UNIVERSITI MALAYSIA SARAWAK

UNIMAS

Faculty of Computer Science and Information Technology

UTILISING SOCIAL MEDIA THROUGH CROWDSOURCING FOR MORPHOLOGICAL RESOURCES ACQUISITION OF UNDER-RESOURCED LANGUAGE (U-RL): MELANAU

Voon Mei Wei

Bachelor of Computer Science with Honours
(Information Systems)

UTILISING SOCIAL MEDIA THROUGH CROWDSOURCING FOR MORPHOLOGICAL RESOURCES ACQUISITION OF UNDER-RESOURCED LANGUAGE (U-RL): MELANAU

VOON MEI WEI

This project is submitted in partial fulfilment of the
requirements for the degree of
Bachelor of Computer Science with Honours
(Information Systems)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2014/2015

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE UTILISING SOCIAL MEDIA THROUGH CROWDSOURCING FOR MORPHOLOGICAL RESOURCES ACQUISITION OF UNDER-RESOURCED LANGUAGE (U-RL): MELANAU

ACADEMIC SESSION: 2014/2015

VOON MEI WEI
(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

□ CONFIDENTIAL (Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)

□ RESTRICTED (Contains restricted information as dictated by the body or organization where the research was conducted)

☑ UNRESTRICTED

Validated by

(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address

NO 7, JALAN 6, TAMAN USAHA JAYA,
KEPONG 52100 KUALA LUMPUR
W.P.K.L.

Date: 25TH JUNE 2015

Date: 25/6/15

Note * Thesis refers to PhD, Master, and Bachelor Degree

** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

DECLARATION

I hereby declare that this project is my original work. I have not copied from any other student's work or from any other sources except where due references or acknowledgement is not made explicitly in the text, nor has any part had been written for me by another person.

....

(VOON MEI WEI) (JUNE 2015)

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest appreciation to all those who have directly and indirectly provided me the possibility to complete this report. A special gratitude I give to my final year project supervisor, Ms. Suhaila Saee, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report. Besides, I would like to thank Dr. Nadianatra Musa for her useful comments and remarks through the examination process of this bachelor degree project report.

Furthermore, I have to appreciate Faculty of Science Computer and Information Technology, Universiti Malaysia Sarawak for providing me an opportunity to apply my knowledge received in these four years and new exploring on my final year project topic today. Not to forget also the guidance given by other supervisors as well as the panels especially in our project presentation that has improved our presentation skills and thanks to their comment and advices.

I would as well like to thank my loved ones, who have supported me throughout the four years of studies, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for your love.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
U-RL	Under-Resourced Languages
UPM	Unified Process Model
OOA/D	Object-Oriented Analysis/Development
POS	Part of Speech
OOV	Out-of-Vocabulary
LD	Levenshtein Distance
SED	Spelling Edit Distance
HD	Hamming Distance
(Full) DLD	(Full) Damerau- Levenshtein Distance
JWD	Jaro-Winkler Distance
UML	Unified Modelling Language
UC	Use Case
MRDS	Morphological Resources Development System
(G)UI	(Graphical) User Interface
DB	Database
UAT	User Acceptance Test
P	Precision
R	Recall

ABSTRACT

Morphological analyser is the first processing tool required in Natural Language Processing To analyse structure of a word, the analyser needs morphological resources The resources are from dictionary, grammar book(s), and written texts. Yet, how to acquire morphological resources for under-resourced languages knowing that the languages are critically lacking of materials? In current approach. morphological resources are acquired from hardcopy versions whereby one needs to digitise the documents into softcopy versions. Due to difficulty in digitisation as it is time consuming and expensive, this project is proposing a workflow of acquiring morphological resources for under-resourced languages, in the case of Melanau language, by utilising social media. Three main stages in the work are: i) crowdsourcing the social media by using a web crawler Spider 3.3 and Jsoup method; ii) performing hybrid normalisation to transform the crawled data with informal and noisy nature into a cleaned wordlist; iii) validating the wordlist, is a crucial stage due to languages mixing that causes uncertainty of spelling standard. At this stage, edit distance similarity algorithms, Jaro-Winkler distance, Levenshtein-based distance, and N-gram distance, are applied to identify the spelling standard between a source word from the wordlist and a target word in the dictionary. The results show that Jaro-Winkler performs the best compared to the other two algorithms because it returns the highest F-score and the longest validated wordlist. The validated wordlists are then considered as the Melanau morphological resources that can be applied by computational linguists in the computational morphology. Indirectly, the proposed workflow can also be used to acquire morphological resources for other under-resourced languages in Sarawak

