History of
Natural Language Processing

Source: Jyunichi Tsujii (2000) "Language & Computer"
Language, Monthly Issue

History of Linguistics: Ancient Times

I. Logos is the faculty of human reason that underlies all speeches of men and thus becomes a source of life in the universe

II. Language study in ancient Greece (highly sophisticated Greek language):

  I. Language went through various transitions. Linguistics is the study of seeking for a true meaning which may have lost over the years.

  II. Issues debated were:

    I. Do languages develop according to the laws of nature or human custom?

    II. Do languages inherit underlying orders as the first principle of all things?

    III. How many part of speech (POS) exist?

III. Names for all things: Socrates

IV. Logic behind language: Aristotle

V. Rhetoric studies: Quintilianus

  I. Hierarchy in arts: grammar, logic and rhetoric

➢ Spoken language to written language

➢ Shift from being conceptual to practical.

History of Linguistics: Mid Centuries

I. Latin continued to be the common language in Europe for more than 1000 years.

  - I. Realist = Universals (human, horse, etc) exist in reality, independently and prior to physical reality.
  - II. Nominalist = Various objects have independent existence. By universals abstract objects (symbols) are labeled.

II. Constantinople's fall in 1453

  - I. Latin linguists returned to Italy.
  - II. Greek and Roman classics were revitalized.
  - III. The European continent was politically divided. Centralized governments used local languages as national languages to rule nations. -> Latin was on the decline.
  - IV. A global economy has emerged. Technical developments have influenced the world.

History of Linguistics: Mid Centuries

I. Grammar (Accidence, Syntactics, and Pragmatics): Port-Royal Grammar

II. Concepts and expressions: Locke

III. Breakdown into minimal elements that bear meaning: Condiak

➢ Structure & meaning ->Modern issues presented.

I. Adoption of a standardized language for printing technology: Caxton

➢ Controlled English due to practical and technical issues with regard to printing: The invention of printing by Gutenberg greatly influenced linguistic studies. The impact cannot be matched by that of all philologists and linguists together.

History of Linguistics: Modern times

I. Seeking for a true language->
   I. Sanskrit (inflective language): an ancient language sophisticated more than Greek -> A high productivity pertaining to inflective languages
   II. Proto-Indo-European (PIE): Humboldt
   III. Darwinism increased momentum for the study of ancestral languages.

➢ The evolution occurred.

Saussure

- According to Saussure, ideas are like the Galaxy, which nothing can be requisitely distinguished from.
- Nothing can be identified before language.
- Arbitrary of signs
- Study of synchronicity
- For Saussure, what could be discovered by tracking back languages would be only an afterthought of researchers.
- More central to his study was the clarification of language structure.
- Saussure argued that questions regarding the relation between language and the world were not the core part of the study of language.

Science of Natural Language
-Saussure Evolution

- Saussure:
  - Langue refers to the whole system of synchronic (or same time) language system.
  - Parole descries the concrete use of language.
- Linguistics studies the structure of langue as a science.
- Most of modern natural language processing using computers are the extension of Saussure’s theory. Based on the study of langue, the area of research has been expanded to parole.

Science of Natural Language

- Before Saussure, language was considered to name natural objects. (Theory of Names)
- Saussure assumed that the nature in chaotic state could only be distinguished when language described reference (so that each object was recognized). (A 180 degree conversion from conventional views.)
- In other words, the autonomy of language was proposed. Language study became a science on the subject independent to the natural world.

- Autonomy of Language->
  - Signifant <- signe -> signifie
  - Pronunciation & spelling sign concept (object)
  - Signifant and signifie are inherent in language, not external constituents. = autonomy of language
- Arbitrary Nature
  - Sign, spelling, pronunciation, and concept are determined arbitrarily. (This framework is understandable. But how does it work?)

