Table 4. List of Core Arguments on Propbank **[10]**

Tag	Description
ARG0	Agent, Operator
ARG1	Thing, Operated
ARG2	Explicit Patient
ARG4	Explicit Argument
ARG5	Explicit Instrument

Table 5. List of Adjunct Arguments on Propbank **[10]**

Tag	Description	Example
ARGM-LOC	Locative	The museum, in Westborough, Mass
ARGM-TMP	Temporal	Now, by next summer
ARGM-MNR	Manner	Heavily, clearly, at a rapid rate
ARGM-DIR	Direction	To market, to Bangkok
ARGM-CAU	Cause	In response to the ruling
ARGM-DIS	Discourse	For example, in part, Similarity
ARGM-EXT	Extent	At $38.375,50 points
ARGM-PRP	Purpose	To pay for the plant
ARGM-NEG	Negation	Not
ARGM-MOD	Modal	Can, might, should, will
ARGM-REC	Reciprocals	Each other
ARGM-PRD	Secondary Predication	To become a teacher
ARGM	Bare ARGM	With a police escort
ARGM-ADV	Adverbials	(none of the above)

2.1.3.2. English Translation of The Holy Quran

In its original Arabic, referring to [11], there are some uniqueness of the Quran sentence compared to the general sentence, among others:

1. Arabic Verbs.

In general, classical Arabic follows Verb-Subject-Object (VSO) order. The majority of Arabic verbs are trilateral, which can be derived to 15 different forms. Each derivation signifies some semantic variation over the original form.

2. The Quran Linguistic Style.

2.1. Literal vs Technical Sense Of Word

The Quran borrows an Arabic word and specializes it to indicate a technical term. For example, the word "jannah" meaning literally "a garden", but as a technical term in the Quran whenever this word is used to refer "the paradise".

2.2. Grammatical Shift

The Quran often draws the attention of the reader by shifting grammatical agreement is a statement. For example in verse [3:133], *"when you are in the ships and they sail with them with a fair breeze"*. The mode changed from "you" to "they" and "them" moving from the second person to the third person.

2.3. Verbs associated with different proposition

The Quran exhibits many examples where a certain verb is associated with a preposition which is unusual to the verb, but common with a different verb. For example the verb *"khala"* that means 'be alone'. This verb is usually followed by the preposition 'with' like 'John was alone with Mary'. However in verse [2:14], the Quran choose to use the preposition 'to', which sounds unusual to say, 'John was alone to Mary'.

2.4. Metaphors and Figurative

The Quran uses heavily metaphors and figurative. Verse [9:14] used the verb "shine", but the Arabic verb *ishtala* means "to flare" and shows the analogy of 'old age symptom by many gray hair' wit a 'fire burning a bush'.

2.5. Metonymy

In many verses the Quran uses metonymy. In [12:82] the Arabic verse literally means 'ask the town' which means (and was translated so) 'ask the people who live in the town'.

2.6. Imperative and non-Imperative

Arabic verbs are classified into the past, present and imperative. thus, in Arabic, the imperative structure can be understood from the type of the verb used.

The Quran English translation is a translation of the original Arabic Quran. Hence the composition of grammar and the sentence structure, English-Quran is still influenced by the original languages, namely Arabic. In this thesis, English Translation of The Holy Quran is used as the testing data. The data is compiled into an XML file with the same structure as the corpus PropBank.

2.1.4. PractNLPTools

Practical Natural Language Processing Tools for Humans or practNLPTools is a pythonic library over SENNA and Stanford Dependency Extractor [12]. This research proposes a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling.

This thesis uses practNLPTools to perform argument identification on Quran domain data. The argument identification process generates the Qur'an data that has been labeled semantically and in this study is referred as auto labeled data.

2.1.5. Data Preparation

One important task in text mining is in preparing the data. It is because of no existing structured text data, therefore it is necessary a process transform the data into a structured space-vector model. The necessary steps are generally known as data preprocessing process. Some data preprocessing steps commonly used are:

1. Selection: Decides of the text that will be processed (sentences, paragraphs, and so on).
2. Tokenization: create a token of a text sentence into discrete words.
3. Stopwords: deletion of the words that are considered unimportant or will affect the data processing such as; a, the, of, and others.
4. Stemming: elimination of prefixes and suffixes to change a word into its basic form.

