Introduction to Natural Language Processing

"Language or Math?:
Statistical Linguistic Processing goes with Computing Power"

Hiroshi Nakagawa

(Information Technology Center; Mathematical Informatics, Graduate School of Information Science and Technology; Graduate School of Interdisciplinary Information Studies, The University of Tokyo)

nakagawa@dl.itc.u-tokyo.ac.jp

http: //www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

Linguistic Processing of a Large-scale Corpus

- Desire to obtain useful information from a large-scale corpus.
- Conventionally, the processing of corpus was carried out manually for statistical data analysis. However, the size of a corpus became too large for manual processing. Today, a corpus is processed with computer.
- To process a 100MB (GB) corpus is a time-consuming task which slavishly performs CPU-intensive operations such as morphological analysis and syntactic analysis that involve unnecessary jobs.
- A less CPU-intensive statistical analyses is currently under consideration.
- Various statistical approaches are being studied, many of which take words as a useful observable linguistic unit.

List of Available Statistical Linguistic Approaches

- Information Retrieval from Text Data
  - N-gram statistics
  - Word, term extraction, frequency mapping, and indexing
  - Alignment of multiple corpora
  - Alignment of bilingual corpora; automatic selection of translation
- Text conversion processing
  - Document categories
  - Automatic paragraph break
  - Automatic summarization; topic extraction
  - Machine translation
  - Alignment of multi-modal corpora (e.g. video indexing)

Why Statistics for Language?

- Reason to use statistics value:

  1. To know about the language that we use.

  2. To create a linguistics model (grammar; used-based linguistics (pragmatics); rules of document structure)

  3. To build a dictionary or e-dictionary.

  4. To quote for computing resources required for linguistic processing.

Total & Type

- *Ta-yo-ri /no/na-i/no/ha/yo-i/ta-yo-ri* 'No news is good news.'
- Morpheme = the smallest meaningful unit of language, which cannot be broken into even a smaller unit

What is “words”?

- Total number of morpheme = 7, Total number of types of morpheme = 5
- Frequency: *ta-yo-ri* = 2, *no* = 2, *na-i* = 1, *ha* = 1, *yo-i* = 1
- What is “words”?
  1. Is *no-de* 'because, since' one word?
  2. How about interaction marks such as *odoriji* 'repetition mark' in *hi-bi* 'day after day'?
  3. Is *ni-chi/gi-n* 'Bank of Japan' one word?
  4. Is *ni-ho-n/gi-n-ko-u* 'Japan/Bank; Bank of Japan' one word or two? (Apparently two morphemes)
  5. This has long been a challenging question for scholars of language.
  6. From practical point of view, the issue is regarding which word/term should be recorded in a dictionary. Words can be defined based on application and purpose.

Linguistic Unit for Statistical Analysis

- A phoneme is each sound corresponding to an Alphabet character represented in *Kunrei*-system *Romaji*.
- Character depends on a character set given, for instance, upper and lower cases in Alphabet characters; new character styles in Japanese *kanji* system. Is the new character style of *sa-wa* 'stream; bush' different from its old character style?
- Words can be distinguished with spaces between words according to Orthography as in European languages. Is *to-ra-no-ko* 'apple of one's eye' can be treated as one word? -> Issue of collocation
- Compound:
  1. Compound noun:

     nihon gakujyutu kaigi daigobu kaiinn senkyo nittei kentou kekka houkokusyo betuzoe siryou hoi san peiji ni gyoume 'on page 3 in the second row of addendum for attachment to the report by Science Council of Japan regarding the day for the election of members of the fifth group' ????
  2. Compound predicate:

     "*o-i-ka-ke-ru* 'run after' " = "run" + "follow"

     "*fu-ke-ko-mu* 'glow old' " = "glow old" + "turn"