C.S.Pirce

- Pirce studied human cognitive processes while Saussure suggested signifiant against signifie as well as the arbitrary nature.
  - Pirce introduced an "interpretation" to treat language in context.
  - Three frameworks below were proposed.
  - Abstraction: from left to right.

icon	index	symbol
abduction	induction	deduction
term	proposition	artumentation
Saussure excluded this area from his study due to the independency of language.	signifant	signifie

- Deductive Reasoning
  - A conclusion is reached from deductive reasoning only. With the system of axioms provided an argument must be true.
- Inductive Reasoning
  - A valid universal rule is reached based on multiple supporting premises.
    - Human->die, stars->die -> everything->die
- Hypothesis-based Reasoning
  - A hypothesis regarding the real world is reached according to rules and provided results.
    - A dies. Human->die -> A is human.
    - Skeptical, but moral
  - Language, regarded as reasoning and language on a daily basis, and a mirror of the real world.
  - The relationship of language and the real world is yet to be discovered.
    - Does artificial intelligence (AI), such as robots, which can experience the real world, give new knowledge and understanding?

Chomsky - Computing Language

- It is impossible to cover all aspects of synchronic langue.
- Linguists have challenged this insolvable question.
    - Focused research on specific phenomena. E.g. the difference between "*ha"* and "*ga"*
    - "*wa-i-n <u>ga</u> su-ki-da* 'I like wine (in particular).' " v.s. "*wa-i-n <u>ha</u> su-ki-da* 'I like wine (but not necessarily other kinds of alcohol).' "
- Linguists have performed reasoning based on the linguistic phenomena that they observe or hear.
    - A massive database of sentence examples which linguists have stored and organized.

Chomsky - Computing Language

- Chomsky’s view on the ability for language that it is heredity and genetically transmitted. (human innate aptitude) supports the possibility that the nature of langue should be revealed by using our aptitude on language.
  - E.x. John kills him. (him != John)
  - John kills himself.
- Obviously, the area of research is limited to unconscious knowledge of grammar (syntax), and semantics is excluded.

History of Computational Linguistics

- With the birth of computers in the 1940s, computational linguistics started.
  - In the early 1950s, IBM’s Luhn proposed a computer system which extracted abstracts from documents.
- Various research projects were launched to create a machine translation system.
  - In the 1960, the report by ALPAC (Automatic Language Processing Advisory Committee) said that machine translation was an important task…

Cognitive Revolution

- The question laid out before the cognitive revolution was whether or not the science of language could be based on a deductive system like physics. (1950s)
  - Inductive analysis only with data provided. Structuralism denies the instinct orientation of analysis.
  - Machine’s capability back then was very limited. The inductive approach did not develop due to a lack of algorithm models.
- Cognitive Revolution in the 1960s: a top-down model for human language processing as well as information processing.
  - Chomsky’s transformational generative grammar
  - Newell & Simon’s solution: artificial intelligence (AI)
  - The advancement of computer’s performance contributed to a large part of research development.

Turing Test

- How to develop a system for natural language processing which can pass the turning test?
- A huge multiplication table
  - A table matching sentence and meaning, or Japanese and English sentences.
  - This is like cheating, and does not reveal the essence.
  - The table explodes because possible cases are unlimited.
- Algorithms to cope with the unlimited possibilities.
  - Chomsky’s model and AI-based approach
  - Possible sentences and contexts are endless. What an algorithm can deal with them?
  - Personally, I can only come to terms with the "division and governance" approach.

Top down v.s. Bottom up
Rationalism v.s. Empiricism

- Typical pitfalls are…
- Developing theories in which data is not taken care of enough (Top-Down), and
- Storing data without logical approaches (Bottom-Up).
- The following slides will explain the history of machine translation research, through which the struggles between the top-down and bottom-up strategies are described.

