COMBINING LEXICAL CHAIN AND DOMAIN DRIVEN APPROACHES TO ENHANCE LEXICAL CHAIN PERFORMANCE

Lee Wei Jan

Master of Science
2012
COMBINING LEXICAL CHAIN AND DOMAIN DRIVEN APPROACHES TO ENHANCE LEXICAL CHAIN PERFORMANCE

LEE WEI JAN

A thesis submitted in fulfillment of the requirements for the degree of Master of Science (Computer Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2012
Declaration

No portion of the work referred to in this report has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher learning.

LEE WEI JAN

4th September 2012
I would like to dedicate my appreciation first to my family for their unconditional support, encouragement, and understanding. My second word of gratitude goes to my supervisor, Dr Edwin Mit, who patiently and excellently guided me through this progress especially when I was having a hard time with the research studies. Without the financial support from Zamalah Naib Canselor, I might not able to go through the entire process. Therefore I would like to show my gratitude to the UNIMAS Postgraduate Fellowship. Last but not least, my acknowledgment goes to my supportive friends who shared the ideas of my research and support me to pursue my goal when I was having difficult time in the middle of this learning process.
Abstract

Word Sense Disambiguation (WSD) is the process of identifying the meaning of the words in context of computational manner (Navigli, 2009). Every text or discourse is actually a composition of words, phrases, and sentences which tend to describe the similar topic. Therefore Morris and Hirst (1991) proposed an approach, named Lexical Chain approach to disambiguate the words by finding the relationships between words in the given text. Lexical Chain approach is originally used to exploit the lexical cohesion of a text by looking for the semantics relationships between words (relationships provided by the dictionaries).

Over the years, several researches were conducted to improve the performance of Lexical Chain by adapting different knowledge resources and different measurements. However the insufficient of process in determine the level of the semantics similarity of the relationships formed between words based on the text restricted the performance of the Lexical Chain. The ordinary Lexical Chain approach formed the relationships between strongly related words such as car and vehicle, and at the same time, some of the abstract words are sometimes tended to be related such as ballot and resignation. The abstract relationships produce noises in the process of determining the most appropriate sense of the words.

Therefore in this research is to propose a combination approach to improve the disambiguating performance of Lexical Chain approach. The purpose of this research is to improve the sense identification by integrating Lexical Chain approach with Domain Driven approach. This combination approach is derived from the concept of exploiting the lexical cohesion and textual
coherence of any given text. The proposed combination approach will form the relationships between words by using the Lexical Chain approach and determines the semantics similarity based on textual coherence obtained by Domain Driven approach.

Domain Driven approach is proposed to integrate with Lexical Chain approach because Domain Driven approach was proposed to exploit the coherence of any given text by accessing to the domain knowledge of the words in the text. The Domain Driven approach will act as the decision maker for determining the similarity of the related words that established by Lexical Chain approach. Hence the proposed framework does not only relying on the information obtained from either lexical cohesion or textual coherence, but obtains the wellness information from both approaches.

The experiments had been carried out to prove the performance of the proposed combination framework. The results obtained from the experiments indicate an improvement when the Lexical Chain approach is integrated with Domain Driven approach.
**Abstrak**


Beberapa tahun yang lalu, beberapa kajian seperti menggunakan sumber-sumber pengetahuan yang berbeza dan mencadangkan pengukuran yang berbeza telah dijalankan untuk meningkatkan proses mengenal pastikan makna-makna perkataan oleh *Lexical Chain*. Manakala, kekurangan proses untuk menentukan tahap persamaan semantik hubungan antara perkataan berdasarkan teks telah mempengaruhi ketepatan pendekatan *Lexical Chain*. Ini adalah disebabkan oleh pendekatan *Lexical Chain* dapat membentukan hubungan semantik antara perkataan yang dapat dikaitan dengan hubungan yang kuat seperti *car* dan *vehicle*, tetapi, kadang kalang, perkataan seperti *ballot* dan *resignation* juga dapat dihubungkan dengan hubungan yang abstrak. Hubungan abstrak ini mempengaruhi ketepatan untuk mengenal pasti makna-makna perkataan.

Oleh sebab itu dalam kajian ini, dua pendetakan akan dicadangkan untuk meningkatkan proses mengenal pasti makna-makna perkataan oleh *Lexical Chain*. Pendekatan kombinasi adalah
direkakan berdasarkan konsep mengeksploitasi lexical cohesion dan textual coherence oleh teks yang diberikan. Pendekatan kombinasi yang dicadangkan itu akan menghubungkan perkataan dengan menggunakan Lexical Chain dan menentukan persamaan semantik berdasarkan textual coherence yang diperolehi oleh keadah Domain Driven.

