Introduction to Natural Language Processing

"Text Resource Pool"

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/

PC, HDD, CD-ROM, Internet

- Since 20 years ago when word processors spread, e-document has become a popular format. (PC, HDD, etc.)
- Desire to share and utilize e-document. (CD-ROM, the Internet, etc.)
- e-document in standard formats:
  - First, plain text was primarily used.
  - Later developed into tagged structured document, such as SGML, which was further refined as XML.
- A large volume of text data were stored and organized (as corpus, plural *corpora*) for further utilization.
- Linguistic resources to be gathered and structured for the use of corpora.

What are “Linguistic Information Resources”?

- Corpus
- Linguistic Model
- Dictionary
- Grammar
- Thesaurus

Corpus (Corpora)

- Corpus: a large set of digitized sound or text data.
- Sound Corpus:
  1 One person vs. many people
  2 Utterance made for a purpose vs. Utterance naturally spoken
  3 Existence of transcripts
  4 Tags generated for the transcripts
- Text Corpus:
  1 Raw Corpus
  2 Tagged Corpus (with part-of-speech tags based on morphological analysis)
  3 Bracket Corpus
  4 Processed Corpus
- Monolingual Corpus vs. Multilingual Corpus (side-by-side comparison of translation)
- Subject area and genre:
  Newspapers, journals, the Internet, academic disciplines, etc.

Merits and Demerits of Linguistic Processing based on Probability and Statistics with Corpus

- Accurate
- Broad-coverage
- Robust
- Data-driven v.s. Rationalism
- Issues pertaining to the applicability to some areas which a specific corpus does not cover
- Sparseness, or zero frequency issue

Bilingual Corpus

- Bilingual corpus contains text data written and formatted in two languages.
- Comparable corpus:
  - The contents of corpus are comparable.
- (Aligned) Parallel corpus
  - 0 Bilingual corpus whose text data simply has translation in the other language.
  - 1 Corpus formatted and aligned for translation comparison.
  - 2 Corpus aligned even at word level.
- Alignment is a primary research subject in 1990s in the field of corpus linguistic processing.

Dictionary

- Of course, e-dictionaries.
- Dictionaries alternative for books (monolingual or parallel)
- Dictionaries for the *kana-kanji* conversion in Japanese word processing (monolingual)
- Dictionaries for spelling check function in word processing (monolingual)
- Dictionaries containing comparable translation for machine translation (semantic description may be necessary.)
- Dictionaries for natural language processing (with semantic descriptions)
- Dictionaries for search engine (data for referencing keyword and document)
- Dictionaries for cross-lingual search engine (parallel corpus)

Elements in Dictionary

- Headword
- Pronunciation
- Transcription
- Part of Speech (POS)
- Grammatical data (conjugation and sentence pattern):
  - 1 Conjugation is gender, number, person, tense, and aspect in inflective language
  - 2 Tense, aspect, and conjunction for agglutinative language
- Adjacent and co-occurrence data
- Definition (The description of meaning is challenging.) and conceptual structure
- Examples
- Related Term (synonym and antonym)
- Translation
- Selection restrictions, etc.

Sample Description in Dictionary (IPAL Dictionary)

Headword/Pronunciation	*A-ge-ru (hiragana)* (1), *A-ge-ru (katakana)*
Grammatical Information	下1　Base: age
Transcript	*A-ge-ru* 'move up, raise'
Grammatical Information	Transitive verb <<Syntax>>
Sentence Pattern	N1-*ga* N2-*wo* (N3-*kara*) (N4-*ni*) (N5-*e*)
Example	He put up a book on the shelf.
Definition	To move up something
Related Term (RT)	Hypernym = up, Antonym = bring down

Descriptions in Dictionary –Details in Semantics and Grammar Vol.1

- Characteristics of Nouns as a subject or an object of predicate:
  - 1 Person: e.g. OK: “it is hot.” X: “He is hot.”
  - 2 Animate or inanimate: e.g. OK: “It is hot in this room.” (“Room is hot.” ?)
- Voices of predicate (passive, active, causative, reciprocal, etc.)
  - OK: “hit” vs. “get hit”
  - X: “be born” vs. “be able to be born”
  - OK: “be born” vs. “make someone bear”
  - X: “be sick” vs. “make someone sick”
  - X: “be born” vs. “be born each other”
- Semantic features of predicate itself:
  - 1 Grammatical features, such as intransitive and transitive verbs.
  - 2 Movement, change of state, attachment and detachment, change of pattern, etc.
- Semantic features of noun:
  - 1 location, time, animate, concrete, abstract, organization, …
- Semantic features of noun as an object or a subject of predicate:
  - e.g. “to come from N”: N is a location.
- Selection restrictions can be described with semantic features.

