Compiling Bilingual Lexicon for Technical Terms using the Web

Takehito Utsuro
Department of Intelligent Interaction Technologies,
Graduate School of Systems and Information Engineering,
University of Tsukuba, Japan

(joint work with
Masatsugu Tonoike, Mitsuhiro Kida, Yasuhiro Sasaki,
Xavier Robitaille, Yasuo Banba, and Satoshi Sato)

Topics of this Presentation

Compiling Bilingual Lexicon for Technical Terms

- Collecting Technical Terms of a Domain
- Translation Knowledge Acquisition of Technical Terms

Technical terms

- Not included in existing bilingual lexicons for general use
- Found in
  - technical documents
  - Web documents

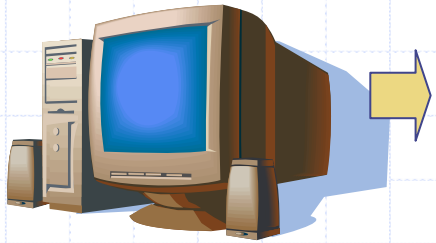
Background

- Needs for (bilingual) lexicons of technical terms
  - Various domains/topics
  - Only some portions are covered in general purpose bilingual lexicons.
  - It is expensive to compile a bilingual lexicon for each domain/topic manually.

The Association for Behavior Analysis
(国際行動分析学会)

Umm… I'm not familiar with this research area….

Input a keyword
"*behavior analysis*"
…



Are you ready?

bilingual terminology in "*behavior analysis*"

applied behavior analysis	応用行動分析
three-term contingency	三項随伴性
stimulus dimensions	刺激次元
stimulus generalization	刺激般化
operant level	自発水準
forward chaining	順行連鎖化
pairing procedure	対提示手続き
deprivation	遮断化

Compiling Bilingual Lexicons for Technical Terms

process

seeds (a word, a document, etc.)

behavior analysis

data

1. collecting terms for headwords

headwords

2. translation estimation

bilingual lexicon

bilingual lexicon of “behavior analysis”

applied behavior analysis	応用行動分析
three-term contingency	三項随伴性
stimulus dimensions	刺激次元
stimulus generalization	刺激般化
operant level	自発水準
forward chaining	順行連鎖化
pairing procedure	対提示手続き
deprivation	遮断化

Compiling Bilingual Lexicons for Technical Terms

Elemental Technologies

- Collecting Technical Terms of a Technical Domain
  - Term (candidates) recognition: existing technologies
    - statistical / grammatical approaches
  - Estimating domain specificity of term candidates
    ⇒ discriminating technical terms and general words
- Translation Estimation of (Compound) Technical Terms
  - Compositional translation estimation
  - Validating translation candidates against technical documents of the domain collected from the Web
  - Japanese-English: PACLING-2005, IJCNLP-2005, EACL-2006-WS
  - Japanese-French: EACL-2006

Compiling Bilingual Lexicons for Technical Terms

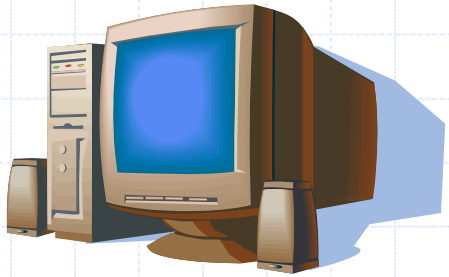
Elemental Technologies

- Collecting Technical Terms of a Technical Domain
  - Term (candidates) recognition: existing technologies
    - statistical / grammatical approaches
  - Estimating domain specificity of term candidates [IEICE 2006]
    ⇒ discriminating technical terms and general words
- Translation Estimation of (Compound) Technical Terms
  - Compositional translation estimation
  - Validating translation candidates against technical documents of the domain collected from the Web
  - Japanese-English: PACLING-2005, IJCNLP-2005, EACL-2006-WS
  - Japanese-French: EACL-2006

Task of Domain Specificity Estimation of a Term *using the Web*

a term「インピーダンス特性」
(“*impedance characteristic*”)

a domain
“*electric engineering*”



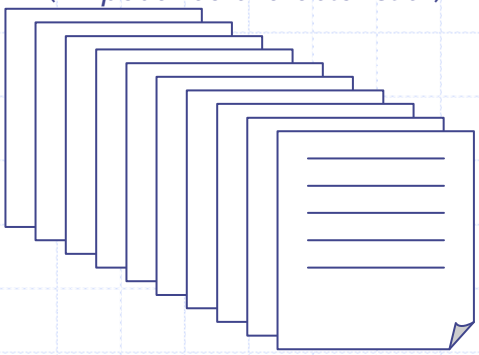
Web

Q: Is the term「インピーダンス特性」(“*impedance characteristic*”)
a technical term of the domain “*electric engineering*” ?

