A Survey of Research on Computing Language in Bahasa Indonesia Conducted at the University of Indonesia

Information Retrieval Lab

Conference on "Policy and Sustainability of Local Language Computing in Developing Asia"
Lahore, 29 Jan to 3 Feb 2012

Faculty of Computer Science
University of Indonesia

INTRODUCTION

Faculty of Computer Science

- Undergraduate (1986), Masters (1988) & Doctoral (1998) programme
- >1400 active students, annual intake ±350
- ±60 teaching staff, mostly MSc & PhD holders
- Research labs:
  - Computational Intelligence
  - Digital Libraries & Distance Learning
  - Formal Methods in Software Engineering
  - Architecture, Networks & High Performance Computing
  - Image Processing & Pattern Recognition
  - **Information Retrieval & Text Processing**
  - IT Governance
  - E-Government

NLP RESOURCES & TOOLS

Computational lexicography

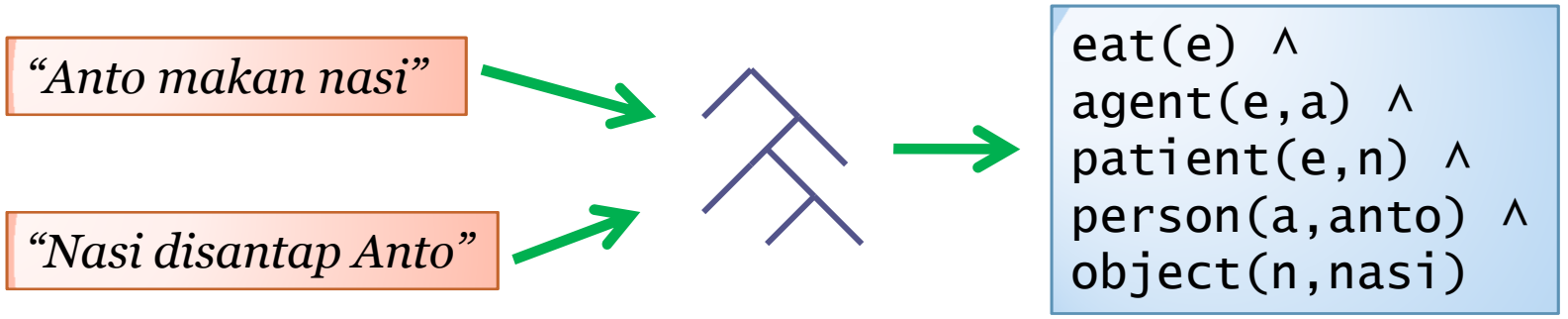
- Corpus-based word-frequency dictionary (1996)
- Electronic *Kamus Besar Bahasa Indonesia* – joint work with *Pusat Bahasa*:
  - Data structures & compression
  - Efficient matching
- Spell-checking (Lotus SmartSuite)
  - Data structures
  - Error-tolerant matching
  - Suggestion

Morphological analysis

- Stemming algorithms:
  - Rule-based
  - Corpus-based
- Morphological parser:
  - Two-level morphology
- Syntactic Parsing
  - Formal modeling:
    - Regular languages
    - Context Free Grammars
    - Feature structure unification grammars
  - Statistical parsing:
    - Probabilistic CFGs

Semantic and discourse analysis

- Lexical semantics
  - Vector-space lexical similarity
  - Indonesian WordNet
- Text semantic analysis
  - Syntax-based, lambda calculus



INFORMATION RETRIEVAL (IR) APPLICATIONS

IR: Cross-language Retrieval

- Retrieving documents in one language using query in another language
  - Retrieving Indonesian documents using English queries
  - Retrieving English documents using Indonesian queries

- Mode of Translation
  - Query Translation (2006)
  - Document Translation (2007)

- CLIR Translation Approaches
  - Bilingual Dictionaries
    - Direct Translation (2006)
      - English-to-Indonesian & Indonesian-to-English
    - Transitive Translation (2007)
      - Indonesian-French-German-English
  - Machine Translations (2006)
    - Transtool, Toggletext
  - Parallel Corpus (2007)
    - Collecting parallel articles from the Internet
    - Translating English documents into Indonesian

IR: Document summarization

- Single document summarization
  - Extract sentences containing cue phrases & important keywords (2006)
  - Query Biased Summary
    - Extract sentences related to query words (2007)

- Multi Document Summarization
  - Extract sentences containing important keywords that occur in the centroid of a cluster (2008)

IR: Question answering

- Finding answers to Indonesian questions in Indonesian documents
  - Using statistical technique and considering the position of a candidate answers in the passages (2006)
- Finding answers to Indonesian questions in English documents (CLEF-QA)
  - Translate Indonesian queries into English
  - Using linguistic knowledge and external resources found on the Internet to find the answer (2008)

IR: Geographic Information Retrieval

- Finding events occurring in certain locations
  - We develop a location parser to identify any location name that appears on the Indonesian query and documents (2007)
  - We use geographic relation words to identify events that happen in certain locations (2007)
  - We use a location-based query expansion technique to improve the retrieval performance of CL-GIR (2007)

IR: Information extraction

- Extracting important information from Indonesian documents
  - Developing named entity tagger to identify person, location, organization names using rules based (2004) and machine learning approach with association rules (2007)
  - Identify whether some named entities found refer to the same object (co-reference resolution).
  - Identify the relationship exist between those named entities

Speech recognition

- Developing ASR for Bahasa Indonesia
- Using open-source ASR systems
  - Sphinx-4
  - Julius
- Intended for telephone applications
- Building a speech corpus
  - 5000 speakers
  - Each speaker spends 15 minutes to record a list of sentences
- Indonesia has many local languages and dialect (> 600)
  - Need to identify various pronunciation for words

Indonesian WordNet

- Indonesian WordNet using expand approach based on Princeton WordNet (PWN) & KBBI
- Automatic PWN-KBBI mapping using LSA

Legal Information System

- The Indonesian law document is written in natural language
- Standardizing Indonesian Law document using XML format
- Indonesian legal document search engine
- Recapitulation System for the Indonesian Law document

Publications

- Malindo Workshop (http://malindo.org) : 6th workshop will be held in Malaysia
- ICACSIS Conferences http://icacsis.cs.ui.ac.id
- Workshop on Technologies and Corpora for Asia-Pacific Speech Translation
- etc

Resources

- The Resources can be seen in: http://fws.cs.ui.ac.id
- It includes lexical resources (Electronic KBBI, Indonesian WordNet), NLP Tools (stemmer, parser, POS, etc), and IR Application (machine translator, speech recognition, etc)

- Furthermore, UI collaborates with Kyoto University in Language Grid (LangGrid) Project, UI will become **'Jakarta Operation Center'**. UI will provide language resources and tools for Bahasa Indonesia that can be accessed through web service.

Further Information

- Further information can be found in:
  - http://ir.cs.ui.ac.id
  - http://bahasa.cs.ui.ac.id

Thank You