Pendekatan Domain Driven dicadangkan untuk digambung dengan Lexical Chain kerana Domain Driven dicadangkan untuk mengeksploitasi textual coherence yang diberikan dengan mengakses kepada pengetahuan domain dalam teks yang diberikan. Pendekatan kombinasi ini direkabentuk untuk menentukan persamaan semantik hubungan yang dibentuk oleh pendekatan Lexical Chain untuk mengurangkan kesilapan yang mungkin berlaku dalam proses mengenalpasti makna-makna perkataan.

Pendekatan Domain Driven akan digunakan sebagai keputusan untuk mengenal pasti persamaan perkataan-perkataan yang berkaitan yang diperolehi oleh pendetakan Lexical Chain. Oleh sebab itu, rangka kerja yang dicadangkan tidak bergantung hanya maklumat yang diperolehi sama ada daripada lexical cohesion atau textual coherence, tetapi memperolehi maklumat yang sepenuhnya daripada kedua-dua pendetakan.

Experimen-experimen telah dijalankan untuk membuktikan prestasi rangka kerja yang dicadangkan. Keputusan yang diperolehi daripada eksperimen-eksperimen telah membuktikan peningkatan proses mengenal pastikan makna-makna perkataan apabila pendekatan Lexical Chain digabungkan dengan pendekatan Domain Driven untuk menentukan persamaan perkataan yang berkaitan,
# Table of Content

Declaration ................................................................................................................. I  
Acknowledgments ................................................................................................. I  
Abstract ................................................................................................................. III  
Abstrak .................................................................................................................... V  
Table of Content ..................................................................................................... VII  
List of Published Papers ....................................................................................... X  
List of Figures ......................................................................................................... XI  
List of Tables .......................................................................................................... XIII  
List of Abbreviations ............................................................................................. XIV  

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Problem Statement</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.3 Hypothesis</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1.4 Research Objectives</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1.5 Scopes</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1.6 Significance of the Project</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.7 Thesis Structure</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Literature Review</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2.1 Introduction</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2.2 WordNet 2.0 and WordNet Domains 3.2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2.3 Lexical Chain</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>2.4 Domain Driven Approach</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>2.5 Semantic Similarity</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>2.6 Other Word Sense Disambiguation Approaches</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>2.7 Summary</td>
<td>39</td>
</tr>
<tr>
<td>3</td>
<td>Research Methodology</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>3.1 Introduction</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>3.2 Requirement</td>
<td>41</td>
</tr>
</tbody>
</table>
3.3 Analysis................................................................................................................... 42
3.3.1 Lexical Chain Approach ..................................................................................... 44
3.3.2 Domain Driven Approach .................................................................................. 54
3.4 Conceptual Design ............................................................................................... 61
3.4.1 Preprocessing Stage .......................................................................................... 63
3.4.2 Disambiguation Design Module ........................................................................ 65
3.4.2.1 The Global Combination Approach Design ................................................. 72
3.4.2.2 The Local Combination Approach Design .................................................... 73
3.5 Summary.................................................................................................................. 75

Chapter 4 Implementation of Disambiguation Process ................................................. 76
4.1 Introduction ............................................................................................................ 76
4.2 Data Set ................................................................................................................. 78
4.3 Preprocessing Module ......................................................................................... 81
4.4 Disambiguation Implementation and Evaluation Module ...................................... 85
4.4.1 Lexical Chain Approach Evaluation ................................................................. 85
4.4.2 Domain Driven Approach Evaluation ............................................................... 90
4.4.3 The Combination Approach Evaluation ........................................................... 96
4.4.3.1 The Global Combination Approach Evaluation ........................................... 98
4.4.3.2 The Local Combination Approach Evaluation ............................................ 100
4.5 Summary................................................................................................................ 101

Chapter 5 Evaluation and Discussion ........................................................................ 102
5.1 Introduction .......................................................................................................... 102
5.2 Results of Preprocessing Module ......................................................................... 103
5.2.1 Discussion ........................................................................................................ 104
5.3 Results of Lexical Chain approach .................................................................... 105
5.3.1 Depth Stability Evaluation ............................................................................... 105
5.3.2 Lexical Stability Evaluation ............................................................................ 108
5.3.3 Discussion ........................................................................................................ 109
5.4 Results of Domain Driven approach .................................................................. 111
5.4.1 Discussion ........................................................................................................ 113
List of Published Papers