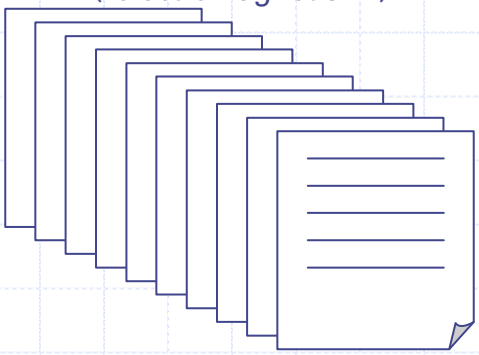
Answer: Yes or No

Domain Specificity Estimation of a Term based on the Domains of Sample Documents collected from the Web

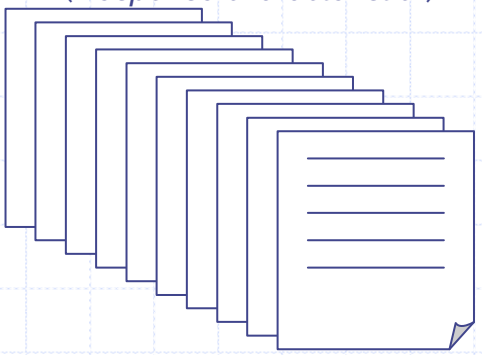
documents set with
「インピーダンス特性」
(“*impedance characteristic*”)



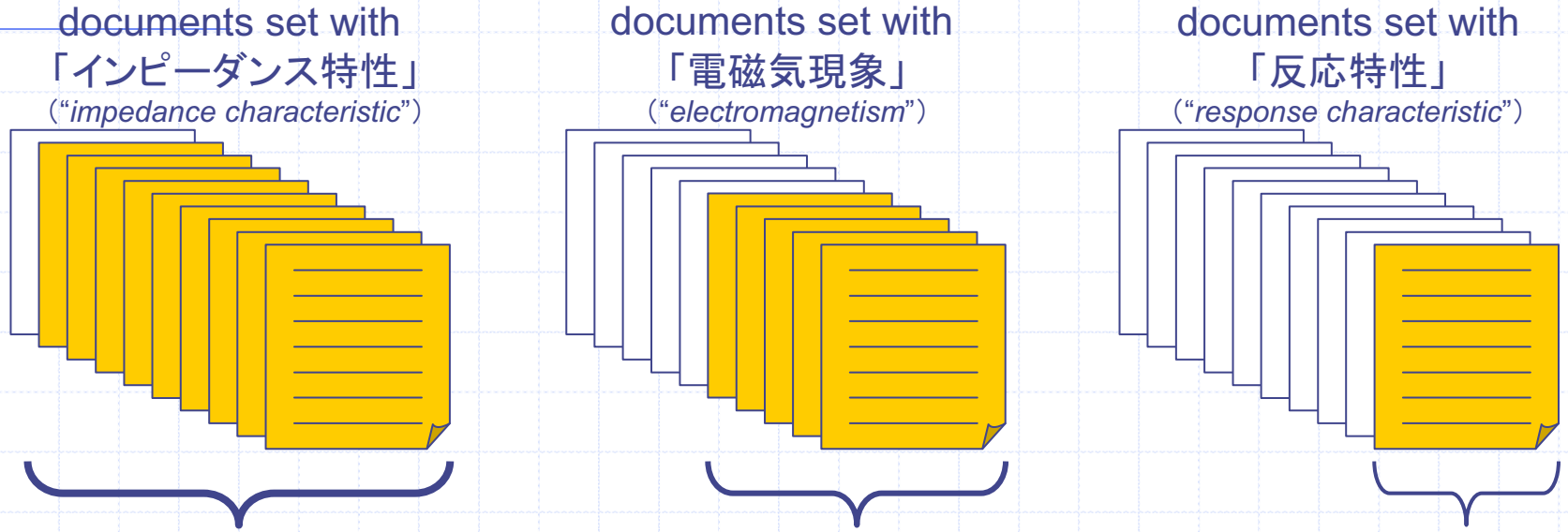
documents set with
「電磁気現象」
(“*electromagnetism*”)