ABSTRAK

Penganalisis morfologi ialah alat pemprosesan pertama yang diperlukan dalam pemprosesan bahasa asli. Untuk menganalisis struktur bagi sepatah perkataan, penganalisis tersebut memerlukan sumber morfologi. Sumber tersebut adalah daripada kamus, buku tatabahasa, dan teks bertulis. Namun, bagaimana untuk memperoleh sumber morfologi bagi bahasa yang bawah pembangunan mengetahui bahasa tersebut dalam keadaan kritikal kekurangan sumber? Dalam pendekatan semasa, sumber morfologi diperolehi dari versi bercetak di mana seseorang itu perlu mendigitalkan dokumen bercetak ke dalam versi salinan lembut. Oleh kerana kesukaran dalam proses pendigitalan atas sebab pengambilan masa yang lama dan kos yang tinggi, projek ini menyusulkan sebuah aliran kerja untuk memperoleh sumber morfologi bagi bahasa yang bawah pembangunan, di mana tertumpu pada bahasa Melanau, dengan menggunakan sepenuhnya media sosial. Tiga proses utama dalam projek ini ialah: i) mengumpul data dari orang ramai dalam media sosial dengan kaedah perangkak web *Spider 3.3* dan *Jsoup;* ii) mengaplikasikan pemulihan hibrid untuk menukar data tersebut yang mempunyai sifat tidak rasmi dan bercampuradukan bahasa kepada set data yang bersih; iii) mengesahkan set data yang bersih tersebut, ialah peringkat penting kerana percampuran bahasa menyebabkan ketidaksatuan piawai ejaan. Pada peringkat ini, algoritma persamaan jarak, *Jaro-Winkler, Levenshtein-based,* dan *N-gram,* digunakan untuk mengenal pasti piawai ejaan antara perkataan sumber dari set data bersih dan perkataan sasaran dalam kamus. Keputusan menunjukkan bahawa *Jaro-Winkler* adalah terbaik berbanding dengan dua algoritma yang lain kerana *Jaro-Winkler* memberi nilai F-skor yang tertinggi dan senarai perkataan sah yang terpanjang. Senarai ini maka akan dijadikan sumber morfologi Melanau untuk digunakan oleh ahli-ahli linguistik perkomputeraan dalam morfologi perkomputeraan. Secara tidak langsung, aliran kerja yang diusulkan mampu juga diperluaskan aplikasinya untuk memperoleh sumber morfologi bagi bahasa-bahasa yang bawah pembangunan lain di Sarawak.

CHAPTER 1 INTRODUCTION

Presently, only a small number of the world's languages are enjoying latest welfares such as morphological analysers and morphological resources. Morphological resources are needed in morphological analysers for words analysing. Morphological analysers, in contrast are crucial first processes in natural language processing (NLP) application. As currently there is lacking of morphological resources in under-resourced languages (U-RL) which with limited resources for development, they are starting to be discovered and developed.

Melanau language spoken widely in Central Sarawak, Borneo, Malaysia (Rensch, 2012), is an example of the U-RL where its data is hardly reachable elsewhere than dictionary and written text files. In process of digitising the hardcopies, it is time consuming and costly. Thus, this work is proposing utilise social media through crowdsourcing approach to acquire morphological resources for Melanau by employing a large number of dispersed users from *Twitter* and *Blogs* to do a massive collective intelligence on Melanau raw data text.

1.1. Problem Statements

This research project is initiated due to the following reasons:

1. Difficulty in obtaining Melanau data to acquire morphological resources due to time consuming and expensive cost in digitizing hardcopy documents. Thus, *crowdsourcing* approach is proposed to speed up the acquisition process.
2. Informal and noisy nature of social media has led to vague data in morphological resources acquisition. Hence, *normalisation* is required to handle the obtained data to be in a standard format.