List of Figures

Figure 2.1 Example of the semantics relations between words (Hirst & St.Onge, 1998) 18
Figure 2.2 Example of the semantics medium-relations between words 19
Figure 2.3 The patterns of the allowable path for medium-strong relation (Hirst & St.Onge, 1998) 19
Figure 2.4 Example of Yarowsky’s WSD method 22
Figure 2.5 The steps for the heuristic domain approach (Kolte & Bhirud, 2008) 27
Figure 2.6 The noun taxonomy of Car, Bicycle and Fork in WordNet 30
Figure 3.1 The example of lexical chain of the text (words in bold) 45
Figure 3.2 Example of relation between words Machine, Car and Accelerator (Morris & Hirst, 1991) 48
Figure 3.3 Semantic graph for candidate word car and its related synsets 49
Figure 3.4 Example of relations between words Machine, Car and Accelerator 49
Figure 3.5 Conceptual design of the proposed framework 62
Figure 3.6 Format of the index file generated after preprocessing stage 65
Figure 3.7 The example text (bold words are related) 66
Figure 3.8 The example of Hypernym semantics relations from WordNet 67
Figure 3.9 Framework for combination approach 69
Figure 3.10 Example of local combination approach for legislature 74
Figure 4.1 The process flow of Chapter 4 76
Figure 4.2 Example text from one single document of SemCor 2.0 corpus 79
List of Tables

Table 2.1 Definition and example of semantics relation types in WordNet 13
Table 2.2 Domain distribution across the synsets in WordNet (Bentivogli et al., 2004) 16
Table 2.3 The senses of the words (Lesk, 1986) 33
Table 2.4 The combination between approaches 40
Table 3.1 Semantics relationship weight scheme 52
Table 3.2 WordNet senses and domains for the word “bank” 56
Table 4.1 Total of words for some documents from SemCor 81
Table 4.2 Punctuation list for sentence splitting process 81
Table 4.3 Punctuation and symbol list to be removed during tokenization process 82
Table 4.4 A set of predefined thresholds for the scenarios of combination approach 98
Table 5.1 The accuracy of POS tagging 103
Table 5.2 The accuracy of lexical chain with and without the depth stability 108
Table 5.3 The accuracy of lexical chain with and without the lexical stability 109
Table 5.4 The accuracy of the Domain Driven approach 112
Table 5.5 Results of the Global Combination approach 118
Table 5.6 Results of the Local Combination approach 123
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Sense Disambiguation</td>
<td>WSD</td>
</tr>
<tr>
<td>Part-of-Speech</td>
<td>POS</td>
</tr>
<tr>
<td>Noun</td>
<td>NN</td>
</tr>
<tr>
<td>Verb</td>
<td>VB</td>
</tr>
<tr>
<td>Adjective</td>
<td>ADJ</td>
</tr>
<tr>
<td>Adverb</td>
<td>ADV</td>
</tr>
<tr>
<td>Lexical Chain</td>
<td>LC</td>
</tr>
<tr>
<td>Domain Driven</td>
<td>DD</td>
</tr>
<tr>
<td>Domain Relevance Score</td>
<td>DR score</td>
</tr>
<tr>
<td>Domain Frequency Score</td>
<td>DF score</td>
</tr>
</tbody>
</table>
Chapter 1  Introduction

1.1  Introduction

WSD is the process of identifying the meaning of the words in context of computational manner (Navigli, 2009). WSD is essential for most of the Natural Language understanding applications such as Machine Translation, Information Retrieval, Information Extraction and Content Analysis as the knowledge about the meanings of the word and an accurate WSD could significantly improve the precision of these Natural Language applications (Villarejo, 2006).

According to Navigli (2009), there are four elements of WSD, which are: the selection of word senses, the use of external knowledge sources such as external dictionaries or corpora, the representation of context, and the selection of an automatic classification method. Therefore, in the general terms, a WSD task can be described in two steps which are first, the process of determining of all the senses that are relevant to the text or discourse, and second, the process of assigning the appropriate sense to each word (Ide & Veronis, 1998).

From years to years, there are several different types of approach had been introduced to WSD, and they are mainly distinguished as supervised WSD, unsupervised WSD and knowledge based WSD. Each of these approaches provides a significant finding in determining the sense for polysemous words.
Supervised WSD approaches are the approaches that apply the machine-learning techniques to identify the senses from labeled training sets, which a few sets of examples that had been assigned together with the features and appropriate label of sense (Navigli, 2009). There are supervised WSD approaches such as decision lists (Rivest, 1997; Yarowsky, 1994), decision trees, neural networks (Yatsaronis et al., 2007) and support vector machine (Boser et al., 1992). Supervised based approaches always provide the most accurate performance compared to unsupervised based approaches. However, in order to obtain a corpus that is sufficient to assist the supervised approaches performs a better result, Ng (1997) estimate that approximately 3.2 million sense tagged corpus might be required. The human effort in constructing such kind of corpus could be large and it is expensive.

