IBAN LANGUAGE PARSER USING RULE BASED APPROACH

Chia Yong Seng

Master of Advanced Information Technology
2010
IBAN LANGUAGE PARSER USING
RULE BASED APPROACH

CHIA YONG SENG

A dissertation submitted
in partial fulfillment of the requirements for the degree of
Master of Advanced Information Technology

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2009
ACKNOWLEDGEMENT

The author wishes to express sincere appreciation to Dr. Edwin Mit, Ms. Suhaila, and Dr. Alvin Yeo for their assistance in the preparation of this dissertation. In addition, special thanks to those whose familiarity with the needs and ideas of this research project was helpful during the early programming phase of this undertaking. Thanks also to the members of the school council for their valuable inputs.

And finally thanks to my family members for their faithful supports.
# TABLE OF CONTENTS

ACKNOWLEDGMENT ........................................................................................................................................ ii
TABLE OF CONTENTS ................................................................................................................................... iii
LIST OF FIGURES ................................................................................................................................. vii
LIST OF TABLES ....................................................................................................................................... ix
ABSTRACT ................................................................................................................................................... x
ABSTRAK ................................................................................................................................................... xi

CHAPTER 1: INTRODUCTION ................................................................................................................... 12

1.1 Introduction ........................................................................................................................................ 12
1.2 Research Background ....................................................................................................................... 13
1.3 Scope Of The Research ....................................................................................................................... 13
1.4 Objectives Of The Research ............................................................................................................... 13
1.5 Significances Of The Research .......................................................................................................... 14
1.6 Problem Statements .......................................................................................................................... 14
1.7 Propose Solution ............................................................................................................................... 15
1.8 Chapter Summary .............................................................................................................................. 15

CHAPTER 2: LITERATURE REVIEW ....................................................................................................... 16

2.1 Introduction ....................................................................................................................................... 16
2.2 The Parser ........................................................................................................................................ 16

2.2.1 The Parsing Process ................................................................................................................... 17
2.2.2 Word Tokenizing ......................................................................................................................... 18
2.2.3 Word Tagging .............................................................................................................................. 19
2.2.4 Word Aligning ............................................................................................................................ 19

2.3 Computer Perception On Linguistic ............................................................................................... 19
2.4 Different Approaches Of Parser

2.4.1 The Top Down Approach Parser

2.4.2 The Bottom Up Approach Parser

2.5 Reviews On Language Parsers

2.5.1 Apple Pie Parser

2.5.2 LingSoft’s ENGCG Parser

2.5.3 Parser A Sentence (Phrase Parser)

2.5.4 SalingWika (A Top Down Parser)

2.5.5 Overview Comparisons

2.6 Chapter Summary

CHAPTER 3: METHODOLOGY

3.1 Introduction

3.2 Development Methodology

3.2.1 Spiral Methodology Cycles

3.3 Parser’s Process Flow

3.4 Iban Formal Grammar

3.5 Rule Based Grammar Applied

3.6 The Top Down Approach Parser

3.7 The Bottom Up Approach Parser

3.8 Chapter Summary

CHAPTER 4: IMPLEMENTATIONS

4.1 Introduction

4.2 Implementing The Parser

4.2.1 The Secondary Word Tagger
4.2.2 The Source Of Iban Dictionary ................................................................. 70
4.2.3 Database Design ....................................................................................... 71
4.2.4 Tagset Used .............................................................................................. 72
4.2.5 Finding Object And Subject In Sentence .................................................... 73
4.2.6 Finding Subject And Object In Multiple Sentence ................................. 74
4.2.7 Iban Tree Structure .................................................................................. 75
4.3 System Development .................................................................................... 77
4.4 System Input Output .................................................................................... 78
4.5. Chapter Summary ....................................................................................... 80
CHAPTER 5: DISCUSSION (RESULTS & TESTING) .............................................. 81
5.1 Introduction .................................................................................................... 81
5.2 Test Samples .................................................................................................. 81
5.3 Conditional Coverage Testing ....................................................................... 83
5.4 Predicate Coverage Testing ......................................................................... 85
5.5 Permutable Predicate Coverage Testing ....................................................... 89
5.6 Lengthy Predicate Coverage Testing ............................................................ 93
5.7 Permutable And Lengthy Predicate Coverage Testing ................................. 109
5.8 Multiple Predicates Coverage Testing ......................................................... 115
5.9 Performance Metric ...................................................................................... 120
5.10 Total Words Not Available From Dictionary ............................................... 124
5.12 Analysis Results ......................................................................................... 126
5.13 Iban Parser Limitations ............................................................................. 126
5.14 Chapter Summary ...................................................................................... 130
CHAPTER 6: FUTURE WORKS/EXTENTIONS ..................................................... 132
LIST OF FIGURES