3. Uncertainty of spelling standard due to languages mixing in social media data. Therefore, *validation* is required to ensure a *similarity* in term of spelling standard with existing Melanau dictionary.

1.2. Objectives

1. To acquire morphological resources for Melan'au language through crowdsourcing approach using a web crawler method.
2. To normalise the social media data using hybrid method of rule-based and corpus-based.
3. To validate normalised results by measuring with edit distance similarity algorithms.

1.3. Methodology

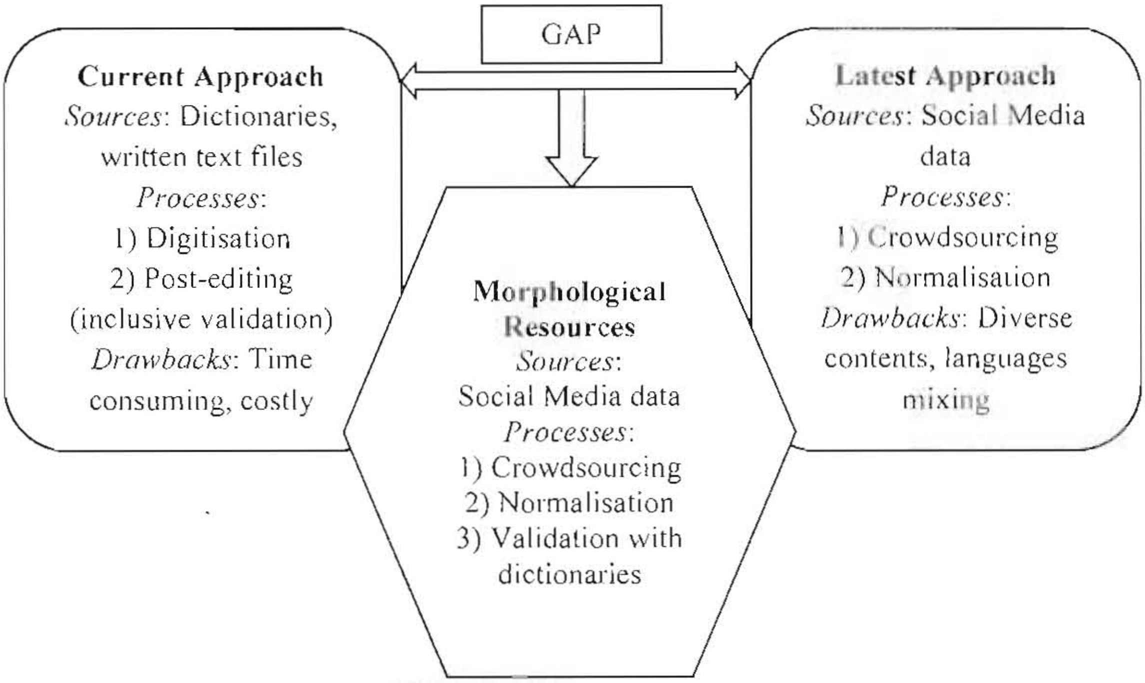
GAP
Current Approach
Sources: Dictionaries, written text files
Processes:
1) Digitisation
2) Post-editing (inclusive validation)
Drawbacks: Time consuming, costly
Latest Approach
Sources: Social Media data
Processes:
1) Crowdsourcing
2) Normalisation
Drawbacks: Diverse contents, languages mixing
Morphological Resources
Sources:
Social Media data
Processes:
1) Crowdsourcing
2) Normalisation
3) Validation with dictionaries

Figure 1.1 Current scenario

From Figure 1.1, current approach for acquiring morphological resources is from dictionaries and written text files. The processes involved are digitisation and post-editing inclusive of validation while the drawbacks are time consuming and costly. Afterward in latest approach, morphological resources are acquired through social media data. The processes involved are crowdsourcing and normalisation while the drawbacks are diverse contents and languages mixing.

From the current approach to the latest approach, there is a gap. The current is acquiring morphological resources from hardcopies and the latest is from softcopies. The digitisation process of the current is not necessary in the latest while the normalisation process of the latest is not necessary in the current either. The drawbacks from both approaches significantly will impact the quality of the acquired morphological resources. So to overcome the gap and eliminate the drawbacks, both of the approaches are combined. The new approach for acquiring morphological resources is through social media data in softcopies. The processes involved are crowdsourcing, normalisation for standard texts and validation for true words.