Unsupervised WSD approaches assigned the most appropriate sense for a word without referring to the well labeled training sets as supervised WSD approaches. The example of unsupervised WSD approaches are context clustering (Schutze, 1992), word clustering (Lin, 1998), and co-occurrence graphs. Unlike Supervised WSD approaches, Unsupervised WSD approaches are invented based on the idea that same sense of word will have similar neighboring words. These approaches determine the word senses by clustering word occurrences in the given input text, and classifying the new occurrences into the induced clusters (Navigli, 2009). Since the senses were obtained based on the clustered results but not from the traditional dictionaries, the evaluation is usually more difficult as human experts are required to determine the accuracy of the approaches.

Knowledge-based WSD approaches are relying on the external lexical resources such as dictionaries, thesauri and ontology instead of well labeled discourse or corpus. Knowledge-based
approaches obtain the sense for words by exploiting the knowledge sources and some statistically or heuristically methods were used to determine the word senses. By accessing these lexical resources, some of the external syntax and semantic information about a word could be obtained. For example, WordNet (Miller et al., 1990) is a computational lexicon of English words, and the words are grouped into synset, (a set of synonymys). Besides, by exploiting the WordNet, not only senses of a word could be obtained but also its semantic relations. With these richness information obtained, a wider coverage of WSD can be carried out.

Instead of disambiguating a word by using single WSD approach, researchers started to integrate different approaches to achieve a better result in disambiguate an ambiguous word. In the process of integrating, some components will be eliminated, augmented or adopted to overcome the shortages or restrictions of a particular approach.

In this research, an integration of approaches will be introduced to improve the performance of the WSD approaches. This research starts with reviewing few knowledge based approaches, identifying the limitations and finding the solutions. Lexical Chain approach is selected as the approach to be enhanced because it had gone through several augmentations by different researchers to improve the performance in disambiguation process. However it is still having limitations in determining the semantic similarity of the formed semantic relationships between words. Therefore in this research, Domain Driven approach is adapted to integrate with Lexical Chain approach to improve the performance of WSD process because Domain Driven approach is able to determine another type of relationship between words and it is able to determine the similarity between formed relationships.
1.2 Problem Statement

Every text or discourse is actually a composition of words, sentences, and phrases which each tends to describe the similar topic. Therefore in the real life environment, human experts tend to understand and identify the most appropriate meaning of every word in a given text based on the understanding according to the context or the neighboring words. The neighbor words can appear whether within the same sentence or within the same paragraph. In the text linguistic perspective, the relatedness of words can be described as a cohesion and a coherence.

Cohesion of a text is a property where a text is not simply only a composition of a set of sentences and phrases, but each sentence and phrase in the text tends to discuss about similar things or concepts. Human understands the text by reading through every sentence and each word in the sentences that contributes a simple understanding to human in order to draw a complete idea about the text. In the other words, the relatedness between words help human to define the meaning of most of the words in the text. Words are tending to occur in similar environment because they describe the similar situations or context (Morris & Hirst, 1991). For instance, for a given context of \{gin, alcohol, sober, drinks\}, narrowed down the meaning of drinks to alcoholic drinks (Morris & Hirst, 1991). The semantic relations formed between words were then known as the lexical cohesion. However, unlike human expert, machine unable to understand the semantic relations between the words unless a knowledge resource which contains the information about the relationships between words is adopted.
Even though the knowledge based approaches are able to identify the word relationships, these approaches are having limitations in identifying the similarity of the related words in the given context. In each of the knowledge resources, strong semantically related words such as *car* and *vehicle* are related, but at the same time, some of the abstract words are sometimes tended to be related such as *ballot* and *resignation*.

Therefore, the process to determine the semantic similarity of every formed related word becomes a crucial issue for the approaches that applied lexical cohesion as the basic foundation in WSD, such as Lexical Chain (Morris & Hirst, 1991).

1.3 Hypothesis

In this research, an idea of integrating the Lexical Chain with a semantic similarity determination approach is believed to be able to improve the performance of the original Lexical Chain approach.

