Natural Language Processing II (SC674)

KAIST, EECS
Prof. Feng Zhiwei

Ch.1 Introduction of NLP

1.0 What is NLP?

1.0.1 Definition of NLP

“NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the result systems.” (Bill Manaris, <Natural language processing: A human-computer interaction perspective>, Advances in Computers, Volume 47, 1999).

1.0.2 Knowledge levels in NLP

In NLP models, the knowledge levels including following aspects:

1. Acoustic/prosodic knowledge: rhythm and intonation of language; how to form phonemes.
2. Phonologic knowledge: Spoken sounds; how to form morphemes.
3. Morphologic knowledge: Sub-word units; how to form words.
4. Lexical knowledge: Words; how to derive units of meaning.
5. Syntactic knowledge: Structural rules of words (or collection of words); how to form sentences.
6. Semantic knowledge: (Situation) Context-independent meaning; how to derive sentence meaning.
7. Discourse knowledge: Structural roles of sentences (or collection od sentences); how to form dialogs.
8. Pragmatic knowledge: (Situation) Context-dependent meaning. How to derive sentence meaning relative to surrounding discourse.
9. World knowledge: General knowledge about the language user and the environment, such as user beliefs and goals; how to derive belief and goal structure, this is a catch-all category for linguistic processes and phenomena that are not well understood yet. Based on past evolutionary trends, this knowledge level may be further subdivided in the future to account for new linguistic/cognitive theories and models.

There are different school of thought for the knowledge levels in NLP, but, in general, researchers agree that linguistic knowledge can be subdivided into at least lexical, syntactic, semantic, and pragmatic levels. Each level conveys information in a different way. For example, the lexical level might dea; with actual words (i.e., lexemes), their constituents (i.e., morphemes), and their inflected forms. The syntactic level might deal with the way words can be combined to form sentences in a given language. The semantic level might deal with the assignment of meaning to individual words and sentences. The pragmatic level might deal with monitoring of context/focus shifts within a dialog and with actual sentence interpretation in the given context.

Following is the commonly used classification of knowledge levels in NLP.

