EBMT Tutorial

TMI-2002

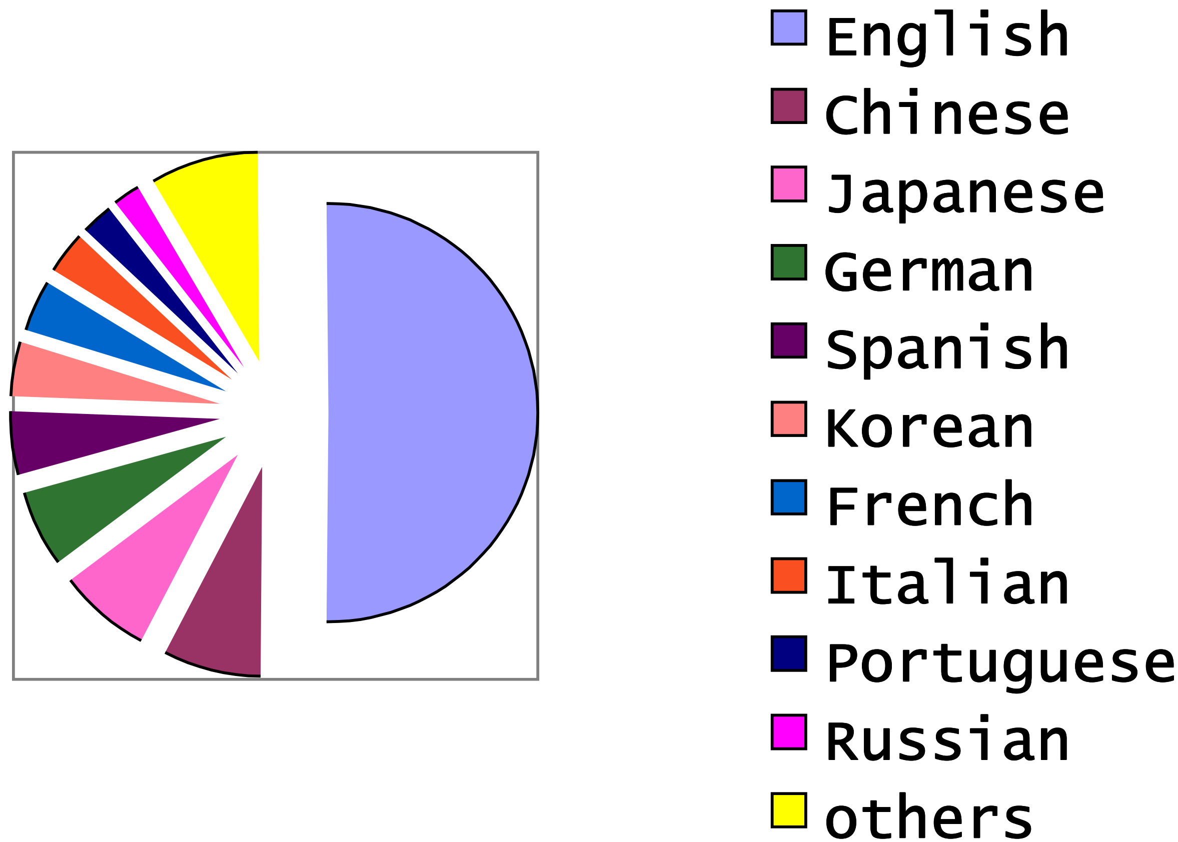
March 17, 2002

Eiichiro SUMITA and Kenji IMAMURA

{eiichiro.sumita, kenji.imamura} @ atr.co.jp

Increasing needs for wider languages and domains

Top ten languages on WEB
(source: Global Reach)



The total number of languages on our planet: around 6,000.

Current State of the Art (1)

- Machine translation is **growing**.
  - **Many systems** have been
    - Commercialized for **PC**s
      - (Visit http://homepage2.nifty.com/oto3/)
    - Available on the **WEB**
      - (Visit http://mason.gmu.edu/~aross2/mtgrid.htm)
  - Most machine translation systems provide a large vocabulary and **broad coverage**.
  - They translate literally and produce a **moderate quality** translation.

A series of translations by **a bi-directional system on the WEB**

(1) [Input] I'd like to reserve a table
(2) [EJ] 私はテーブルを確保することを望む
(3) [JE] I desire the fact that the table is guaranteed
(4) [EJ] 私はテーブルが保証されるという事実を望む
(5) [EJ] I desire the fact that the table is guaranteed

A loop!

A series of translations by **our bi-directional EBMT**

(1) [Input] I'd like to reserve a table

(2) [EJ] 席を予約したいです

(3) [JE] I'd like to reserve a seat

(4) [EJ] 席を予約したいです

A loop :-)

Current State of the Art (2)

- Machine translation is **spreading**.
  - High-quality translation is achieved by
    - Carefully **domain targeted** systems.
    - **Control language based** systems.
  - **Speech-to-speech** translator has emerged.
    - Eg., ATR, CMU, DFKI, NEC, Matsushita, Hitachi

Remaining **problems**

1. **Knowledge building**
   - Handcrafted → Expensive and snail-paced
2. **Translation Quality**
   - Structure-preserving → Not always high quality
3. **Quality Evaluation**
   - No evaluation → Self evaluation

EBMT is **attacking** these problems.

What is EBMT?

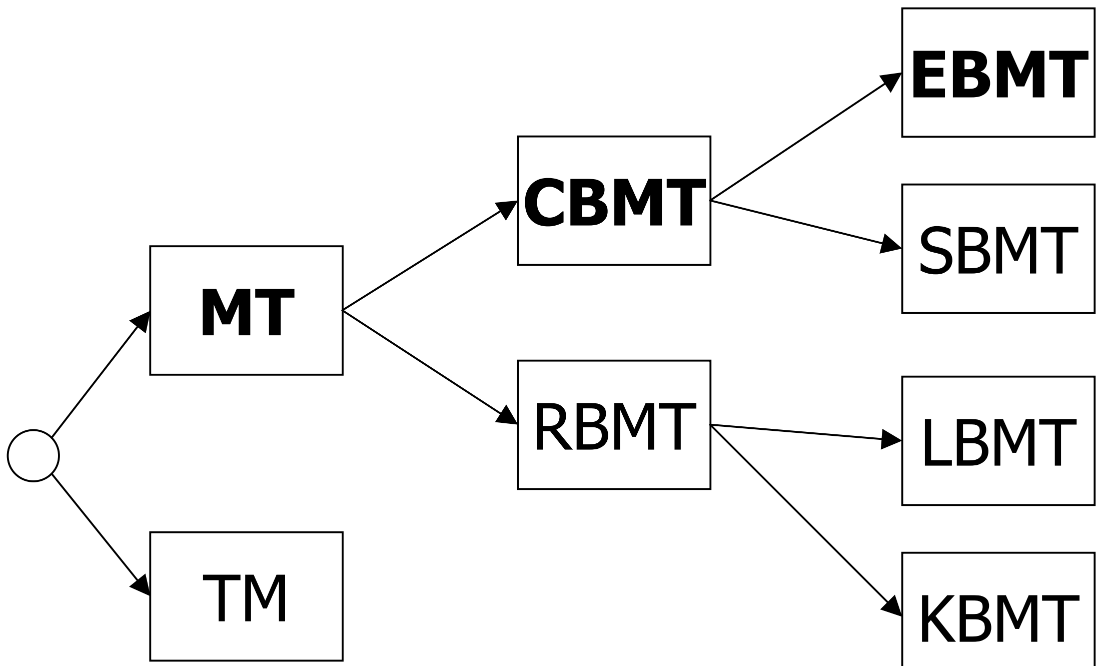
EBMT is an acronym for

Example-Based Machine Translation.

Analogy-based,
Memory-based,
Pattern-based,
Case-based,
Similarity-based,
,
,
,
,

EBMT in the hierarchy of translation technology

- **EBMT** is **a major approach** among **corpus-based approaches.**



TM ≠ EBMT

≠

TM: an **interactive tool for** bilingual **professional translators**

EBMT: an **automatic translator** for monolingual **ordinary people**

=

- the idea of **reusing past translation examples**
- the technology of **storing and retrieving** a large translation example collection

Good Reviews and Books

1. H. SOMERS, "Review Article: Example-based Machine Translation," *Journal of Machine Translation*, pp. 113-157, 1999.
2. N. Uramoto, Chap. 8 of *Natural Language Processing and Its Application*, H. Tanaka (ed.), IEICE (in Japanese), 1999.
3. M. Carl and A. Way, *Recent advances in Example-Based Machine Translation*, Kluwer Text, Speech and Dialog series, summer of 2002.
4. S. Sato, *Machine Translation by Analogy*, Kyoritsu-syuppan, p. 130, (in Japanese), 1997.

Outline

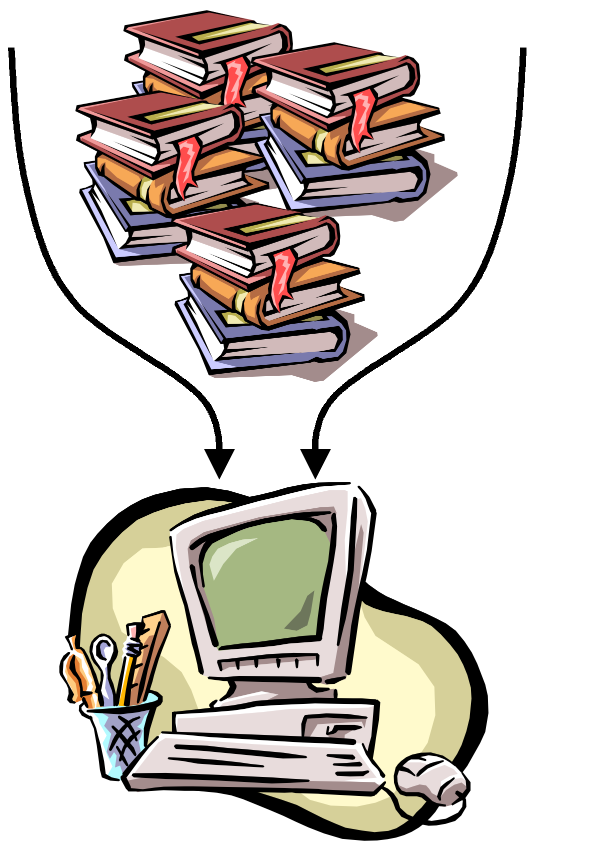
I. Concepts & Features
II. Elements
III. Case studies
IV. Remarks

Heinrich Shliemann, 19th century

- The discoverer of the remains of Troy.
- A born linguist.
  - His method of language study
    - He spent no time on grammar.
    - He learned fifteen foreign languages by simply memorizing textbooks.
    - Too hard for ordinary people.

Shliemann's method based on **memory fits the computer**.