Figure 2.1 Example of parse tree in Apple Pie Parser .......................................................... 25
Figure 2.2 Score calculation formulae in Apple Pie Parser .................................................. 27
Figure 2.3 Screenshot taken from Apple Pie Parser ............................................................. 28
Figure 2.4 Screenshot taken from ENGCG Parser ................................................................. 29
Figure 2.5 Screenshot taken from Phrase Parser ................................................................. 32
Figure 2.6 Phrase Parser's connector connections ................................................................. 34
Figure 3.1 Architecture of proposed Iban Parser System ..................................................... 41
Figure 3.2 Spiral methodology taken in building Iban language Parser ............................... 42
Figure 3.3 Process flow for parsing an Iban sentence ............................................................ 52
Figure 4.1 Example of Iban word in Iban dictionary ............................................................. 70
Figure 4.2 Subject and Object in sentence ......................................................................... 73
Figure 4.3 Basic construction of conjunction for multiple sentences ................................. 74
Figure 4.4 Iban Tree Structure ......................................................................................... 76
Figure 4.5 Interface layout of input interface ...................................................................... 78
Figure 4.6 Interface layout of Output interface ................................................................... 79
Figure 5.1 Top Down approach separation point .................................................................. 127
Figure 5.2 Bottom Up approach separation point ............................................................... 128
Figure 5.3 Pronoun on first parse in Top Down approach .................................................. 129
Figure 5.4 Pronoun on first parse in Bottom Up approach ................................................ 129
Figure 6 Iban Parser's input interface ............................................................................... 143
Figure 7 Apache Tomcat's console display ......................................................................... 143
Figure 8 Iban Parser's result ............................................................................................. 144
Figure 9 Iban Parser's Top Down Tree structure ............................................................... 145
LIST OF TABLES

Table 2.1 Comparison between Parsers ................................................................. 37
Table 4.1 IBAN_ENG_LEXICON database schema .................................................. 71
Table 5.1 Test sample for testing ........................................................................... 82
Table 5.2 Conditional coverage testing ................................................................. 84
Table 5.3 Predicate coverage testing ...................................................................... 89
Table 5.4 Permutable Predicate coverage testing .................................................... 92
Table 5.5 Lengthy Predicate coverage testing ......................................................... 107
Table 5.6 Permutable and Lengthy Predicate coverage testing ............................... 114
Table 5.7 Multiple Predicates coverage testing ....................................................... 120
Table 5.8 Iban Parser’s performance metric ............................................................ 121
Table 5.9 Iban Parser’s performance metric on Regular sentences ......................... 122
Table 5.10 Iban Parser’s performance metric on Irregular sentences ..................... 124
Table 5.11 Total words not available from Iban dictionary .................................... 125
ABSTRACT

There is a need for documentation or studies on Iban language in Natural Language Processing (NLP), because tools or Parser for Iban language is not available. In order to understanding and learning Iban language, an Iban Parser is required to generate Iban sentence structure, which allow computer scientist to study Iban language in academic ways.

The purpose of this research project is to propose an Iban Parser, a Parser that will parse Iban sentence. The Parser will recognize sentence's part of speech with Rule Based Grammar. Upon recognize all Iban words in a sentence; the Parser will present that sentence in Tree data structure presentation. Proposed Iban Parser is design to parse sentence with Top Down approach and Bottom Down approach.