Descriptions in Dictionary –Details in Semantics and Grammar Vol.2

- OK: *ha-ya-ku-ha-shi-ru* 'run fast'
  X: *ha-ya-i-ha-shi-ru* 'fast run'
  Organizing these rules makes phrase structure grammar.

- Suffix juncture:
- The connectivity of predicate suffix is also important.
  - OK: *shi-n-de-a-ru* 'have dead'
  - X: *ko-ro-shi-te-a-ru* 'have and killed'
- Japanese suffix:
  - *te-i-ru, te-a-ru, te-o-ku, te-shi-ma-u, te-ku-ru, te-i-ku, ga-ru, na-i*
  - *te-ku-re-ru, te-ya-ru, te-a-ge-ru, te-mo-ra-u*

Thesaurus

- Thesaurus is a system in which every term is listed in a given domain of knowledge (concept) and in a set of related terms and concepts.
- *Bunrui Goi Hyo* (a Japanese lexicon/thesaurus developed by the National Institute for Japanese Language):

1. 1 Abstract relations
    1. 100 *kosoado* 'demonstrative pronouns'

1. 2 Entity of human activity
    1. 200 *wa-re* 'we'*, kare* 'he'

1. 3 Human activity - spirit and action
    1. 300 heart

1. 4 Product and tool
    1. 400 goods - money, extraneous substance, actual goods, cheap goods, special goods
    1. 471 road and bridge - national road, underground path, railroad, bridge

1.5 Natural object and natural phenomenon
    1.500 stimulus
    1.501 light - brightness, luminous, saturation

Relations of Concept

- Relations of superior and subordinate concepts:
  - Goods > animate beings > animal > vertebrate> mammals > human
- Called “is-a” or “kind-of” relationship in artificial intelligence.
- Brother nodes:
  - Limited express, express, local, local trains which stop at every station
- Hyponym:
  - OS -> Linux
- Antonym: (In many cases, antonyms hold brother nodes.)
  - Departure <-> arrival, north <-> south
- Synonym:
  - *Ni-chi-gi-n* 'BOJ' = *Ni-ho-n-gi-n-koh* 'Bank of Japan'
  - Syu-soh 'chancellor, premier, prime minister' = ? *So-u-ri-da-i-ji-n* 'prime minister'
- Meronym vs. holonym -> part of
  - Engine roof, tire <- automobile

WordNet

- WordNet is an English thesaurus developed under the direction of Dr. George A. Miller at Princeton University.

- EuroWordNet is created as a multilingual thesaurus for European languages which are structured in the same way as WordNet. “Synsets” in each European language (semantic class and sets of synonyms) are interconnected to corresponding English synsets in WordNet. Synonym search is available for any language pair. In addition, upper structures in the thesaurus are completely shared. Danish, Italian, Spanish, German, French, etc. are available.

WORDNET

Sample:

The **noun** "human" has 2 senses in WordNet.

1. person, individual, someone, somebody, mortal, **human**, soul -- (a human being; "there was too much for one person to do")
2. homo, man, human being, **human** -- (any living or extinct member of the family Hominidae)

WORDNET

Sample (cont.):

- The **adjective** "human" has 3 senses in WordNet.

  1. **human** -- (characteristic of humanity; "human nature")
  2. **human** -- (relating to a person; "the experiment was conducted on 6 monkeys and 2 human subjects")
  3. human (vs.. nonhuman) -- (having human form or attributes as opposed to those of animals or divine beings; "human beings"; "the human body"; "human kindness"; "human frailty")