- **Computers remember quickly and never forget data unless they are broken.**
- **Semiconductor price/performance is continuously doubling every eighteen months (Moore's Law).**
- **A tremendous number of documents are being input into computer networks.**



History

- The progress of the computers boosted EBMT.

	EBMT	Computer	**Cost/Performance**
1981	**Birth**	Mainframe	**1**
1989-	**Small-scale**	Workstation	**100**
2000-	**Large-scale**	PC	**10,000**

The Birth of EBMT (**1**)

Prof. **Nagao Makoto**'s seminal paper "Translation by analogy" **in 1981**.

Machine translation systems developed so far have a kind of inherent contradiction in themselves. **The more detailed a system has become by the additional improvements, the cleaner the limitation and the boundary will be made as for translation ability**. To break through this difficulty **we have to think about the mechanism of human translation**, and have to build a model based on the fundamental function of the language processing in human brain.

The Birth of EBMT (**2**)

"Translation by analogy."

(1) Man does **not** translate a simple sentence by doing deep **linguistic analysis**, rather,

(2) Man does translation, first, by properly decomposing an input sentence into certain fragmental phrases……………….. The translation of each fragmental phrase will be done by the **analogy** translation principle **via proper examples** as its reference.

Nagao's sample

A selection of **Japanese** translations for the English word "eat"

1.	A man	eats	vegetables
	Hito-wa	yasai-o	**taberu**
2.	Acid	eats	metal
	San-wa	kinzoku-o	**okasu**

input	He	eats	potatoes
output	kare-wa	poteto-o	**taberu**

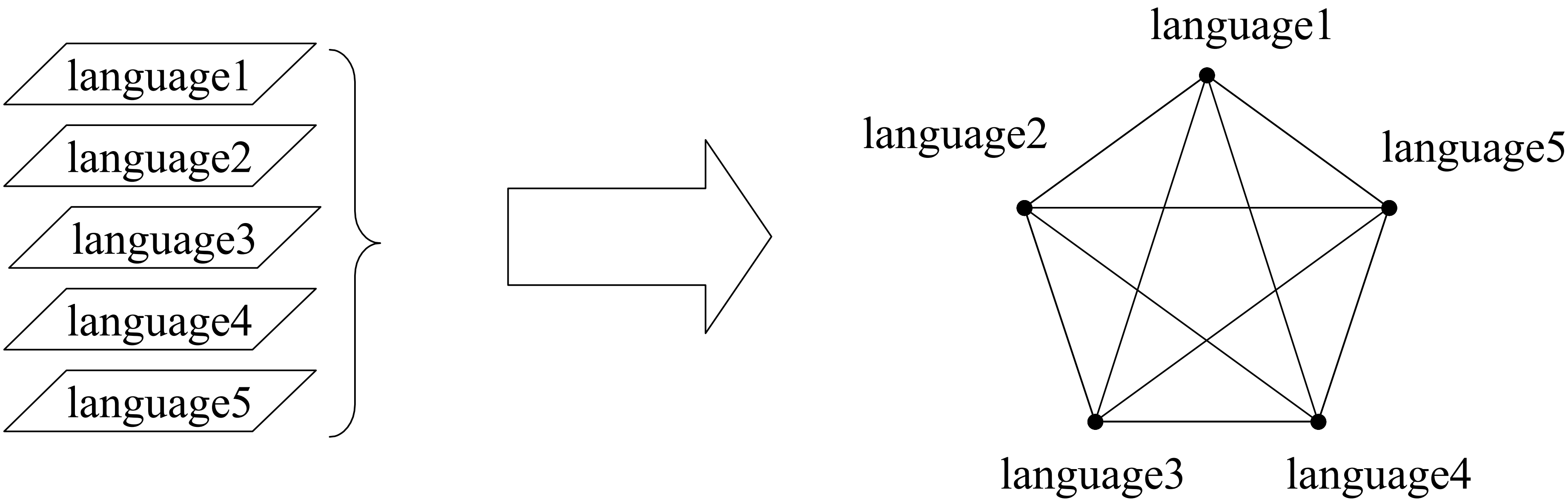
Suitable problems for EBMT

- EBMT is **solving** problems.
  1. **Knowledge building**
  2. **Translation quality**
  3. **Quality evaluation**.

- EBMT **is suitable** for
  A) **Multi-language** translation
  B) **Sub-language** translation
  C) **Non-literal** translation
  D) **Self-confident** translation

A) EBMT is suitable for **Multi-language** translation

- Knowledge is acquired automatically, so, EBMT is expandable by simply adding text for a new language.



n-lingual texts → n(n-1) MTs
(n=6000 → 36 million MTs)

B) EBMT is suitable for **sub-language translation**.

- For certain text types and subject domains, the language used is ***naturally*** **restricted in vocabulary and structures, therefore less ambiguous**.
- Defined by corpus.

  Weather bulletins, stock market reports, **instruction manuals**,

  ……………………………………………,

  travel conversation like phrase books

  ……………………………………………,

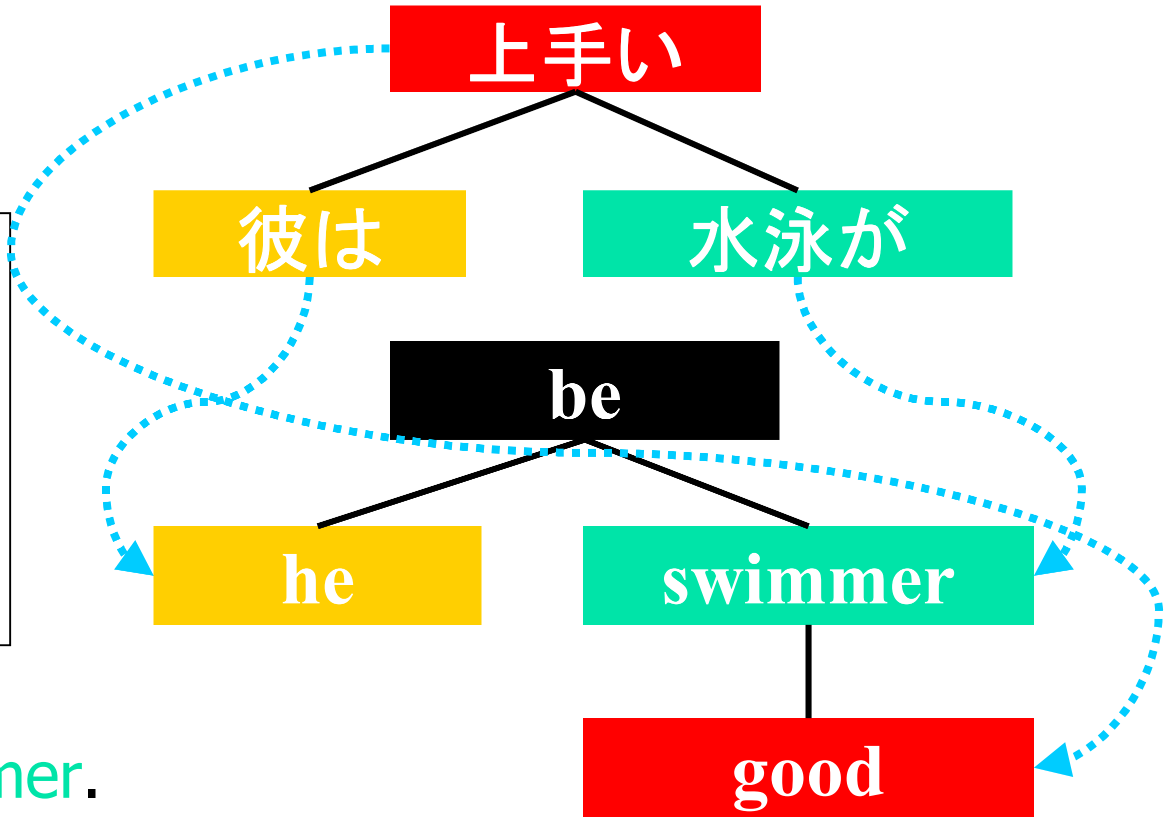
  legal contracts, patents.
- However, **high-quality translation** is often **required**.

C) EBMT is suitable for ***non-literal*** translation

彼は水泳が上手い。

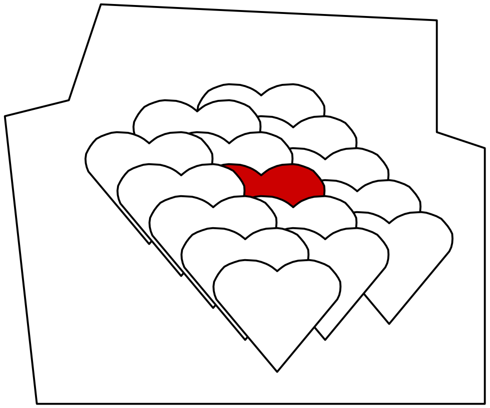
- difficult to deal with in a *structure-preserving* way.

He is a good swimmer.



D) EBMT is suitable for **self confident** translation

1. Output of conventional MTs = **A jar of cookies, some of which are poisoned.**
2. People want cookies to be often **required to** marked safe and delicious.
3. EBMT can attach a **reliability** value to each **translation.**
4. People can cooperate with EBMT.



