Development of Computational Linguistics Research: a Challenge for Indonesia

Bobby Nazief, Ph.D.
Computer Science Center, University of Indonesia
Jakarta, Indonesia
nazief@cs.ui.ac.id

1 Introduction

The emergence of Internet as a global information repository, where information of all kind is stored, requires intelligent information processing tools (i.e., computer applications) to help the information seeker to retrieve the stored information. To build these intelligent information processing tools, we need to build computer applications that understand human language since most of those information is represented in human language. This is where computational linguistics becomes important, especially for countries like Indonesia that hosts more than 200 million people. We need to develop a systematic understanding of the Bahasa Indonesia (the Indonesian national language) to enable us develop the needed computer applications that will help us manage information intelligently.

However, until recently, there is only a few research activities in computational linguistics conducted in Indonesia. The establishment of Computer Science departments in Indonesian universities that did not start until the beginning of 1980's may be partially responsible for this[1]. In addition, the Indonesian linguists seem to be keen on working "manually" instead of using computers in conducting their linguistics researches as stated in Muhadjir (1995), only few of them really make use of the technology. While, on the other hand, most computer scientists tend to use the practical approach rather than constructing a complete framework to understand the language when building related applications such as a specific information retrieval system.

In the following, I will describe past research activities in computational linguistics on Bahasa Indonesia. This description is by no means exhaustive since it is very difficult to find out research activities in computational linguistics in Indonesia.

2 Past Research Activities

2.1 Corpus Analysis

Corpus analysis is an important means as a way to understand the evolution of language usage by its people. In the case of Bahasa Indonesia, research activities on corpus analysis were almost none. There was one work by R. R. Hardjadibrata (1969) from Monash University, who conducted word frequency analysis of Indonesian newspapers. There was also similar work conducted the MMTS project (will be described later, in the following section); however, the result of the group's corpus analysis was not made public.

Given this condition, with a group of colleague both from the Faculty of Computer Science and the Faculty of Letters, I conducted an Indonesian corpus analysis using newspapers as the text source. We collected 52 editions of *Kompas*, a national newspaper with a large number of readers, published in the year of 1994. Each of the 52 editions corresponds to a particular week of the year and was taken randomly from the 7 daily editions of that given week. From this collection, we constructed a corpus consisting of 2.200.818 words that were formed by 74.559 unique words. Of these more than 2 million words, 1.826.740 words that were formed by 27.738 unique words are actually words that matched with the KBBI[2] entries, while the rest are either names or foreign words. Detailed analysis can be found in Muhadjir (1996).

[1] Bandung Institute of Technology was the first among public universities that established Computer Science department in 1980.

[2] KBBI (Kamus Besar Bahasa Indonesia), the standard word dictionary for Bahasa Indonesia, contains a little more than 70.000 word entries.

2.2 Morphological Analysis

Everyone who has used a word processor understands the importance of a spelling checker in helping him/her to produce an error-free document. To develop a spelling checker, we need to understand the morphological structure of words especially how derived-words are constructed from their root-words and the addition of affixes.

We have conducted research to analyze the morphological structure of Indonesian words and based on this analysis we have developed a stemming algorithm suitable for those words. Unlike English, where the role of suffix dominates the generation of derived-words, Bahasa Indonesia depends on both prefix and suffix to derive new words. Therefore, to stem a derived Indonesian word in order to obtain its root-word, we have to look at the presence of both prefix and suffix in that derived-word (Nazief, 1996). In addition, similar to English, multiple suffixes can also be present on a given derived-word.

Based on this stemming algorithm, we have developed a spelling checker and spelling-error corrector utilities as part of the Lotus Smartsuite[3] package.

2.3 The MMTS Project

One notable research activity among the few computational linguistics research activities in Indonesia is the Multilingual Machine Translation System (MMTS) project conducted by the Agency for Assessment and Application of Technology (BPPT) as part of multi-national research project between China, Indonesia, Malaysia, Thailand, and lead by Japan (see http://www.cicc.or.jp/homepage/english/about/act/mt/mt.htm, http://www.aia.bppt.go.id/mmts). Unfortunately, there are very few publications about this work that could have benefited the computational linguistic community in the country. One of the few publications that the MMTS project made available for public is the Indonesian Word Electronic Dictionary (KEBI), which could be accessed on-line on http://nlp.aia.bppt.go.id/. The dictionary contains 22.500 root-word and 43.500 derived-word entries.

3 Understanding Indonesian Grammar

Currently, I am concentrating my work on developing syntax analyzer for sentences written in Bahasa Indonesia. The approach taken initially was to use the context free grammar with restriction such as that used in *the linguistic string analysis* (Sager, 1981). Using this approach, we have developed grammar that understands declarative sentences (Shavitri, 1999). However, our experience shows that we need to have a more detailed word categories than is currently available in the standard Indonesian word dictionary (KBBI) before the grammar can be used effectively.

This finding really shows us the importance of collaborating with the linguists who understand this field better. But before we do this, we need to educate our linguist-fellows the importance of computer in their fields.

4 Acknowledgements

I would like to thank Mirna, bu Multamia, pak Muhadjir, bu Kiswartini, and all of my students who have collaborated with me in these efforts to understand Bahasa Indonesia better.

5 References

R. R. Hardjadibrata (1969) *An Indonesian Newspaper Wordcount*. Department of Indonesian and Malay, Faculty of Arts, Monash University, Cayton, Victoria.

Muhadjir (1995) *Menjaring Data dari Teks*. Lembaran Sastra Universitas Indonesia, edisi khusus:Tautan Sastra & Komputer, Faculty of Letters, University of Indonesia, Depok, pp. 81--91.

Muhadjir, et. al. (1996) *Frekuensi Kosakata Bahasa Indonesia*. Faculty of Letters, University of Indonesia, Depok, 207 p.

Bobby A. A. Nazief and Mirna Adriani (1996) *Confix Stripping: Approach to Stemming Algorithm for Bahasa Indonesia*. Internal publication. Faculty of Computer Science, University of Indonesia, Depok.

Naomi Sager (1981) *Natural Language Information Processing: A Computer Grammar of English and Its Aplications*. Addison-Wesley Publishing Company, Massachusetts.

Shelly Shavitri (1999) *Analisa Struktur Kalimat Bahasa Indonesia dengan Menggunakan Pengurai Kalimat Berbasis Linguistic String Analysis.* Bachelor's Thesis. Faculty of Computer Science, University of Indonesia, Depok, 88 p.

[3] Lotus Smartsuite is an office automation package consisting word processor, spreadsheet, presentation editor, and database applications developed by Lotus Development Corporation.