Introduction to Information Extraction

Term Extraction

Hiroshi Nakagawa

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

Term Definitions

- Definitions
  - Term: a word or combination that has a precise meaning in a peculiar domain
  - Terminology: vocabulary used in a particular domain, or a group of terms
- Difference between general and technical terms
  - Terms used in different types of documents
  - A group of people using terms (general v.s. professionals)
  - Terms bearing a single meaning (for a smooth communication)
- Vocabulary of the technical terms used to represent a concept of a certain domain
  - *“Bu-do-ma-ri* ‘material yield; success rate?’ ”, “*Yo-ko-mo-ti* 'side take, or transportation between a main station or warehouse to the destination’ ”, “*Ta-ma-ga-ke* 'ball hinge, or to operate a power shovel’ ”, “*Ha-ra-ku-ku-ri* ‘bridge' "???
- These terms defined based on technical documents and corpora they appear in.

Termhood v.s. Unithood

- Unithood: the degree of stability of syntagmatic combinations (word order, syntax, and semantics) or collocations
- Termhood: the degree to which a linguistic unit such as combination and compound is related to domain-specific concepts

Framework of Automatic Term Extraction

1. Parse a set of documents in a specific domain (corpus), and attach POS tags.
2. Extract a sequence of words if they are considered to be appropriate terms based on the POS indexes. List them as possible terms.
3. Weigh each possible term in according to termhood.
4. Extract appropriate terms from the list of prioritized possible terms. (for instance, select up to a defined number)

Possible Term Extraction based on Unithood

- Character N-gram is not suitable because words must be extracted. This architecture must be employed in Chinese and German whose word extraction is relatively hard.
- Parse a set of documents in Japanese and segment into the unit of words.
- Attach POS tags to each word.
- First define POS strings for (technical) terms (specific words can be included). List them as possible terms if they match the pre-defined terms.
- What types of POS strings should be selected?
- What types of POS strings and word structure do exist for technical terms?

Grammatical Structure of Terms

- Japanese
  - Noun^{+}, E.x. information system, square-well potential model, Chomsky's hierarchy
  - Noun^{+}*no* Noun^{+}, E.x. "*ge-n-go no bu-n-se-ki* 'linguistic analysis' "
  - Adjective Noun^{+}, E.x. global language
  - Number Noun, E.x. type-3 language
- Adjective: *i*-adjective: "*o-o-ki-I* 'large' "

  na-adjective: "*ze-tta-i-te-ki* 'absolutely' "

Grammatical Structure of Terms

- English
  - ->less than | or, A^+ :repeat A more than once, A: repeat A $^?$ 0 or 1 time
  - Noun$^+$, E.x. computer network
  - Noun$^+$ “of” Noun$^+$, E.x. lack of stimulus
  - Noun Preposition Noun, E.x.
  - Adjective Noun$^+$, E.x. global data, balancing act
  - Number Noun$^+$, E.x. first order logic
- In short,
  - ((Adjective|Noun) * | (Adjective|Noun) * (Noun Phrase) $^?$) (Adjective|Noun) *) Noun

Term Weighting in terms of Termhood: Quantifiable Measure

- dj: document set of domain, $D=\{d_1,\ldots,d_j,..d_{n\,(D)}\}$
- wj: word appeared in *D*, $W_D=\{w_1,\ldots,w_i,..\}$
- wij: word (*wi*) appeared in d_j
- f (wij) = occurrence frequency of w_i in d_j
- g (wij) = 1 when w_i is in d_j,
  0 when w_i is not in d_j
- Two types of measures based on the occurrence frequency in document:
  - average word freq in d_j = $$If_{ij} = \frac{f(wij)}{\sum_i f(wij)}$$

- Appearance of word *i* in document *j* in a special way:

$$Iij = \frac{f(wij)}{\sum_i f(wij)} - \frac{\sum_j f(wij)}{\sum_i \sum_j f(wij)}$$

- Similar idea: tf × idf

$$tf \times idf_{ij} = f(wij) \times \log\left(\frac{n(D)}{\sum_j g(wij)} + 1\right)$$

- In these two equations, word w_i which appears frequently only in document d_j is prioritized.

Summary: Calculation Method of Termhood

- All methods illustrated in the slides are for measuring termhood. In short,
  - Terms are some particular words which frequently appear in documents.
  - Terms appear only in specific documents.
  - Terms (or index terms) pertaining to a document are those that substantially appear only in the particular document.
  - Terms are some particular words of specific distribution across the whole document.
  - Etc.
- Depends on the characteristics in terms of a set of documents. ->Document space based method
  - A different viewpoint (lexical space) will be introduced later in the course.

Unithood & Termhood of Compund & Collocation: Word-Vector Space Model

- Clarify the degree of stability of compounds and collocations (unithood).
  - Determine whether basic words (words other than compound words) co-occur at a substantially high level.

  -> Contingency matrix
  - Architectures utilizing statistical characteristics found from word relations in document space.

Contingency Matrix

- Statistical significance of the conjunctive occurrence of two words.

	W1	no W1
W2	a	b
no W2	c	d

- Reciprocal information

$$MI = \log \frac{p(w1, w2)}{p(w1)p(w2)} = \log \frac{aN}{(a+b)(a+c)}, N = a+b+c+d$$

- χ 2 test

$$\chi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(B+c)(b+d)}$$

- Log likelihood ratio

Contingency Matrix (Reciprocal information & Example)

- Statistical significance of the conjunctive occurrence of two words.