This project is using unified process model (UPM), one of the examples using object-oriented analysis/development (OOA/D) as its main methodology. Figure 1.2 shows the phases of UPM in the first column, their mapping with the proposed workflow in the second column, also the respective outputs from each stage of each row in the third column.

There are four phases of UPM which phase inception is to pitch the research concept, phase elaboration is to add details to early understanding of what the research should do, phase construction is to develop the prototype, and the last phase of transition is for final packaging of the system.

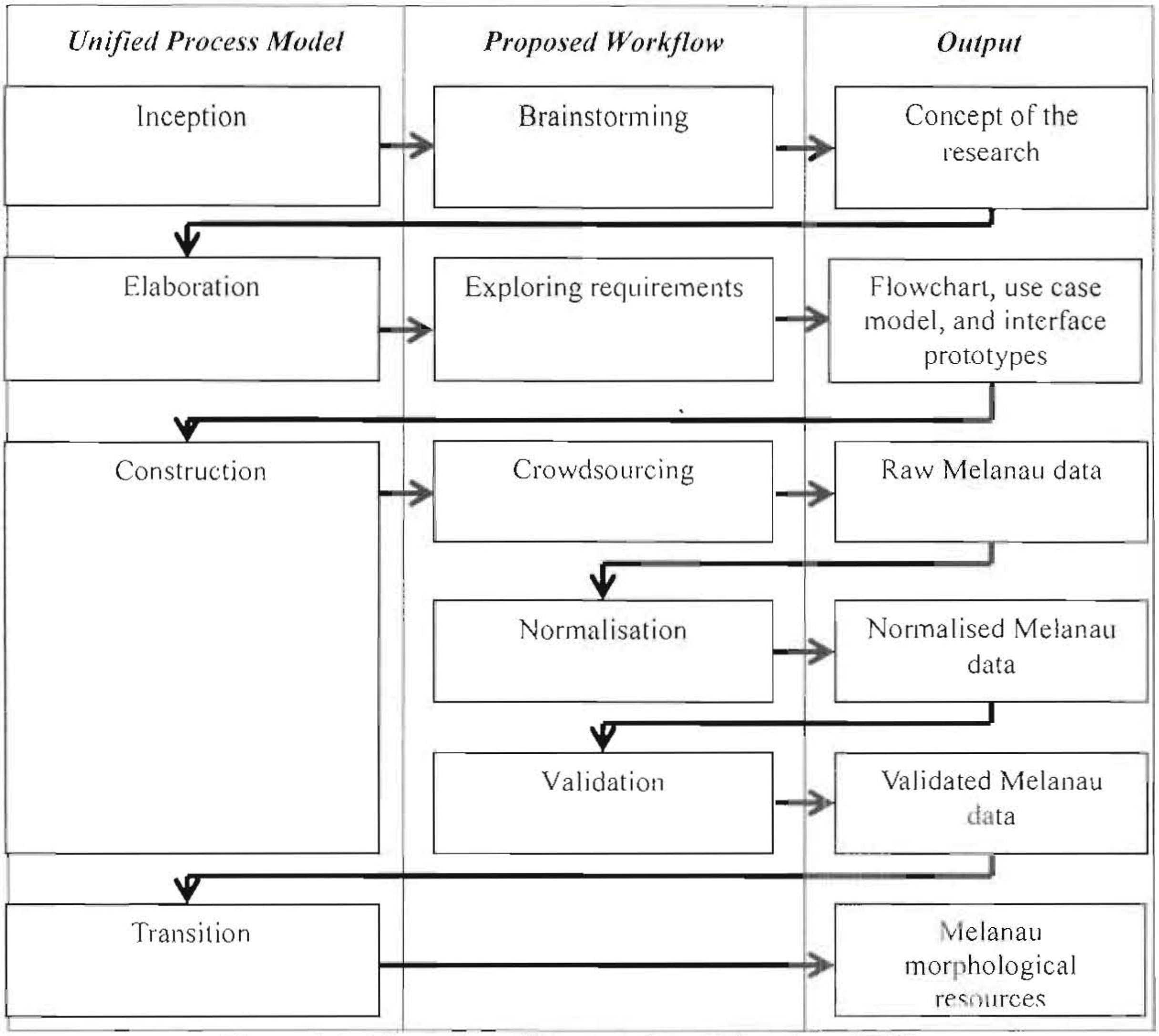

Figure 1.2 Adaptation of proposed workflow into UPM

As illustrated, each phase is mapped with one or at least one proposed stage while each stage does produce an output which is then taken as input for next stage. For instance, phase inception is mapped with brainstorming and concept of the research is developed; phase elaboration is mapped with exploring requirements then flowchart, a use case model, and interface prototypes are produced; phase construction is mapped with crowdsourcing to obtain raw Melanau data, with normalisation to normalise the crawled Melanau data, and lastly with validation to validate the normalised Melanau data; lastly phase transition producing Melanau morphological resources.

In the first stage of brainstorming, it is to get ideas of concept for the research on its purposes, inputs, processes, outcomes, and contributions. Next in the stage of exploring requirements, with the input from the first stage, three outputs which are a flowchart, a use case model, and interface prototypes are produced. A flowchart is used to establish the idea design in a diagram. A use case model is used to establish the possible activities that can be carried out. Interface prototypes are used to establish the imaginable interfaces of the system.

Moving to stage crowdsourcing, it is to obtain Melanau data from social media since the data is hardly available to get from elsewhere of only limited hardcopy documents resources possibly available. Though, it is time consuming and costing a lot to digitise them for NLP practicing. To ease the acquisition process, crowdsourcing approach is performed using a web crawler method which is installed and being customised to crawl only for Melanau language data in two main social media which are *Twitter* and *Blogs*.

As the data is obtained from publicly available sources, there will be unavoidable informal and noisy nature present which may lead to vague data. So an extensive normalisation is required to transform the crawled data to a standard format. In normalisation stage, the input crawled data is being standardised to its typical format which are those in a dictionary using a hybrid normalisation approach. For instance, when 'kmn' is crawled, it will be normalised into 'keman' (meaning: eat).

