Refugees (CPRSR) was the last dataset for this experiment. It was preprocessed according to the priorities written above. The characteristics of the CPRSR and CRC were similar. Figure 3.4 shows the sample of the CPRSR structure. The CPRSR comprised of introductory note, final Act, preamble, chapter level, sentence level, and sub-sentence level. In this experiment, three parts which were the introductory note, final act, and preamble part were excluded. There was a total of 46 articles in this convention. Table 4.1 demonstrates the benchmark of CPRSR. The preprocessing result showed that there was a total of 124 sentences. The gold CPRSR dataset, which was manually tagged, contained 84 sentences with mention and 40 without mention. There were 115 mention pairs according to CPRSR gold mentions.

Table 4.1 Benchmark of the Sample Datasets

Description	CRC	CRPD	CPRSR
Total Number of Sentences	217	233	124
Sentences With Gold mention	108	99	84
Sentences Without Gold Mention	109	134	40
Gold Mention	136	115	115

4.2   Part of Speech Tagging and Dependency Parsing Result

The Spacy was used in this experiment to detect and pre-screen the Part of Speech in the datasets. The Part of Speech exploration was completed to view the characteristic of the datasets. The Part of Speech tagging for three datasets, including CRC, CRPD, and CPSR, was completed. Seven types of POS, which were Noun, Verb, Conjunction, Determinator, Proper Noun, Punctuation, and Pronoun, were counted in this exploration.

Starting with the CRC, Figure 4.1 shows the bar chart of the Part of Speech tagging of the CRC, CRPD, CPRSR in a form of frequency. The result showed that this corpus had 1,887 tokens of noun, 519 tokens of proper noun, and 67 tokens of pronoun. In the CRPD, Figure 4.1 shows that the number of nouns were significantly high at 2,754 tokens while the number

of pronouns were at 48 tokens. Figure 4.1 shows that CPRSR shared a similar pattern with the other datasets, which were 1240 tokens of noun, and 92 tokens of pronoun.
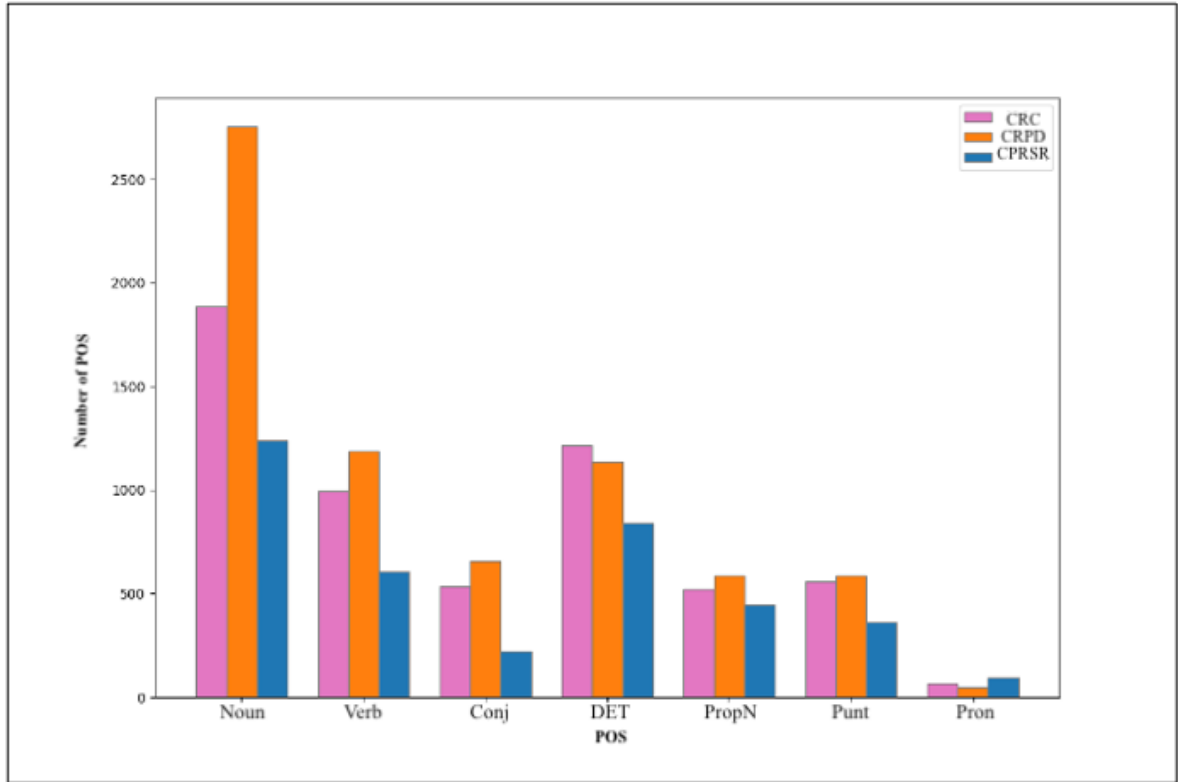
Figure 4.1 Part of Speech Tagging Exploration in CRC, CRPD, CPRSR

Dependency Parsing was conducted on three datasets with the facilitation of Spacy Library. The parsing images were converted to a table to comprehensively view the relationship among tokens.