	University	¬University
Reform	a=10	b=5
¬Reform	c=5	d=980

- Reciprocal information

$$MI = \log \frac{aN}{(a+b)(a+c)}$$

$$= \log \frac{10 \times 1000}{(10+5) \times (10+5)}$$

$$= \log 3333 = 8.38$$

Contingency Matrix (Reciprocal information & Example -1)

- Statistical significance of the conjunctive occurrence of two words.

	University	¬University
Reform	a=10	b=100
¬Reform	c=90	d=800

- Reciprocal information

$$MI = \log \frac{aN}{(a+b)(a+c)}$$

$$= \log \frac{10 \times 1000}{(10+90) \times (10+100)}$$

$$= \log 9.09 = 3.18$$

Issue with Reciprocal Information

- Statistical significance of the conjunctive occurrence of two words.

	University	¬University
Reform	a=1	b=0
¬Reform	c=0	d=999

- Reciprocal information

$$MI = \log \frac{aN}{(a+b)(a+c)} = \log \frac{1 \times 1000}{(1) \times (1)} = 9.96$$

- Assessed greater than real terms ->dice coefficient (weighted)

$$Dice = \log(a \times \frac{2a}{(a+b)+(a+c)}) = \log(1 \times \frac{2}{1+1}) = 0$$

$$\textit{compare previous} \quad Dice = \log(10 \times \frac{20}{15+15}) = \log 6.7 = 2.74$$

Contingency Matrix (χ 2 test& Example)

- Statistical significance of the conjunctive occurrence of two words.

	University	¬University
Reform	a=10	b=5
¬Reform	c=5	d=980

- χ 2 test

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} = \frac{1000(9800-10)^2}{15\times 15\times 985\times 985} = 489$$

- The rejection rate in χ 2 test with 1 d.o.f. is less than 0.1%. ->Statistical significance is observed in co-occurrence.

Contingency Matix (χ 2 test& Example-1)

- Statistical significance of the conjunctive occurrence of two words.

	University	¬University
Reform	a=10	b=100
¬Reform	c=90	d=800

- χ 2 test

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} = \frac{1000(8000-9000)^2}{110\times100\times890\times900} = 0.11$$

- The rejection rate in χ 2 test with 1 d.o.f. less than 75%. ->Statistical significance is observed in co-occurrence.

Likelihood ratio

- Hypothesis H1: *p (w2|w1) =p (w2|¬w1)*
- Hypothesis H2: *p (w2|w1) >p (w2| ¬w1)*
- Given H1,H2 likelihood as L (H1),L (H2),
- if $\log\lambda = \log\frac{L(H1)}{L(H2)}$ is less than threshold C,

  w1 and w2 are collocation with statistical significance.
- The calculation of *L (H1),L (H2)* is complicated.

Sample Equations

$$H1: p(w2|w1) = p(w2|\neg w1) = p = \frac{a+b}{N}$$

$$H2: p(w2|w1) = p1 = \frac{a}{a+c},$$

$$p(w2|\neg w1) = p2 = \frac{b}{b+d} = \frac{b}{N-a-c}$$

$$b(k,n,x) = \binom{n}{k} x^k (1-x)^{(n-k)}$$ Binominal distribution

$$L(H1) = b(a, a+c, p)b(b, b+d, p)$$

$$L(H2) = b(a, a+c, p1)b(b, b+d, p2)$$

Sample Equations:

$$\boldsymbol{H1: p(w2 \mid w1) = p(w2 \mid \neg w1) = p = \frac{10+5}{1000} = 0.015}$$

$$H\,2:\ p(w2 \mid w1) = p1 = \frac{a}{a+c} = \frac{10}{10+5} = 0.67,$$

$$p(w2 \mid \neg w1) = p2 = \frac{b}{b+d} = \frac{b}{N-a-c} = \frac{5}{980+5} = 0.005$$

$$b(k,n,x) = \binom{n}{k} x^k (1-x)^{(n-k)}$$ Binominal distribution

$$L(H1) = b(a, a+c, p)b(b, b+d, p) = b(10,15,0.015)b(5,985,0.015)$$

$$L(H\,2) = b(a, a+c, p1)b(b, b+d, p2) = b(10,15,0,67)b(5,985,0.005)$$

$$\Rightarrow \frac{L(H1)}{L(H\,2)} = \frac{1.39 \times 10^{-34}}{1.60 \times 10^{-18}} << 1$$

$$\Rightarrow \log\left(\frac{L(H1)}{L(H\,2)}\right) = -53 \Rightarrow$$ Substantial Co-occurrence

Sample Equations-1

$$H1: p(w2 \mid w1) = p(w2 \mid \neg w1) = p = \frac{10+90}{1000} = 0.1$$

$$H2: p(w2 \mid w1) = p1 = \frac{a}{a+c} = \frac{10}{100} = 0.1,$$

$$p(w2 \mid \neg w1) = p2 = \frac{b}{b+d} = \frac{b}{N-a-c} = \frac{100}{900} = 0.11$$

$$b(k,n,x) = \binom{n}{k} x^k (1-x)^{(n-k)}$$ Binominal distribution

$$L(H1) = b(a, a+c, p)b(b, b+d, p) = b(10,100,0.1)b(90,900,0.1)$$

$$L(H2) = b(a, a+c, p1)b(b, b+d, p2) = b(10,100,0.1)b(90,900,0.11)$$

$$\Rightarrow \frac{L(H1)}{L(H2)} = \frac{6.58 \times 10^{-142}}{4.10 \times 10^{-142}} \approx 1$$

$$\Rightarrow \log(\frac{L(H1)}{L(H2)}) = 0.68 \Rightarrow$$ No Substantial Co-occurrence

Unithood & Termhood of Compound & Collocation (Lexical Space)