Proposed Iban Parser comes with Top Down approach and Bottom Up approach, both approaches perform sentence parsing differently. This research projects had ran multiples tests which are (1) Conditional coverage testing, (2) Predicate coverage testing, (3) Lengthy Predicate coverage testing, (4) Permutable Predicate coverage testing, (5) Lengthy and Permutable Predicate testing, and lastly (6) Multiple Predicates coverage testing to test the Iban Parser. Overall test results showed that Iban Parser can recognize the Part Of Speech in Iban sentence. The design and multiple tests conducted were recorded in this research project would serves as stepping stone for related research fields in Iban language.
ABSTRAK

Adanya keperluan untuk dokumen atau belajar tentang bahasa Iban dalam “Natural Language Processing” (NLP) kerana alat “Parser” untuk memahami bahasa Iban yang tidak tersedia ada. Dalam rangka untuk memahami dan belajar bahasa Iban, sebuah alat “Parser” Iban diperlukan untuk menghasilkan struktur ayat Iban, yang memungkinkan ilmuwan komputer untuk belajar bahasa Iban dari segi akademik.

Tujuan dari projek penelitian ini adalah untuk mencadangkan sebuah alat “Parser” Iban yang akan “tokenize” ayat Iban. Alat “Parser” Iban akan mengenali bahagian pidato dengan berdasarkan Peraturan Nahu Iban. Setelah alat “Parser” Iban mengenali semua kata-kata dalam sebuah ayat, ianya akan menghasilkan ayat dalam presentasi struktur data Pohon. Alat “Parser” Iban yang dicadangkan akan “tokenize” ayat dengan pendekatan “Top Down” dan pendekatan “Bottom Up”.

Alat “Parser” Iban yang dicadangkan dengan pendekatan “Top Down” dan “Bottom Up” pendekatan akan melakukan “tokenizing” yang berbeza. Projek penelitian ini telah melakukan satu Siri ujian untuk menguji pendekatan tersebut untuk alat “Parser” Iban. Secara keseluruhan hasil ujian menunjukkan bahawa alat “Parser” Iban dapat mengenali bahagian ayat Iban dari segi pidato. Reka bentuk dan siri ujian yang direkod dalam dokumen projek penelitian ini akan berfungsi sebagai batu loncatan untuk bidang penelitian yang berkaitan dalam bahasa Iban.
CHAPTER 1: INTRODUCTION

1.1 Introduction

A Parser is Natural Language Processing tool for generating sentence structure; different language will have a different Parser. A Parser role is to break a sentence (input) into atomic form (which is also known as tokens), to enable computer to recognize each word grammatical representation.

The purpose of this research project is to present the basic conceptual design, the parsing process flow, and parsed data presentation of Iban Parser. This research project would serves as reference for audiences such as computer scientist and researchers in related research study field in Natural Language Processing for Iban language.

Dissertation written for this research project was organized in the following manner; Chapter 1 (Introduction) introduces the background and objectives of this research project. Chapter 2 (Literature Review) reviews existing Parsers and their approaches. Chapter 3 (Methodology) describes some of design aspects of Iban Parser. Chapter 4 (Implementation) records Parser's construction procedure or steps taken. Chapter 5 (Discussion) analyzes testing results on the Iban Parser and reviews the its limitations. Chapter 6 (Future Works) concludes this dissertation with achievements and recommendations for future works.
1.2 Research Background

According to the research projects list compiled by John Hutchins (2009) of European Association for Machine Translation on behalf of the International Association for Machine Translation, there is no documented works on translating English to Iban language or vice versa. Research fields related to Iban language is not listed and not available for references.

Therefore this dissertation (or research project) would also acts as stepping stone for further research works or any related researches.

1.3 Scope Of The Research

This research project deals with Iban sentences (5 to 10 words) as inputs, constructs a Parser for parsing these sentences and recognizes the sentence structure based on author defined Rule Based Grammar. This project also utilies a small Iban dictionary (with 10,000 entries).

1.4 Objectives Of The Research

Objectives of this research project are listed as below:

(1) Develop a prototype of Iban language Parser.
(2) Automate the generation Iban sentence structure.

(3) Recognize Iban language's Part Of Speech (e.g., RJN (Rambai Jaku Nama), RJA (Rambai Jaku Adjektif), and RJP (Rambai Jaku Pengawa)).

1.5 Significances Of The Research

This research project will be very useful as reference in learning and understanding Iban language structure. Possible benefits foreseen from this research project are listed as below;

(1) Assist human translator work in translating Iban language documents.

(2) Act as foundation in applications such as concordance and grammar checker.

(3) Serve as reference for other related researches in Natural Language Processing field.