Old Bottom-Up Generation: Structuralism

- An attempt to transform speculative linguistics into a science of language.
- Subjective speculation should not be provided in the observation of data collection so that the real nature of language should be clarified.
- Verb suffixes *te* v.s. *de*
  - They are the same phonetic segments although *te* becomes *de* after a nasal verb such as *si-n-de* 'die'.
  - The relationship between "nasal sound v.s. non-nasal sound" becomes complementary distribution.
  - Idea of minimal pairs:
- Is there no subjectivity involved when we put *si-n-de* 'die' and *i-ki-te* 'live' in the same category?

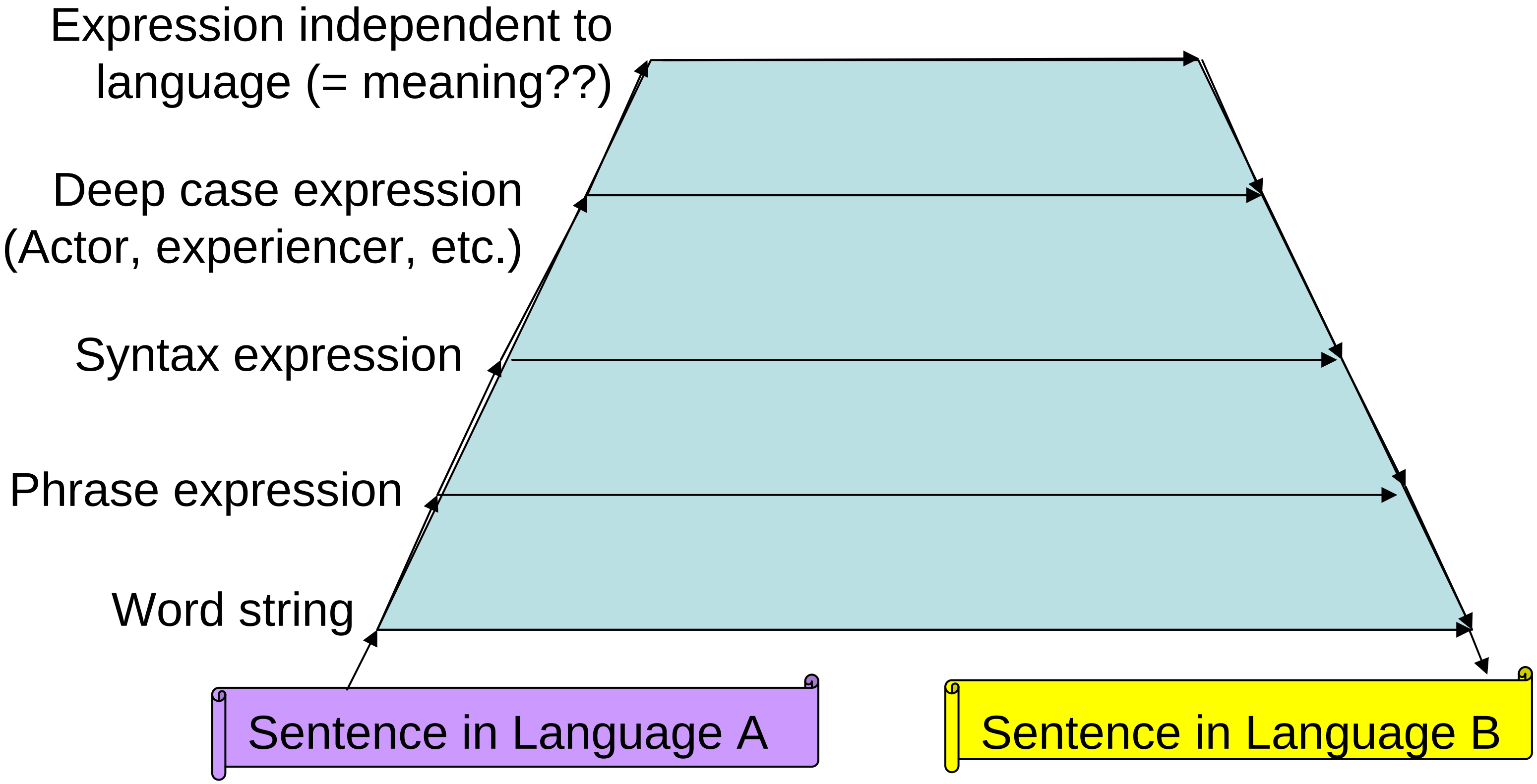
Rationalism

- Starting point: Computational models, which are independent to language, are presumed.
- Models are better accepted when they are as simple as possible.
- Saussure’s idea is to separate language from the real world.
- An initial performance is not good with this approach; however, as it develops, the performance exceeds that of the bottom-up systems (BT). The prediction ability of BT is still low because they look at actual data only.
  - What if the initial model does not work?

- Chomsky’s language theory (of universal grammar) that provides a schema independent to rules within an individual language faculty.
- Syntax can best correspond to the models of independent languages when only languages are taken into account.

Transfer Fundamentalist

- Transit from language A to language B at a level in the figure below.
- At the transit level, a conversion list of expressions for language A and B can be created. (belief)
  - Even if the conversion list is very long…


Expression independent to language (= meaning??)
Deep case expression (Actor, experiencer, etc.)
Syntax expression
Phrase expression
Word string
Sentence in Language A
Sentence in Language B

Issues with Transfer Fundamentalist

  - As the levels go up, the structure becomes large. Even so, the transition from language A to B is achieved because…
- This notion is based on formal semantics that once a part of the meaning is determined, the meaning of the whole part is determined by combining the parts. But…
- The pair of words in language A and B bring more than one meaning.