- Based on the internal structure of compound and collocation
  - Xtract
- Structure in lexical space:
  - Issues when the co-occurrence of structures are combined.
  - Statistics reflecting the lexical structure
  - C-value and junctures

Collocation:

- A sequence of two or more consecutive words
- regarded as a syntactic and semantic unit,
- Non-compositinality: its meaning cannot directly be derived grom its components
  - kick the bucket
- Non-substitutability: cannot substitute other word into its component
  - white wine ≠ yellow wine
- Non-modifiability: cannot freely modify its component
  - *“O-ku-ba ni mo-no ga ha-sa-ma-tta yo-u-na* 'not frank; mealy-mouthed' " ≠ “*O-ku-ba ni o-o-ki-na mo-no ga ha-sa-ma-tta yo-u-na* 'have something in my back tooth' "

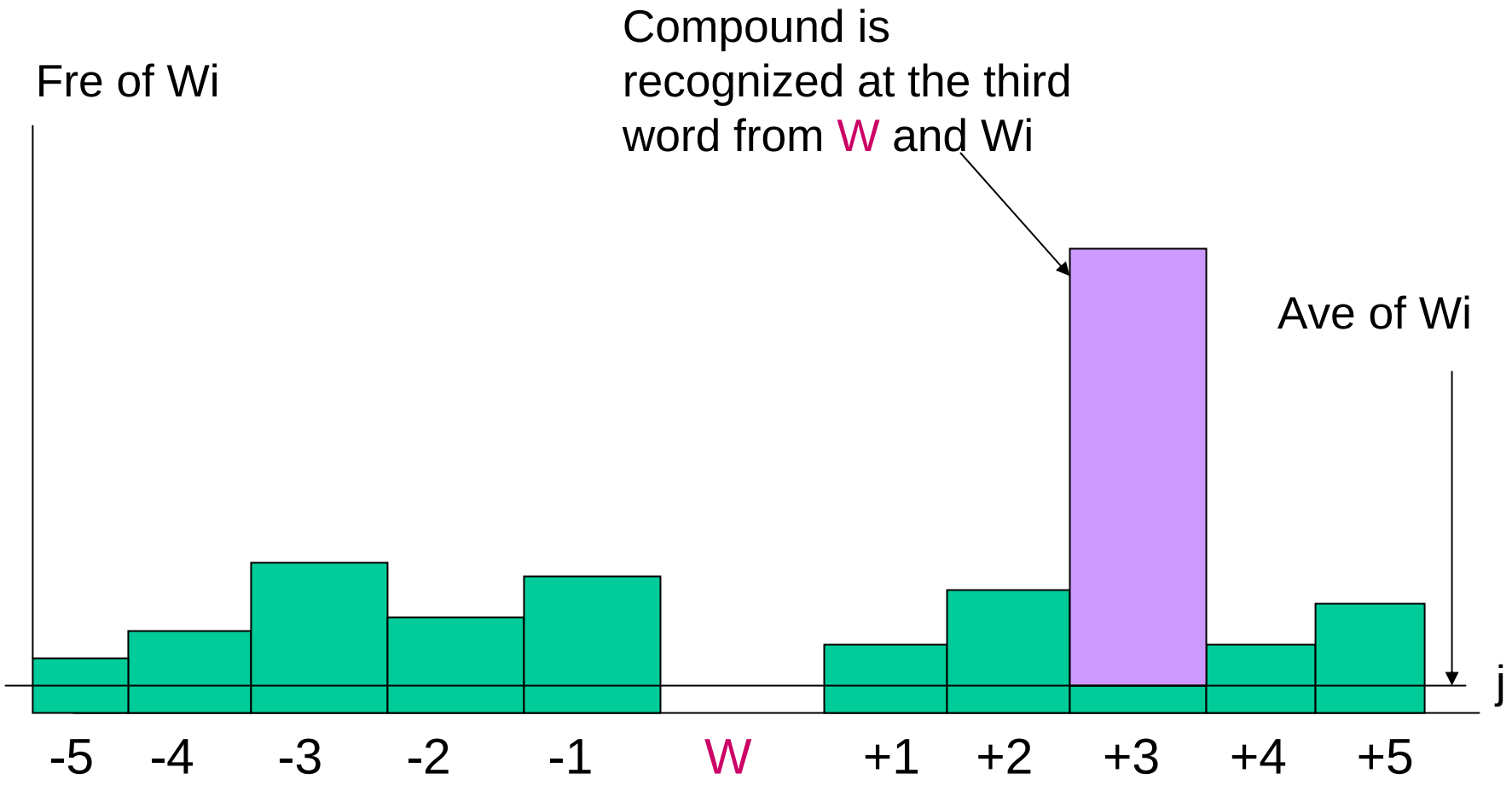
Noun, Compound Noun & Collocation

- Categories of possible terms = Noun, compound noun, and collocation
- Noun: A type of nouns which are unbreakable into smaller units. App. 10% of all technical terms.
- Compound noun: A combined form of nouns. App. 85% of all technical terms.
- Collocation:
  - The meaning of collocation is not obtained from that of constituents.
  - United states, “*to-ra-no-ko* 'a kid of a tiger' " ->idiom
  - Collocations do not need to come side by side (a broad interpretation).

