CHAPTER 1

INTRODUCTION

This chapter includes the following subtopics, namely: 1) Rationale; 2) Theoretical Framework; 3) Conceptual Framework/Paradigm; 4) Statement of the problem; 5) Hypothesis; 6) Assumption; 7) Scope and Delimitation; 8) Importance of the study.

1.1 Rationale

Languages with rich morphology or high of inflection word such as Arabic, German, Spanish, have a core raises problem. The problem is difficulties to find a single word form [1][2]. Where in language with rich morphology a lexical item can appear as a form with highly variation in a corpus. This large variation can reduce the possibility of finding a single word form and reduce the effectiveness of natural language processing tasks [3][1].

In general, many languages use morphological marking in the form of affixes (i.e., suffixes, prefixes, and infixes) to show syntactic and semantic differences. For example, in English have additions to indicate the singular and plural forms (e.g., read and reads). The other example in Bahasa Indonesia there is additions to indicate active or passive words (e.g., masak -> me-masak (active) or di-masak (passive), these are known as inflected forms. English and Bahasa Indonesia have a quite a bit of inflectional morphology. Arabic with rich morphology have more complex way to inflecting words or generating new words by given some rules.

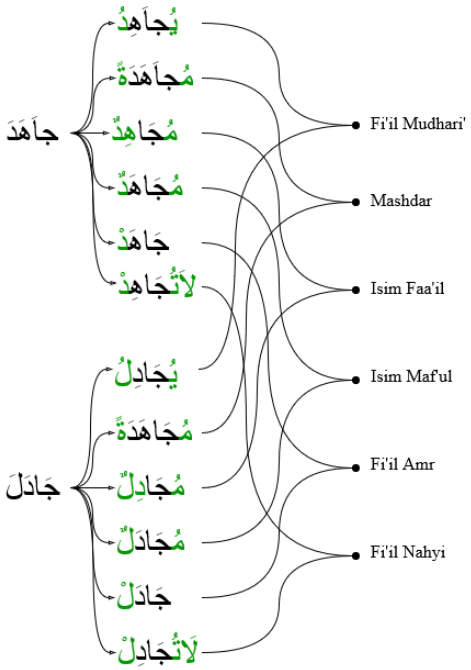

Figure 1: The relatedness of inflection forms in Arabic. The green colored word shows the difference with the previous words. The additions of green color must follow some of the rules in Arabic morphology. On the far right is the similarity of tenses or forms.

To handle the problem of finding a single word form, some techniques like lemmatization [4] [5] or stemming could be done for some languages. These techniques are not very suitable for Arabic. Arabic need system that can accurately learn and capture the mappings of the rules and the principles, to maximizing the capability of most human language technologies. Before that, we must know the principle and the linguistic of word formation in Arabic.

There is study of the rules to forming and changing words in Arabic called Sharaf or *shorof*. Sharaf discuss the basics of word formation and provides rules known as Morphology before they are combined or assembled with other words [6]. Word formation in Arabic is very diverse and flexible, in Sharaf the way forming word called *tashrif.* In *tashrif*, we could generate word by *tashrif lughawi* (inflection) and *tashrif istilahy* (derivation). Arabic has three important aspects or elements for word formation, they are class words (*fi'il*, *isim*, hurf), *wazan* (patterns), and *dhamir* (pronouns) [7][8][9].

The word class in Arabic consists of three type: *fi'il* / verb, *isim* / noun, and *hurf* / conjunction. For example, there is sentence "يَذْهَبُ الْمُعَلِّمُ إِلَى الْفَصْلِ " (The teacher goes to class) then يَذْهَبُ (*yaḏhabu:* 'goes') is *fi'il*, الْمُعَلِّمُ (*al-muʿallim:* 'teacher') and الْفَصْلِ (*al-faṣli:* 'class') are *isim* and إِلَى (*ʾilā:* 'to') is *hurf.* Generally, nouns and adjectives include in *isim*. Not all adjectives are isim. Because, there are adjectives in Arabic that include in the group of verbs (*fi'il*) [7]. Each word class has one *dhamir*, that is why *dhamir* is one of the important elements in Arabic word mapping [9]. *Wazan* is verb-form or pattern for word formation in Arabic[7], and the patterns can affect the meaning or semantics of the words. Therefore, it would be good if there is a system that can correctly doing Arabic word formation by given some rules and the important elements in Morphological Arabic.

Research of the mapping of infected forms has been purposed by Kann and Schutze [10]. They propose morphological reinflection techniques to handling the single word form problem called MED. The experiments were conducted in 10 languages (Arabic, Finnish, Georgian, German, Hungarian, Maltese, Navajo, Russian, Spanish, and Turkish). It means that the system is multilingual or general. The reinflection process is process where one input word is given in the form of an inflected word, and the system need to generate another form of the word based on certain rules. They get more than 95 percent accuracy for Arabic. However, the model that was built did not include *wazan* (verb-form) which is one of important aspects in Arabic word mapping.

Other research on word mapping for Arabic has been purposed by Khaled and Shaalan [11], [12] using rule-based method. This research results are pretty good by including *wazan* on the system but it has problems on number-counted result. This rule-based approach also use Arabic without the diacritics (or wolves) and it is essential for the disambiguation of words. For example, the word بر هان can become بُرْهَانٌ (*burhān:* 'a convincing proof') or بُرْهَانَ (*burhāna: '*the proof'). Therefore, the diacritics (or wolves) is important to reduce disambiguation.