Firstly, Table 4.2, Table 4.3, and Table 4.4 show the parsing of the CRC, CRPD, and CPRSR respectively. To explain, Table 4.2 shows the dependency of "pobj", which was the root of dependency, and the object of a preposition. The "nsubj" was the syntactic object of the noun clauses whereas the "mean" was the root dependency of the noun chuck named "a child." The "nsubjpass" was the passive nominal subject, and the "dobj" was the direct object of the verb in the noun phrase form. The result of dependency parsing was shown in Table 4.3. The result showed that there was the "dobj" relationship between "the full and equal enjoyment" and "respect.". The "conj" represented the conjunction which showed the relationship

between the two tokens that were connected together. The conjunction examples were "and", and "or." The noun chunk, the root of the tokens, and the relationship between the tokens were shown in the dependency parsing. However, when mapping dependency parsing with the AMR, both objectives were different. Therefore, dependency parsing was not used with AMR experiments.

Table 4.2 Sample of Dependency Parsing in CRC

Convention on the Right of the Child			
For the purposes of the present Convention, a child means every human being below the age of eighteen years unless under the law applicable to the child, majority is attained earlier.			
Chunk	**Root**	**Dependency**	**Root Dependency**
the purposes	purposes	pobj	For
the present Convention	Convention	pobj	Of
a child	child	nsubj	Means
the age	age	pobj	Below
eighteen years	years	pobj	Of
the law	law	pobj	Under
the child	child	pobj	To
majority	majority	nsubjpass	Attained

Table 4.3 Sample of Dependency Parsing in CRPD

Convention on the Rights of the Person with Disabilities			
The purpose of the present Convention is to promote, protect and ensure the full and equal enjoyment of all human rights and fundamental freedoms by all persons with disabilities, and to promote respect for their inherent dignity.			
Chunk	**Root**	**Dependency**	**Root Dependency**
The purpose	purpose	nsubj	Is
the present Convention	Convention	pobj	Of
the full and equal enjoyment	enjoyment	dobj	Ensure
all human rights	rights	pobj	Of
fundamental freedoms	freedoms	conj	Rights
all persons	persons	pobj	By
disabilities	disabilities	pobj	With
respect	respect	dobj	Promote

Chunk	Root	Dependency	Root Dependency
their inherent dignity	dignity	pobj	For

Table 4.4 Sample of Dependency Parsing in CPRSR

Convention and Protocol Relating to the Status of the Refugees			
The emphasis of this definition is on the protection of persons from political or other forms of persecution.			
Chunk	**Root**	**Dependency**	**Root Dependency**
The emphasis	emphasis	nsubj	Is
this definition	definition	pobj	Of
the protection	protection	pobj	On
persons	persons	pobj	Of
political or other forms	forms	pobj	From
persecution	persecution	pobj	Of

4.3   Coreference Resolution from SpanBERT Output

In this section, the coreference resolution experiment was conducted on three natural datasets, which were CRC, CRPD, and CPRSR. The outputs of the SpanBERT, including the sample of outputs, and marked mentioned in terms of precision, recall, and F1 scores were described,.

To simplify the corpus, SpanBERT was used as a tool to resolve the mention pairs in the sentences and to replace the proforms indexes with the antecedent. The outputs from SpanBERT in CRC, CRPD, and CPRSR are shown in Table 4.5, Table 4.6, and Table 4.7 respectively. The preprocessed sentences of CRC dataset were fed into the SpanBERT to gain the outputs including resolved sentences, detected mentions, and number of detected mentions. To illustrate, "States Parties recognized the right of the child to education, and with a view to achieving this right progressively and on the basis of equal opportunity, they should, in particular Encourage the development of different forms of secondary education, including general and vocational education, make them available and accessible to every child, and take appropriate measures such as the introduction of free education and offering financial assistance in case of needed" was fed to SpanBERT, and the detected mentions were shown below:

- "The right of the child to education" was antecedent, and "this right" was proform,

- "State Parties" was antecedent, and "they" was proform,
- "Different form of the secondary education including general and vocational education" was antecedent, and "Them" was proform.

The SpanBERT model extracted the mention pairs from the preprocessed sentences. Then, the outputs were consolidated into CSV files for manual verification. The coreference pairs of SpanBERT were manually counted and verified. Detected mention pairs and correctly detected mention pairs were counted and summarized for further evaluation tasks.

In the second legislative natural dataset, the CRPD output from SpanBERT was shown in Table 4.6 and it also shows the sample sentence before and after the resolution process. Sample sentence consisted of two mentioned pairs, one was "The committee" and "the specialized agencies and other competent bodies" as antecedents and the other was "its" and "their" as proforms. After recognizing the index of coreference pairs, the SpanBERT model replaced the index of proform with the antecedent for simplification.

The last experiment on coreference resolution was set up on the CPRSR. The preprocessing outputs were fed into the SpanBERT. The sentence sample before and after resolution process, mention pairs and number of mentioned pairs are shown on Table 4.7. SpanBERT detected three mention pairs which were "The contracting state", "their territory", and "travel documents for travel outside their territory" as antecedents, and "their", "their", "such document" as proforms. The SpanBERT replaced the proform with the antecedent.