Statistics-based Extraction of Compounds

- Smadja, Xtract System Computational Linguistics, 1993
- Categories of Compound:
  1. Subject-predicate relation: make-decision, hostile-takeover
  2. Fixed noun phrase: stock market, foreign exchange
  3. Templates of phrases: The Dow Jones average fell NUMBER* points to NUBER*

Identification of Compound in Xtract



Xtract: stage1: Extracting Significant bigrams

1. Producing concordance (term index): Extract all sentences including *W* from tagged corpus + word :*W*.
2. Compile and sort: *W* co-occur with W_i. Find *freq* $(W_i)^j$ *(-6<j<6)* at which W_i appears at *j* number of words away from *W*.
3. Extract word pairs that co-occur at statistical significance. Screen out based on the following three criteria:

Stage 1

- *p (Wi) j (j=-5,..+5)*: Appearance of *Wi* at locations between –5 and +5 from *W*.
- Condition 1
  - $$k_0 < \frac{freq\ (Wi\ ) - E_{Wi}\ [\ freq\ (w_i\ )]}{\sigma_{W\ i}[Wi\ ]} = ki$$
- *k0* (predefined threshold):Frequency of w*i* high enough
- Condition 2

θ is a threshold. $$Ui = (\sum_{j=-5}^{+5} (p(Wi)j - \overline{p(Wi)})^2)/10 > \theta$$

  - :Frequency distribution in the vicinity has a peak. The expression (compound) is fixed.

Stage 1

- Condition 3

$$p((W_i)\,j) \geq \overline{p(W_i)} + k1 \times \sqrt{Ui}$$

- Condition to extract compound in which W_i co-occur at location *j.*

Stage 2: From 2-grams to N-grams

- In the vicinity of bigram based on Stage 1, words which appear at a high frequency at m words apart (m<N) are extracted to detect compound consisting more than 3 words. In some cases, they are replaced for part of speech (POS) to extend to N-gram.
- E.x. composite index -> The NYSE’s composite index of all its listed common stockes fell *NUMBER to *NUMBER

Stage3: Adding syntax to Compounds

- By Stage2, compounds are identified. Next, POS tags are attached to each element of the compounds.
- Original sentences are parsed to assign grammatical roles (S, V, O, etc.).
- If the same grammatical roles are assigned at a statistical substantial rate, the grammatical roles are employed as compound.
- When grammatical roles are not assigned after parsing, they are not recognized as compound.
  - E.x.:O savings fell: SV,
  - X savings failing: undefined

Structure of Possible Terms & Statistic Approach

- Method to directly measure unithood and termhood.
- C-value method (unithood)
- Statistical data for conjoined nouns (termhood)

C-value Method

- In Xtract at stage2, compound of bigram is extended to N-gram. The opposite direction can be considered.
- Frantzi & Ananiadou 96 said: ” (they try to extract) substring of other longer one (=compounds) .” “ they (including Xtract) try not to extract unwanted substrings of compounds.”
- Thus, they try to extract parts of compound as long as they are qualified as compound.
- E.x.: They try to extract Wall Street from Wall Street Journal as long as they are a useful compound.

C-value

- A long compound: a part of *C1*: when *C2* and *C1* appear at the same frequency, *C2* is not considered a compound.
- If *a* is not a part of known compound, C-value *(a) = (length (a) -1) n (a). n (a)* is the frequency of *a*.
- If *a* is a part of known compound, *C-value (a) = (length (a) -1) (n (a) -t (a) /c (a))*.
- t (a) is the frequency of *a* to appear in a long compound. *c (a)* is the total number of different long compounds.

C-value Calculation: Sample

- E.x.: The number of appearance in the corpus is now clarified.
- word tri gram (3 times), tri gram statistics (2 times)
  class tri gram (1 time), tri gram catch (1 time)
  characteristic tri gram (1 times), tri gram (4 times)
- Obtain C-value of trigram.
- length (tri gram) =2 n (tri gram) =12 times
- t (try gram) =8 times c (tri gram) =5 types
- C-value (tri gram)
  = (length (tri gram) -1) (n (try gram) -t (tr..) /c (tr..))
  = (2-1) (12-8/5) =10.4
- C-value (word trigram) = (3-1) 3=6

Sample Extraction of C-value

- Examples:
  - WALL STREET JOURNAL,
  - Staff Reporter of The Wall Street Journal,
  - Wall Street,
  - of its, it is, because of
- C-value is proportional to length (a). Thus, longer compounds are preferred.

“Gensen Web”
Automatic Domain Terminology Extraction System

- This system extracts valued domain specific terms from short texts.
- Multilingual ready

->Candidate translations to be generated

- Output to be scored and sorted

Method for Single Noun Compounds

- The C-value method has been used for extracting a part of long compounds. On the other hand, this method is used for ranking single nouns according to the number of conjoined nouns in compounds, which is a different approach from Xtract that utilizes bigram.
- This method is unique with regard to the ranking of single nouns and the use of data combinations. (One demerit is that terms in the same domains only can be extracted.)
- This method utilizes information on compounds in a specific lexicon rather than probabilities of lexical groups. It can analyze short texts about one web page.

Terminology Extraction Sys...

Japanese | Western Europe language, Japanese, Chines | Chinese

Automatic Domain Terminology Extraction System
Welcome to "Gensen Web"

You can extract valued domain specific terms from Web pages or text you input. The extracted terms are sorted and displayed in descending order of their importance in other words, the extracted terms are well selected ones: thus the name of this system is "Gensen" which means "well selected."

"Gensen Web" system is a Web version of the original term extraction system "TermExtract" written in Perl. The function is a little bit limited compared to the original stand-alone version.

Usage

1. Input URL of Web page written in HTML or PDF from which you want to extract terms. Or input, probably copy and paste document. Or select your local PC file (text file or PDF only).
2. Choose **POS tagger version**: highquality but slow or **high speed version**: but a little bit less quality
3. Click the "start" button.
4. Wait a while, then the extracted and sorted terms are displayed.