Rather than a general system, this research aims to create a system that is more specific to one language Arabic. MED [10] have good accuracy on Reinflection process without *wazan* for multilingual, while in this study *wazan* is added to the feature and use full diacritics to avoid the disambiguation.

The method will use Recurrent Neural Network (RNN). Based on SIGMORPHON 2016[1], 2017[13] and 2018 [14] RNN give the best results compared to other methods for Reinflection process. In other words, this study purposes a model that could accurately generate morphological reinflection word in Arabic by an emphasis on the morphological aspects of Arabic.

1.2 Theoretical Framework

This research purpose two tasks for reinflection process. The first task (Task 1) is given three inputs namely source form, MSD-source, and MSD-target. The second task (Task 2) is given only two inputs namely source form and MSD-target without MSD-source. The source form is Arabic word, and the output for both of the tasks is reinflected Arabic word. The example of the input-output using English and Arabic are on the Table 1.1.

Table 1.1 Example of Input-output System

		Task 1	**Task 2**
Input	***Source form***	writing	writing
	MSD-source	pos=V,tense=PRS	-
	MSD-target	pos=N,num=PL	pos=V,tense=PST
Output	***Target form***	writer	wrote
Input	***Source form***	عَبِّدُو (*abbidū)*	عَبِّدُو (*abbidū)*
	MSD-source	pos=V,mood=IMP,gen=MASC,per=2, voice=ACT,num=PL,form=2	-
	MSD-target	pos=V,mood=IND,tense={PRS/FUT}, gen=FEM,per=2,voice=PASS,num=SG, aspect={IPFV/PFV},form=2	pos=V,mood=IND,tense={PRS/FUT}, gen=FEM,per=3,voice=ACT,num=DU, aspect={IPFV/PFV},form=2
Output	***Target form***	تُعَبَّدِينَ (*tuʿabbadīna)*	تُعَبِّدَانِ (tuʿabbadāni)

The MSDs are the specific information about the words. Table 1.2 shows the MSDs of type of words, *wazan* (pattern), and *dhamir* (pronoun) in Arabic. More explanations about the MSD will explained later on 2.2.6.

Table 1.2 The contents of the MSDs in Arabic

Aspects	**MSDs**
Type of words	pos={V/N},mood={IND/CON/IMP/SBJV},aspect={IPFV/PFV},voice={ACT/PASS}
Dhamir	GEN={M/F},PER={1/2/3},NUM={SG/DU/PL}
Wazan	form={I-13} / form={Iq/2q/3q/4q}

1.3 Conceptual Framework/Paradigm

The previous study or the MED by Kann and Schutze [10] did not include *wazan* (verb form or pattern) to their model, which is *wazan* is one of the important elements to word formation in Arabic[7]. And the previous research build the model as a multilingual model. This research wants to specific to one language, and it is Arabic. To specific the model, *wazan* added together with the type of word and *dhamir* as the feature (MSD) of the dataset. Additional *wazan*

Table 1.3 Wazan must be known before generate new word

No.	Input			Output
	Source form	MSD-Source	MSD-Target	Target form
1.	أَبْرِقِي *(ʾabriqī)*	pos=V mood=IMP gen=FEM per=2 voice=ACT num=SG form=4	pos=V mood=IND tense=PST gen=FEM per=2 voice=ACT num=DU aspect={IPFV/PFV/PRF} form=4	أَبْرَقْتُمَا *(ʾabraqtumā)*
2.	اِكْتَحِلَا *(iktaḥilā)*	pos=V mood=IMP gen=MASC per=2 voice=ACT num=DU X	pos=V mood=SBJV gen=FEM per=3 voice=ACT num=SG X	?

Word formation in Arabic need to use all three important elements (type of words, *wazan, dhamir*). On the Table 1.3 the word with green color is the *wazan* (verb-form). Therefore, *wazan* must be added to the feature (MSD) before the new word is generated to get the pattern of the form. All the *wazan* on dataset added manually by the author based on guide by book [7], Wiktionary and Corpus Quran, also confirmed by the expert.

1.4 Statement of the Problem

Based on the background and the lack of the previous research, the problem of this research is how to build model that can accurately generate morphological reinflection word in Arabic by the MSD from the word and also its *wazan* (verb form).

1.5 Objective and Hypothesis

The purpose of this study is to build a model that could accurately generate word in Arabic by an emphasis on the morphological aspects of Arabic. The objective is to handling the problem of finding single word form.

To support the reinflection model specific to Arabic, *wazan* added together with type of word and *dhamir* as the feature (MSD) of dataset. Recurrent Neural Network (RNN) approach used to capture the sequence information and hope can help get better generated word.

1.1 Assumption

Using RNN sequence-to-sequence (seq2seq) can accurately generate morphological reinflection word in Arabic by the MSD of the word. Using *wazan* as the additional feature can make the model better than previous study.

2.1 Scope and Delimitation

This study follows some rules, namely:

1. Type of words that can be process include *fi'il* (verb) and *isim* (noun)
2. Use all 14 *dhamir*
3. Use only 17 *wazan* from 35 *wazan*.
4. The Arabic word full of diacritics (*harakat*)

3.1 Importance of the Study

The results of this study can be used to facilitate Muslims who want to study the Quran or those who want to learn Arabic. In the field of linguistics, reinflection results can also be used for the development of morphological analysis or other study in the field. The addition of *wazan* to the dataset can also be used as additional resources for the future studies.