Natural Language Processing
The saara Approach

Kavi Narayana Murthy

School of Computer and Information Sciences
University of Hyderabad

NLP

- NLP Today: a data driven empirical science. NLP systems are built by training language independent and generic machine learning algorithms on large scale language data.
- Original goals of NLP: NL understanding, NL generation and NL learning
- Meaning has gradually lost focus, almost forgotten?

The saara Approach

- Given a sentence, how to compute its meaning.
  - Given a word, how to compute its meaning.
- Problem: Computers do not understand meanings.
- Solution: Structure indicates meaning. Use appropriate structures and manipulations.
- Grammar: Mapping Structure to Meaning
- Main Focus: Development of Computational Grammars
- Philosophy: Do it right, no short cuts!

Grammar

- Morphology: Relating word internal structure to word meaning
- Syntax: Relating sentence structure to sentence meaning
- What exactly does a sentence mean?
  - Can be computed. Provided
    - Speaker knows what exactly to say
    - Speaker knows how exactly to say
- Universal / language-independent

Grammar and Usage

- People are not always very careful. They may also not know what to say or how to express it. Thus, actual usage may not indicate the intended meaning directly, precisely and unambiguously.
- Layered Approach: Grammar should be designed for carefully usage only. This forms the core. We can then build layers or wrappers to handle all the variations we find in actual usage. This way, we can get a simple, neat, elegant, efficient grammar and cater to practical needs at the same time.

The saara System

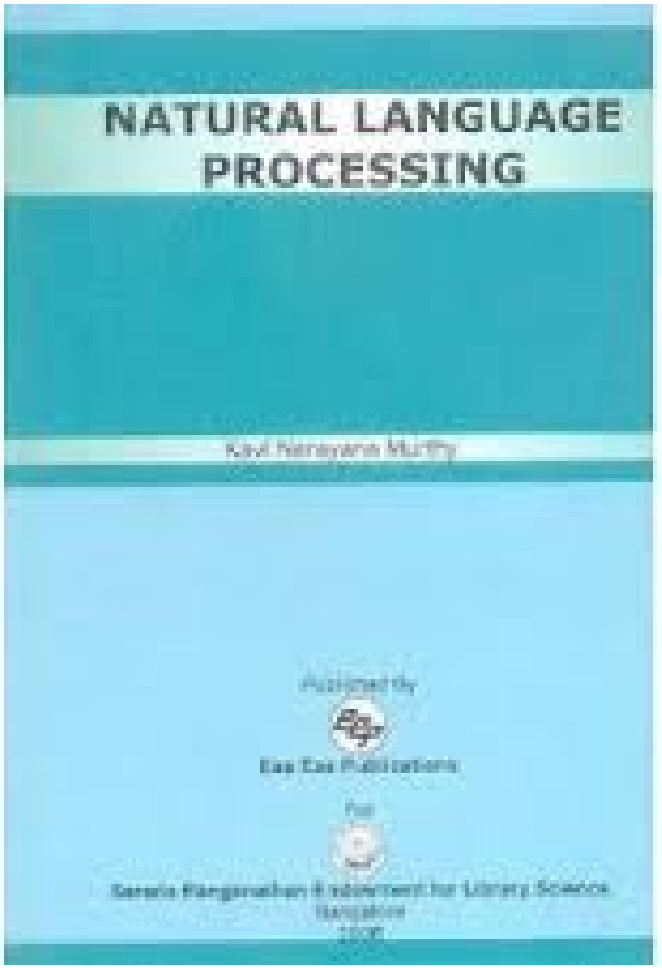
- Computational Models for
  - Analysis
  - Generation
  - Translation and other Applications
- kannaDa-saara Alpha Ver 3
- telugu-saaramu Alpha Ver 3
- Lexical Resources and Tools
- The ***saara*** Translator System

The saara Approach

- Correct Analysis of Source Language
  - Can be translated into any other language
    - Only bilingual dictionary and transfer grammar required
- Correct Generation
  - Can take the analysis produced by any other system for any other language and translate to our language
- How to guarantee correctness?
  - Machine Learning cannot!
- NLP is a technology. Q: What is the scientific foundation? A: The saara Approach!