Input URL

Input (or copy and paste) document

Select local file(text file or PDF owritten by utf8 only)

ファイルを選択 ファイルが選択されていません

high speed version ○**English** ○**French** ○**German** ○**Italian** ○**Spanish** ○**Finnish** ○**Swedish**

POS tagger version ○**Japanese** ○**Chines-simple** ○**English**: highquality but slow

⊙**Auto** (Powerd by Perl module Lingua::LanguageGuesser)

start clear

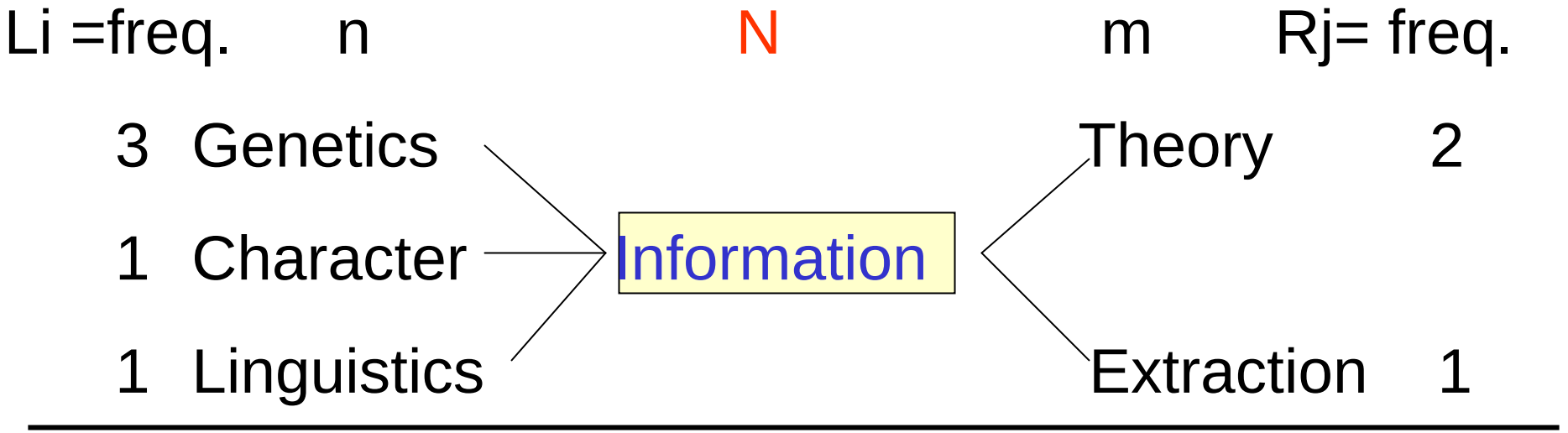
Key word only ☐Perplexity mode

The "Perplexity mode" score importance of terms in context based on "Diversity of information".

Gensen-Web Algorithm

- Importance are attached more to basic nouns (single nouns) than compound nouns because basic nouns are the elements of many compound nouns.
  - The idea is that many compound nouns (= many domain specific concepts) are consisted of more important single nouns (= basic concepts).
- Terms are weighted in a vocabulary space separated from texts. The analysis is not dependent on the size of texts.

Scoring of Single Nouns



LN (Information) =5 n=3 m=2 RN (Information) =3

Genetics Information, Genetics Information Extraction, Character Information, Information Theory,

Genetics Information, Information Theory

Scoring of Compound Nouns -Geometric Mean (GM)

$$LR(CN) = \left\{ \prod_{i=1}^{L} \left[\left(LN(N_i) + 1 \right) \cdot \left(RN(N_i) + 1 \right) \right] \right\}^{\frac{1}{2L}}$$

$$CN = N_1 N_2 \ldots\ldots N_L$$

GM(CN) is independent to the length of compound nouns (CM), or the number of basic words (elements).

Scoring Frequency: FGM (CN)

if *CN* occurs independently

then $FLR(CN) = f(CN) \times LR(CN)$

where *f (CN)* represents the number of independent occurrences of noun *CN*

(= the number of the CN that is not a part of longer CN)

Ex. *LR* (Information) = ((5+1) x (3+1)) $^{1/2}$=4.9

if *f* (Information) = 5

FLR (Information) =24.5

Gensen-Web can extract words of importance from short texts.

- It can extract keywords even from a single web page.

Application to Chinese

- Which should Gensen-Web algorithms be applied, word or character?

「言选Ｗｅｂ」（中文·停止语方式版）

本网页能从文章中抽出专业用语（关键字）。

本系统提供在线专业用语自动抽出Perl模块"TermExtract"的功能。其与单独运作(stand-alone)版相比，虽然利用功能有所限制，但有检索方便的

1. 直接输入（或剪贴）文章，或以URL指定网上的html文档。
2. 选择中文（GB）或中文（UTF-8）。
3. 点击专业用语（关键字）抽出按钮。
4. 专业用语（关键字）将以重要度的顺序抽出。

请输入URL

请输入（或剪贴）文章

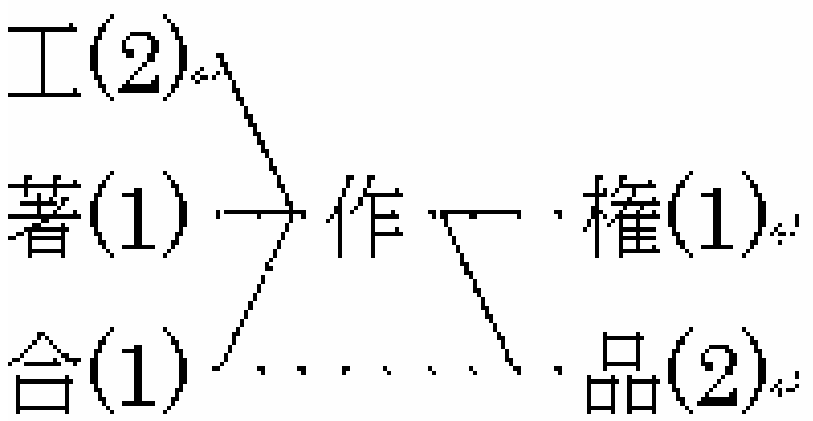
◉ **中文**（GB) ○ **中文**（UTF-8)