Table 4.5 Sample of SpanBERT Output in CRC

Convention on the Right of the Child	
Sentence before Coreference Resolution	States Parties recognize the right of the child to education, and with a view to achieving this right progressively and on the basis of equal opportunity, they shall, in particular Encourage the development of different forms of secondary education, including general and vocational education, make them available and accessible to every child, and take appropriate measures such as the introduction of free education and offering financial assistance in case of needed.

Convention on the Right of the Child	
Sentence after Coreference Resolution from SpanBERT	States Parties recognize the right of the child to education, and with a view to achieving the right of the child to education progressively and on the basis of equal opportunity, States Parties shall, in particular Encourage the development of different forms of secondary education, including general and vocational education, make different forms of secondary education, including general and vocational education available and accessible to every child, and take appropriate measures such as the introduction of free education and offering financial assistance in case of need.
Mention detected by SpanBERT in the sentences	[['the', 'right', 'of', 'the', 'child', 'to', 'education', 'this', 'right'], ['States', 'Parties', 'they'], ['different', 'forms', 'of', 'secondary', 'education', ',', 'including', 'general', 'and', 'vocational', 'education', 'them']]
Number of Mentioned Detected	3

Table 4.6 Sample of SpanBERT Output in CRPD

Convention on the Right of Person with Disabilities	
Sentence before Coreference Resolution	The Committee may invite the specialized agencies and other competent bodies as it may consider appropriate to provide expert advice on the implementation of the Convention in areas falling within the scope of their respective mandates
Sentence after Coreference Resolution from SpanBERT	The Committee may invite the specialized agencies and other competent bodies as The Committee may consider appropriate to provide expert advice on the implementation of the Convention in areas falling within the scope of the specialized agencies and other competent bodies's respective mandates
Mention detected by SpanBERT in the sentences	[['The', 'Committee', 'it'], ['the', 'specialized', 'agencies', 'and', 'other', 'competent', 'bodies', 'their']]
Number of Mentioned Detected	2

Table 4.7 Sample of SpanBERT Output in CPRSR

Convention and Protocol Relating to the Status of the Refugees	
Sentence before Coreference Resolution	The Contracting States shall issue to refugees lawfully staying in their territory travel documents for the purpose of travel outside their territory, unless compelling reasons of national security or public order otherwise require, and the provisions of the Schedule to this Convention shall apply with respect to such documents
Sentence after Coreference Resolution from SpanBERT	The Contracting States shall issue to refugees lawfully staying in The Contracting States's territory travel documents for the purpose of travel outside their territory, unless compelling reasons of national security or public order otherwise require, and the provisions of the Schedule to this Convention shall apply with respect to travel documents for the purpose of travel outside their territory
Mention detected by SpanBERT in the sentences	[['The', 'Contracting', 'States', 'their', 'their'], ['their', 'territory', 'their', 'territory'], ['travel', 'documents', 'for', 'the', 'purpose', 'of', 'travel', 'outside', 'their', 'territory', 'such', 'documents']]
Number of Mentioned Detected	3

The Table 4.8 shows a number of mentions detected by SpanBERT and the gold mentions of each dataset. The experiment showed that the SpanBERT detected 147 mentions in CRC, 149 mentions in CRPD, and 122 mentions in CPRSR. While correctly detected mentions in CRC, CRPD and CPRSR were 115, 108 and 92 respectively, gold mentions were 136, 115 and 115.

Table 4.8 Mention Detected in Three Datasets

Experiment Result Description	CRC	CRPD	CPRSR
Number of SpanBERT Mentions	147	149	122
Number of SpanBERT Correct Mentions	115	108	92
Number of SpanBERT Detected and Resolved Mentions	102	103	82
Number of Gold Mentions	136	115	115

The Table 4.9 shows total sentences in four different groups of result: sentences with and without gold mentions, and sentences with and without SpanBERT mentions. There was a total of 217, 233, and 124 sentences in CRC, CRPD, and CPRSR respectively. It could be seen from Table 4.9 that the total number of sentences with gold mentions and sentences with SpanBERT were slightly different, 108 on gold mentions and 106 on SpanBERT mentions. In CRPD, there were 99 sentences with gold mention, and 112 sentences with SpanBERT mention. In CPRSR, there were 84 sentences with gold mention, and 74 sentences with SpanBERT mention.

For calculating the CRC, CRPD, and CPRSR on precision, recall, and F1, the formulas from resolution in legal text were used. The formula used to calculate are listed below:

$$Precission = \frac{\#correctly\ detected\ metnions}{\#detected\ mentions} \qquad (4.1)$$

$$Recall = \frac{\#correctly\ detected\ mentions}{\#gold\ mentions} \qquad (4.2)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (4.3)$$

Table 4.9 Sentences with Detected Mention

Experiment Result Description	CRC	CRPD	CPRSR
Sentences With Gold Mention	108	99	84
Sentences Without Gold Mention	109	134	40
Sentences With SpanBERT Mention	106	112	74
Sentences Without SpanBERT Mention	111	121	50
Total Sentences	217	233	124

After investigating coreference resolution, precision, recall, and F1, SpanBERT scores are shown on Table 4.10. The precision score in CRC, CRPD, CPRSR were 0.782, 0.725, 0.754 respectively. Comparing the precision scores with the recall scores, the recall scores were significantly higher which meant that the model classified and marked the coreferences correctly. SpanBERT proved that it seized the coreferences pairs (mentions pairs) in an acceptable criterion. However, the precision scores were lower than the recall scores because detected mentions were divided by the total number of predicted mentions.