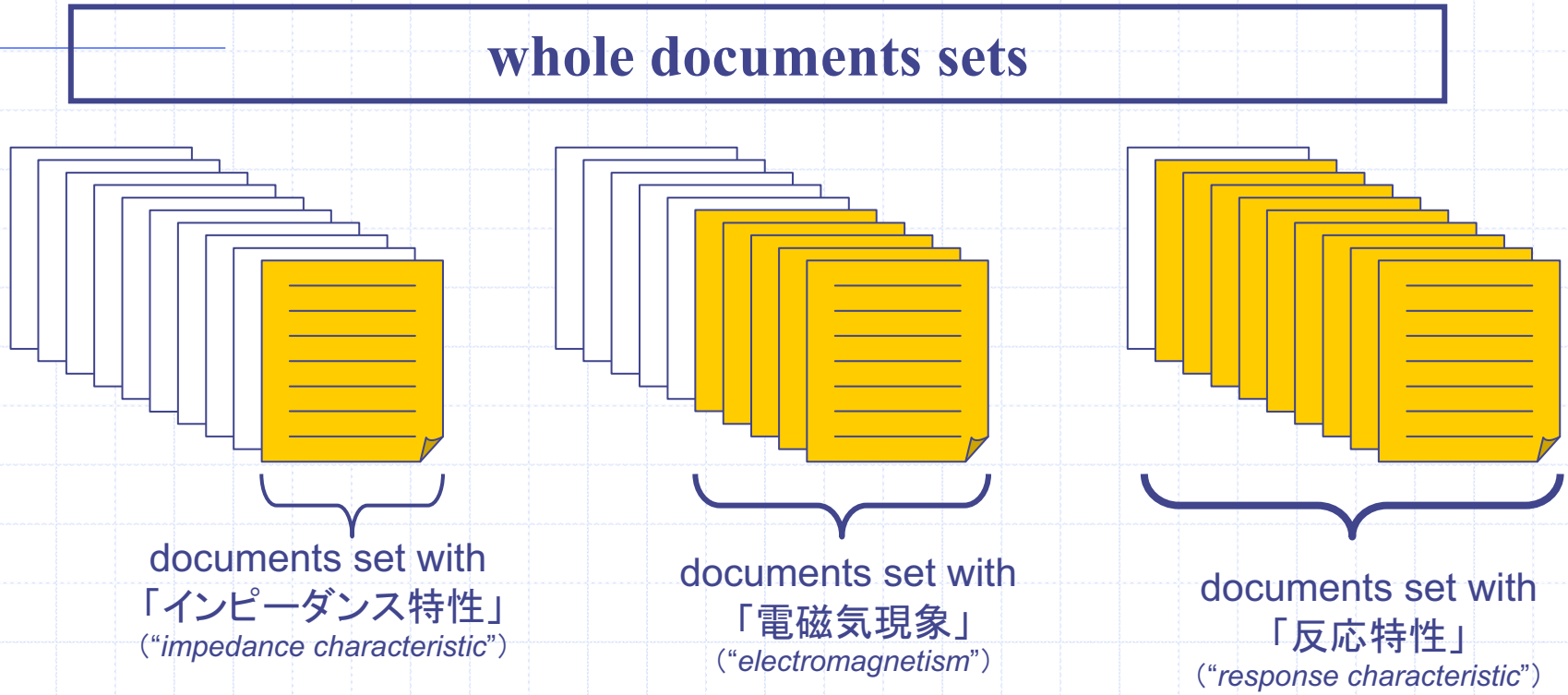
documents set with
「反応特性」
(“*response characteristic*”)



Domain Specificity Estimation of a Term based on the Domains of Sample Documents collected from the Web