专业用语（关键字）抽出 清除 ICTCLAS版

Key word only

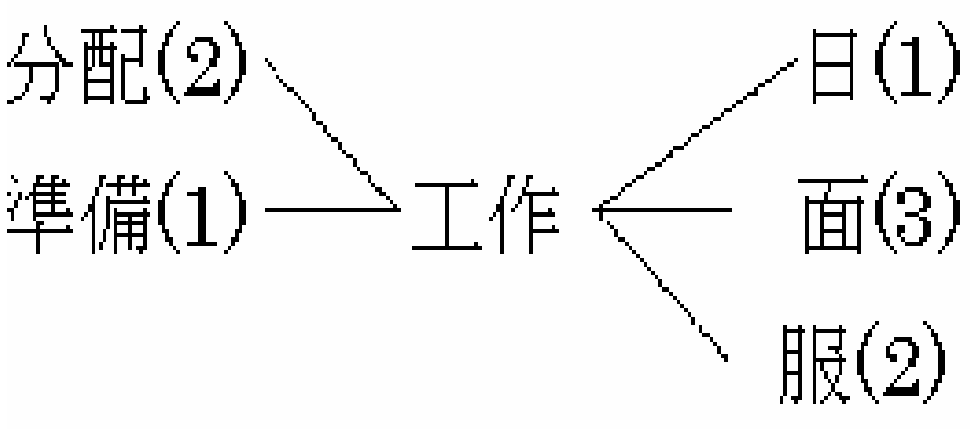
Sample: Character-based FLR

工作 ‘operation; woodwork’, 著作権 ‘copyright’, 合作 ‘joint work’,
工作 ‘operation; woodwork’, 作品 ‘art; work’, 作品 ‘art; work’



LN (作) =4　　RN (作) =3

Sample: Word-based LR



- LN (工作) =3, RN (工作) =6
- Word extraction by morphological analysis is required.

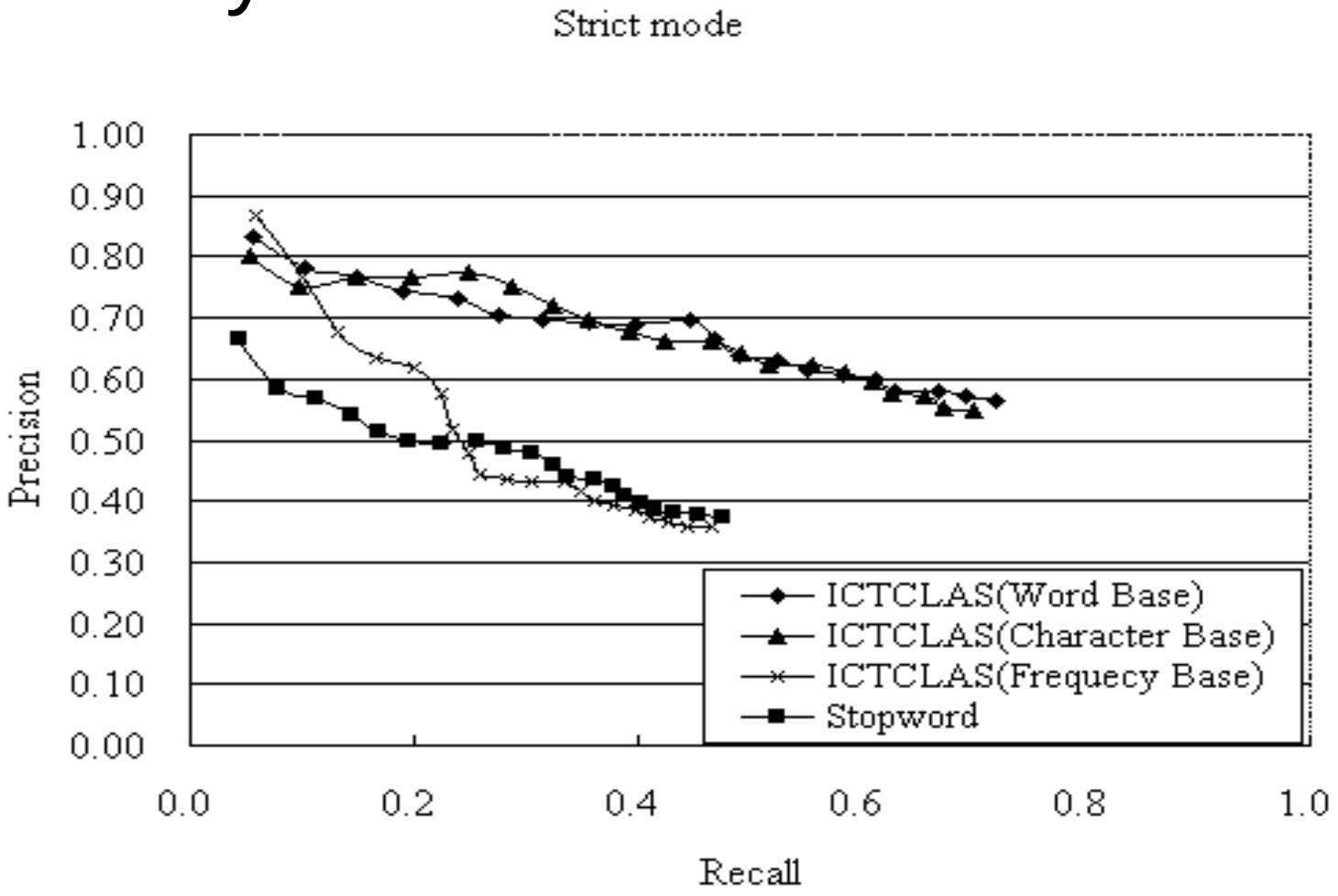
Morphological Analysis & Extraction Rules for Candidate Term in ICTCLAS