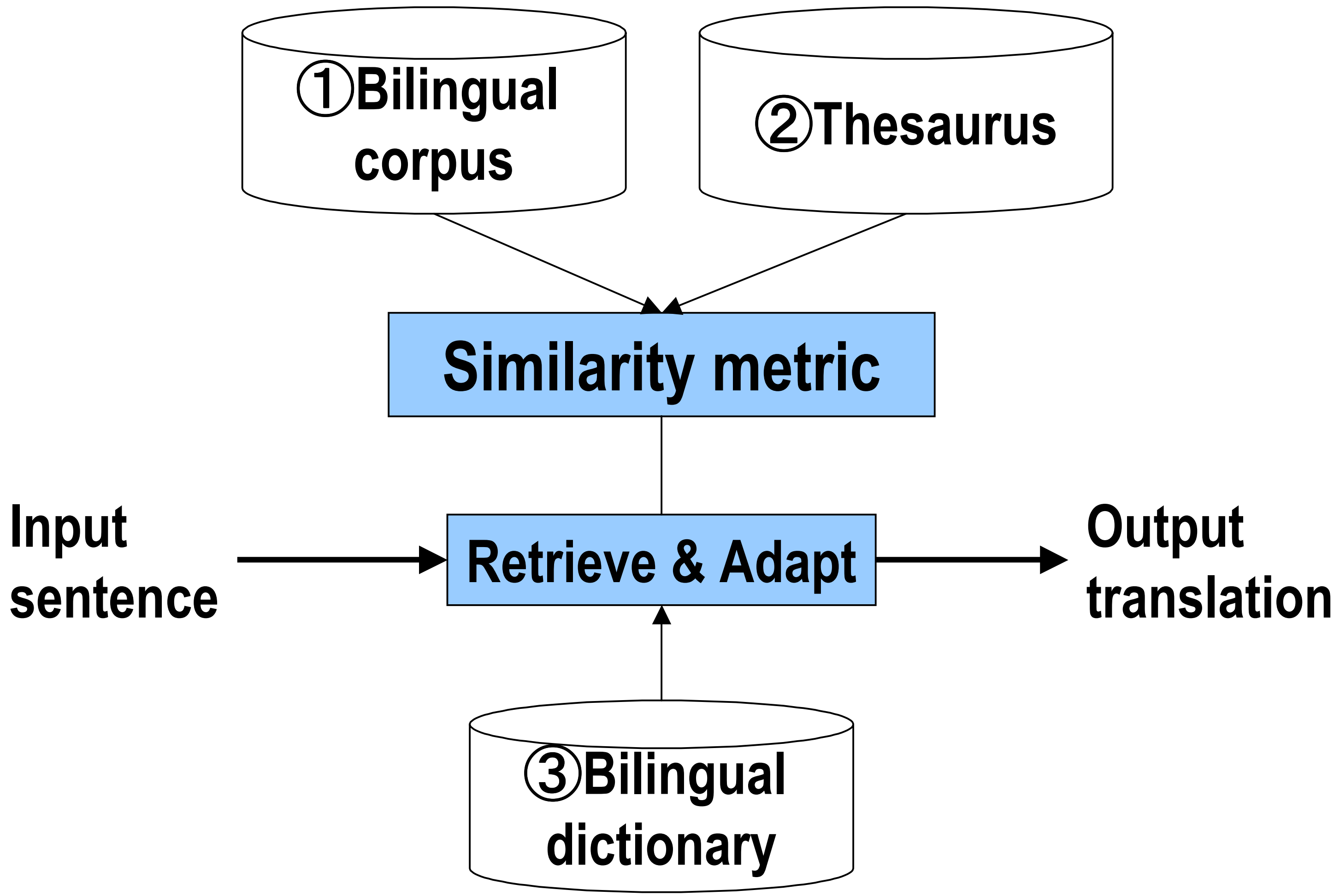
Outline

I. Concepts and Features
II. Elements
III. Case studies
IV. Remarks

Elements

- **Configuration**
- **Resources**
  - **Bilingual Corpus**
  - **Thesaurus**
- **Processes**
  - **Example Storage**
  - **Matching**
  - **Alignment**
  - **Acceleration**
- **Hybrid**

The basic Configuration of EBMT



An EBMT (Sumita, 1991)

- A notoriously tough problem, a Japanese NP of the form "A no B" into an English NP
- EBMT solved this translation problem accurately.

youka no gogo	*B **of** A*	*the afternoon **of** the 8th*
kaigi no sankaryou	*B **for** A*	*the fee **for** the conference*
kyouto no kaigi	*B **in** A*	*the conference **in** Kyoto*
issyuukan no kyuuka	*A **s**' B*	*one week'**s** holiday*
mittsu no hoteru	*A B*	*three hotels*

Bilingual Corpora (Types)

Easy to use

1. Comparable
   - Share the topic
2. Parallel
   - Translated
     - Documents in an international company
     - Canadian parliament proceedings
   - Aligned
     - Paragraph-Aligned
     - Sentence-Aligned
     - Word-Aligned

Easy to get

Bilingual Corpora (Sentence count)

- **Small-scale**
  - **10^1~10^3**
  - **Many systems**
- **Large-scale**
  - **10^4 ~10^5**
  - **PanEBMT@CMU, D^3@ATR, EBMT@VerbMobil, Candide@IBM,**
- Ultra large-scale
  - WEB (Grefenstette 99)

Thesauri (1)

- Used for **similarity or distance** calculation
  - eg., **distance** calculation in (Sumita, 91)

Level of MSCA	0	1	2	3 (same class)
Distance	1	2/3	1/3	0

- **Hand-made**
  - [E] WordNet, Roget
  - [J] Bunrui-Goi-Hyou, Kadokawa, EDR, NTT

Manning, D., C. and Schuetze, H., 1999. *Foundations of statistical natural language processing*, MIT Press, p. 680.

Thesauri (2)

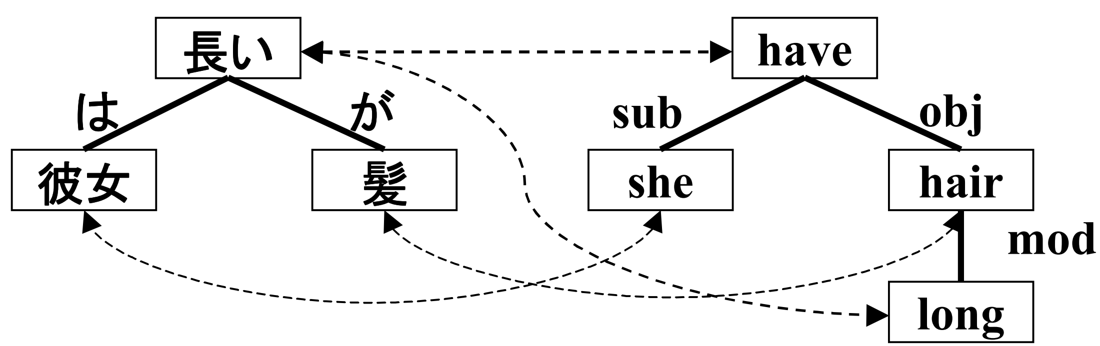
- **Computer-made**
  - Many methods have been based on **word distribution** in the corpus
    - Tanimoto, **Dice**, Overlap, Matching coefficient, Cosine,,
    - Eg. ***wine*** **~** ***beer***
      - *Wine* **co-occurs** for ***drink***, *grape*, ***bottle***, *red*, ***white***, *sweater*, ***bar***,,,,,.
      - *Beer* **co-occurs** for ***drink***, *grain*, ***bottle***, *belly*, *lager*, *black*, ***white***, ***bar***,,,,,.
  - **Not good with low-frequency words**

Storage

- Character sequence
  - 彼女は髪が長い⇔**She has long hair.**
- **Word sequence**
  - 彼女/は/髪/が/長い⇔**She/has/long/hair**
- **Syntactic / Semantic structure**

More informative

Easy to get



(Watanabe 92)

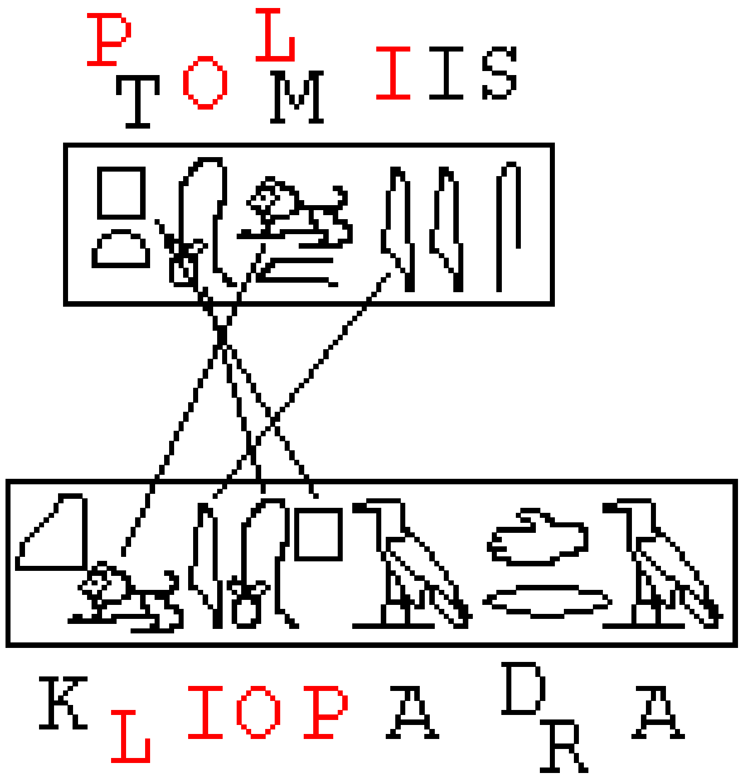
Matching

- Character-based
  - **EDIT DISTANCE** between character sequence
  - Eg. trans**la**ion～trans**al**ion
- Word-based
  - SEMANTIC DISTANCE based on **THESAURUS** (Eg. **translation～interpretation**)
- Structure-based
  - Constituent Boundary Parsing (Furuse 94)
  - TREE COVER SEARCH during transfer (Maruyama 92)
  - TREE EDIT DISTANCE (Zhang 97)

Alignment

1. Manning, 1999
2. Veronis, 2000
3. Melamed, 2001

- **Many** papers
  - Parallel vs. comparable
  - Statistics-based vs. lexicon-based
  - Sentence, Subsentence, and Word alignment



An alignment on the Rosetta Stone

Acceleration

- Can EBMT retrieve Mega examples quickly?
- Yes, definitely.
  - IR techniques
    - Indexing and compression
    - Clustering [Cranias 97]
  - Parallel processing
    - [Kitano 91, Sumita 93]

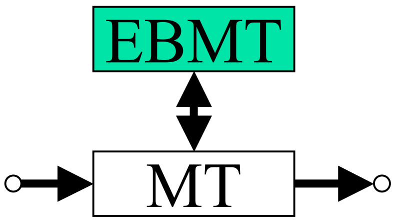
Hybrid (1)

- EBMT is not necessarily an all-around approach. It is complementary with other MT in coverage and quality.

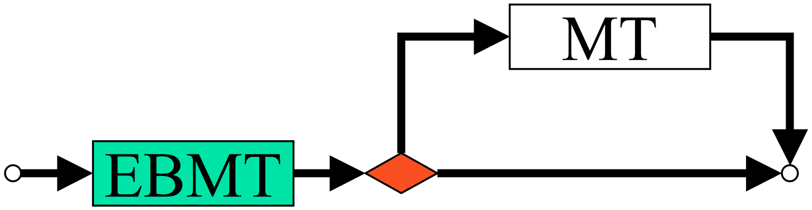
- **A hybrid architecture** is often adopted to improve performance.
  - Subroutine
  - Bypass
  - One engine of a multi-engine MT

Hybrid (2)

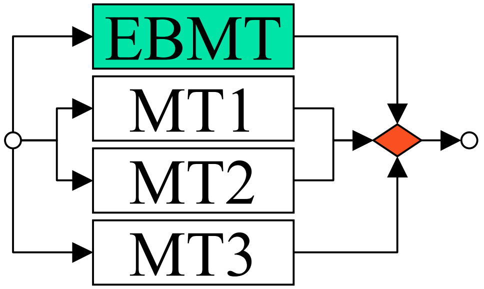
- Subroutine
  - (Sumita 91)(Sato 93)



- Bypass
  - (Katoh 94)



- Multi-engine
  - (Brown 96)



Outline

Sumita, E. 2001 "Example-based machine translation using DP-matching between word sequences," DDMT workshop of 39th ACL, pp. 1-8 .