  - *Yu* 'hot water', *mizu* 'water' -> ‘water’
- Some grammatical characteristics and function words exist in either language.
  - Articles and the persons of noun
  - A detailed, complex conversion list may be helpful with handling the situation.

Issues with Transfer Fundamentalist

  - The most serious issues were…
- Context dependency of meaning:
  - To convert nouns from language A to language B in which language A does not have singular and plural forms, the understanding of context is required. However, the variety of contexts is unlimited.
  - At default, all nouns are handled in singular forms. They are converted into plural forms with the indication of contexts.
  - *"Ke-kkō-de-su."*-> 'thank you' or 'no thank you'
    - Impossible to translate properly with the default setting!?

Sings

-Widen the scope: Artificial Intelligence (AI)-

- Contribution of signs and theorem to the edifice of knowledge (Wittgenstein as previously mentioned)
  - The existence of sign lies in itself. Reasoning by using signs can be defined as an operation for any assumed group. (Extension theory)
  - The study of AI had taken this direction until the 1980s. Computers back then had poor input/output interfaces, and could not communicate with the external world.

- However, the system was only be able to meet limited purposes (expert system).
- The interaction with external world has been emphasized since 1980’s.
  - Subsumption architecture of robotics
  - Distributed Intelligence
  - Agent (Current software engineering)
- Trend for signal processing with context information taken into considerations.

Signal Processing with Context Information

- Sings are consisted of:
  - a. core meaning, and
  - b. meaning in terms of language use based on context.
- Thus, a large volume of examples were collected to develop sample-based translation so that (b) be taken into accounts.
  - Translation sample:
    - *Ta-ro ha syō-se-tu wo yo-n-da.* vs. "Taro read a novel."
    - The context such as *Ta-ro* = human and *syō-se-tu* = character media has the capability to define *yo-mu* as ‘read’.
  - Nevertheless, the core meaning of each word must be predefined to achieve a satisfactory process.

