A survey of bahasa Indonesia NLP research conducted at the University of Indonesia

Mirna Adriani and Ruli Manurung
Faculty of Computer Science
University of Indonesia
mirna@cs.ui.ac.id, maruli@cs.ui.ac.id

Abstract

The Faculty of Computer Science, University of Indonesia, has been conducting research concerning the computational modelling and processing of various aspects of bahasa Indonesia since the mid-1990s. This paper serves as a historical account of that work, as a convenient bibliographic reference, and as a platform from which to identify future research.

1. Introduction

Bahasa Indonesia (hereinafter simply Indonesian) is the official language of Indonesia, spoken by well over 100 million people. Given this fact, we believe it is underrepresented in terms of Indonesian natural language processing research. There is a great need for Indonesian-aware NLP resources and applications, i.e. lexicons, parsers, dialogue systems, machine translation, information retrieval systems.

This paper serves as a historical account of work into this area, specifically work carried out at the Faculty of Computer Science, University of Indonesia, henceforth simply Fasilkom[1]. Thus, whilst acknowledging the work carried out at other institutions, this paper is not intended as a comprehensive bibliography of all Indonesian NLP research. See [33], in which Bobby Nazief presents a snapshot of the state of Indonesian computational linguistics circa 2000.

In Section 2 we discuss work in the development of basic natural language processing resources and tools, covering lexicography, morphological analysis, syntactic parsing, and semantic analysis, whereas in Section 3 we present a wealth of research work that applies these tools for Information Retrieval applications, including cross-language retrieval, information extraction, document summarization, and question answering. We then discuss the current and future research agenda at Fasilkom.

[1]This is the widely used abbreviation of *Fakultas Ilmu Komputer*

2. Development of NLP resources and tools

2.1. Computational lexicography

Work on Indonesian NLP resources at Fasilkom began in 1995, with research into the construction of various Indonesian lexical resources, i.e. dictionary, thesaurus [10]. One of the earliest lexical resource created was a word-frequency dictionary based on a corpus of newspaper articles [29], developed in conjunction with the Linguistics department at the Faculty of Arts, University of Indonesia. This dictionary contained word frequencies of basic words (lemmas) and their affixed words that appeared in the newspaper.

The work was expanded with the construction of an electronic database version of the *Kamus Besar Bahasa Indonesia* (KBBI), considered by many to be the authoritative Indonesian lexicon, in a collaboration with the KBBI developers, i.e. *Pusat Bahasa*, the Indonesian government's centre for language development.

These various resources spawned subsequent research work concerning Indonesian spelling correction. Experiments to identify the most efficient data structures and compression techniques to support this task are reported in [34, 5]. Other aspects researched include lexical organization [15, 16]. An algorithm to identify proper names in a text to improve the spelling checker performance has also been studied [42]. Suggestion words were given by identifying the stem word of misspelled words using the vocal-consonant pattern and syllable pattern [9] and using word error tolerance [7]. This work also included a collaboration with Lotus Corporation to develop an Indonesian spelling checker for the Lotus SmartSuite office productivity application.

2.2. Morphological analysis

Closely related to lexicography is the area of Indonesian morphological analysis, which has also received much attention at Fasilkom. One strand of work focused on various

Surface form	Formation	Tags
memukul	meN- + pukul	pukul +V +Act +Trans
pukulan	pukul + -an	pukul +N
berpukul-pukulan	ber-an + pukul + redup	pukul +Verb +Recip

Table 1. Example morphological analysis

stemming algorithms and techniques [43, 3], which seek to recover the stem word from a word that has undergone an affixation process. For example, the words *memukul*, *pukulan*, and *berpukul-pukulan* all share the same stem *pukul*. The stemming algorithm was developed based on Indonesian morphological structure, and made use of a dictionary containing only lemmas. If a word being searched was not found in the dictionary then the prefix, suffix, and infix of the word was removed according to the morphology rules. On the other hand, a word generator was also developed to produce affixed words based on a lemma [34]. Stemmers are a crucial element in most information retrieval systems (Section 3).

A corpus-based stemming approach is presented in [17, 18]. A word was stemmed based on word variants that appeared in a corpus to get the correct stem words in certain domains.