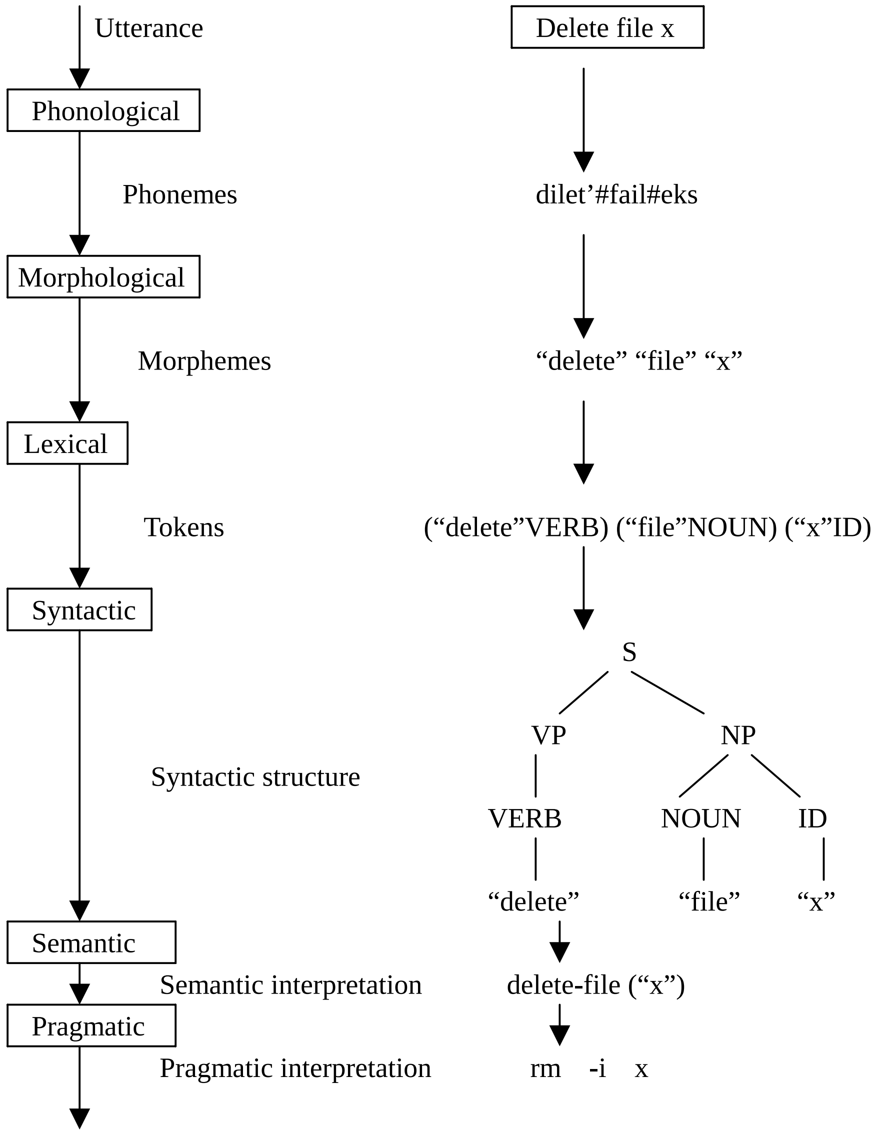

Fig 1. Knowledge levels in NLP system

In this classification, each level is defined in term of the declarative and procedural characteristics of knowledge that it encompasses (includes).

Based on the application domain, NLP systems may require only subsets of this knowledge levels to meet their application requirements. For example, Eurotra MT system focuses on documents dealing with telecommunications, and covers nine languages (Danish, Germany, Greek, English, Spanisg, French, Italian, Dutch, Portuguese), this MT system may require only knowledge levels 3 to 7 (or possible 8). A speech recognition system, such as Dragon systems' Naturally Speeking and Kurzweil Voice", may require only knowledge levels 1 to 5, although it would benefit from having access to knowledge levels 5 to 7.

A speech understanding interface to UNIX (UNIX Consultant) may use knowledge levels 1 to 6 to produce a semantic representation of a givem input, e.g., **delete-file ("x")**, and then knowledge level 8 to convert that semantic interpretation to the corresponding realization in the underlying command language, e.g., "**rm –i x**". Such an interface could benefit from having access to specific knowledge about dialog in the context of communication with an operating system UNIX.

(Wilensky, 1988).

1.0.3 NLP is an interdisciplinary area

The study of NLP is related with following areas:

- Computer science, which provides techniques for model representation, and algorithm design and implementation;
- Linguistics, which identifies linguistic models and processes;
- Mathematics, which contributes formal models and methods;
- Psychology, which studies models and theories of human behavior;
- Philosophy, which provides theories and questions regarding the underlying principles of thought, linguistic knowledge, and phenomena;
- Statistics, which provides techniques for predicting events based on sample data;
- Electrical engineering, which contributes information theory and techniques for signal processing;
- Biology, which explores the underlying architecture of linguistic processes in the brain.

1.1 Main Domains of NLP

1.1.1 Spoken Language Input

- Speech Recognition
- Signal Representation (voice signal analysis)
- Robust Speech Recognition
- HMM (Hidden Markov Model) Methods in Speech Recognition
- Language Representation (Language Model)
- Speaker Recognition
- Spoken Language Understanding

1.1.2: Written Language Input

- Document Image (format) Analysis
- OCR (Optical Character Recognition) Print
- OCR: Handwriting
- Handwriting as Computer Interface (e.g. pen computer)
- Handwriting Analysis (e.g. signature verification)

1.1.3 Language Analysis and Understanding

- Sub-Sentential Processing (Morphological analysis, Morphological disambiguation)
- Grammar Formalisms (e.g. CFG, LFG, FUG, HPSG)
- Lexicons for Constraint-Based Grammars
- Semantics
- Sentence Modeling and Parsing
- Robust Parsing

1.1.4. Language Generation

- Syntactic Generation
- Deep Generation

1.1.5 Spoken Output Technologies

- Synthetic Speech Generation
- Text Interpretation for Text-to-Speech Synthesis
- Spoken Language Generation (Conception to Speech)

1.1.6 Discourse and Dialogue

- Discourse Modeling
- Dialogue Modeling
- Spoken Language Dialogue

1.1.7 Document Processing

- Document Retrieval
- Text Interpretation: Extracting Information
- Summarization (e.g. text abstraction)
- Computer Assistance in Text Creation and Editing
- Controlled Languages in Industry

1.1.8 Multilinguality

- Machine Translation
- (Human-Aided) Machine Translation
- Machine-aided Human Translation
- Multilingual Information Retrieval
- Multilingual Speech Processing
- Automatic Language Identification