Meaning of Words

- The meaning of a word was expressed with the segmented meanings of the term (in the 1980s).
  - Kill = cause (someone (alive -> death))
- What basic elements were necessary for an adequate expression?
  - A mainstream since the 1990s?
- The meaning of a word is disambiguated based on words which co-occurred in the context where the target term is used.
  - A large corpus (NY Times for twenty months) exhibited 90% precision in a disambiguation process for the meaning of "capital", which can mean financial capital or a capital city.
  - Currently an approach is being proposed in which the translation of unseen words can be estimated according to the similarity of the terms that co-occur in the same context.

Empiricism v.s. Data-oriented Parsing

- Empiricism introduces architectures more radical than data-oriented approaches in terms of semantics along with contexts and the use of language.
- IBM's statistical machine translation (SMT) systems (in early 1990s).
- In this system, IBM discovered some English and French expressions, which might not have been manually recognized, by taking a pure machine translation methodology (statistical machine-learning algorithms).
    - EM, Viterbi search, etc.
    - A large volume of memory and high-spec computers
    - A large volume of quality translation pairs (teaching data)
        - Quality data is hard to be built up.

Computational Linguistics in the Late 20th Century

- Due to the improvement of computer performance and capabilities in the 1970s, the development of machine translation systems was made into a reality.
- The first systems developed based on linguistic knowledge.
  - Linguistics failed to cover a wide range of actual phenomena of the use of language.
  - Analysis was performed for limited cases. E.x. "*ha"* v.s. "*ga"*
  - In the 1980s, computer scientists initiated the description of grammars.
- Grammars observed from formal, well-written texts were not practical.
  - Only 60% of grammatical rules articulated by linguists were said to be working rules…
  - In reality, different varieties in language were too wide.

What is a science of natural language?

- What can we do, other than articulating the relation between language and the real world?
- Machine translation (MT) handles original and target language. Both sides are language in nature, and the architecture of MT can be concluded in the study of language. Current MT systems are incorporating such an idea.
- Activities such as document categoritization, information search, summarization, and paraphrasing are completed in the realm of language.
- ➢ The combination of pictures/movies and texts presents various hard issues.
- ➢ Linguistics, as a closed discipline in the world of language, cannot create a computer system which can influence the natural world and human society.
  - ➢ For instance, it is difficult to built an linguistic interface between machine and human.
  - ➢ A robot needs an environmental model to understand an imperative statement (e.g. "Put away this trash into that can."

Computational Linguistics since the 1990s

- Is it a real science when researchers depend on their instincts?
  - Only 60% of grammatical rules articulated by linguists were said to be working rules…
  - In reality, different varieties in language were too wide.
- How can we acquire a huge collection of real linguistic data to perform comprehensive analyses and automatically discover grammar?
  - Statistical natural language processing (a mainstream from the 1990s)

Computational Linguistics

- Sound recognition:
  - A grammar of spoken language is necessary in addition to that of written language.
- A large scale corpora emerged:
  - A huge volume of electric texts (GB class) prepared for computer processing = corpus
  - A variety of new issues have been recognized as machines become capable of processing newspapers from the past ten years.
- New issues are being identified.
- Is it possible to describe a grammar that corresponds to such a wide variety of language phenomena?
- Is it possible to acquire enough linguistic data?

Modern Issues - Computational Linguistics

- New issues are being identified.
    - Is it possible to describe a grammar that corresponds to such a wide variety of language phenomena?
    - Is it possible to acquire enough linguistic data?

- In many cases, researchers are lack of data for their targeted linguistic phenomena.
    - The issue of data sparseness:
    - For instance, when they try to determine the probability for every two consecutive terms, they cannot find linguistic data of such two terms in sequence.
    - One approach is to improve the precision of estimation based on small sample work in statistics.
    - The knowledge developed by linguists is referenced if appropriate and necessary.

- However, in reality, researchers need to work with low quality data on aligned translation pairs.
  - When the data is not paired, bilingual corpora on the same subject or topic are used.
  - Dictionaries for basic lexicon are at least available.
  - Computers are speedy with a large memory.
  - But machine learning paradigms almost seem run out.
  - Then, a collaboration with human??