Other work explored more complex models which recover not just the stem but also various syntactic information. In the latter case, one particular approach is two-level morphology, discussed in [12]. Whereas stemmers discard the affixes, a two-level morphology analyser transduces them into syntactic tags that would be useful for subsequent syntactic and semantic processing. Table 1 shows an example of such an analysis.

This work has recently been revived, with the intention of constructing a wide-coverage and linguistically-motivated working system (see Section 4).

2.3. Syntactic Parsing

The first work at Fasilkom on parsing Indonesian dates back to 1995 [37], in which an initial grammar for Indonesian was written using LL and LR-regular languages, formalisms more commonly used for defining computer programming languages, and widely acknowledged to lack the expressiveness required for natural language.

This work was continued in [14, 36], where context-free grammars were constructed based on prescriptive rules defined in the *Tata Baku Bahasa Indonesia*, or official grammar of Indonesian [6].

In [28, 40, 45], and latterly [21], the previous work was extended with more sophisticated linguistic analysis, namely an attempt to account for subcategorization and selectional restrictions. The grammars utilized feature structures in the PATR-II formalism [41].

However, these works by and large contained many linguistic infelicities, as the overall goal of the research seemed to be to achieve high accuracy on a given corpus, with little regard to true wide-coverage and linguistic generality.

For example, a single SUBCAT feature was used to account for all linguistic issues. Whilst succeesful in correctly modelling Indonesian sentences (with a reported $> 90\%$ accuracy for certain domains in [21]), the phrase structure rules were often duplicated for various combinations of valid SUBCAT features between the constituents, thus defeating the purpose of the unification approach to elegantly capture linguistic regularities. Another issue was that the lexicon was directly derived from the corpus, with large clauses sometimes being analysed as a single lexical entry[2].

2.4. Semantic and discourse analysis

Less attention has been paid to semantics than for the previous aspects. Some initial work has been done in Indonesian lexical semantics, i.e. the study of associating meaning representations to Indonesian words. For example, [24] reports initial experiments in developing an Indonesian thesaurus, where semantic similarity between words was obtained using a vector space model derived from a corpus. [20] presents a design for an Indonesian semantic lexicon. This work is now being continued by an ongoing project whose goal is the construction of an Indonesian WordNet (see Section 4).

Semantic analysis of Indonesian sentences is reported in [22] and [23], where a Montagovian compositional semantics approach is taken to construct simple logical representations of Indonesian declarative and interrogative sentences. Unification-based inference is then applied on these representations to answer simple factoid questions based on the discourse history.

Research into statistical 'shallow' techniques for discourse and semantic analyis has also been done, e.g. named entity recognition [48, 19] and coreference resolution [46].

3. Information retrieval applications

3.1. Cross-Language IR

We have done some work on Cross-Language Information Retrieval using Indonesian queries in retrieving English documents. In order to solve the language barrier, we experimented with translating the Indonesian queries using bilingual dictionaries, machine translation, and parallel corpus techniques [13]. Translating Indonesian queries into

[2] a particularly extreme example is the following 'noun': *"keadilan berdasarkan ketuhanan yang maha esa"*.

English using bilingual dictionary and machine translation worked well, however using parallel corpus did not perform very well. A parallel corpus was used to find word association between Indonesian and English, then it was used to translate Indonesian queries into English [38]. We also applied a transitive translation between Indonesian and English which made use of pivot languages, where Indonesian queries were translated into German and/or French first before they were translated into English [39].

3.2. Document Summarization

Our work on summarization studied how to extract sentences from Indonesian documents based on cue phrases and the weight of a word in a sentence [44]. Summarization was also done by considering the query that has been submitted. The results showed that summaries containing a group of sentences related to the query words are better than producing summaries without considering any words of interest to the user [27].

3.3. Question Answering

In Question Answering topic, we identified question types based on the question words used in Bahasa Indonesia. We extracted the answer from the paragraph. We scored each paragraph containing the candidate answer based on the location of the candidate answers and query words that appear in the paragraph [31, 32]. In Cross-Language Question Answering, we used Indonesian queries and extract the answers from English documents. We combined linguistic knowledge and external resources found on the Internet to get the answer [50, 1, 2].