1.6 Problem Statements

This research project was initiated due to several factors, these factors are listed as below;

(1) There is lacking documented or related (similar with this research project) works on Iban language made available. Proper documentations are important and act as references for related works in Iban language translation.

(2) Natural Language Processing tools or Parser for Iban language is not available, Parser is needed for recognizing Iban language sentence structure.
(3) Lack of documented computational defined grammar rule for Iban sentence in Natural Language Processing.

1.7 Propose Solution

To tackle problems identified in section 1.6, the following solutions are proposed in this research project.

(1) This research project will provide a write-up document on studies done Iban Parser. This research project will be documented as dissertation, and be anchors as reference in related research fields.

(2) This research project proposes an Iban Parser's design. The proposed Iban Parser will automated generate Iban language sentence structure.

(3) This research project proposed defined Iban sentence grammar rules for Natural Language Processing field.

1.8 Chapter Summary

As mentioned in this Chapter 1, currently there is no Iban Parser developed for this purpose. In order to translate and learn Iban language (based on sentence structure), an Iban Parser is required. This research project on Iban language will propose and present a suitable and experimental Iban Parser.
CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter discusses about language Parsers that have been made available and studies that have been done on Parser's parsing process. Parsers chosen for review are Apple Pie Parser, ENGCG Parser, Phrase Parser and SalingWika. Reviewing their parsing process and recognizes their distinctive features.

This chapter will discuss studies on Parser's parsing process which involves Word Tokenizing, Word Tagging, and Word Aligning.

2.2 The Parser

A Natural Language Parser (NLP) is a program constructed to recognize the grammatical structure of a sentence. The Parser breaks the sentence into small parts, and later regroups them in generated sentence structure as Object or Subject of a verb (James Allen, 1995).

Generated sentence structure is represented as lexical symbols (will be refer as Key in this research project), each symbols is used for representing a sentence in computer linguistic manner. Putting lexical symbols together will form grammatical sentence presentation.
Below are list of common Part Of Speech syntax used;

NP  Noun Phrase, for referring to things, place, qualities, concepts, events or objects.

S   Sentence, a sentence that is used for assert, query or command purpose.

VP  Verb Phrase, a Predicate.

VP[inf]  VP starting with infinite form.

S[inf]  Sentences in infinite form.

PP  Preposition Phrase, verb that involves specific Preposition Phrase.

ADJP  Adjective Phrase, consisting of single Adjective.

ADV VP  Adversial Phrase.

ART  An article.

N  A common noun.

2.2.1 The Parsing Process

Parsing a sentence can be done in two ways, Syntactic parsing and Semantic parsing. According to Wikipedia (2009), the Syntactic parsing check sentence based on token and create expression (or recognitions) that is usually ruled by Context-free grammar (CFG). Context-free grammar is used to describe structure of language. While Semantic parsing
(Wikipedia, 2009) took place after Syntactic parsing, it will try to work out the implications of expression.

This research study will only involved *Syntactic parsing* in Iban Parser, where an Iban sentence will be broken into small tokens and go through parsing processing. In a generic Parser, the parsing process of a sentence will involved Tokenizing, Tagging and Aligning.

2.2.2 Word Tokenizing

A Tokenizer is a NLP tool for scanning a string of characters (James Allen, 1995), such as added line of text from command prompt, and converting these character strings into a list of words and punctuation marks. Each item in this list is called a “token”.

When parsing a sentence, the whole “chunk” (which is the entire added sentence) will not be parsing by Parser; instead the Parser will work with tokens, which is faster and easier. Without Tokenizer, Parser would need to go through steps such recognizing word boundaries, skipping whitespace, and finding delimiters (such as quotes and parenthesis). Tokenizer would perform all this in advance when a string is tokenizes, so these steps would not be repeated in parsing process.
2.2.3 Word Tagging

Tagging is a process handled by Tagger for giving a Key (part of speech such as Noun, Verb, Adjective, etc.) to a string of word, in many cases a string of word can be Noun or Determinant (James Allen, 1995). This is done by matching a string of word with huge predefined tag library, usually tagging will come after tokenizing a sentence of word.

2.2.4 Word Aligning

Aligning is a process of matching a string of word with another string of word (James Allen, 1995); this is usually done with pre defined source of lexicon (dictionary).