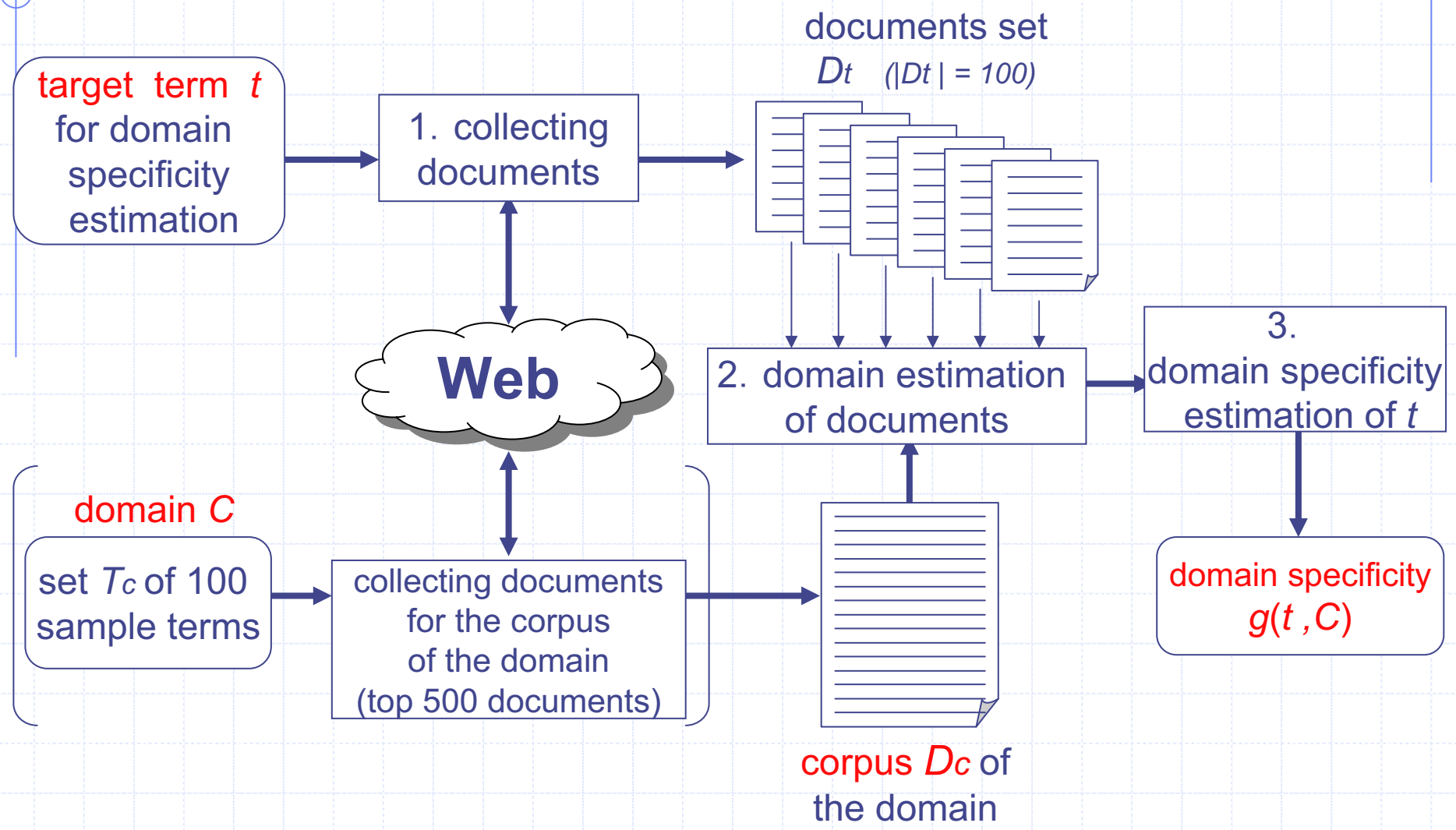
Cf. Inverse Document Frequency (IDF)



- *IDF* is higher if the term appears in fewer documents in the whole documents set.
- ⇔

  Domain specificity of a term is higher if its documents sampled from the whole documents set (the whole Web) are closer to the target domain.

Domain Specificity Estimation of Terms based on Web Documents



Experimental Evaluation

- For Japanese technical terms
- Five domains
  - "*electric engineering*", "*optics*", "*aerospace engineering*", "*nucleonics*", and "*astronomy*".
- Mostly 90% precision/recall
- Discovering novel technical terms
  - not included in existing lexicons of technical terms
  - out of 1,000 candidates (per a domain) selected from Web documents
  - 150-200 novel terms
  - with 75% precision and 80% recall

Compiling Bilingual Lexicons for Technical Terms

Elemental Technologies

- Collecting Technical Terms of a Technical Domain
  - Term (candidates) recognition: existing technologies
    - statistical / grammatical approaches
  - Estimating domain specificity of term candidates
    ⇒ discriminating technical terms and general words

- Translation Estimation of (Compound) Technical Terms
  - Compositional translation estimation
  - Validating translation candidates against technical documents of the domain collected from the Web
  - Japanese-English: PACLING-2005, IJCNLP-2005, EACL-2006-WS
  - Japanese-French: EACL-2006

An Issue:
With/without a target language corpus/Web?