Table 4.10 Precision Recall and F1 Score in Three Datasets for Detected Mentions

Dataset	Precision	Recall	F1
Convention on the Rights of the Child	0.782	0.846	0.813
Convention on the Rights of the Person with Disabilities	0.725	0.939	0.818
Convention and Protocol Relating to the Status of the Refugees	0.754	0.800	0.776

SpanBERT had a disadvantage which was it could not resolve the mention pairs inside the antecedent. As shown on Table 4.7, SpanBERT could not detect the mention pairs in this long antecedent, "['travel', 'documents', 'for', 'the', 'purpose', 'of', 'travel', 'outside', 'their', 'territory', 'such', 'documents']." It was noticeable that there were more one than coreferences in this antecedent. In this case, the "travel documents" and "their territory" were two antecedents. This was the exploitation of the SpanBERT algorithm.

$$Precision\ on\ End\ to\ End\ = \frac{\#correctly\ detected\ and\ resolve\ metnions}{\#detected\ mentions} \quad (4.5)$$

$$Recall\ on\ End\ to\ End\ = \frac{\#correctly\ detected\ and\ resolve\ mentions}{\#gold\ mentions} \quad (4.4)$$

$$F1\ on\ End\ to\ End\ = \frac{2*Precision\ on\ End\ to\ End\ *Recall\ on\ End\ to\ End}{Precision\ on\ End\ to\ End + Recall\ on\ End\ to\ End} \quad (4.6)$$

Table 4.11 Precision Recall and F1 Score on Three Datasets for End-to-End Performance

Dataset	Precision	Recall	F1
Convention on the Rights of the Child	0.694	0.750	0.721
Convention on the Rights of the Person with Disabilities	0.691	0.896	0.780
Convention and Protocol Relating to the Status of the Refugees	0.672	0.713	0.692

Table 4.11 shows the End-to-End performance of the coreference resolution in three datasets. The F1 scores of CRC, CRPD, and CPRSR were 0.721. 0.780, and 0.692 respectively. The equations of precision, recall, and F1 on End-to-End performance score are shown above. F1 scores on every dataset in End-to-End performance dropped significantly compared to the F1 score on detected mentions. CRPD dataset achieved the highest F1 score on End-to-End

performance. On the other hand, the CPRSR dataset achieved the lowest score among all datasets. The main reason that affected the performance of SpanBERT was the complexity of the datasets. The CPRSR was a complex and ambiguous dataset; as a result, the SpanBERT did not solve the coreference. This issue was expected to happen due to the SpanBERT algorithm that comprised of Mask Language Model and Span Boundary Objectives. However, when dealing with the compound antecedent with another coreference pair, the SpanBERT was more likely fail to detect and resolve compound coreference. The neural network only ran through the first layer of the coreference pair, so it hardly detected the coreference inside the antecedent.

Automated resolution in legal text work used four-step framework for the coreference resolution, and achieved 61.01% on F1 score in Japanese National Pension corpus. With the deep learning advancement, the coreference resolution work from SpanBERT achieved the F1 score of 0.812 in CRC, 0.818 in CRPD, and 0.776 in CPRSR. The result proved that the SpanBERT resolved and replaced the coreference with a significant F1 score over 77%. The SpanBERT pre-trained model, "coref-spandbert-large-2020.0" trained by the OntoNotes dataset, was another proposed methodology for coreference resolution.

4.4   Abstract Meaning Representation, Smatch and BLEU Evaluation

Figure 4.3 shows AMR graphs after resolving the coreference. "In particular, where the person having financial responsibility for the child lives in a State different from that of the child, States Parties shall **promote** the accession to international agreements or the conclusion of such agreements, as well as the making of other appropriate arrangements" was the sample sentence. Figure 4.3 shows the logical semantic meaning of this sentence. The main verb of this sentence was "promote" while the meaning was "to promote the accession and conclude the agreements." Figure 4.2 shows the PENMAN logic format of the mentioned sentence. AMR graphs and PENMAN conversion were completed on the rest of three datasets.

51

```
(p / promote-02
      :ARG0 (s / state
                :ARG0-of (h / have-org-role-91
                          :ARG1 (p2 / party)))
      :ARG1 (a / and
                :op1 (o / or
                          :op1 (a2 / accession-02
                                    :ARG1 (a3 / agree-01
                                    :mod (ii / international)))
                :op2 (c / conclude-01
                          :ARG1 (a4 / agree-01
                                    :mod ii)))
                :op2 (m / make-01
                          :ARG1 (a5 / arrange-01
                                    :ARG1-of (a6 / appropriate-02)
                                    :mod (o2 / other))))
      :condition (l / live-01
                :ARG0 (p3 / person
                          :ARG0-of (r / responsible-01
                                    :ARG1 (c2 / child)
                                    :mod (f / finance)))
                :location (s2 / state
                          :ARG1-of (d / differ-02
                                    :ARG2 (s3 / state
                                                :poss c2)))
                :mod (p4 / particular)))
```