- ✓ MWU <-- [ag a]* [ng n nr ns nt nz nx vn an i j]$^{+}$
- ✓ MWU <-- MWU$^{?}$b [ng n nr ns nt nz nx vn an i j]$^{+}$
- ✓ MWU ← [ag a]$^{+}$ [u k] MWU
- ✓ MWU ← MWU (u|k|*he-2*|*yu-3*) MWU

Precision Evaluation -Methodology

- (1) Extract candidate terms according to stop-words + Character-based FLR.
- (2) Extract candidate terms according to POS tag + Character-based FLR.
- (3) Extract candidate terms according to POS tag + Word-based FLR.
- (4) Simple frequency calculation

30 articles from the People's Daily were used for evaluation.



Bootstrap Method

- Demetriou ’00 (LREC2000)
  1. Initialisation
     1. Extract initial set of contextual patterns (left and right n-grams n=6) using seeds
     2. Identify significant patterns (χ^2 test 0.5%)
     3. Terminate if no significant patterns exist
  2. Name extraction
     1. Apply significant patterns and extract new names
     2. Terminate if no new names are found
  3. Pattern generation
     1. Match the new names and extract new patterns
     2. Identify significant patterns
     3. Terminate if no significant patterns exist
  4. Go to step 2

Bootstrap Method (Cont.)

- Identification of protein names from medical documents
- Sample patterns: of human, 3-dimensinal structure of, the bacterial,….
- Result:
  - Original data:Recall=31%, precision=97%
  - Original+new terms: R=38%,P=96%
  - Original+newt terms+grammar rules: R=90%,P=96%
- Observation:
  - Technical documents have lower perplexity than BNC.
  - Technical documents have less grammatical difference.
- Similar to NE task.

Method using Bilingual Corpus

- Beatrice Daille‘94
- Identify single noun pairs from an English (French) corpus according to statistical analysis.
- Generate compounds from the extracted pairs.
- Generation of sample compounds: (interference, level) -> interference level(s) , level of interference(s)
- An attempt to associates several statistical measures (Log-likelihood, MI, etc) have failed.

Dalli's Method (Cont.)

In aligned corpora in English and French, if candidate terms generated from a monolingual corpus are correct, significant terms can be found in aligned sentences from the counterpart corpus. Based on this hypothesis, correct terms can be identified.

- Top 500 -> 80%, Top 1000 -> 70% Precision
- This method of preparing candidate terms before working on bilingual corpora is a new idea. English-Japanese aligned translations are extracted using NTCIR1&2 corpus by a similar method (Nakagawa).

Extension of Terms

- Extracted terms are not enough in some cases.
- For example, keyword search can work more effectively if keywords are provided in an extended form.
  - 1 word: sorting -> sorting algorithms, array sorting
  - 2 words: Japanese, analysis -> Japanese syntax analysis, Japanese lexical analysis, Japanese morphological analysis

Extension of Terms

- 1 word: ‘sorting’ -> sorting algorithms, array sorting
- 2 words: ‘Japanese, analysis’ ->Japanese syntax analysis, Japanese lexical analysis, Japanese morphological analysis

- What kind of terms should be used to extend keywords?
  - To use thesaurus and extend relevant terms.
  - To extend based on grammars:
    - “Ha-I-re-tsu-so-u-to ‘array sorting’ “ -> This compound phrase is an abbreviation of ‘to sort arrays’.

Methodology for Structure-oriented Formation

- Derivational morphology
  - Linking elements of known terms (single nouns) to find variants and build a list of terms.
  - Generating new terms which are grammatically correct according to known grammatical relations.
    - Example (French):

      N1 de N3 + N2 de N3 -> N1 et N2 de N3

      Assemblage et deassemblage de paquet

FASTER

- Jacquemin&Rayoute ’94 (SIGIR ’94)
- To describe rules to detect compounds according to context-free grammar.
- To expand and refine rules using positive meta-rules.
  - Coordination: (X1->X2 X3 X4) = X1 -> X2 C5 X6 X3 X4
    - Inflammatory ***and erosive*** joint disease
  - Insertion: (X1->X2 X3 X4) = X1 -> X2 X5 X3 X4
    - Impaired intravenous glucose tolerance
  - Permutation: (X1->X2 X3 X4) = X1 -> X4 X5 X6 X7 X2 X3
    - Disease of the central nervous system [Nervous system disease]
- To exclude (inappropriate) candidates created by negative meta-rules.
  - Coordination :

    (X1 -> X2 X3) = X1 -> X2 C4 X5 X3: <X2 number> = plural

    (5% of terms were successfully excluded in system evaluation.)
  - × cells or fatal cultures <-- Cell cultures
- 31,428 terms are identified in a medical corpus (9MB). FASTER is used to extract correct variants for additional 8,747 terms. Breakdown of terms is Permutation 48%, insertion 43%, and coordination 9%.