In this thesis, the selected text as the data are sentences that are derived from PropBank and English Quran's translation. Furthermore, the sentence will be converted into a parse tree with the help of a parser for information extraction.

2.1.6. Features

The commonly NLP task is to label the words. As well as SRL aims at delivering a semantic role to a syntactic constituent of a sentence. Traditional NLP approach is by extract a set of manually designed features from a sentence which is then fed to a standard classification algorithm, such as the Support Vector Machine (SVM), often with a linear kernel. The choice of features is a completely empirical process, mainly based first on linguistic intuition, and then trial and error, and the feature selection is task dependent, implying additional research for each new NLP task [12]. Complex tasks such as SRL suppose a large number of potentially complicated features (e.g., taken from a parse tree).

2.1.6.1. SRL Basic Features

Features commonly used by SRL system are called SRL basic features introduced by Gildea and Jurafsky [6]. Basic features are a set of features that are used for labeling the semantic argument.

1. Predicate: Lemma predicate is used as a feature. For example in a sentence "The Lecturer went to classroom", the predicate of the sentence is: "went".
2. Path: Syntactic path passing through parse tree from parse constituent towards predicate classified. Figure 4 illustrates the tree with NP path NP↑S↓VP↑VBD. ↑ presenting the upward movement in the tree and ↓ presenting downward movement in the tree.

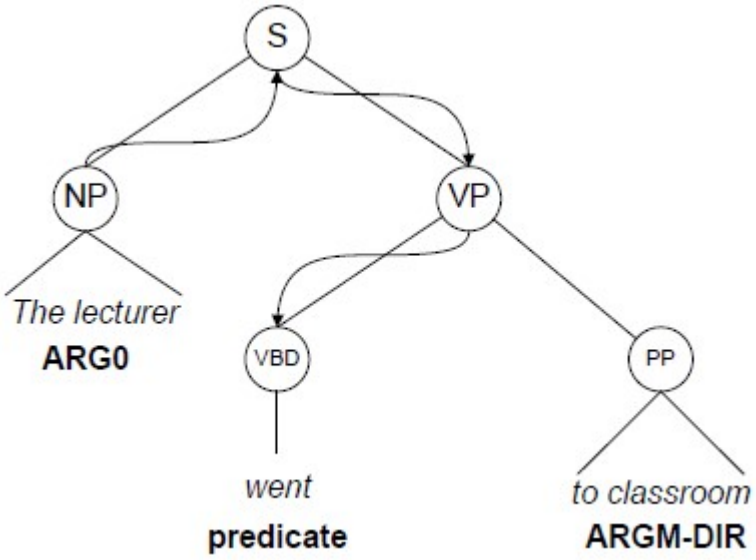

Figure 5. Illustration of Baseline Features **[6]**

3. Phrase Type: Syntax category of correspondence phrase is based on a semantic argument. Example: (NP, VP, S, etc.). Example: Phrase type of the phrase "The lecturer" of the sentence "The lecturer went to classroom" in figure 4 is NP.
4. Position: Features binary identification are based on the position of the phrase whether before or after the predicate. This feature is highly correlated with grammatical function because usually, the subject will appear before the verb or after the object. The Position usually represented in binary like 'L' or 'R' (left or right) or 'before' and 'after'.
5. Voice: The feature determines the predicate in a sentence whether active or passive predicate. The difference between active and passive verbs play an important role in the relationship between semantic roles and functions of grammar because the direct object of active verbs usually has a semantic relationship with the subject of passive verbs.
6. Head Word: The keyword of a phrase is calculated based on the table Head Word compiled by Magerman (1994) and modified by Collins (1999). Head words in a noun phrase can be used to specify the limits of choice of the semantic role.

7. Sub-Categorization: This feature is a phrase structure which expands the parent node of a predicate in a parse tree. For example in figure 4 illustration of Sub-Categorization features from predicate "went" is VP → VBD-PP.