While there is always uncertainty of spelling standards in open texts caused mainly by languages mixing, the normalised data is validated to ensure their accuracy of spelling. In stage of validation, each of the normalised word will be paired with each dictionary word and computed the likelihood. For example, 'umei' is matched with dictionary word 'umai' and the likelihood between them is 75%. Automatically, 'umei' is not validated as a Melanau word and will be marked as invalidated.

1.4. Scopes

The scopes of this research are mainly focusing on five areas. Firstly, it is fully utilising social media of *Twitter* and *Blogs* to collect Melanau data. Secondly, it is crowdsourcing on general topics. Thirdly, it is normalising on crawled data to get a cleaned data. Fourthly, it is validating Melanau cleaned data for acquiring morphological resources of Melanau. Lastly, linguists will be the chief users to use the final prototype.

1.5. Significance of Project

There are three core contributions from this work towards the society especially in the further development of NLP. Firstly, the morphological resources can be used by computational linguists and NLP practitioners as a reference for future works. Secondly, the normalised social media data of Melanau language can be used to speed up the progression of data pre-processing because the data set already contains most of the existing and new pairs of nonstandard words with their standard words. Lastly, the validated Melanau words list can be served as an input for morphological processing as well as Melanau preserving.

1.6. Project Schedule

Please refer to Appendix A for project schedule 2014/2015.

1.7. Expected Outcomes

From this paper, there are three outcomes expected to be composed at the end of the research. First, they are morphological resources for Melanau language which will be served as input for morphological processing in further stage. Second, it is a set of normalised social media Melanau data which can be constantly expanded if more pairs of nonstandard words and their standard words are being identified. Third, it is a validated Melanau language words list which can be constantly expanded as well and be contributed to the Melanau morphological resources. Moreover, both of the normalised data and validated words list can be served as input for morphological analyser in NLP applications.

1.8. Thesis Organisation

This whole paper is divided into six chapters and being organized as followed.