Generation of Compound's Variants based on Morphological Structures

- Jacquemin's SIGIR'94,'97,'99
- From multiple single nouns provided…
  - morphological and semantic variants of single nouns, and
  - syntactic structures including both types of variants
- are utilized to generate variants including such single nouns. (An European approach for the western European languages that are inflection rich.)
  - This system was published as "FASTER".

Derivational Morphology

- Jacquemin ’97 (SIGIR)
  1. Detach a rear part of terms to find a common part. E.x. immuniz- (ation,ed)
  2. Generate terms from each two-word terms based on the common part identified in process 1. In this process, suffixes are min three characters, and the number of terms in compounds is two. E.x. continue (optimum value in experiments)
     - E.x. active immunization , actively immunized
  3. Extract common suffixes (called “signature”) from multiple compounds (classes).
     - E.x. (continuous measure-ment) (continuous-ly measure-d)

       -> { (ε ,ment) , (ly,d) }

       This rule will be used for tern generation.
     - E.x. { (ε , ing) , (ly,ed) }

       According to this rule, ‘diffuse scattering’ -> ‘diffusely scattered’

Derivational Morphology (cont.)

3. Filtering out:
   - F= (the mean value of the numbers of stem characters within single nouns in the classes) / (a max value of the signature)
   - Keep classes as long as they are F>1. Filter out classes associated to long suffix changes (= signature).
4. Clustering of classes:
   - Cluster classes when the distance between each class is small. The distance is defined as such that the last characters of the signatures are weighted more when they have a match.

Generation of Collocation's Variants Jacquemin '99 (ACL '99)

- Morphological family's example:
  - F_M (measurement) ={commensurable,countermeasure, tape-measure,measure, ….}
- Semantic family's example:
  - WordNet: F_{SC} (speed) ={speed,speeding, hurrying,velocity,amphetamine,..}
  - Word97: F_{SL} (speed) ={speed, rapidity, celerity,..}
- Generation rule's example:
  - N1Prep N2 -> F_M (N1) $Adv^?$ $A^?$ Prep $Art^?$ $A^?$ F_{SC} (N2)
  - E.x.: composition du fruit -> compse'chimiques de la graine (chemical compound of the seed)
  - Other rules are, for example, coordination, modification, permutation, VP, and NP.

Generation of Collocation's Variants

- E.x.
  - Pressure decline -> pressure rise and fall
  - Angular measurement -> angles measure
- App. 80% precision for morphology and syntax.
- Less than 50% precision when semantic rules are employed. (Generation variants in this approach is less than a few percentage points.)
- This approach reported three-fold increase in the number of variants generated to the number of collocations appeared in texts.

Approaches using Reciprocal Information -1

- Su-Wu-Chan (ACL94)
- Define word 2-grams and word 3-grams as candidate compounds.
- The criteria for selection is reciprocal information (MI), relative frequncy (RFC), and POS pattern L_i .
- MI is a bi-gram. $$I(x;y) = \log_2 \frac{P(x,y)}{P(x) \times P(y)}$$
- 3-gram $$I(x;y;z) = \log_2 \frac{P(x,y,z)}{P(x)P(y)p(z) + P(x)P(y,z) + P(x,y)P(z)}$$

Approaches using Reciprocal Information -2

- POS patter: L_i=[n,n]
- Context taken into account: L'_{ij}=[adj (n n) n]
- The three types of information are used for an overall analysis. $M_{c\ (nc)}$ represents an event in which an n-gram is generated from (non) collocations:

$$P(\mathbf{x}|M_c)\times P(M_c) \approx \prod_{i=1}^{n}[P(MI_{Li}, RFC_{Li}|M_c)\prod_{j=1}^{ni}P(L'_{ij}|M_c)]\times P(M_c)$$

- MI and RFC in question can be estimated by a normal probability distribution based on the mean and distribution of test-sets.

Approaches using Reciprocal Information -3

- POS pattern in the case of a bigram:

$$P(L_{ij}^{'} \mid M_c) = P(C_0, C_1, C_2, C_3 \mid M_c)$$
$$\approx P(C_3 \mid C_{2,} M_c) \times P(C_2 \mid C_{1,} M_c) \times P(C_1 \mid C_{0,} M_c) \times P(C_0 \mid M_c)$$

- Similar in the case of a 3-gram.
- Acquire probabilities from the test data sets. Unseen data is evaluated according to its likelihood. *C* for compounds; *nc* for non-compounds.

$$\lambda = \frac{P(\mathbf{x} \mid M_C) \times P(M_C)}{P(\mathbf{x} \mid M_{nc}) \times P(M_{nc})}$$

Approaches using Reciprocal Information -4

- Su, et. al. performed an experiment with a corpus which was manually modified and morphologically analyzed.
- Bigram: recall=0.977, prec=0.445 (training set) recall=0.962, prec=0.482 (test set)
- 3-gram: recall=0.976, prec=0.402 (training set), recall=0.966, prec=0.396 (test set)
- Extraction samples: dialog box, mail label, main document, datafile, file menu, World User's guide, Microsoft Word User's, Template option button, new document base, File name box

Extension of C-value

- NC-value (Frantzi, et. Al., 2000; Maynard, et. Al., 2001)
  - A linear combination of the C-value of a term in question *a* and *CF(a)*: C-value*0.8 + *CF(a) *0.2*

$$CF(a) = \sum_{w \in Ca} F(a,w) F_w / n_w$$

  - *Ca*: the class for terms which appear in the context of *a*
  - *F(a,w)*: the number of occurrence of *w* in the context of *a*
  - *Fw*: the number of terms which appear in the same context of *w*
  - n_w: the total number of occurrence in the corpus of w