Figure 4.2 Sample of PENMAN Logic Format in CRC
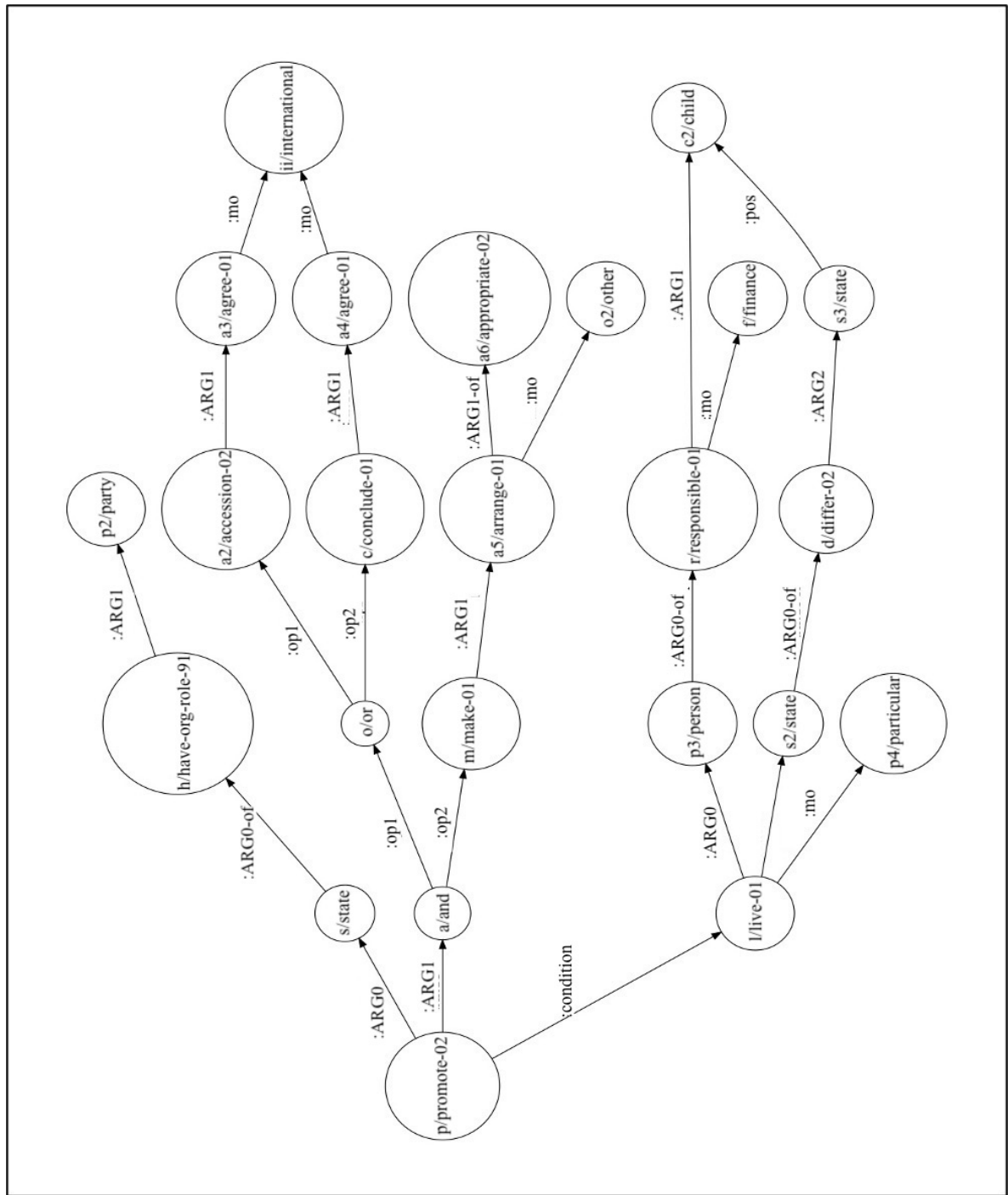
Figure 4.3 Sample of Abstract Meaning Representation in CRC

Table 4.12 Average CRC Smatch Score on 106 Sentences

Experiment	Precision	Recall	F1
Smatch	0.871	0.883	0.874
Unlabeled	0.882	0.895	0.886
No WSD	0.872	0.884	0.875
Non_sense_frames	0.963	0.939	0.949
Wikification	0.000	0.000	0.000
Named Ent	0.189	0.179	0.182
Negations	0.198	0.195	0.196
IgnoreVars	0.834	0.832	0.831
Concepts	0.947	0.938	0.941
Frames	0.958	0.934	0.943
Reentrancies	0.830	0.824	0.821
SRL	0.867	0.866	0.863

Table 4.13 Average CRPD Smatch Score on 112 Sentences

Experiment	Precision	Recall	F1
Smatch	0.849	0.874	0.855
Unlabeled	0.866	0.892	0.872
No WSD	0.850	0.876	0.857
Non_sense_frames	0.937	0.945	0.934
Wikification	0.000	0.000	0.000
Named Ent	0.147	0.140	0.141
Negations	0.241	0.234	0.237
IgnoreVars	0.802	0.817	0.805
Concepts	0.920	0.934	0.921
Frames	0.930	0.937	0.926
Reentrancies	0.706	0.730	0.712
SRL	0.842	0.871	0.849