Therefore, the basic features extracted from the sentence "The Lecturer went to classroom" are as follows:

Table 6. Examples of Baseline Features Extraction

Pr	**Vo**	**Sc**	**Pt**	**Hw**	**Pa**	**Po**	**Ar**
went	active	VP: VBD_PP	NP	lecturer	NP↑S↓VP↓VBD	before	ARG1
went	active	VP: VBD_PP	PP	to	PP↑VP↓VBD	after	ARG-DIR

Information :

Pr : Predicate
Vo : Voice
Sc : Sub-Categorization
Pt : Phrase Type
Hw : Head Word
Pa : Path
Po : Position
Ar : Argument

2.1.6.2. Additional Features

Throughout the study of the SRL system, some additional features have been developed after SRL basic features introduced by Gildea and Jurafsky. Pradhan et al. [13] use these basic features and designs some additional features, i.e. the part-of-speech tag of the headword, the predicted named entity class of the argument, features providing word sense disambiguation for the verb (they append 25 variants of 12 new feature types as a whole). They use the Propbank data that released on Feb 2004 adn SVM as the classier. It is close to the state-of-the-art in performance.

Xue et al [14] proposed some new additional features to perform an SRL task on by using the Propbank data released in April 2004, they tested the system with maximum entropy classifier and achieved very comparable result with [13].

In Toutanova et al [15] an SRL model over Propbank that effectively exploits the semantic argument frame as a joint structure, is presented. It incorporates strong dependencies within a comprehensive statistical joint model with a rich set of features over multiple argument phrases.

Yang et al. [16] propose some new features to improve SRL performance. The key idea of their work is to make a group of similar arguments activate one feature and another group of similar arguments activate another feature. The experiment conducted on Chinese and English Propbank.

This research uses some additional features that are expected to improve the performance of the classification. The additional features used in this thesis are:

1. Constituent Order

Constituent Order is related to the first/last word/POS in the constituent argument. However, this feature is designed for distinguishing arguments and non-arguments. [7] use a version of this feature. This features calculate the position of each constituent relative to the predicate, which is support the position of its proximity to the predicate [10].

2. Argument Order

Argument Order is introduced by [17], this feature is an integer that indicates the position of constituents in order of argument to the verbs. It is calculated after the initial phase constituents are classified as an argument or a non-argument. Because this feature does not use syntax parse tree, this can help create a strong semantic role labeling without being influenced the error parser [10].

3. Syntactic Frame

Syntactic Frame is introduced by [14]. This is a feature to complete the path and Sub-cat features. This feature refers syntactic predicates and NP as "pivots" and other elements are defined in relation to them. It describes a sequential pattern of noun phrases and predicate in a sentence. As example for constituent "classrooms" in Figure 2-1, the syntactic frame is np_v_NP, while "the lecturer" syntactic frame is

NP_v_np. The current constituent is expressed in capital letters on the syntactic frames produced, but can also be generalized to other terms such as CUR, X, etc that declare the constituent's position.

4. Noun Head of PP

Noun Head of PP is introduced by [13]. When the argument verb is a prepositional phrase (PP), head of the word is a preposition. This can often be a reliable indicator of the semantic role (e.g. in, across, and toward generally indicate the location), some prepositions can be used in various ways, and meaning can be determined by the object of the preposition in this case a noun [10]. For example, in March indicates time, while in Indonesia indicates location. So to figure 2-1, at "The lecturer" Noun Head PP produced is null because the lecturer is not a PP but NP. While for the head to classroom noun PP produced is "to".

5. First/Last Word/POS In Constituent

First/Last Word/POS In Constituent is introduced by [13]. This feature takes the first/last word/POS (Part-Of-Speech) in constituent no matter what the type is. This feature is obtained in from general way so that it is free from parser errors and applies to all types of its compilers.