3.4. Geographic Information Retrieval

We studied how to find information on events occurring in certain locations in Geographic Information Retrieval (GIR). We developed a location parser to identify any location name that appears in Indonesian queries and documents. Then we used geographic relation words to identify events that happen in certain location based on the geographic coordinates [35, 4]. For Cross-Language GIR we applied a query expansion technique to improve the performance of finding documents containing geographic events [30].

3.5. Information Extraction

Important information in a document could be identified and extracted for some purposes using various techniques in the field of information extraction. For the first step in our information extraction work for Indonesian documents, we developed a named-entity tagger to identify person, location and organization names. The named entity tagger was developed using a machine learning approach with association rules [49, 26]. The rules used morphological and part-of-speech information [48] to recognize the named entity word. Words that have named-entity tags can have relations defined in relation templates. The relation is identified using a machine learning technique and the vector space model [25].

4. Current work, and the road ahead

Currently, the IR lab at Fasilkom is actively engaged in several research projects, including collaborations with other institutions, both within and outwith Indonesia. Some of the aspects being studied are:

- **Speech recognition.** Leveraging our years of experience with Indonesian language models, we are currently developing acoustic models trained on a large multi-speaker speech corpus, and investigating the suitability of applying these models to existing open-source speech recognition systems such as JULIUS[3] and SPHINX[4].
- **Corpus-based NLP tools for Information Retrieval.** In a joint collaboration with NUS and USM, various resources and algorithms are being researched for large-scale Malay and Indonesian information retrieval using corpus-based methods. Interim results include the development of a statistical part-of-speech tagger.
- **Construction of an Indonesian WordNet.** An ongoing project is concerned with the development of an Indonesian WordNet. Using the expand model approach discussed in [47], we first map Princeton WordNet [11] synsets to existing word sense definitions in the KBBI, which defines semantic equivalence classes between KBBI senses. These classes are collapse to form Indonesian synsets, and semantic relations between them are transferred over from Princeton WordNet.
- **Finite state morphological analysis.** Given that Indonesian (and Malay) morphology is fairly complex, including prefixes, infixes, affixes, confixes, and reduplication, there is still work to be done. Additionally, reduplication is an example of non-concatenative morphology, which formally cannot be modelled by finite state techniques, the prevalent paradigm for modern morphological analysis. An ongoing collaboration with the University of Sydney seeks to develop a wide-coverage morphological analyser using two-level morphology, which outputs syntactic information required

[3] http://julius.sourceforge.jp/en
[4] http://cmusphinx.sourceforge.net

for an Indonesian grammar using Lexical Functional Grammar [8].

5. Summary

This paper has given a historical account of research work into Indonesian natural language processing and information retrieval applications carried out at the Faculty of Computer Science, University of Indonesia. Together with other institutions, both within and outwith Indonesia, we aim to continue developing resources, tools, and fully working applications, which enable the development of computer software that processes Indonesian language artifacts intelligently. More details about our work can be found at `http://ir.cs.ui.ac.id` and `http://bahasa.cs.ui.ac.id`.

References

[1] S. Adiwibowo. Penemuan jawaban pada sistem tanya jawab bahasa indonesia-inggris dengan pembobotan kata dan informasi dari internet. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: SK-0699.

[2] S. Adiwibowo and M. Adriani. Finding answers using resources in the internet. In *Working Notes of the Workshop in Cross-Language Evaluation Forum 2007*, Budapest, Hungary, September 2007.

[3] M. Adriani, J. Asian, B. Nazief, S. Tahaghoghi, and H. E. Williams. Stemming indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*, 6(4), December 2007.

[4] M. Adriani and M. L. Paramita. Identifying location in indonesian document for geographic information retrieval. In *Proceedings of the Workshop on Geographic Information retrieval (CIKM 2007)*, Lisbon, Portugal, November 2007.

[5] I. A. Agus. Pemilihan strategi dan algoritme pemampatan data teks untuk kamus besar bahasa indonesia elektronik. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1997. Call number: SK-0369.

[6] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. Moeliono. *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, Indonesia, third edition, 1998.