1. Translation using **DP-matching**

D³ is an EBMT system.

Input いろ/が/気/に/入り/ません

RETRIEVE

Example デザイン/が/気/に/入り/ません

I do not like the design.

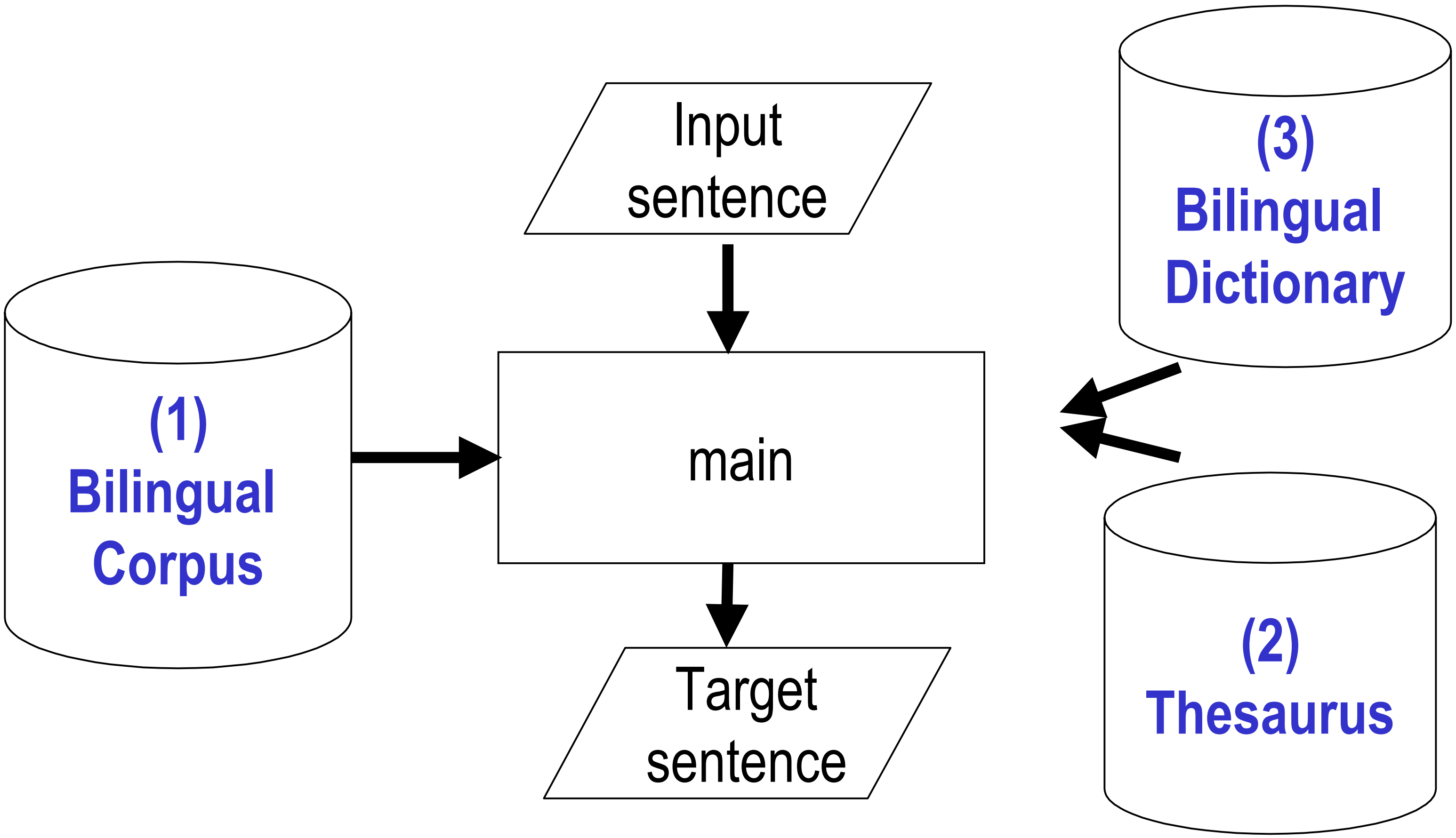
ADAPT

Output I do not like the color.

Characteristics of **D^3**

1. **D^3** assumes **neither syntactic parsing nor bilingual tree banks**;
2. **D^3 generates translation patterns on the fly** according to input and retrieved translation examples.

Three language data of **D**^{3}



Flowchart of **D^3**

(1) **Retrieve** the most similar translation pair by DP-Match

↓

(2) **Generate** translation patterns

↓

(3) **Select** the best translation pattern

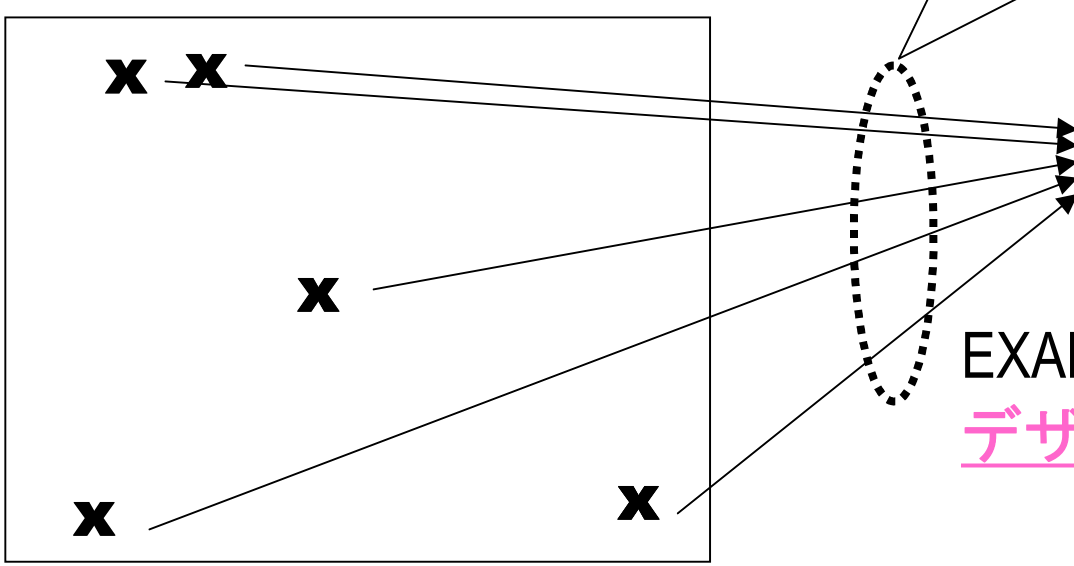
↓

(4) **Substitute** target words for source words

Step (1) **Retrieve** the similar pair

1. **Retrieve** similar example source sentences.
2. **Fail,** if not found.

dist (input sentence, example source sentence) < δ (=1/3)



INPUT:
いろが気にいりません

EXAMPLE SOURCE:
デザインが気にいりません
⋮

Distance between **word** sequences

- Distance, *dist* is computed by **DP-maching**.
- **Semantic distance**, *SEMDIST* is incorporated.

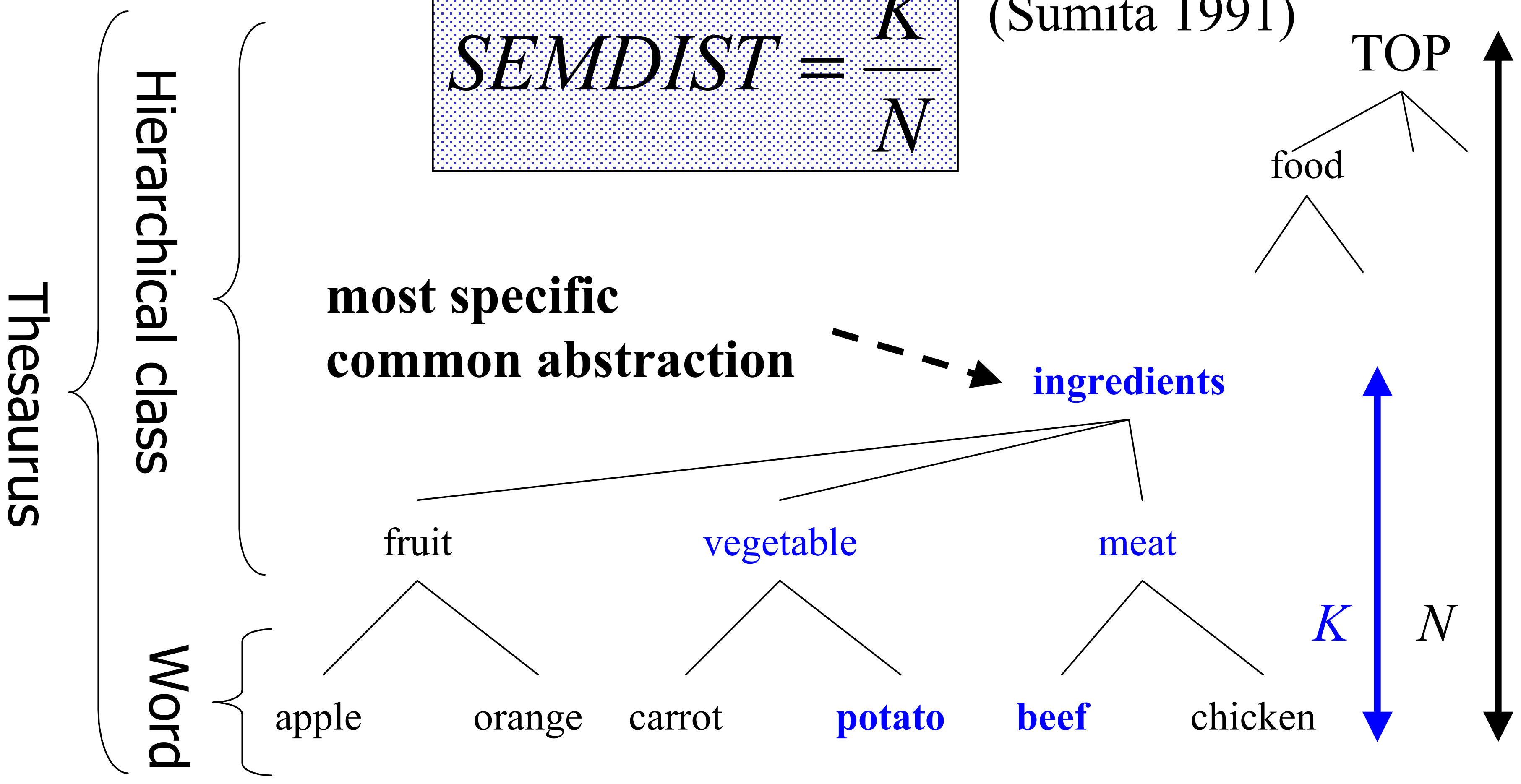
$$dist = \frac{I + D + 2\sum SEMDIST}{L_{input} + L_{example}}$$

Cormen, H. T., Leiserson, C. E. and Rivest, L. R. 1989. *Introduction to Algorithms*, MIT Press, p. 1028.

