Building JATI: A Treebank for Indonesian

David Moeljadi

Nanyang Technological University, Singapore

The 4th Atma Jaya Conference on Corpus Studies (ConCorps 2017), Atma Jaya Catholic University, Jakarta

21 July 2017
Outline

1. What is a treebank?

2. Indonesian treebanks

3. The corpus: Kamus Besar Bahasa Indonesia (KBBI)

4. The parser: Indonesian Resource Grammar (INDRA)

5. Treebank development

6. Summary and future work
Treebank

- A treebank is a linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech level [8]
- Usages:
  - empirical linguistic research, as well as Natural Language Processing (NLP)
  - enables more precise queries
  - in qualitative research, such as finding an example of a certain linguistic construction or a counter-example to a claim about syntactic structure
  - in quantitative research, as a source of information about frequencies and co-occurrences
  - building statistical model, robust broad-coverage parsing
  - developing a broad-coverage grammar, test the grammar
Motivation

- We want to understand natural language
  - it is interesting in and of itself
  - it offers a view into human cognition
  - much knowledge is encoded in natural language
  - we want to make computers understand

- What does it mean for a machine to understand?
  - The system analyses text and grows clever
    - it increase the lexicon
    - it builds up the ontology
    - it changes the stochastic model
Indonesian treebanks

- The Indonesian Dependency Treebank developed by Charles University in Prague [5]
- The Indonesian Treebank developed by the Faculty of Computer Science of University of Indonesia [4]
- The Indonesian Treebank in the Asian Language Treebank (ALT), built by the Agency for the Assessment and Application of Technology (BPPT) [13]
- the Indonesian Treebank in the ParGram Parallel Treebank (ParGramBank), based on LFG “IndoGram” [15]
Other treebanks

- Penn Treebank
- Hinoki [2]
Based on an HPSG grammar of Indonesian: Indonesian Resource Grammar (INDRA) [6]
We want to develop a broad-coverage grammar together with the treebank. Treebanking allows us to immediately identify problems in the grammar and improving the grammar directly improves the quality of the treebank [9]

Parsing (a subset of) dictionary definition sentences: KBBI Fifth Edition [1]

Creating a corpus that can be studied: JATI
The corpus: Kamus Besar Bahasa Indonesia (KBBI)

- The fifth edition of KBBI [1], published by Badan Pengembangan dan Pembinaan Bahasa
- The KBBI database, a machine-tractable dictionary [7]
- 108,240 entries, 126,643 definitions, 29,260 examples (as of 15 June 2017)
KBBI definition sentences

Definitions related to food, drinks, spices, edible things are extracted and edited

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>minuman keras <em>yg</em> dibuat <em>dr</em> nira <em>yg</em> telah disuling</td>
<td>minuman keras <em>yang</em> dibuat <em>dari</em> nira <em>yang</em> telah disuling</td>
</tr>
<tr>
<td>kue kering, dibuat <em>dr</em> sagu dan dibungkus <em>dng</em> daun nipah</td>
<td>kue kering <em>yang</em> dibuat <em>dari</em> sagu dan dibungkus <em>dengan</em> daun nipah</td>
</tr>
<tr>
<td>makanan <em>terbuat</em> <em>dr</em> daging, udang, ikan <em>yg</em> dicincang</td>
<td>makanan <em>yang</em> <em>dibuat dari</em> daging, udang, <em>atau</em> ikan <em>yang</em> dicincang</td>
</tr>
</tbody>
</table>

- Shorter, compared with other commonly used text for corpora, such as newspaper text
- Contain more fragments, especially noun phrases
- Valid examples of naturally occurring texts
open-source Indonesian computational grammar [6]
https://github.com/davidmoeljadi/INDRA

parse and generate Indonesian text

open-source tools in Deep Linguistic Processing with HPSG Initiative (DELPH-IN)
  ▶ Documentation (http://moin.delph-in.net/IndraTop)
  ▶ ITSDB or [incr tsdb()] [10]
  ▶ Full Forest Treebanker (FFTB) [12]

theoretical framework of Head Driven Phrase Structure Grammar (HPSG) [14]

Minimal Recursion Semantics (MRS) [3]

1,885 types, 15,099 lexical items, 38 rules (as of 15 June 2017)
Choosing a Grammar

HPSG is chosen for the following reasons:

- Serious attempt to cover linguistic phenomena both core and periphery
- Unification- and constraint-based context free grammar (phrase structure grammar)
  - consists of a set of rules and a lexicon of symbols (parts-of-speech) and words, surface oriented (no additional abstract structures)
- Integration of syntax and semantics (mono-stratal)
  - we are most interested in semantics
    - tractable representation: MRS
- A vibrant research community
  - well developed open source tools
  - integration with shallow processing

Moeljadi (ConCorps 2017)
Deep Linguistic Processing with HPSG Initiative

- Grammars: English (ERG), Japanese (JACY), Chinese (Zhong), Indonesian (INDRA), ...
- Development Environment: Linguistic Knowledge Builder (LKB)
- Processor: Answer Constraint Engine (ACE)
- Test Environment: ITSDB or [incr tsdb()]
- Treebanking tools: FFTB
- Machine Translation: LOGON
Approaches to Treebanking

- Manual Annotation
- Semi-Automatic
  - Parse and repair by hand: Penn WSJ, Kyoto Corpus
    - 100% cover, reasonably fast
    - Often inconsistent, Hard to update,
      Simple grammars only (prop-bank is separate)
  - Parse and select by hand: Redwoods, Hinoki, JATI
    - All parses grammatical, Feedback to grammar,
      Consistent
    - Both syntax and semantics, Easy to update
    - Cover restricted by grammar

- Discriminant-based treebanking: select or reject discriminants until one parse remains
Grammar development

1. Develop initial test suite
2. Identify phenomena to analyze
3. Extend test suite with examples documenting analysis
4. Implement analysis
5. Compile grammar
6. Debug implementation
7. Parse sample sentences
8. Parse full test suite
9. Treebank

Moeljadi (ConCorps 2017)
Summary and future work

- Refining the analyses
  - Improving INDRA by adding new rules and lexical types
- Automate analysis
  - parse ranking
- Expanding the system
  - Adding non-familiar words (lexical acquisition)
  - Dynamic handling of unknown words
Long Term Goals

- Make text understanding available to everyone
  - Machine translation
  - Question answering
  - Speech recognition
  - Man-machine interfaces

- Link words to meanings for all languages
Acknowledgments

- Thanks to Francis Bond for his inspiration and advice to build JATI
- Thanks to Dora Amalia who gave permission to use a part of the fifth edition of KBBI data
- Some slides use material from:
  - “The Hinoki Treebank: Toward Text Understanding” by Francis Bond, Sanae Fujita, Chikara Hashimoto, Shigeko Noriyama, Eric Nichols, Takaaki Tanaka, and Hiromi Nakaiwa
  - “Treebanking an Open Forest: The Tanaka Corpus” by Francis Bond and Takayuki Kuribayashi


References III


Thank you

terima kasih n rasa syukur;
berterima kasih v mengucap syukur; melahirkan rasa syukur atau membalas budi setelah menerima kebaikan dsb