[7] W. Ariesusetyo. Implementasi algoritme pelacakan string dengan toleransi kesalahan pada modul pengoreksi ejaan bahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1997. Call number: SK-0352.

[8] I. W. Arka, J. Simpson, A. Andrews, and M. Dalrymple. Challenges of developing a balanced indonesian corpus. In *Proceedings of the Eleventh International Symposium on Malay/Indonesian Linguistics (ISMIL 11)*, Manokwari, Indonesia, August 2007.

[9] I. M. Artana. Pengoreksi ejaan bahasa indonesia untuk kata turunan dengan kesalahan pada afiks. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1998. Call number: SK-0376.

[10] Fakultas Sastra Universitas Indonesia. *Tautan Komputer dan Sastra: Prosiding Seminar Penerapan Teknologi Komputer pada Penelitian dan Pengembangan Bahasa (Edisi Khusus Lembaran Sastra Universitas Indonesia)*, Depok, Indonesia, January 1995.

[11] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[12] H. Hartono. Pengembangan pengurai morfologi untuk bahasa indonesia dengan model morfologi dua tingkat berbasiskan pc-kimmo. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2002. Call number: SK-0516.

[13] H. Hayurani, S. Sari, and M. Adriani. Evaluating language resources for indonesian-english clir. In *Working Notes of the Workshop in Cross-Language Evaluation Forum 2006*, Alicante, Spain, September 2006.

[14] I. Hendrawan. Pengurai sintaks kalimat untuuk bahasa indonesia dengan metode linguistic string analysis. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1999. Call number: SK-0388.

[15] Hendri. Algoritma pelacakan teks dengan toleransi kesalahan dan implementasinya. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1996. Call number: SK-0307.

[16] A. F. Horasi. Rancangan kamus kata untuk pemeriksa ejaan elektronis. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1997. Call number: SK-0349.

[17] M. Ichsan. Pemotong imbuhan berdasarkan korpus untuk kata bahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2005. Call number: SK-0615.

[18] M. Ichsan and M. Adriani. Corpus based stemmer for bahasa indonesia. In *Prosiding Seminar Nasional Teknologi Informasi (SNTI)*, Jakarta, Indonesia, November 2006.

[19] A. Irawan. Pengenalan entitas bernama menggunakan hidden markov model dengan fitur morfologi dan kelas kata. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2005. Call number: SK-0575.

[20] M. H. Ismail. Perancangan kamus kata konsep untuk sistem ekstraksi informasi berbasis bahasa alami (indonesia). Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1999. Call number: SK-0395.

[21] Joice. Pengembangan lanjut pengurai struktur kalimat bahasa indonesia yang menggunakan constraint-based formalism. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2002. Call number: SK-0487.

[22] S. D. Larasati. Pengembangan awal analisis semantik bahasa indonesia dengan metode syntax-driven semantic analysis dan penerapannya dalam sistem tanya jawab. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: SK-0657.

[23] S. D. Larasati and R. Manurung. Towards a semantic analysis of bahasa indonesia for question answering. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, Melbourne, Australia, September 2007.

[24] R. Magdalena. Pemrosesan teks: penyusunan tesaurus secara otomatis dalam dalam bahasa indonesia dengan analisis kluster hubungan lengkap (complete link). Undergraduate

Thesis, Faculty of Computer Science, University of Indonesia, 1996. Call number: SK-0325.

[25] K. Marjuki. Penggunaan model ruang vektor dalam proses pengenalan relasi antar entitas pada sistem ekstraksi informasi. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2006. Call number: SK-0629.

[26] Markus. Pengenalan entitas bernama menggunakan metode association rules pada dokumen berbahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: SK-0660.

[27] A. Melani. Peringkas otomatis untuk dokumen dalam bahasa indonesia menggunakan metode statis dan metode query-biased. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: SK-0681.

[28] S. Meliana. Perancangan penalisis struktur kalimat bahasa indonesia, khususnya frasa verbal, dengan menggunakan constraint-based formalism berbasis feature structures dan unification, dan penerapannya pada dokumen resmi. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2001. Call number: SK-0457.