Natural Language Engineering at SCIS, UoH

- Foundations: Word, Word Classes, Sentence, Syntactic Relations, Parsing, ...
  - Universal and Precise Definitions
  - Lexical Resources and Tools
- Core Research and Development:
  - Telugu, Kannada, Indian Languages, English
- Applications
  - Text Categorization, Text Summarization, NERC
  - Language ID, WSD, Spell Checking, Anaphora Reso.
  - ASR, TTS, OCR, Machine Translation, IR, IE

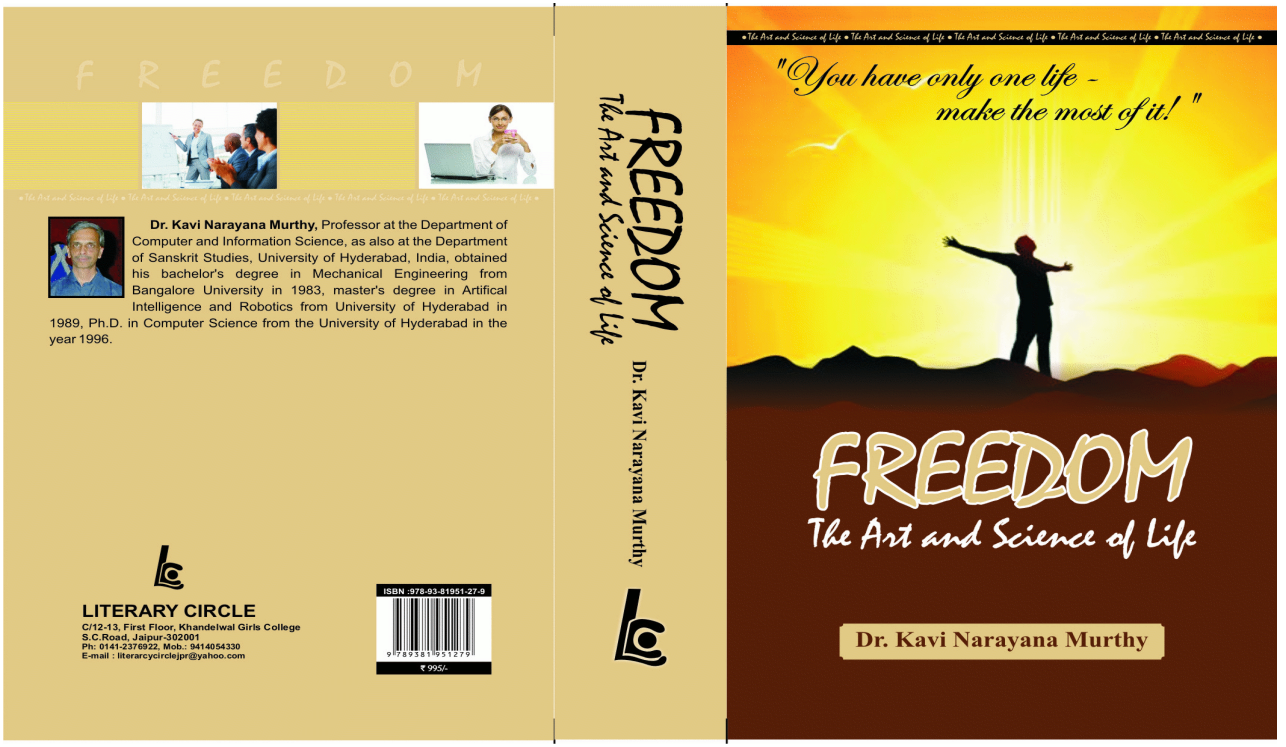

NATURAL LANGUAGE
PROCESSING
Kavi Narayana Murthy

If You are Interested

- Yoga, Ayurveda, Holistic Healing
- Vedanta
- Classical Music

- Books:
  - Brahmacharya
  - Ahimsa
  - Freedom

Just Released!



Freedom

See what the previewers have said:

"Here is one book which œvers all aspects of life, simple, clear, written with a scientific bent of mind, a great book, very much unlike many other books I have seen on similar topics"

"The book is a masterpiece - no doubt"

" The book is simply awesome and I feel it is a 'must read' for all human beings"

"This is a book I love to read again and again"

Thank You

Visit

202.41.85.68

email: knmuh@yahoo.com

Words

- What Exactly is a Word?
- Universal Word Classes
- Sub-Categorization
- Tag-Set
  - Hierarchical, Extensible, Fine-Grained
  - More Than POS
- Languages are NOT as ambiguous as they seem to be!

Words

- Lexicon
- Morphological Analysis and Generation
- Stemming and Lemmatization
- Spelling Error Detection and Correction
- Lexical Resource Toolkit, Glossing
- Tagging
  - No ML, No Training Data, No Manual Work
  - High Performance

Sentence

- What exactly is a Sentence?
- What exactly is Syntax?
- Universal Syntactic Relations
- How to identify them?

The saara Architecture

- Purely Linguistic, no ML
- Simple Pipe-Line Architecture
  - Phonemes, Words, Sentences, Discourse
  - No tagging, chunking, local-word-grouping, ...
- Perl and Java – Platform Independent
  - Synchronization
  - Stand Alone, No other dependencies

Status, Plans

- Word Level: Alpha Versions Released, Beta soon
  - > 90% Analyzed, Mostly Correct
  - First Complete Morph for Kannada/Telugu
- Sentence Level: Going On
  - To be Ready in about an Year
- Workshop Series
- No Funding taken from any source so far.
  TDIL/DeitY has now sanctioned a project.

References

- Pl. visit our website!