Chapter 2 of literature review discovers on previous studies and works on acquisition of morphological resources through crowdsourcing, normalisation, and validation. This chapter identifies various ways to conduct the three processes brought up and the significance results of each way to be taken as references and guides for the current research work.

Chapter 3 of methodology explains the construction of the system prototype for the research to carry out the crowdsourcing, normalisation and validation at a single location. The construction is being clarified from the perspectives of requirement analysis on inputs, outputs, and tools needed; planning, design, and procedures developing.

Chapter 4 of implementation illuminates the steps and tools to develop this system and illustrates the application of the developed prototype with real data of Melanau language.

CHAPTER 2 LITERATURE REVIEW

Until recently, there are more and more studies have been done on U-RL in NLP field particularly of building corpus and WordNet, machine translation, morphological disambiguation, and utilising of social media on this field. Nonetheless, almost all of the researchers are concerning on the post-working part and fewer is really emphasising on the pre-working part like the origin of the input data, in terms of what, where, when, and how to obtain it. In this chapter, a sequence of steps on how to acquire morphological resources of Melanau will be discussed. Although the literature covers a wide range of such studies, in this chapter will only focus on three major activities which are crowdsourcing, normalisation, and validation. Before going in deeply to each of the activity, an overall sum-up for this chapter is presented in Figure 2.1 for the first insight of readers.

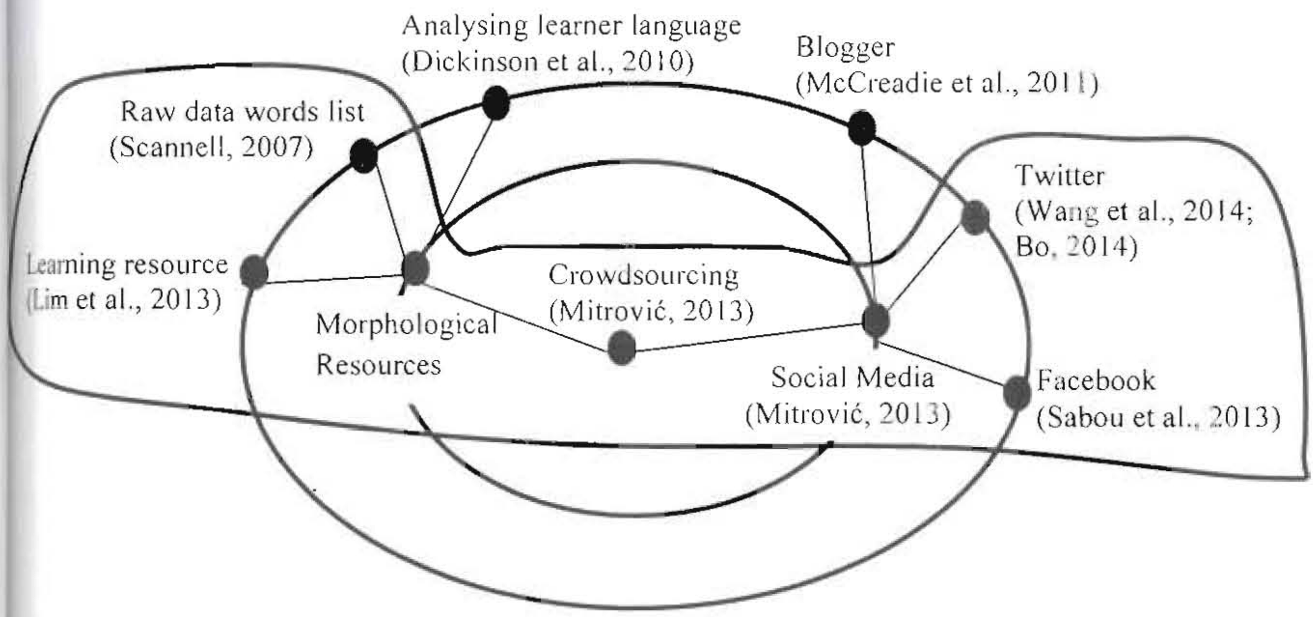

Figure 2.1 Summary of literature review

First of all, Figure 2.1 summarises how other researchers contributed in the similar works. The idea was starting from crowdsourcing on social media to acquire morphological resources. The researchers crowdsourced on *Facebook* and *Twitter* mostly for their widespread and strong influences among their users. They believed crowdsourcing on these two websites should return more valuable data than other social media. On the other hand, the morphological resources acquired are in the form of raw data words list and serves as learning resources. In the simplest form of raw data words list, it can be easily passed on to other NLP applications for further processing.