Performance Comparison

- With the whole Web (through search engine)
- With a domain/topic-specific corpus collected from the Web
- Without the corpus/Web

Collecting a monolingual domain/topic specific corpus from the Web

technical terms of *medicine*

synapse
transplanted liver
seventh cranial nerve
sensory cell
vastus intermedius
oculomotor nerve
osseous system
pleural mesothelium
skin capillary
palatal mucosa
maxillary gingiva

found in the existing bilingual lexicon

synapse : シナプス
transplanted liver : 移植肝臓
seventh cranial nerve : 第7脳神経
sensory cell : 感覚細胞
vastus intermedius : 中間広筋
oculomotor nerve : 動眼神経

used for collecting a domain/topic specific corpus

domain/topic specific corpus

existing bilingual lexicon

translation estimation

osseous system
pleural mesothelium
skin capillary
palatal mucosa
maxillary gingiva

not found in the existing bilingual lexicon

estimate translations of these terms

having only one translation in *Eijiro* :

Collecting a domain/topic specific corpus

Translations of terms found
in the existing bilingual lexicon

synapse : シナプス
transplanted liver : 移植肝臓
seventh cranial nerve : 第7脳神経
sensory cell : 感覚細胞
vastus intermedius : 中間広筋
oculomotor nerve : 動眼神経

⋮

the Web

search engine

Collected web pages

これらの線維は主として前庭神経上核と内側核から起こる。

上核からの線維は主として滑車神経核と動眼神経核に同側性に投射する。

Japanese corpus of *medicine* domain

- queries
  - "*x*"
  - "*x* は"
  - "*x* とは"
  - "*x* の"
  - "*x* という"
- download top 100 pages per query

of translation pairs for evaluation

(estimating Japanese translation of English terms)

10 categories for evaluation	having only one translation in Eijiro	translation pairs for evaluation (not found in Eijiro)	collected Japanese corpus size
Electromagnetics	36	33	28MB
Electrical engineering	34	45	21MB
Optics	42	31	37MB
Programming language	37	29	34MB
Programming	29	29	33MB
computer	91	100	67MB
Anatomical Terms	91	100	73MB
Disease	91	100	83MB
Chemicals and Drugs	94	100	54MB
Physical Science and Statistics	88	100	56MB
Total	633	667	482MB