Table 4.14 Average CPRSR Smatch Score on 74 Sentences

Experiment	Precision	Recall	F1
Smatch	0.876	0.895	0.883
Unlabeled	0.889	0.909	0.896
No WSD	0.878	0.897	0.885
Non_sense_frames	0.915	0.901	0.905
Wikification	0.000	0.000	0.000
Named Ent	0.270	0.270	0.268
Negations	0.318	0.321	0.318
IgnoreVars	0.834	0.837	0.834
Concepts	0.953	0.938	0.944
Frames	0.899	0.886	0.890
Reentrancies	0.643	0.654	0.645
SRL	0.772	0.792	0.778

Table 4.15 Average BLEU Score Among Three Datasets

Result Description	CRC	CRPD	CPRSR
Number of Sentence on BLEU Score	106	112	74
BLEU Score	0.876	0.879	0.856

The Smatch and BLEU evaluation were applied to evaluate the AMR graphs from three datasets. Smatch score revealed overlapped meaning when the sentences were converted to the graph structure, as shown in Table 4.12, Table 4.13, and Table 4.14. The average F1 scores of Smatch on, CRC, CRPD and CPRSR were 0.874, 0.855 and 0.883 respectively. The name entity's experiment revealed that there were some name entities in the corpus. For example, the United Nations was the common entity among three corpora. The corpus that received the highest name entity's score was CPRSR with an F1 score of 0.318. In addition, the negation experiments showed that among three legal corpora, there was the polarity in the sentences; however, the corpus that received the highest score with negation experiments was CPRSR with 0.318 F1 scores. This indicated that the CPRSR had the most negative polarity among those three corpora. In addition, the Wikification experiment showed 0.000 on F-score because there was no Wikification role among the three datasets.

Moreover, the BLEU evaluation was conducted to evaluate the level of human understanding on the sentences after the Abstract Meaning Representation and coreference resolution

process. The result of BLEU is shown in Table 4.15. The CPRD obtained the highest BLEU score at 0.879 while the remaining datasets obtained more than 0.800. This indicated that these sentences could be understood by human.

Table 4.16 Nodes Comparison After the Coreference Resolution

Node Description	CRC	CRPD	CPRSR
Decreased Node Sentences	30	29	13
Increased Node Sentences	35	42	26
Same number of Node Sentences	41	41	35
Total sentences	106	112	74

After the coreference resolution process, the complexity and ambiguity of the sentences were resolved and reduced to the minimum level. There were changes of AMR nodes in the process of replacing the proform indexes with antecedent. Table 4.16 shows the total number of nodes after the coreference resolution. To summarize, most of the nodes in CRC, and CRPSR remained stable at 41 and 35 sentences respectively. On the other hand, there was a slight increase in nodes in CRPD, which was 42 increased node sentences compared to 41 same number of node sentences.

After the investigation, the number of nodes decreased because the tokens in an antecedent were more than the number of proform tokens. The sample sentence was "States Parties shall ensure the implementation of these rights by their national law and their obligations under the relevant international instruments in this field, in particular where the child would otherwise be stateless." The proform "their" was replaced by the antecedent "State Parties." This event reduced the nodes in the AMR graphs. In addition, some sentences had only one short coreference pairs that caused AMR nodes reduction. On the other hand, the number of nodes increased because the compound set of an antecedent was significantly larger than the proform indexes. The sample sentence was "The Committee might invite the specialized agencies, the United Nations Children's Fund and other competent bodies as it may consider appropriate to provide expert advice on the implementation of the Convention in areas falling

within the scope of their respective mandates." The SpanBERT detected that the antecedent was dramatically long. The antecedents were "the committee" and "the specialized agencies, the United Nations Children's Fund and other competent bodies." The proform mentions were "it" and "their". It was apparent that the length of the antecedent was excessively longer than a proform. Moreover, the node's consistency occurred due to two possible reasons: false negative of SpanBERT, or the same length of antecedent and proform. The false-negative of SpanBERT occurred due to the incorrect placement of mentions or incorrectly detected mentions. Identical mention length happened when more than one mention in the specific sentence had the same length. The average length of antecedent and proform replacement caused the node of the AMR to be consistent.

The Smatch evaluation, BLEU evaluation and the AMR node changes indicated three findings. Firslty, the simplified corpus with human understanding from resolving mention were obtained from coreference resolution. Secondly, the coreference resolution could cause changes in AMR nodes. Thirdly, logical meanings were preserved from resolution. In addition, the coreference resolution could shorten or lengthen the sentences, depending on the complexity of the coreference pairs. Although some of the AMR nodes increased, the semantic sense of the sentences was simplified. The changes in nodes could have positive or negative impacts on machine processing.