However, instead of applying some existing semantic similarity approaches proposed by several researchers over the years such as Resnik (1995), Jiang and Conrath (1997), and Seco et al. (2004), a method that is able to determine the semantic similarity based on the context of the text is believed to perform a better integration with Lexical Chain approach. It is reasonable to believe that determines the semantic similarity of the related words based on the same context words that formed the relationships is able to produce a better accuracy.

---

1 In WordNet, *ballot* and *resignation* are related because both inherit from the parent of *document*. 
Domain Driven approach by Magnini et al. (2001) is an approach that determines the most appropriate sense based on the domain distribution across the text. It holds a few properties both from the lexical and textual point of view such as represent the lexical coherence, reduce the polysemy (Gliozzo et al., 2004).

Therefore in this research, Domain Driven approach is proposed to integrate with Lexical Chain approach because Domain Driven approach is able to determine the likeliness of the words to be related in the given text based on the domain distribution across the text. It determines the semantic similarity by using the content of the text instead of other extended resources such as the machine readable dictionaries, ontologies or thesauri.

1.4 Research Objectives

Even though knowledge acquisition becomes a bottleneck in the development of most of the Natural Language Processing systems, some approaches suffered from the well-defined structure of the knowledge resources. Instead of taking every pieces of information that provided by the knowledge resources, this research is focusing on identifying the reliability and usability of the information obtained from the knowledge resources.

The prime focus of this research is to develop a framework that is able to exploit the lexical cohesion and coherence of the text by accessing to the information provided by adopted knowledge resources. A hybrid approach is proposed. Two necessary steps in this approach are first identifying the lexical cohesion in the given text by using Lexical Chain approach, and then
second, the relations that found between words will be then evaluated using Domain Driven approach before establishing the relations. Hence, the semantic relationships between words that obtained from the knowledge resources will be evaluated in order to improve the accuracy.

The objectives of this research are listed as below:

i) To integrate Lexical Chain and Domain Driven approach to perform a better WSD process
ii) To identify the strengthens and weaknesses, and the gaps of weaknesses the proposed existing WSD approaches
iii) To exploit the lexical cohesion and coherence relations in the text so that the proposed approach behaves more likely toward a human perspective

1.5 Scopes

This research focuses on defining a new framework for WSD that incorporates a combination of approaches. This research employs the Lexical Chain and Domain Driven approaches by integrating the necessary functionalities and then later proposed a better WSD approach.

The scopes of this research are listed as below:

i) In this research, only one knowledge resource will be adopted to provide the semantic relationships, and domain information that is Wordnet 2.0.
ii) The experiments will be conducted by using Semcor 2.0 corpus to collect the accuracy, and there is no human tester involved.
iii) This proposed WSD approach will disambiguate only noun words.

iv) This proposed WSD approach will not disambiguate the compound words.

v) This research focuses on proposing an integrated concept to improve Lexical Chain approach, hence less significant effort will be done in preprocessing module which includes the sentence splitting, tokenization, and Part-of-Speech (POS) tagging techniques.

vi) This research focuses on proposing a combination approach to enhance the performance of Lexical Chain approach and the applicable of the proposed combination approach in other Natural Language Processing field is not discussed.

1.6 Significance of the Project

This research proposes a combination disambiguate approach which will inherit the strength of the existing Lexical Chain approach and improves the weakness of Lexical Chain approach by integrating Domain Driven approach. This combination disambiguate approach does not rely on only the results from the existing approaches and filter the most appropriate sense by filtering the results like the combination approach proposed by Stevenson and Willks (2001) but integrates the approaches to obtain one result.

In this research, it proposes a combination knowledge based approach which will not only rely on the knowledge resource such as machine readable dictionary to disambiguate words, but also determines the similarity of the related words based on the content of the text. It proposes an idea
which will bring the WSD approach to behave more likely towards human experts to identify the
senses by exploiting lexical cohesion and text coherences.

Besides that, this research proposed an idea to break through the bottleneck situation of Lexical
Chain approach. Even though by increasing the number of adopted knowledge resources in the
approach might increase the performance of Lexical Chain, this research proposed the
combination approach which only relies on one single knowledge resource, WordNet in this case
which believes will increase the speed of WSD approach.

At the end of this research, it will introduce a combination approach which proposes a better
performance in WSD matter.

1.7 Thesis Structure

This chapter provides an overview of this dissertation. In the following chapter, Chapter 2, some
reviews on the current works in the WSD areas will be discussed.

Chapter 3 presents the conceptual design of the proposes framework for WSD process. The
processes and formulas that are used in the proposed framework will be discussed in details.

Chapter 4 discusses the environments requirements for setting up the experiments and the
implementation of the proposed framework.