1.1.9 Multimodality

- Representations of Space and Time (Automatic abstraction fo space and time from text)
- Text and Images
- Modality Integration: Speech and Gesture (using data-gloves)
- Modality Integration: Facial Movement & Speech Recognition
- Modality Integration: Facial Movement & Speech Synthesis

1.1.10 Transmission and Storage

- Speech Coding (speech compression)
- Speech Enhancement (speech quality Improvement)

1.1.11 Mathematical Methods

- Statistical Modeling and Classification
- DSP (Digital Signal Processing) Techniques
- Parsing Techniques
- Connectionist Techniques (e.g. Neural Network)

- Finite State Technology
- Optimization and Search in Speech and Language Processing

1.1.12 Language Resources

- Written Language Corpora
- Spoken Language Corpora
- Lexicons
- Terminology

1.1.13 Evaluation

- Task-Oriented Text Analysis Evaluation
- Evaluation of Machine Translation and Translation Tools
- Evaluation of Broad-Coverage Natural-Language Parsers
- Human Factors and User Acceptability
- Speech Input: Assessment and Evaluation
- Speech Synthesis Evaluation
- Usability and Interface Design
- Speech Communication Quality
- Character Recognition

1.2 Brief History of NLP

1.2.1 Foundational insights: 1940s and 1950s

- Automaton theory:

  --Turing's work (1936);

  --McCulloch-Pitts neuron (1943): neuron is a kind of computing element that could be described in terms of proposition logic.

  --Finite automaton and regular expressions (Kleene, 1951, 1956).

  --Application of probabilistic models of discrete Markov processes to automaton for language (Shannon, 1948).

  --Formal language theory (Chomsky, 1956)

  --Backus-Naur normal form in ALGOL programming language (Backus, Naur, 1959)

- Probabilistic or information-theoretic model:

  --Noisy channel and decoding (Shannon)

  --Entropy of language (Shannon)

- Machine Translation

  --Weaver's Memorandum on MT:

  Weaver (1955) states:

  When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols, I will now proceed to decode it".

  --First MT experiment (Georgetown University, IBM)

  --ALPAC (Automatic language Processing Advisory Council) report (1966)

- Speech technique
  --Sound spectrograph (Koenig, 1946);
  --First machine recognizer (1950s);
  --10 digit recognizer (Davis, Bell Labs, 1952)

1.2.2 The two camps: 1957-1970

- Symbolic approach: The symbolic approach in NLP is best formulated by the physical symbol system hypothesis. Although it originated in the late 1950s. This hypothesis states that intelligent behavior can be modeled using a physical symbol system consisting of physical patterns (symbols) which may be used to construct expressions (symbol structures); additionally, this system contains a set of processes that operate on expressions through creation, deletion, reproduction, and arbitrary transformation. A great portion of the work in computational linguistics is based on this hypothesis.

E.g. The sentence “Mary will not take the apple” in case grammar can be described as follows:

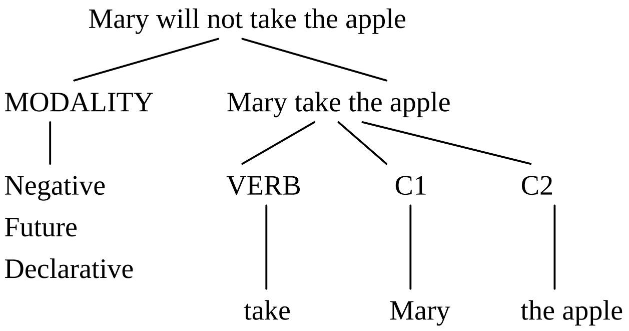

Fig. 2

Main works in symbolic approach:

--Parsing algorithm (top-down, bottom-up, dynamic programming, 1950s to mid 1960s, based on the Chomsky theory)

--Earlist complete parsing system – Transformations and Discourse Analysis Project [TDAP] (Zelig Harris, 1962).

--Discussion on Artificial Intelligence (John Mccarthy, Marvin Minsky, Claude Shannon, Nathaniel Rochester, 1956).

--Logic Theorist and General Problem Solver (Newell, Simon).

The strengths of symbolic formalism:

--They are well understood in terms of their formal descriptive/generative power and practical application.

--They currently provide the most effective approach for modeling long-distance dependencies, such as subject-verb agreement and wh-movement.

--They are usually perspicuous (clear, obvious), in that the linguistic facts being expressed are directly visible in the structure and constituents of the model.

--They are inherently non-directional, in that the same linguistic model may be used for both analysis and generation.