[29] Muhadjir, B. A. A. Nazief, M. Adriani, K. Mangkudilaga, and M. R. Lauder. *Indonesian Word Frequency Dictionary*. Faculty of Arts, University of Indonesia, 1996 1996.

[30] Nasikhin and M. Adriani. Location identification for the geographic information retrieval. In *Working Notes of the Workshop in Cross-Language Evaluation Forum 2007*, Budapest, Hungary, September 2007.

[31] D. Natalia. Penemuan jawaban pada dokumen berbahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2006. Call number: SK-0632.

[32] D. Natalia and M. Adriani. Penemuan jawaban pada dokumen berbahasa indonesia. In *Prosiding Seminar Nasional Teknologi Informasi (SNTI)*, Jakarta, Indonesia, November 2006.

[33] B. A. A. Nazief. Development of computational linguistics research: A challenge for indonesia. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, October 2000.

[34] A. Nugroho. Strategi penyusunan kamus referensi dan analisis kinerja metode pelacakan kata pada pemeriksa ejaan bahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1995. Call number: SK-0298.

[35] M. L. Paramita and M. Adriani. Location identification from indonesian document contents in geographic information retrieval. In *Proceedings of the Seventh National Conference on Computer Science and Information Technology*, Depok, Indonesia, January 2007.

[36] S. Salvitri. Pengembangan lanjut pengurai sintaks bahasa indonesia dengan metode linguistic string analysis. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1999. Call number: SK-0417.

[37] S. Sari. Prototipe pemeriksa tata bahasa baku bahasa indonesia: sebuah program yang dikembangkan dengan alat bantu lex dan yacc pada sistem operasi unix. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1995. Call number: SK-0287.

[38] S. Sari. Perolehan informasi lintas bahasa indonesia-inggris berdasarkan korpus paralel dengan menggunakan metoda mutual information dan metoda similarity thesaurus. Master's thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: T-0617.

[39] S. Sari, H. Hayurani, and M. Adriani. Improving query and document translation using query expansion in cross-language information retrieval. In *Proceedings of the Seventh National Conference on Computer Science and Information Technology*, Depok, Indonesia, January 2007.

[40] A. J. Shidqie. Perancangan pengalisis struktur kalimat bahasa indonesia dengan menggunakan constraint-based formalism berdasarkan features structures dan unification dan penerapannya pada makalah ilmiah. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2001. Call number: SK-0458.

[41] S. Shieber. An introduction to unification-based approaches to grammar. CSLI Lecture Notes 4, Center for the Study of Language and Information, Stanford, USA, 1986.

[42] H. M. A. Sibarani. Alternatif peningkatan kemampuan pemeriksa ejaan bahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1998. Call number: SK-0363.

[43] N. E. Siregar. pencarian kata berimbuhan pada kamus besar bahasa indonesia dengan menggunakan algoritma stemming. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 1995. Call number: SK-0297.

[44] H. D. Sitawati and M. Adriani. Peringkas otomatis untuk dokumen berbahasa indonesia dengan metode frase penunjuk dan metode tf-idf. In *Prosiding Seminar Nasional Sistem dan Teknologi Informasi (SNASTI)*, Surabaya, Indonesia, Agustus 2006.

[45] I. M. D. Sulastra. Perancangan penganalisis struktural kalimat bahasa indonesia, khususnya frasa nominal, dengan menggunakan constraint-based formalism, dan penerapannya pada media massa elektronika. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2001. Call number: SK-0501.

[46] A. K. Sumantri. Perbandingan decision tree, maximum entropy, dan association rules pada resolusi koreferensi untuk bahasa indonesia. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: SK-0697.

[47] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1998.

[48] G. Wahyudi. Pengenalan entitas bernama berdasarkan informasi kontekstual morfologi dan kelas kata. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2004. Call number: SK-0564.

[49] B. Wibowo. Pengenalan entitas bernama menggunakan metode association rules dengan fitur berganda. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2005. Call number: SK-0594.

[50] S. H. Wijono, I. Budi, L. Fitria, and M. Adriani. Finding answers to indonesian questions from english documents. In *Working Notes of the Workshop in Cross-Language Evaluation Forum 2006*, Alicante, Spain, September 2006.