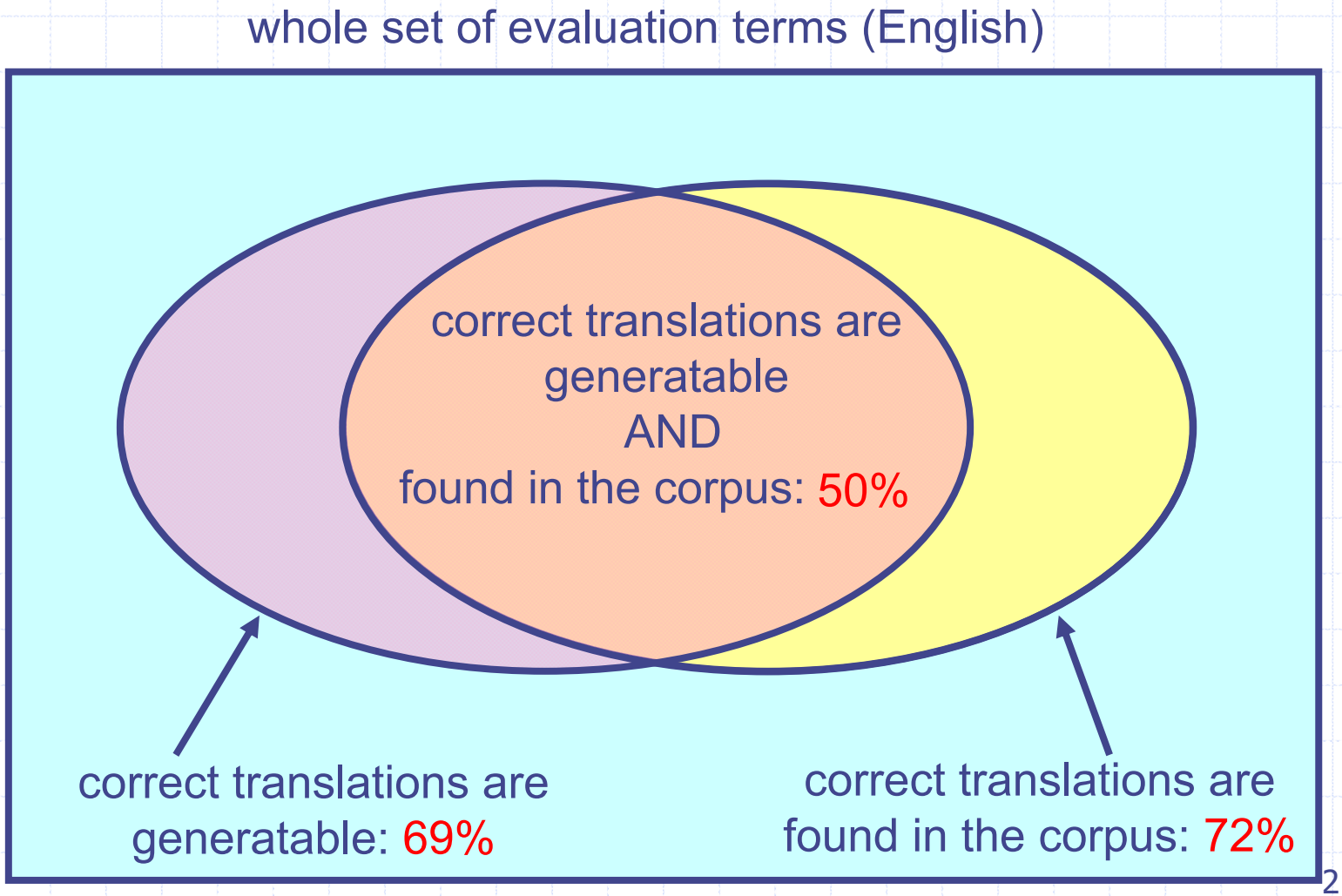
Statistics on Compositionality

- rate of compositional translation:

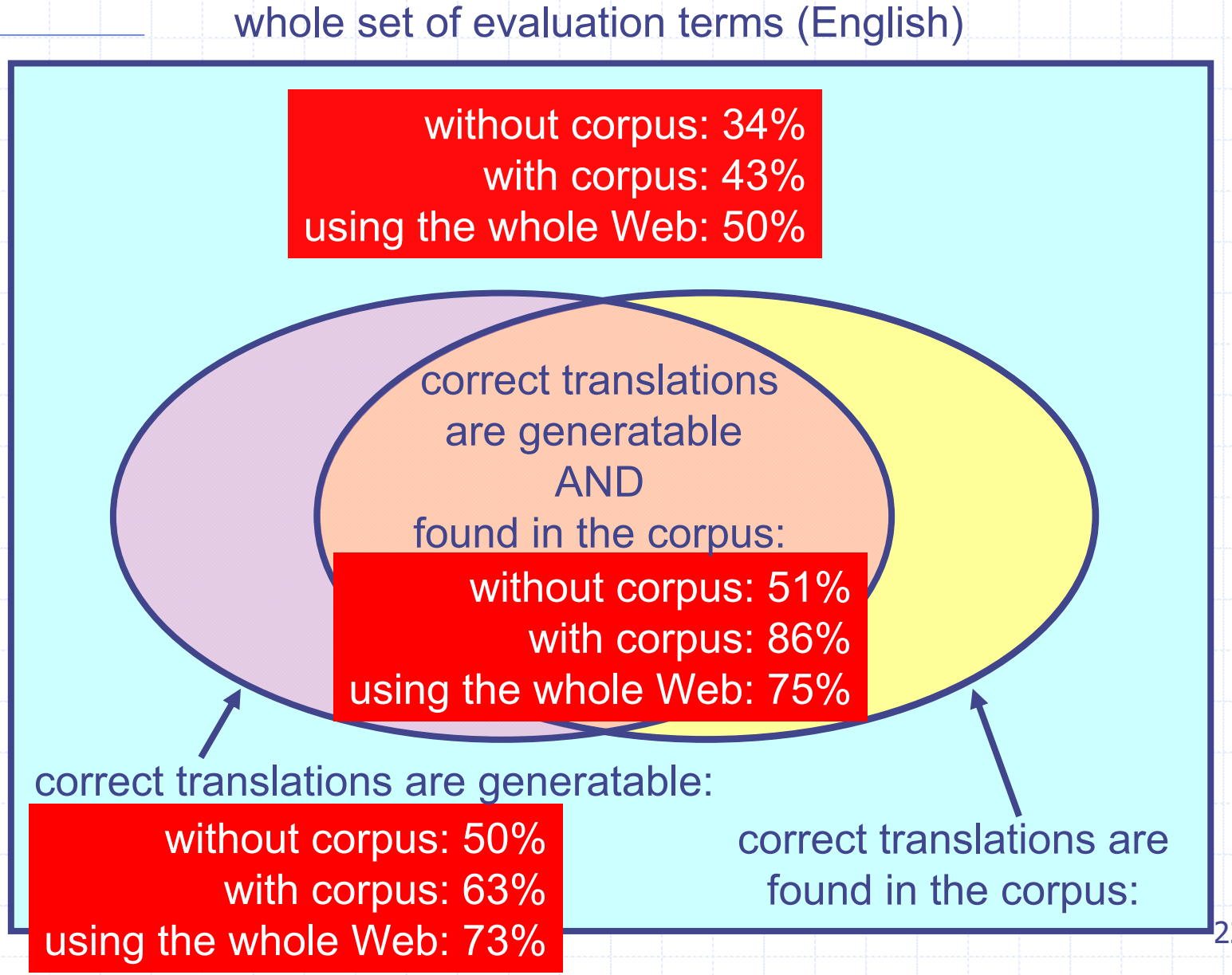
  the rate of translation pairs that are compositional