Semantic distance

$$SEMDIST = \frac{K}{N}$$

(Sumita 1991)



Sample of *dist* calculation

input:

<u>いろ</u>が気にいりません

example source:

<u>デザイン</u>が気にいりません

Deletion=0
Insertion = 0
Substitution = 1

SEMDIST = 1.0

dist
=(0+0+2*1.0) / (6+6)
=0.167

Step (2) **Generate** Translation Patterns

INPUT: いろが気にいりません

EXAMPLE_1
SOURCE: デザインが気にいりません
TARGET: **I do not like the design**

① Store input word

①X=色

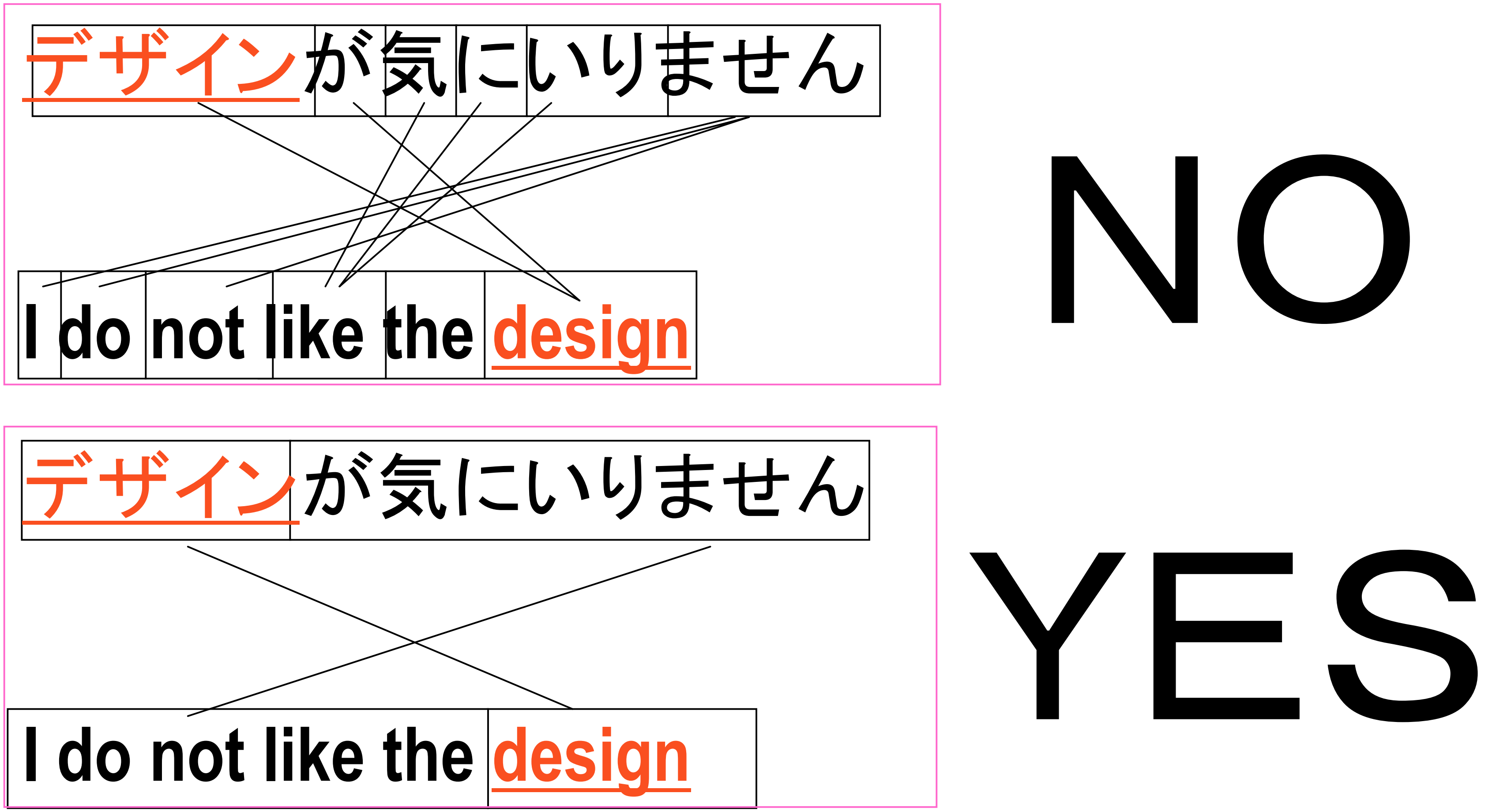
②
PATTERN_1
SOURCE: Xが気にいりません

TARGET: **I do not like the X**

Bilingual Dictionary
デザイン=design

② Align source and target

Large translation UNITS

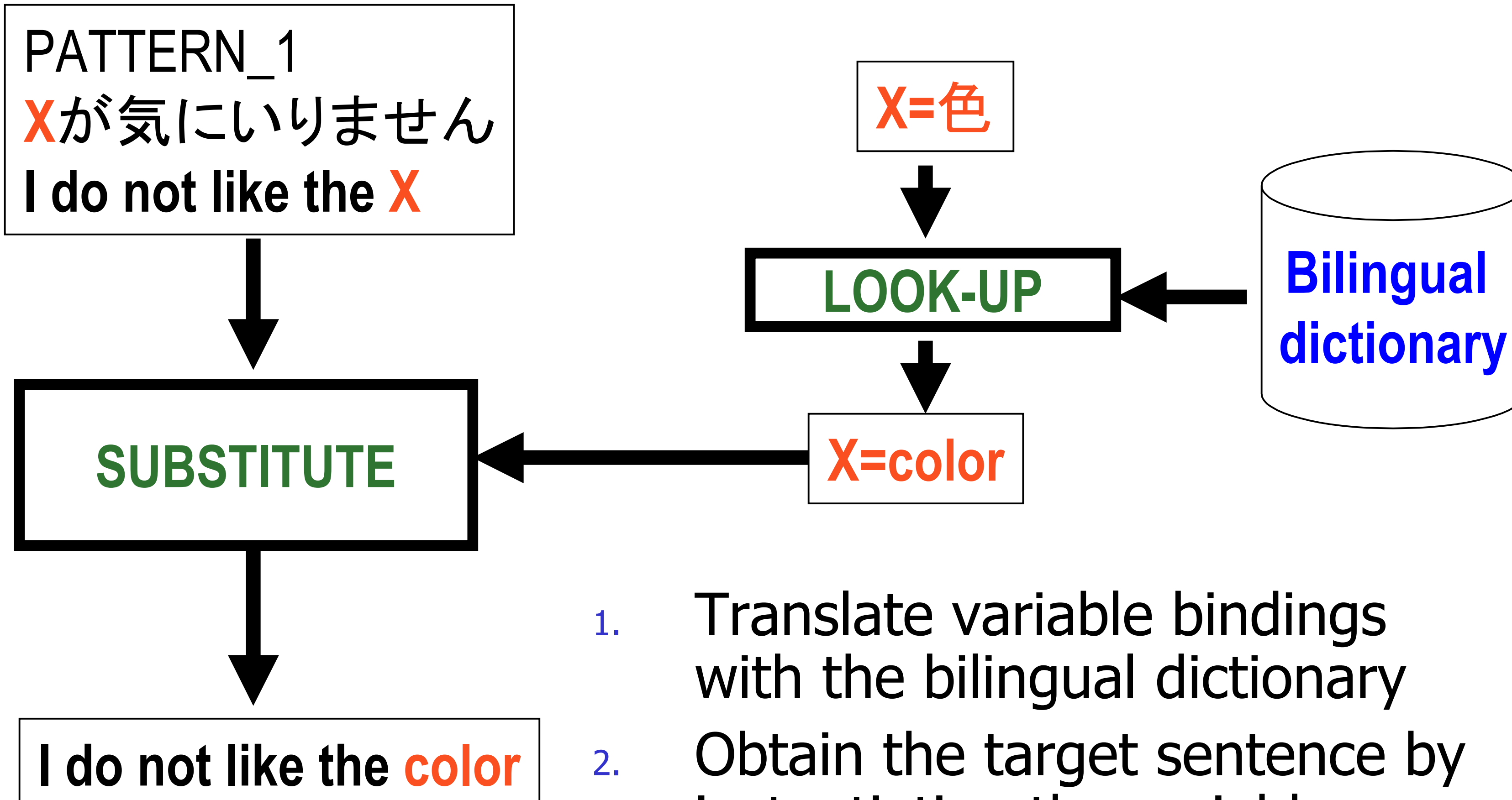


Step (3) **Select** the Best Translation Pattern

- There can be **multiple** translation patterns if translation examples have the same distance.

- Pick out the **most commonly used pattern** according to the next heuristic rule.
  - Maximize the frequency of the pattern.
    - Maximize the sum of frequencies of words in the generated patterns.
      - Select any one randomly as a last resort.

Step (4) **Substitute** target for source



1. Translate variable bindings with the bilingual dictionary
2. Obtain the target sentence by instantiating the variable.

Experiment with **200,000** sentences

1. **Preprocessing of Phrasebook**:
   - **Sentence-aligned**
   - **Morphologically tagged** on both sides
2. **Evaluation Procedure**:
   - **Test set (randomly-selected): 500**
   - **Example pairs: 200,000 – 500 = 199,500**
   - **The translation quality is ranked A,B,C,D from *good* to *bad*.**
3. **Bilingual dictionary**:
   - **20,000 words (from our spoken language translation system, TDMT)**
4. **Thesauri**:
   - **20,000 words (from our spoken language translation system, TDMT)**

Randomly-sampled pairs from our **Japanese** and **English** phrasebook corpus

J: フィルムを買いたいです。

E: I want to buy a roll of film.

J: 8人分予約したいです。

E: I'd like to reserve a table for eight.

J: 紅茶はありますか。

E: Do you have some tea?

J: 自動車を返したいのですが。

E: I 'd like to return the car.

J: そこに行くには橋を渡らねばなりません。

E: You need to cross the bridge to go there.

J: 友人が車にひかれ大けがをしました。

E: My friend was hit by a car and badly injured.

Coverage

	Sentences (%)
EXACT (0=dist)	46.4
DP (0<dist≦1/3)	43.4
No output	10.2

Covers about 90%

Coverage vs. sentence length

	%	average	min	max
EXACT	46.4	5.6	1	13
DP	43.4	7.7	2	22
No output	10.2	**11.0**	3	30
ALL	100.0	7.0	1	30

- **Non-covered sentences are LONGER.**

Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K., and Shirai, S.: 1999, 'Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach', Proc. of 7th MT Summit, pp. 229-235.