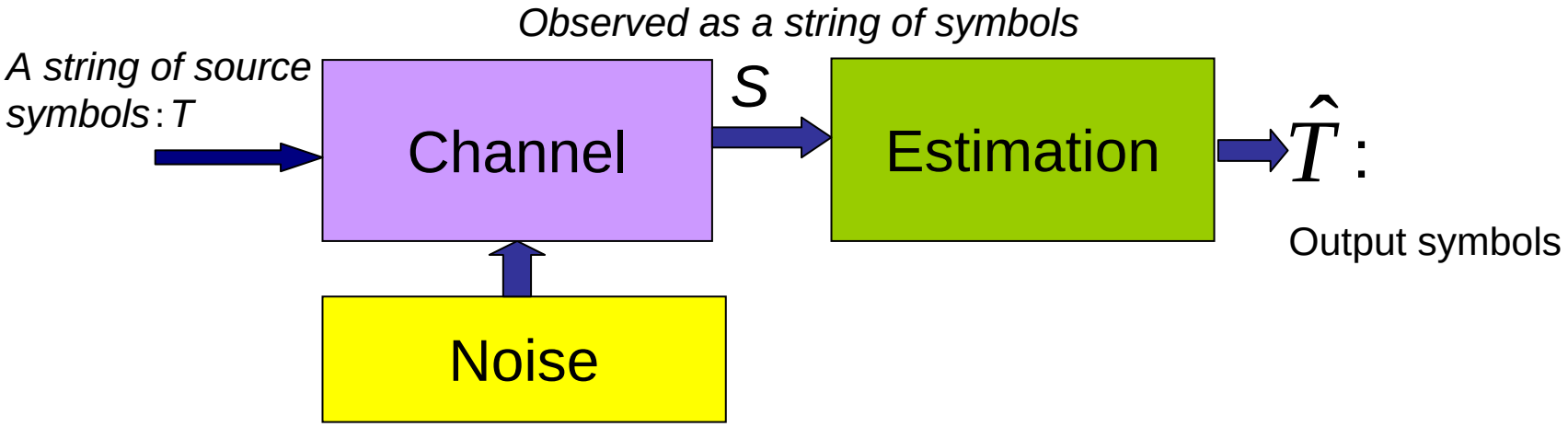
     "*I-ta-ga-ru* 'have pains' " = ??
- Phrase: noun phrase and verb phrase: phrase is gathering more attention than sentence in linguistic informatics.
- Sentence: In Japanese writing, *kuten* (sentence point, or period) is used. How about spoken language?
- Dialogue: a current issue is how to recognize dialog structure and domain.

Lexical Category vs. Functional Category & POS

- Lexical category includes the syntactic elements that are part of a lexicon, such as predicates and nouns. Also called “content word”.
- Functional category includes the elements with little lexical meaning such as predicates and nouns that modify or represent the relations of each content word. Also called “function word (functor)”.
- The distinction of content words and their statistical characteristics are important in the efforts to capture the meaning from sentences such as information search. Therefore, the research in these areas are being proceeded. On the other hand, linguistics and machine translations take authentic approaches toward language and utilize function words to identify grammars and meanings. It is recognized that these two approaches are not completely in line with each other.
- POS describes the grammatical characteristics of content words (and of course function words). In some cases, these words (especially, content words) do not frequently appear. For a statistical analysis, it is advised to group different words of the same POS into the same category.

Text Conversion Model

- Text as a sequence of symbols (words and strings)
- Noisy Channel Model:



The string of source symbols *T* is approximated to find $\hat{T}$ based on the observed *S*.

Estimation Process

- Communication channels represented by conditional probability-based model: *P (T|S)*
- Calculate $\hat{T}$ to maximize the posterior probability *P (T|S)* that *T* is found after *S* is provided.

$$\hat{T} = \arg\max_{T} P(T \mid S)$$

According to Bayes' theorem

$$= \arg\max_{T} P(S \mid T) P(T)$$

- The characteristics of source channel can be used to calculate *P (T).*
- *P (S|T)* is the probability that the source channel *T* changes into *S* due to the noisy channel. Recompute the value based on the data (a corpus) including many *(T,S)* pairs.

Sample Applications of Noisy Channel Model

- Machine translation
  - *P(S|T)*: the probability that text *T* in the original language is translated into text *S* in the target language
  - *P(T)*: the degree of naturalness of original text *T*. For example, the probability that word 3-gram appear in the corpus containing N word sequences.
  - $\hat{T}$: the output of machine translation
- Document summarization
  - *P(S|T)*: the probability that summarized text *T* is created from original long text *S*
    - The probability can be computed when pairs of original texts and summarized text are provided as teaching data.
    - However, it is substantially difficult to statistically determine how the texts are summarized.
  - *P(T)*: the degree of naturalness of the summarized text
  - $\hat{T}$: the summarized text produced by the document summarization system

- Document Category
  - Denote *S* as provided texts and *T* as a category in *P(T|S).*

    Approximated category $\hat{T} = \arg\max_{T} P(S \mid T) P(T)$
  - *P(T):* Occurrence probability of documents in category *T*
  - *P(S|T):* Occurrence probability of text *S* within the documents in category *T*
  - Various methods for the modeling of *S* are available. Simple one is to use words $w_1,\ldots w_n$
  - *P(S|T) = P($w_1,\ldots w_n$|T).* Define $w_1,\ldots w_n$ as independent members to simplify the equation… $P(w_1,..., w_n \mid T) = \prod_{i=1}^{n} P(w_i \mid T)$
  - Called “naïve Bayse” category.