  - **88%** (examined manually)

Rate of evaluation terms whose correct translations are generatable / found in corpus



Precision of 1st ranked translation candidate



Comparison: the corpus and the w

search engine hits

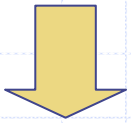
frequency in the corpus

categories	terms	1st ranked translation candidate	
		the whole Web	corpus
Optics	Newtonian focus	ニュートン的 / 526 / 0	ニュートン焦点 / 84 / 10
Programming Language	bit type	少し型 / 890 / 0	ビット型 / 1160 / 1
Programming	system macro	システム大規模 / 242 / 0	システムマク / 147 / 1
Computer	usage bit	使用部分 / 23300 / 0	使用ビット / 319 / 2
Anatomy	outer cap	外帽 18 / 0	外冠 / 209 / 1
Anatomy	greater wing	大袖 / 3260 / 0	大翼 / 1240 / 68
Disease	iron store	鉄屋 / 1680 / 0	鉄貯蔵 / 173 / 26
Disease	shock position	衝撃位置 / 54 / 0	ショック体位 / 123 / 1

↑ incorrect ↑ correct

Discussion

- Without corpus/Web → the lowest precision
- Against the whole evaluation set
  - Using the whole web achieves the highest precision.
- Against the terms whose correct translations are generatable and exist in the domain/topic-specific corpus
  - Using the corpus achieves the highest precision.



Further issue:

- How to collect
  - *large enough.* and
  - *domain/topic-specific*

  corpus from the Web.

Summary: Compiling Bilingual Lexicons for Technical Terms

Elemental Technologies

- Collecting Technical Terms of a Technical Domain
  - Term (candidates) recognition: existing technologies
    - statistical / grammatical approaches
  - Estimating domain specificity of term candidates
    ⇒ discriminating technical terms and general words

- Translation Estimation of (Compound) Technical Terms
  - Compositional translation estimation
  - Validating translation candidates against technical documents of the domain collected from the Web
  - Japanese-English: PACLING-2005, IJCNLP-2005, EACL-2006-WS
  - Japanese-French: EACL-2006

Future Plans

- Domain Estimation of Technical Terms
  - Large scale experimental evaluation with much more domains
  - Developing a technique for domain/topic estimation of terms for much narrower domains/topics
  - Soft domain estimation with multiple target domains

- Translation Estimation:
  Integration of elemental technologies
  - Compositional translation estimation
  - Translation candidates validation directly through the search engine
  - Translation estimation using partially bilingual Web pages
    - e.g., “応用行動分析(applied behavior analysis)”
  - Transliteration of proper names
    - e.g., “ニューヨーク・タイムズ/New York Times”