Analysis of Language Resources

Fifth Lecture

Hiroyuki Akama

Review:
Association or Embodiment

- Lexical Information vs. Bodily Reaction
- Statistical Linguistics vs. Cognitive Linguistics
- Written and stored things vs. Body action and perception
- However, there is a basic Latin word that designates both of these opponent concepts. →**CORPUS**! (simultaneously meaning " a set of documents", "body (physical existence)" and "principal component".) (The word "corps" means "body" in French.)

New Linguistics for Association

- We need a new type of linguistics that would harmonize the two conflicting traditions of semantics in a wider sense of “corpus”.
- Some tentative plans have been made in this direction (especially in the field of the graph theory).
- The necessary condition of this task might be : *a systematization and application of large-scale knowledge resources*. (Mission of COE21-LKR)

Investigating Semantic Networks

- The new field of semantics will emerge from a future interdisciplinary research on various types of association.
- 1) Association at the micro-level (Free, Perceptual, Sensory-Motor…) : the co-occurring words data taken from the spontaneous speech of an individual ; his (or her) bodily or neural reaction to the word usage ("*Parole*" in French)
- 2) Association at the macro-level (research on world of words): the overall instances of word pairs taken from a large corpus. ("*Langue*" in French)
  - There will be many possible indexes allowing us to perceive the "shape" of the world of language.

Purpose of the Latter Half of the Lectures

- In order to realize a fruitful cooperation in these two research fields, we have to know beforehand the status quo of the corpus linguistics.
- Before tackling such an ambitious project, we need to learn some basic concepts of computer linguistics.
- Quantitative linguistics is considered as the first door to know the world of the language, particularly its *instinctive shape*.

The Fundamental Indexes of Quantitative Linguistics

- Term Frequency, Degree…
- Frequency distribution, Degree distribution…
- Needless to say, statistical measures such as Mean, Variance, Standard Deviation…
- Each word in a document can now be considered as a variable or an observation instance.

Term Significance

- Review: According to the modern linguistics, the possibilities of defining the semantics can be founded only upon the **association** of words (Saussure, father of the modern linguistics).
- Review: Gathering the lexical co-occurrence data is indispensable for information retrieval→Basis of LSA, LSI (Landauer et al), HAL (Burgess). Word Space (Shütze et al)…
- The ***signification*** of the term is also its **relative *significance*.**
- The term **ranking** and the term **weighting**

Indexing Problem : Trade-off

- Trade-off between significance and frequency for Information retrieval
- How to find the important words for retrieval
- High Frequency →Noise Words, Functional words
  - High Recall, ○ Exhaustiveness
  - Low Precision, ×Specificity
- Low Frequency →Rare words, Unusual words
  - High Precision, ○ Specificity
  - Low Recall, × Exhaustiveness

Zipf's First and Second Laws

- Let ***f*** be the term frequency of the word ***w*** in a document ; let ***r*** be the rank of the word ***w*** in the descending order of frequency.
- According to empirical observations,…
- **First Law**: ***r·f=C C:Const***
  - Pretty good for frequent words (in the zone where there is only one word for one rank).
- Let F_f signify how many kinds of words are found at the rank ***f*** (whose word frequency is ***f***).
- **Second Law**: $\frac{F_1}{F_f} = \frac{f(f+1)}{2}$
  - Applicable to the least frequent words

Words of Intermediate Frequency

- Calculate the frequency value which would satisfy both Zipf's two laws in order to know their limits of applicability.
- Let F_f be 1, because according to the first law, there has to be only one word at the same rank level.
- Substitute 1 for F_f in the expression of the second law, and you will get $f = \frac{\sqrt{8F_1 + 1} - 1}{2}$

Summary

- Enlarge the concept of corpus to bridge cognitive linguistics and corpus linguistics
- Corpus linguistics, computer linguistics, quantitative linguistics
- Word Frequency and term weighting methods
- Trade-off for retrieval (high frequency and low frequency)
- Zipf’s first and second Laws (intermediate frequency)