2.3 Computer Perception On Linguistic

Unlike human, a computer cannot recognize a string like “The quick brown fox jumps over the lazy dog”; the computer only understands this string, is built of 43 characters string array which includes whitespaces. For a computer to learn and understand this string array, a new presentation is required (James Allen, 1995). With Word Tokenizing process, this 43 characters string array will be recognized as 8 words based on word boundaries and whitespaces.
One of the common ways for storing persist form of tokens is using XML (Extensible Markup Language) document. This is due to XML simplicity, usability over Internet, and supports via Unicode for languages around the world.

The following is an example a sentence that was converted into XML format, which later can be recognized by a computer during parsing process. Computer can now understand each word as separate entity instead of whole "chunk" (in this case, the entire added sentence) in string array. The sentence "The quick brown fox jumps over the lazy dog", can be represented (in XML format) as,

```
<sentence>
    <word>The</word>
    <word>quick</word>
    <word>brown</word>
    <word>fox</word>
    <word>jumps</word>
    <word>over</word>
    <word>the</word>
    <word>lazy</word>
    <word>dog</word>
</sentence>
```
Computer recognizes word by word (e.g., "The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", and "dog") in XML document by distinguishing content between <word> markup and </word> markup.

2.4 Different Approaches Of Parser

The two common strategies used in parsing a sentence are Top Down approach Parser and Bottom Up approach Parser (James Allen, 1995). The Top Down approach Parser generates sentence structure in expansive manner (from first to last word) while the Bottom Up approach Parser used the reductive approach (begin from last word and end with first word). Each strategy has its strengths and weakness depending on how they are use. Tokenization involved demarcating and classifying sections of an input string.

2.4.1 The Top Down Approach Parser

The Top Down approach Parser breaks the sentence (S) into atomic form (which are token) from left to right (left most derivation) manner, which is starting from first word to last word in S. This approach is known as goal oriented, because symbol hypothesis is made based on unit will be found in the sentence (James Allen, 1995).

Top Down approach Parser involved using stack data structure; 2 strategies available for this Parser are Depth First strategy and Breadth First strategy. Depth First strategy used
“Last In First Out” (LIFO) stack and Breadth First strategy used “First In First Out” (FIFO) stack.

The Depth First strategy searches the main interpretation and expands it; if that interpretation failed to be found, it will consider and search the alternatives. While Breadth First strategy searches the main interpretation and alternatives all together before proceed to the next interpretation searching. The Depth First strategy may be faster in concluding the result if compare to Breadth First strategy, but may take a lot time if pursuing the wrong interpretation.

2.4.2 The Bottom Up Approach Parser

The Bottom Up approach Parser matches word in right to left (right most derivation) manner. Unlike the Top Down approach Parser; it searches from known word in sentence which is the last word in sentence (S) (James Allen, 1995). This Bottom Up approach Parser rewrites a word by its possible Key (Part Of Speech attributes like Noun, Verb, Adjective, etc) and replaces a symbol that matches its right hand in sequence based on grammar rule. Stack data structure is also used to store partial result for searching process.

Parsing process in this Parser is based on Key (Part of Speech attributes like Noun, Verb, Adjectives, etc). Key is used for a string is based on rule that start with the Key itself, or
rule that had already started with previous Key and presence of the current Key in completing or extending the rule.

2.5 Reviews On Language Parsers

To understand about Parser, this research project reviews some made available English Parsers based its features, and techniques. Selected Parsers are;

1. Proteous Project - Apple Pie Parser
2. LingSoft’s ENGGC
3. Parse a sentence (Phrase Parser)
4. SalingWika (A Top Down Parser)

2.5.1 Apple Pie Parser

Apple Pie Parser is a Bottom Up approach Parser type from Proteous Project, its using best first search algorithm. The Parser (Proteous Project, 2009) finds the best Parser tree based on score given by the search algorithm. It generate syntactic tree similar to PennTreeBank (PTB) bracketing. The later version of PTB (version 2.0) includes argument structure label which is not available in APP generated syntactic tree. This Parser is developed for parsing simple English sentences. Unlike most PTB Parser that searches the whole sentence for Part Of Speech complete match, APP Parser searches the sentence partially.