--They can be used in multiple dimensions of patterning that is they can be used for modeling phenomena at various linguistic knowledge levels:

--They allow for for computationally efficient analysis and generation algorithms, as in Earley (1970) and Marcus (1978)

The weaknesses of symbolic approach:

--Symbolic linguistic models tend to be fragile, in that they cannot easily handle minor, yet non-essential deviations of the input from the modeled linguistic knowledge – nevertheless, various flexible robust parsing techniques have been devised to address this weakness. Such techniques may recover from parsing failures.

--Development of symbolic models requires the use of experts such as linguists, phonologists, and domain experts, since such models cannot be instructed to generalize (learn from examples).

--Symbolic models usually do not scale up well. For instance, Slocum (1981) discusses how after two years of intensive development, LIFER's knowledge base grew so large and complex that even its original designers found it hard and impractical to perform the slightest modification. Specially, the mere act of performing a minor modification would cause "ripple effects" throughout the knowledge base. These side-effects eventually became almost impossible to isolate and eliminate.

--In many practical cases, symbolic approach technique perform worse than stochastic approach that tuned by real-life training data. Symbolic approach can not model certain local constraints, such as word preferences that can be very useful for effective part-of-speech tagging and other application.

The symbolic approach is the most well-studied technique for NLP system development. It is still valuable and powerful mechanism in NLP. It is especially useful in case where the linguistic domain is small or well-defined, and where modeling of long-distance dependency is essential.

- Stochastic approach: The driving force behind stochastic (statistical, probabilistic) approach is their ability to perform well even in the presence of incomplete linguistic knowledge about an application domain.

Main work in stochastic approach

--Bayesian method was applied to the problem of OCR.

--Bayesian system for text-recognition (Bledsoe and Browning, 1959)

--Authorship distribution based on Bayesian method (Mosteller and Wallace, 1964).

--First on-line corpora – Brown Corpus: 1 million words (Kucera, Francis, 1963)

--Probabilistic context-free grammar (PCFG). This approach is to assign probabilities based on the rule use; that is, using a set of training data, the probability of each rule's "significance" can be determined based on the frequency of this rule's contribution to successful parses of training sentences.

S → NP VP	0.85
S → VP	015
NP → N	0.40
NP → N PP	0.35
NP → N NP	0.25
VP → V	0.32

VP → V NP	0.31
VP → V PP	0.19
VP → V NP PP	0.18
PP → P NP	1.00
P → like	1.00
V → like	0.40.
V → flies	0.40
V → jumps	0.20
N → flies	0.45
N → jumps	0.20
N → bananas	0.30
N → time	0.20

(Adapted from Charniak, 1993)

Probabilistic context-free grammars (PCFG) was developed as extensions of CFG that include this information in the form of probabilities for each rule.

In order to see how probabilities are assigned in a PCFG, consider the task of deriving a random sentence in a top-down manner, applying each rule expansion according to its probability.

- Initially the start symbol would be the only symbol to be rewritten, and would therefore have a probability 0.85 of being used in the current derivation.
- Expanding this symbol will involve selecting a rule with S as its LHS and some terminal and non-terminal symbols on the RHS. Since S has been chosen it is certain that it must be expanded.
- After applying this expansion, the probability that a rule with NP on the LHS will be used is 1. This means that the probability of all rules with 'NP' on their LHSs must add up to 1.

  A PCFG consists of a set of rules where the probabilities of rules with the same LHS add up to 1. These probabilities can be estimated from their frequency in a corpus of syntactically analyzed sentences. For example, assume that a rule Ri of the form C → $D_1 \ldots D_n$ is used r times in the corpus. That is, there are r sub-trees in the corpus of the following form

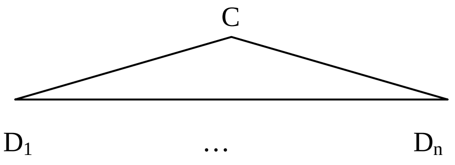

Fig. 3

Then, if c rules with C as their LHS are used in the corpus, the estimate for the probability of Ri is r/c. Thus, if there are 80 occurrences of this rule NP → N and 200 NP rules altogether, then the probability for this rule is 80/200 = 0.40.

- All non-terminals can be thus expanded until there are no more non-terminals.
- The result would be a parse tree whose probability is given by the **product** of the probability of each rule applied.