Then the additional features obtained from the example sentence "The Lecturer went to classroom" are as follows:

Table 7. Examples of Additional Features Extraction

Nh	**Fw**	**Lw**	**Fp**	**Lp**	**Sf**	**Co**	**Ao**
-	the	lecturer	DT	NP	NP_v_np	1	0
to	to	classroom	PP	NP	np_v_NP	1	1

Information :

Nh : Noun Head of PP
Fw : First Word In Constituent
Lw : Last Word In Constituent
Fp : First POS In Constituent
Lp : Last POS In Constituent
Sf : Syntactic Frame

Co : Constituent Order
Ao : Argument Order

2.1.7. Feature Selection/Evaluation

Features or attribute selection/evaluation is a process of selecting or evaluating the most relevant attribute on the entire data with predictive modeling problems are being worked on. Attribute subset selection is mainly an optimization problem, which involves searching the space of possible feature subsets to select the one that is optimal or nearly optimal with respect to the performance measures accuracy, complexity etc. of the application [18].

The problem of Feature Selection can be defined as the process of selecting the best subset of features that describe the hypothesis at least as well as the original set (John, Kohavi, & Pfleger, 1994).

$$F' \in F$$

where F is the set of original 'n' features and F' is the output by a feature selector with m features.

There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods.

1. Filter Methods: This method is applying a statistical measure to assign a score to each feature. The features are sorted by rank scores then selected to be stored or removed from the database. This method is often univariate and considering these features independently or in connection with the dependent variable. Some examples of filter methods are the Chi-squared test, information gain and correlation coefficient scores.
2. Wrapper Methods: Wrapper methods consider the selection of a set of features as a search problem. This method constructs a number of different combinations of features, then evaluated and compared with other combinations. A predictive model is used to evaluate the combination of features and set the value based on the accuracy of the model. The search

process may be methodical as the best-first search, it possibly stochastic such as random hill-climbing algorithm, or may use heuristics, such as forward and backward to add and remove features. The wrapper method's example is the recursive feature elimination algorithm.

3. Embedded Methods: Embedded method learn about the best contribute features to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods is regularization method. Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients). Examples of regularization algorithms are the LASSO, Elastic Net, and Ridge Regression.

This thesis will use Filter Method for evaluating the most relevant attribute on entire data. The most relevant features selected will be developed into new features that expected will accommodate the features of a new argument. The proposed features will construct by detecting the most important features from Quran domain. The process is by finding out the attributes or features that have a high value of correlation with the class. For this process will use Gain Ratio attribute evaluation. Gain Ratio Attribute Evaluator evaluate the worth of an attribute by measuring the gain ratio with respect to the class. The selected features will be developed into a new feature and will be added to the training and testing data sets.

$$GainR(class, attribute) = \frac{H(class) - H(class|attribute)}{H(attribute)}$$

To select a feature set for the experiment, this thesis uses Wrapper Method, that select basic and additional features as in Table 6 and Table 7 then will be evaluated and compared with proposed features combinations.

2.1.8. Classifier

This thesis uses Support Vector Machine (SVM) as a classifier. SVM classification concept is trying to find a hyperplane (decision boundary lines) that separates the two best classes. The basic idea of SVM is to seek the maximum limit hyperplane as illustrated in the following figure:

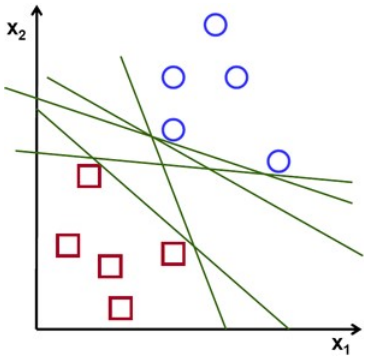

Figure 6. Hyperplane Options Are Possible

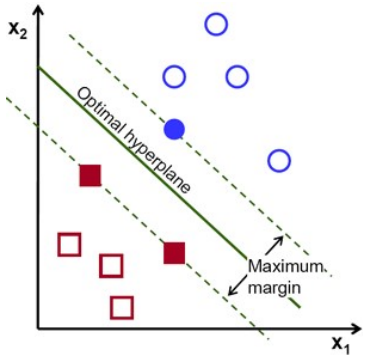

Figure 7. Hyperplane With Maximum Margin

Figure 6 shows the possible selection hyperplane to classify existing data sets. While Figure 7 shows the hyperplane with maximum margin among options that allow. Although it could also be used hyperplane arbitrary, hyperplane with maximum margin will give a better generalization in classification method [19].