Quality

A: Perfect
B: OK
C: Understandable
D: Bad

Better

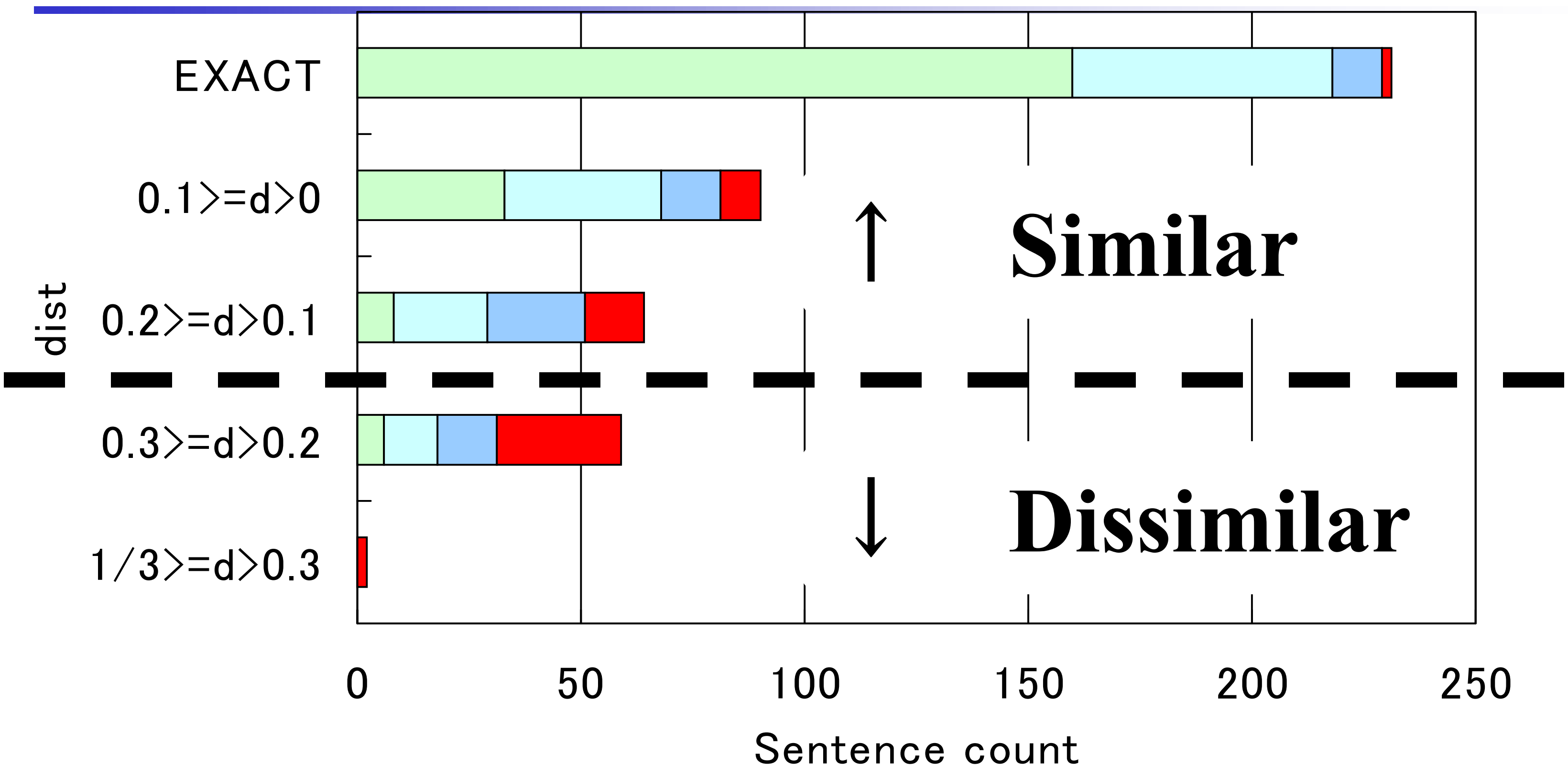
	Rank	%
Good	A	41.4
	B	25.2
	C	11.8
Bad	D	10.8
No output		10.8

Worse

- About 80% are good.

Quality vs. *dist*

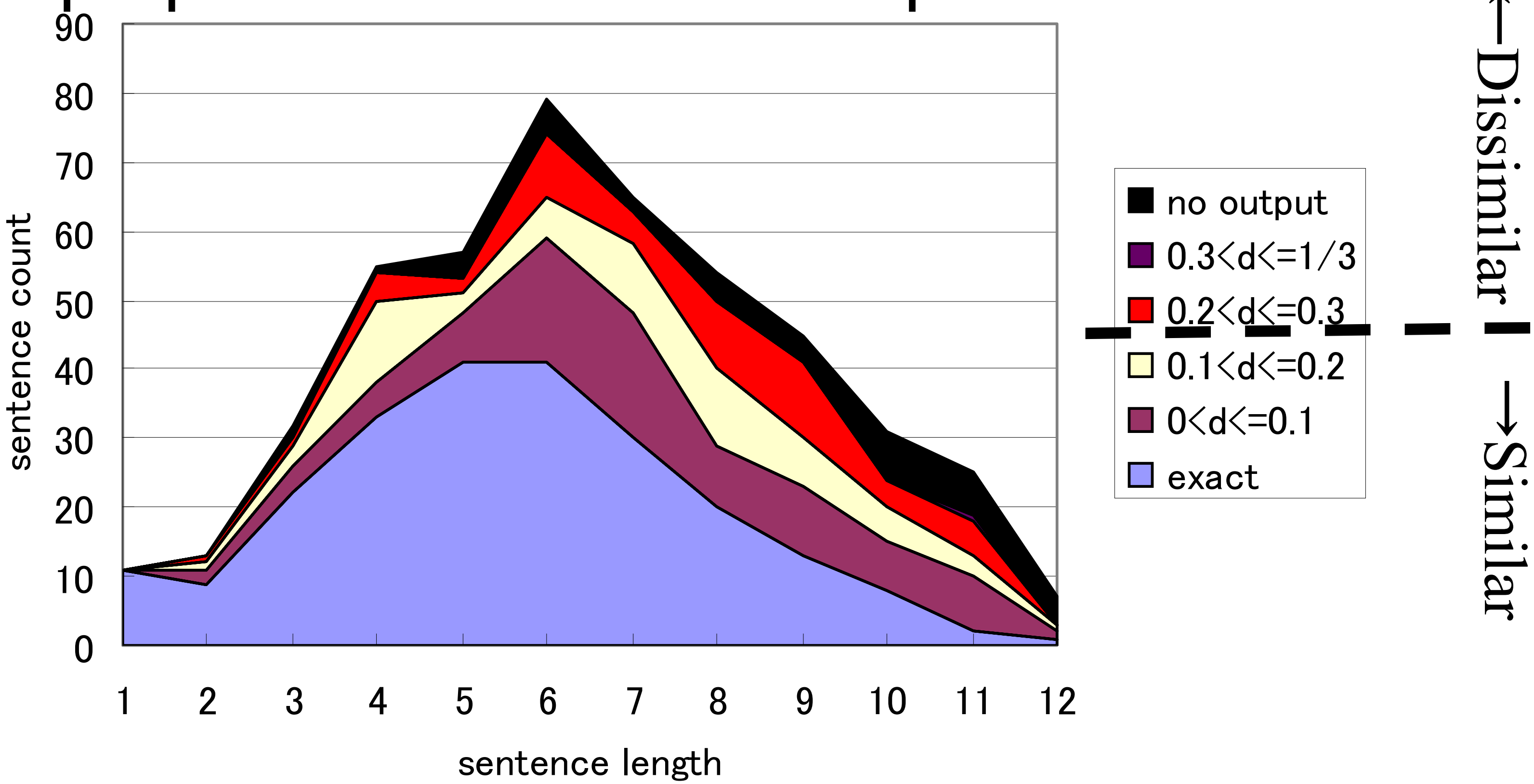
A: Perfect
B: OK
C: Understandable
D: Bad



Outputs reliability values and performs cooperatively with users.

Relationship between length and *dist*

The longer the input is, the larger the proportion of distant examples is.



Less frequent errors - **collocation**

1. 肩/を/つめて/いただけ/ます/か
2. 席/を/つめて/いただけ/ます/か

dist =0.167

1. **Could you tighten the shoulders up?**
2. **Could you move over a little?**

3. コーヒー/一杯/お/願い/し/ます
4. ビール/一杯/お/願い/し/ます

dist =0.056

3. **I'd like a cup of coffee.**
4. **I'd like a glass of beer.**

Less frequent errors - **context** dependency

In response to the question
“Do you have a shuttle bus?

はい/あり/ます

Translation 1.　Yes, we do.
Translation 2.　Yes, we have <u>a shuttle bus</u>.

D^3 Performance as of **Dec. 2001**

- With 200K corpus
  - Processing time
    - (average) **0.04 seconds**/sentence
    - (maximum) 0.66 seconds/sentence
  - Translation quality
    - matches Japanese with **TOEIC (Test Of English for International Communication) SCORE 750**

http://www.toeic.com/

Sugaya, F. *et al. Precise Measurement Method of a Speech Translation System's Capability with a Paired Comparison Method between the System and Humans*, MT-SUMMIT, 2001.

Wrap up of **D^3**

- D^3 uses **DP-matching,** featuring **semantic distance** between words.
- D^3 demonstrates **good quality** and **short turnaround** in a travel conversation such as these in a phrase-book.
- D^3 shows that **distance provides reliability**.

Future work in **D³**

- Methods pursued for improvements
  1. Improving coverage & accuracy
     - Chunking long sentences
     - Weight adjustment of edit operations or words
  2. Automation of constructing resources
     - Thesauri & bilingual lexicons
     - Sentence-alignment
  3. Integration with speech recognizer

No more **rules**.
Only **memory of past translations**.

- A computer **won against the chess world champion, Kasparov** in 1997.
  - Memory-based reasoning surpassed the conventional AI approach of using rules.
- Likewise, **EBMT will compete with a human translator** under some conditions.

(source: http://www.research.ibm.com/deepblue/home/html/b.html)

A **syntax**-based EBMT

Case study

1. Translation using DP-matching (D^3)
2. **Hierarchical Phrase Alignment (HPA)**
3. **HPA-based Translation (HPAT)**

2) HPA (Hierarchical Phrase Alignment)

K. Imamura 2001 Hierarchical phrase alignment harmonized with parsing, In Proc. of NLPRS, pp. 377-384.

Phrase alignment
= extracting **equivalent phrases** from bilingual text.