2.1. Challenges in Morphological Resources Acquisition

Morphological resources are same with corpuses, which amass a series of collection of writings for a language. Both of them serve the same purposes which are to analyse learner language (Dickinson et al., 2010), as learning resources (Lim et al., 2013), and to be further analysed and developed into lexical resources (Wang et al., 2014a, 2014b). While the differences between morphological resources and corpuses are those former undergo stages not beyond part of speech (POS) tagging, definition giving, and sense tags assigning to become advanced resources. As most of the U-RL is in its slowly vanishings among the native speakers, the current research is working on to preserve Melanau language as discussed in Chapter 1.

2.2. Social Media as Possible Sources

There is no doubt that social media are always fantastically popular and vary in form. The phrase 'social media' covers widely which can range from user forums, weblogs (*Blogs*), social blogs, wikis, picture blogs such as *Instagram*, microblogs for instance *Twitter*, and to social networking websites like *Facebook*. Though the posts are short especially the tweets might be the shortest, the contents of the posts are highly diverse to be in different domain and context for extracting varieties of words. Thus taking social media as sources like what Juan et al. (2014) did, for acquiring morphological resources would be beneficial to have a larger collection of words in domains of economy, financial, politic, education, sports, communities, and even daily conversations. The process of utilising the social media through crowdsourcing is being discussed in the next section.

2.3. Common Practices in Morphological Resources Acquisition

2.3.1. Crowdsourcing Social Media using Web Crawlers

Howe (2006) introduced the term crowdsourcing as a business practices that literally outsourcing for lower cost and masses. However, NLP researchers for instance Grier (as cited in Mitrović, 2013, p. 38) explained crowdsourcing was actually "using the Internet to employ large numbers of dispersed workers" for "a special case of such Collective Intelligence" (Buecheler et al., 2010).

Crowdsourcing is on its way to gain more and more popularity in these recent years among NLP researchers for obtaining data freely. Previously, most crowdsourcing was carrying out over Internet to collect text data by externally approaching to crowd through questionnaires (Wang et al., 2014a) or internally crawling over corpora (Wang et al., 2014b), other lexical resources, and social media such as *blogger* (McCreadie et al., 2011), *Facebook* (Sabou et al., 2013) and *Twitter* (Wang et al., 2014; Bo, 2014).

Different from the above, there were also some researchers practiced web crawling using open sourced web crawlers. For instance, *crawler4j* was the most popular and widely applied by researchers in different fields. Examples were: Chang et al. (2014) applied *WebHarvest* and *crawler4j* to collect data from social network, and Kumar Tak and Ojha (2013) used *crawler4j* to detect URL details to further enhance the precision of detection of a compromised site.

Among the many methods for crowdsourcing, one thing to note about was that they were resource hungry. For this research, crowdsourcing is performed on *Twitter* and *Blogs* with the aid of a web crawler to collect general posts for Melanau language as both of them are large enough to obtain necessary data in vary domains and contexts. Even the trend of using digital media to help preserving underrepresented and minority languages has caught attention of Rising Voices (*Twitter to promote and preserve underrepresented languages*, 2011) as well as computer science professor Scannell (2007).

Despite the *crawler4j* used by Chang et al. (2014) and Tak and Ojha (2013), there are many other web crawler tools available such as *Screaming Frog Spider 3.3*, *Web Miner*, and *BrownRecluse*. Meanwhile, different web crawler works in different ways. Here, three web crawlers are suggested:

i. **Screaming Frog Spider 3.3** (*Screaming Frog SEO Spider Tool*, 2015), is an open-sourced web crawler using a small desktop program which anyone can install locally on working machine to crawl websites' links. It fetches key onsite elements, presents them in tabs by type and allows users to filter or slice and dice the data by exporting into Excel. Users can view, analyse and filter the crawl data as it's gathered and updated continuously in the program's user interface.