Consider the following tree:

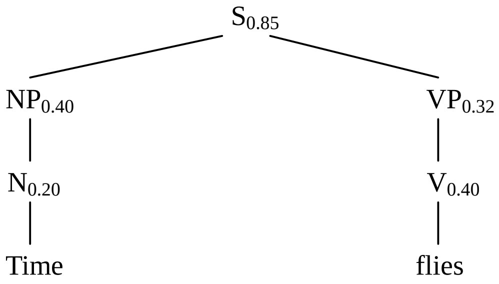

Fig..4

Its probability is given by 0.85 x 0.40 x 0.32 x 0.20 x 0.40 = 8.704 x 10^{-3}

Note that lexical generation probabilities (e.g. P(time | N) = 0.20, P(flies | V) = 0.40) are involved in such a derivation.

Obviously, the PCFG must be based on the corpus linguistics. That is why the corpus linguistics became an important scientific field in NLP.

The strength of stochastic approach:

--Stochastic systems are effective in modeling language performance through training based on most frequent language use. They useful in modeling linguistic phenomena that are not well-understood from a competence perspective, e.g. speech.

--The effectiveness of a stochastic system is highly dependent on the volume of training data available; generally more training data results to better performance.

--Stochastic approach may be easily combined with symbolic model of linguistic constrain, such as dialog structure to enhance the effectiveness and efficiency of an application.

--Stochastic models can be used to model nuances and imprecise concepts such as “few”, “several”, and “many”, that have traditionally been addressed by fuzzy logic.

The weakness of stochastic approach:

--Run-time performance of stochastic systems is generally linearly proportional to the number of distinct classes (symbols) modeled, and such can degrade considerably as classes increase, this holds for both training and pattern classification.

--In general, given the state of the art in corpus development, producing training data for a specific application domain can be a time-consuming and error-prone process. Thus since the effectiveness of stochastic systems is tightly bound to extensive, representative, error-free corpora, the difficulty of developing such system might be similar to that of other approaches.

Stochastic approaches are very effective in addressing modeling problems in application domains where traditional symbolic processes have failed.

1.2.3 Four paradigms: 1970-1983:

- Stochastic paradigm

  --Speech recognition algorithm based on Hidden Markov Model [HMM], noisy channel, decoding theory (Jelinek at IBM Watson Research Center, and Baker at CMU)

  --Speech recognition and synthesis (AT&T’s Bell Labs)

- Logic paradigm
  --Q-systems and metamorphosis grammar (Colmerauer, 1970, 1975)
  --Definite Clause Grammar [DCG] (Pereira and Warren, 1980)
  --Functional Unification Grammar [FUG] (Martin Kay, 1979)
  --Lexical Functional Grammar [LFG] (Bresnan and Kaplan, 1982)

- Natural Language Understanding [NLU] paradigm
  --SHRDLU system (Winograd, 1972)
  --Conceptual Dependency Theory [CD] (Roger Schank, 1977)
  --Network based semantics (Quillian, 1968)
  --Preference Semantics (Wilks, 1975)
  --Case grammar (Fillmore, 1968)
  --Question-Answer system as LUNAR (Woods, 1967, 1973)

- Discourse Modeling paradigm
  --Substructure in discourse and discourse focus (Grosz, 1977, Sidner, 1983)
  --Automatic reference resolution (Hobbs, 1978)
  --Belief-Desire-Intention [BDI] framework based on speech acts (Coehen and Perrault, 1980)

1.2.4 Empiricism and Finite State Models Redux: 1983-1993

- Finite State Model received attention again:
  --Finite-State phonology and morphology (Kaplan and Kay, 1981)
  --Finite-State model of syntax (Church, 1980)

- “Return to Empiricism”
  --Probabilistic model of speech recognition (IBM Thomas J. Watson Research Center)
  --Probabilistic models and other data-driven approaches spread into part-of-speech (POS) tagging, parsing and PP-attachment ambiguities, and connectionist approaches from speech recognition to semantics.

1.2.5 The field comes together: 1994-2002

- Probabilistic and data-driven models became quite standard throughout NLP. Algorithms for parsing, POS tagging, reference solution, and discourse processing all began to incorporate probabilities.
- Speech and language processing algorithms began to be applied to Augmentative and Alternative Communication (AAC).
- The rise of the World-Wide-Web emphasized the needs for language-based information retrieval and information extraction.

Robert K. Merton said:
“All scientific discoveries are in principle multiples, including those that on the surface appear to be singletons.” (<Singletons and Multiples in scientific discovery>, 1961)

This conclusion seems true for the development of NLP.