English: *I have just arrived in New York.*
Japanese: NewYork ni tsui ta bakari desu ga

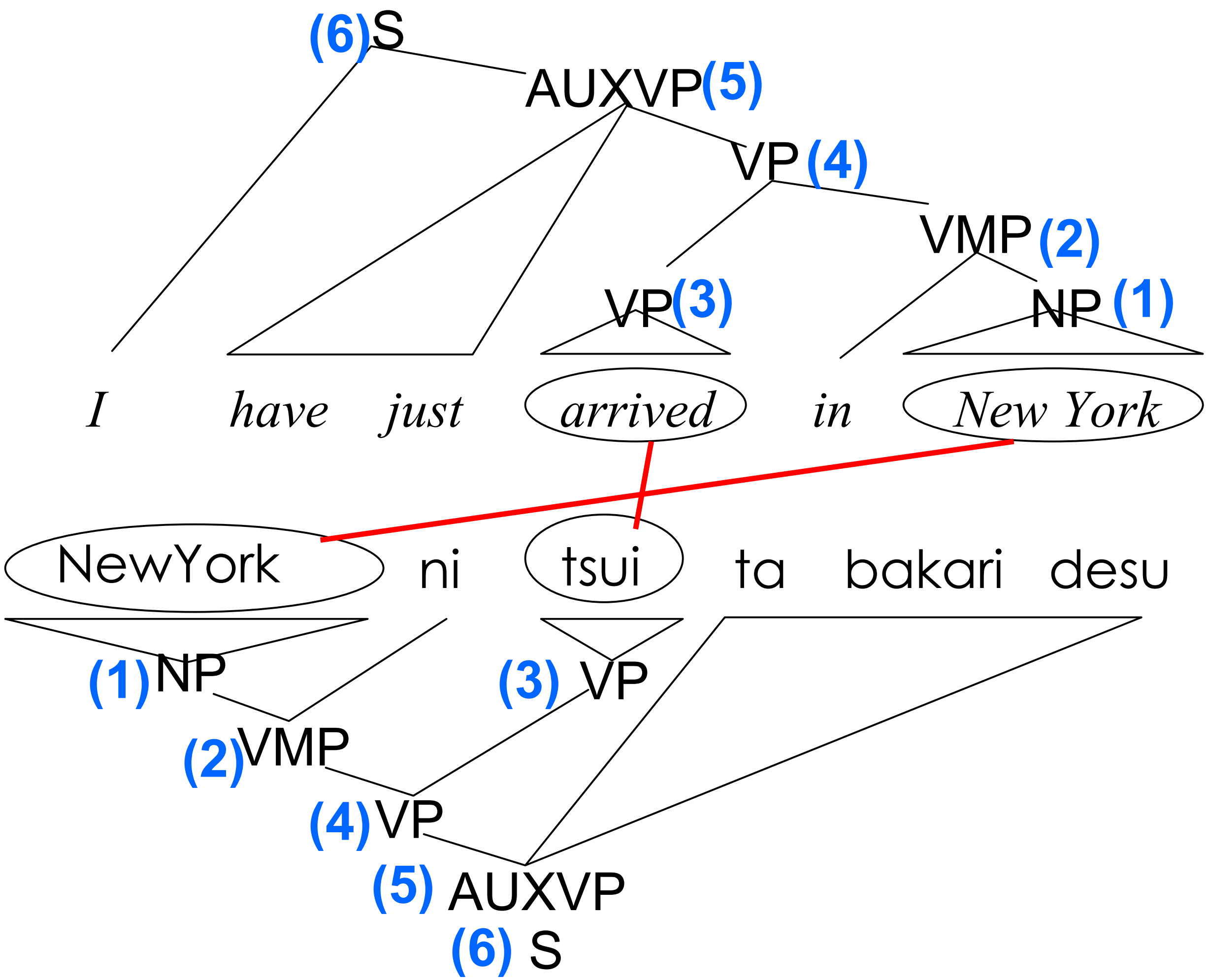
Phrase Alignment

in New York ⇔ NewYork ni
arrived in New York ⇔ NewYork ni tsui
have just arrived in New York ⇔ NewYork ni tsui ta bakari desu

Conditions of **equivalent phrases**

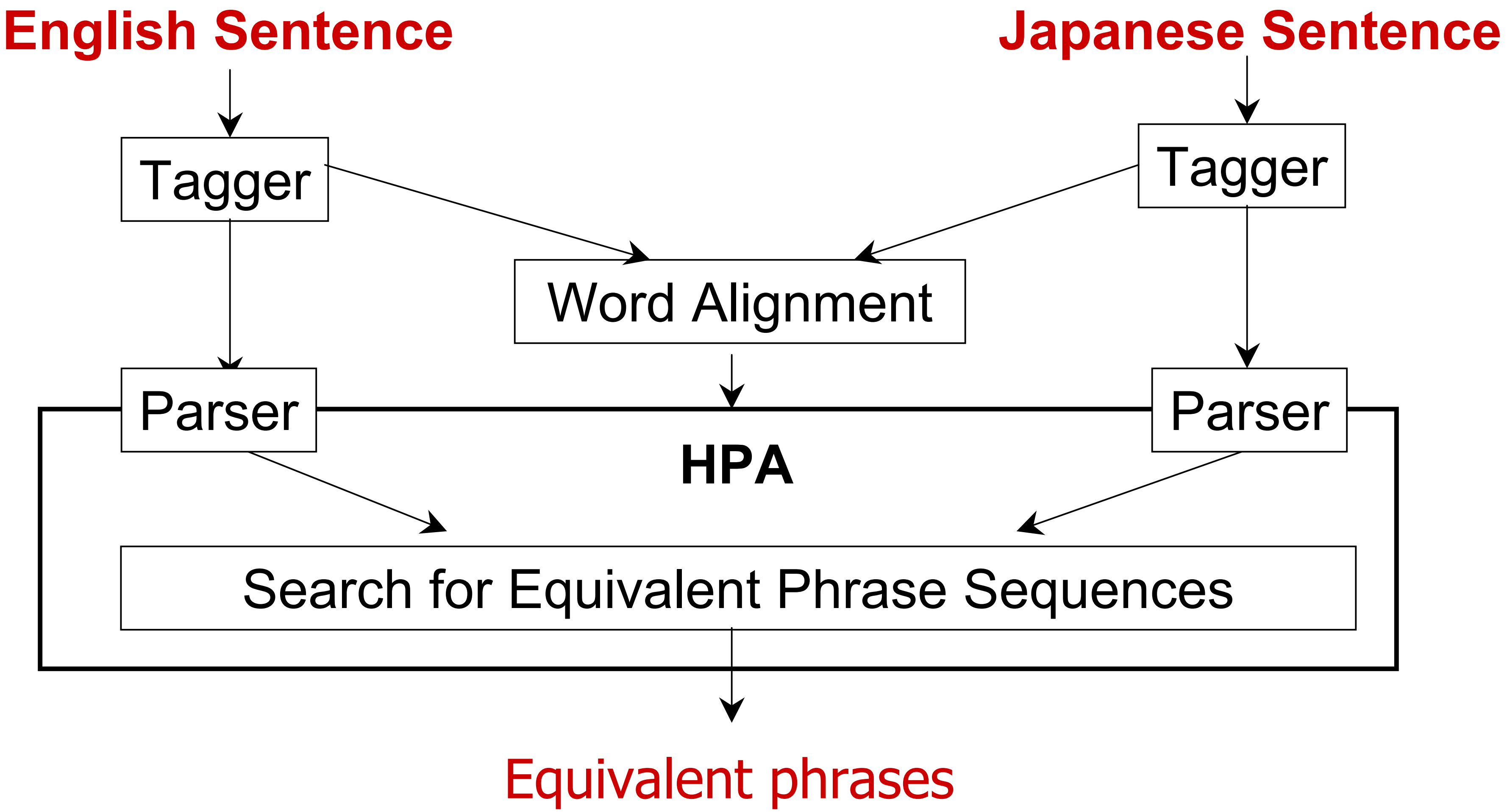
- Condition 1 (Same information)
  ＝**Content words in the pair correspond** with no deficiency and no excess.
- Condition 2 (Same type)
  ＝The phrases are **of the same syntactic category**.

Example of equivalent phrases



Six **equivalent phrases** that satisfy the two conditions.

Flow of **HPA**


English Sentence
Japanese Sentence
Tagger
Tagger
Word Alignment
Parser
Parser
HPA
Search for Equivalent Phrase Sequences
Equivalent phrases

Problem common to previous works

- Previous works of phrase alignment:
  - Between dissimilar language families
    - Kaji et al. (1992)
    - Matsumoto et al. (1993) Kitamura et al. (1995) Yamamoto et al. (2001)
    - Watanabe (2000)
  - Between similar language families
    - Meyers et al. (1996)
    - Menezes et al. (2001)

- They used the **final structures** produced by a parser.
- Problem: **Phrase alignment performance directly depends on parsing accuracy**.

Our **solutions** to the problem

① When the parsing process fails because of incomplete grammar.
- Find the best combination of parts of the unfinished tree

② When the parser selects the wrong candidate for ambiguous input.
- Find the more plausible tree