The result in chapter 4 showed that the AMR could find and illustrate the meaning representation on "who is doing what" in the legal corpus. The AMR was beneficial to this thesis because it proved the generalization of meaning representation in real-world datasets. This parser facilitated the thesis to preserve the semantic meaning in the corpus, especially where the semantic meaning was crucial, such as the legal processing corpus. The coreference resolution and AMR integrated algorithm could simplify the noun, and represent the meaning in the legal corpus. Term of Used (TOU), the Memorandum of Understanding (MOU), Term of Reference (TOR), and other Legal Agreement translations from the English Language to the Thai Language were sensitive to remain the context without "lost in translation" phenomenon. The AMR parser was a potential method for legal machine

translation to help outline the semantic structure of the sentences and keep the meaning of the original message as much as possible.

CHAPTER 5 CONCLUSION

This thesis aims to resolve coreferences phenomenon and to extract meaning representation in three legislative corpora including Convention on the Rights of the Child, Convention on the Right of Person with Disabilities, and Convention and Protocol Relating to the Status of the Refugees. The SpanBERT exploited, detected, resolved the coreference, and replaced antecedent tokens with the proform index. In addition, the meaning representation, or known as the entity's action, was illustrated using the Abstract Meaning Representation method. The Smatch and BLEU evaluations were applied to check overlapping meaning between two sentences, one before resolving coreferences and the other after resolving coreferences. One of the major findings was that the number of AMR nodes changed according to the complexity of coreferences after the process of coreference resolution.

The Part of Speech tagging revealed the significant number of nouns and other types of speech in three legislative corpora. However, the number of pronouns, a simple coreference phenomenon, were deficient, which implied that these three legislative corpora had a deeper level. Moreover, it was shown in the experiments that the dependency parsing, and Abstract Meaning Representations objectives were different. The alignment between AMR and dependency parsing was exceptionally challenging. With the time and resources limitation, this research adopted the AMR method alone to create semantic parsing.

This thesis proposed the methodology called SpanBERT to resolute coreferences in legal engineering domain. The proposed methodology achieved over 69% on End-to-End performance in legal corpora. In addition, the Abstract Meaning Representation method was applied for semantic logical conversion. The Smatch and BLEU evaluations justified that the integration between SpanBERT and AMR could preserve the meaning of original sentences. It proved that AMR was another generic method for meaning representation which converted unstructured legal text into a semantic structure.

It was uncommon for researchers to apply the legal domain with NLP. Nevertheless, this thesis applied two techniques which were SpanBERT for coreference resolution and AMR for meaning representation to justify the application of NLP in the legal domain.

The further research should fine-tune SpanBERT and uses AMR node after complexity reduction for legal application. Firstly, the SpanBERT can be fine-tuned with SBO and MLM length to improve coreference resolution performance. Secondly, fewer complexity datasets from SpanBERT can be used for AMR applications, including the legal question-answering system, machine translation, and primarily legal inference

REFERENCES

1. Chomsky, N., 2006, **Language and Mind**, 3rd ed., Cambridge University Press, Cambridge, pp. 375-408.

2. Djeribia, A., 2017, **Chomsky's Generative Transformational Grammar and its Implications on Language Teaching**, Department of English, Hama Lakhdar Eloued University, pp. 213-220.

3. Sriparna, S. and Ekbal, A., 2013, "Combining Multiple Classifiers Using Vote Based Classifier Ensemble Technique for Named Entity Recognition", **Data & Knowledge Engineering,** 3-5 May 2013, London, pp. 15-39.

4. Jurafsky, D. and Martin, H. James, 2009, **An Introduction to Natural Language Processing and Computational Linguistics and Speech Recognition,** 2nd ed., Prentice-Hall, Inc., Upper Saddle River, pp. 415-441.

5. Oanh, T., Ngo, X. B., Nguyen, L. M. and Shimazu, A., 2014, "Automated Reference Resolution in Legal Texts**",** **Artificial Intelligence and Law**, Vol. 22, No. 1, pp 29-60, DOI: 10.1007/s10506-013-9149-8.

6. Refworld, 1989, **Convention of the Rights of the Child United Nation** [Online], Available: https://www.refworld.org/docid/3ae6b38f0.html [2021, April 17].

7. Refworld, 2007, **Convention on the Rights of Persons with Disabilities United Nations [**Online], Available:https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html. [2021, December 27].

8. Refworld, 2001, **United Nations High Commissioner for Refugees Convention and Protocol Relating to the Status of Refugees UNHCR** [Online]. Available: https://www.unhcr.org/protection/basic/3b66c2aa10/convention-protocol-relating-status-refugees.html. [2021, April 17].

9. Lassalle, E. and Denis, P., 2013, "Improving Pairwise Coreference Models Through Feature Space Hierarchy Learning", **ACL 2013 - Annual Meeting Of The Association For Computational Linguistics**, 4-9 August 2013, Sofia, pp. 497-506.

10. Delvin, J., Chang W. M., Lee, K. and Toutanova, K., 2019, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,** 2-7 June 2019**,** Minneapolis, pp. 4171-4186.

11. Joshi, M., Chen, D., Liu, Y., Weld, S. D., Zettlemoyer, L. and Levy, O., 2020, "SpanBERT: Improving Pre-training by Representing and Predicting Spans**",** **Transactions of the Association for Computational Linguistics**, Vol.8, No.14, pp. 64-77.

12. Flanigan, J., 2018, **Parsing and Generation for the Abstract Meaning Representation**, Degree of Doctor of Philosophy in Language and Information Technologies**,** Language and Technologies, School of Computer Science, Carnegie Mellon University, pp. 7-13.

13. Anikina, T., Koller, A., Roth, M., 2020, "Predicting Coreference in Abstract Meaning Representations", **Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference**, 8-13 December 2020, Barcelona (Online), pp. 33-38.

14. Lehmann, C., 2013, "The Nature Of Parts Of Speech", **STUF, Language Typology and Universals,** Vol.66, No.1, pp.1-35, DOI: 10.1524/stuf.2013.0008.

15. Jurafsky, D. and Martin, H. James, 2009, **An Introduction to Natural Language Processing and Computational Linguistics and Speech Recognition**, 2nd ed., Prentice-Hall, Inc., Upper Saddle River, NJ, pp. 280-304.

16. Spacy, 2016, **Spacy V3**, [Online], Available: https://spacy.io/ [2021, April 12].

17. Lee, S. Z., Tsuji, J. and Rim, H. C., 2000**,** "Part-of-Speech Tagging Based on Hidden Markov Model Assuming Joint Independence", **Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,** 3 October 2000, Hongkong, pp.263-269   DOI.10.3115/1075218.1075252.

18. Meng, W. S., Ng, H. T. and Lim, C. Y. D., 2001, "A Machine Learning Approach to Coreference Resolution of Noun Phrase", **Computational Linguistics**, Vol.4, No.1, pp. 521-544.

19. Banarescu, L. B., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N., 2013, "Abstract Meaning Representation for Sembanking", **Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse**, 8-9 August 2013, Sofia, pp. 178-186.

20. Cai, S. and Knight K., 2013, "Smatch: an Evaluation Metric for Semantic Feature Structures", **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,** 4-9 August 2013, Sofia, pp. 748-752.

21. Goodman, W. M., 2020, "Penman: An Open-Source Library and Tool for AMR Graphs", **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics**, 5-10 July 2020, Washington, pp. 312-319.

22. Amrlib, 2021, **Amrlib**, [Online], Available: https://amrlib.readthedocs.io/en/latest/index.html [2021, December 17].

23. Naseem, T., Abhishek, S., Hui, W., Radu, F., Salim, R. and Miguel, B., 2019, "Rewarding Smatch: Transition-Based AMR Parsing with Reinforcement Learning", **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, 28 July–2 August 2019, Florence, pp. 4586–4592, DOI: 10.18653/v1/P19-1451.

24. Damonte, M., Cohen, S.B. and Satta, G., 2017 "An incremental parser for abstract meaning representation", **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,** 16 September 2016, Valencia, **pp.536-546** , DOI: 10.18653/v1/E17-1051.

25. Garder, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M. and Zettlemover L., 2018, "AllenNLP: A Deep Semantic Natural Language Processing Platform", **Proceedings of Workshop for {NLP} Open Source Software**, 15-20 July 2018, Melbourne, pp. 1-6**.**

26. Henlein, A. and Mehler, A., 2020, "On the Influence of Coreference Resolution on Word Embeddings in Lexical-semantic Evaluation Tasks", **Proceedings of the 12th Language Resources and Evaluation Conference**, 11-16 May 2020, Marseille, pp. 27-339.

27. Vaswani, A., Shazeer, N., Parmar, N., Uszkereit, J., Jones, L., Gomez, N. A. and Kaiser, L., Polosukhin, I., 2017, "Attention Is All You Need", **NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems**, 4-9 December 2017, California, pp. 6000-6010.

28. Kishore, P., Salim, R., Todd, W. and Wei-Jing, Z., 2002, "BLEU: A Method for Automatic Evaluation of Machine Translation", **ACL 02: Proceeding of the 40th Annual Meeting on Association for Computational Linguistics**, 6-12 July 2002, Philadelphia, pp. 311–318, DOI: 10.3115/1073083.1073135.

29. Alfred, V.A., 1991, "Algorithms for finding patterns in strings", **Handbook of Theoretical Computer Science (Vol. A): Algorithms and Complexity",** Vol. A, No. 1, pp. 255-300.

30. Lee, K., He, L. and Zettlemoyer L., "Higher-order coreference resolution with coarse-to-fine inference," 2018, **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Association for Computational Linguistics, 1-6 June 2018, Louisiana, pp. 687-692, DOI: 10.18653/v1/N18-2108.

31. Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J, 2020, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1-67.

CURRICULUM VITAE

NAME	Mr. Surawat Pothong
DATE OF BIRTH	23 April 1998

EDUCATIONAL RECORD

HIGH SCHOOL

High School Graduation
General Education Development, MISSISIPI COMMUNITY COLLEGE BOARD (2014)

BACHELOR'S DEGREE

Bachelor of Science (Information Technology)
Assumption University of Thailand, 2019

MASTER'S DEGREE

Master of Engineering (Computer Engineering)
King Mongkut's University of Technology Thonburi, 2021

PUBLICATIONS

Pothong, S. and Facundes, M., 2021, "Coreference Resolution and Meaning Representation in a Legislative Corpus", 2021, **16th International Joint Symposium on The Artificial Intelligence and Natural Language Processing (iSAI-NLP)**, 21-23 December 2021, Ayutthaya (Online), pp. 1-6, DOI: 10.1109/iSAI-NLP54397.2021.9678168.