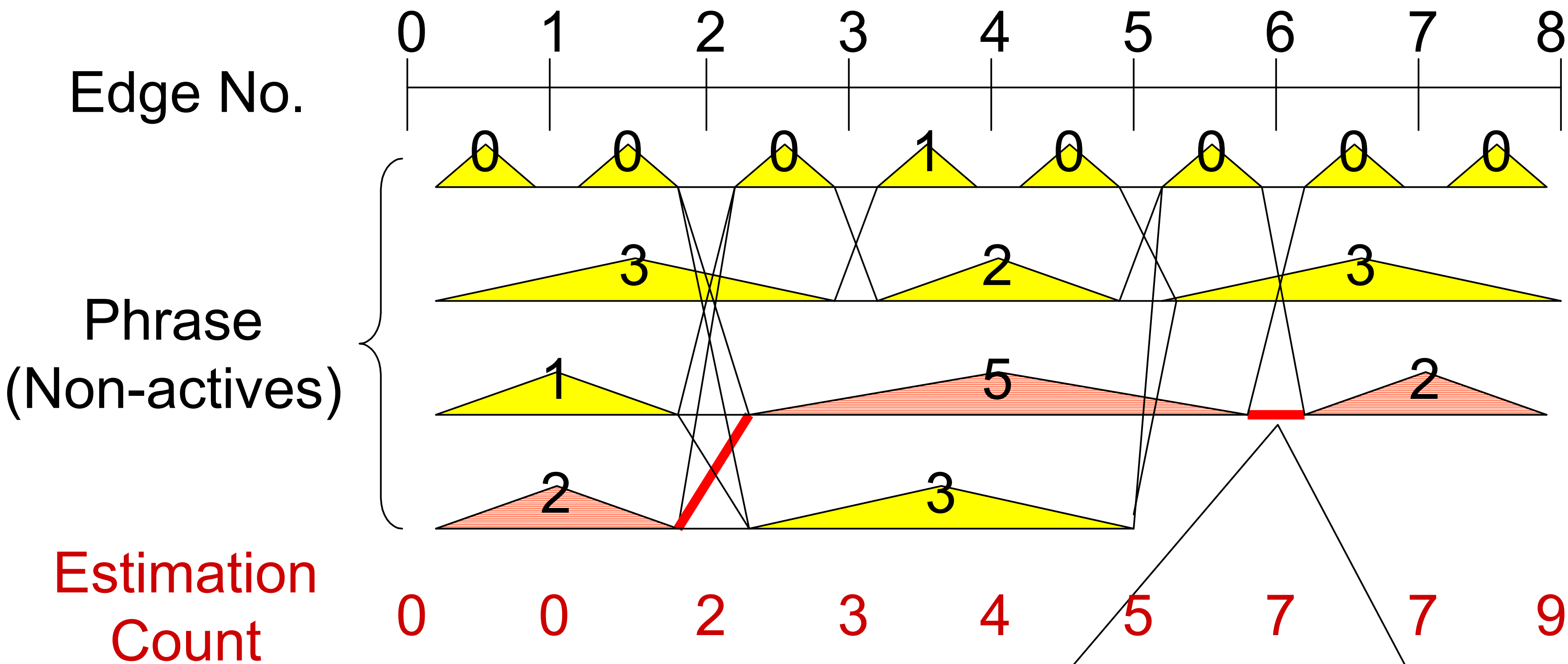
Maximize the count of **equivalent phrases** in combination of partial trees or tree.

① Combination of Partial Trees

- If we **combine partial trees appropriately**, we can overcome brittleness from incomplete grammar or deviations often found in spoken languages.

- To decrease the search time, we employ a **forward DP backward A* search algorithm**.

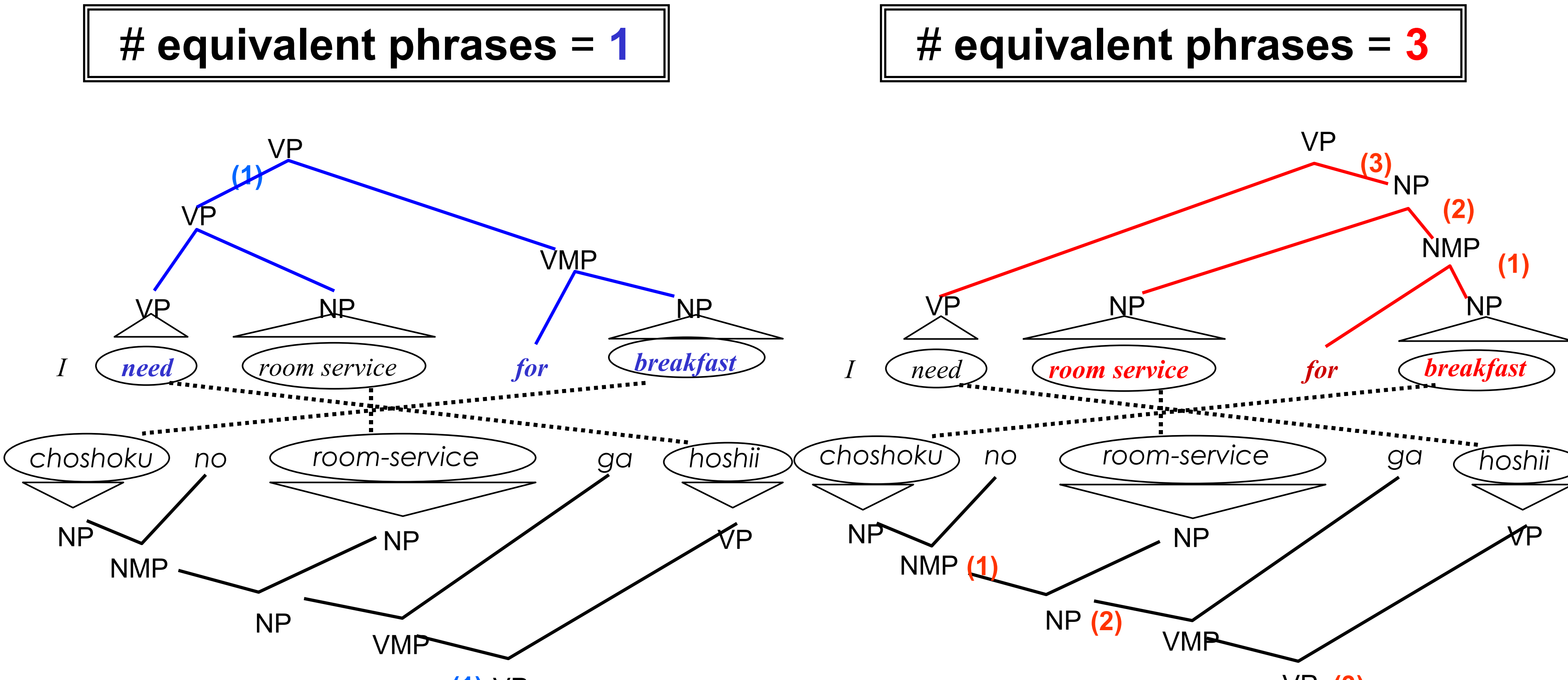
Forward DP Backward A* Search Algorithm



Search for the path that **maximizes the count** of **equivalent phrases** in combination.

② Plausible Attachment ("for breakfast")

- **Maximize** **the count of equivalent phrases** in tree.



Experimental Settings

- A bottom-up chart parser.
- Newly developed grammars.
  - Development cost = 2 person-months

	rule#	coverage	**accuracy**	ambiguity
English	284	67%	**44%**	4.18
Japanese	256	67%	**52%**	1.97

- 300 bilingual sentences used for evaluation.

HPA outperformed previous works

	Equivalent Phrase#	correct	Context-dependent	wrong
HPA	**1,676**	**86.2%**	5.8%	8.0%
Previous work	**726**	**86.5%**	6.3%	7.0%

●Compared with previous work, the proposed method extracted **twice as many equivalent phrases** with almost **no deterioration in accuracy**.

K. Imamura 2001 Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment, In Proc. of TMI, (in print).

3) **HPAT** (HPA based Translation)

- Extract **transfer pattern** from HPAed corpus in advance
- Translate using the **transfer pattern**
  - Parse
  - Transfer
  - Generate

HPAT: **Transfer Pattern**

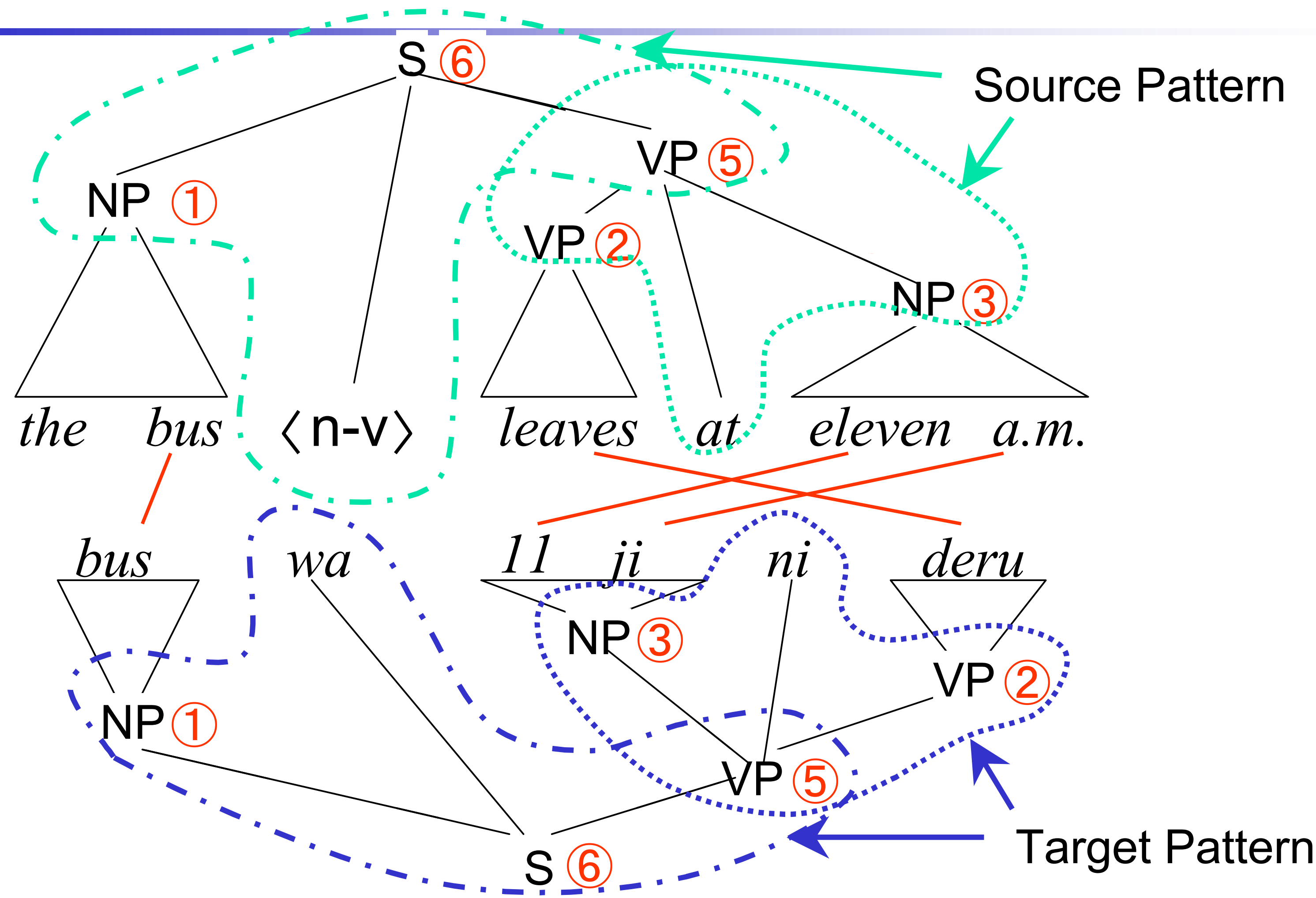
CFG

Syn. Cat.	Source Pattern		Target Pattern	Source Example
VP	X_{VP} *at* Y_{NP}	⇨	Y' *de* X'	(*present, conference*)
			Y' *ni* X'	(*stay, hotel*) (*arrive, p.m.*)
			Y' *wo* X'	(*look, it*)
NP	X_{NP} *at* Y_{NP}	⇨	Y' *no* X'	(*man, front desk*)

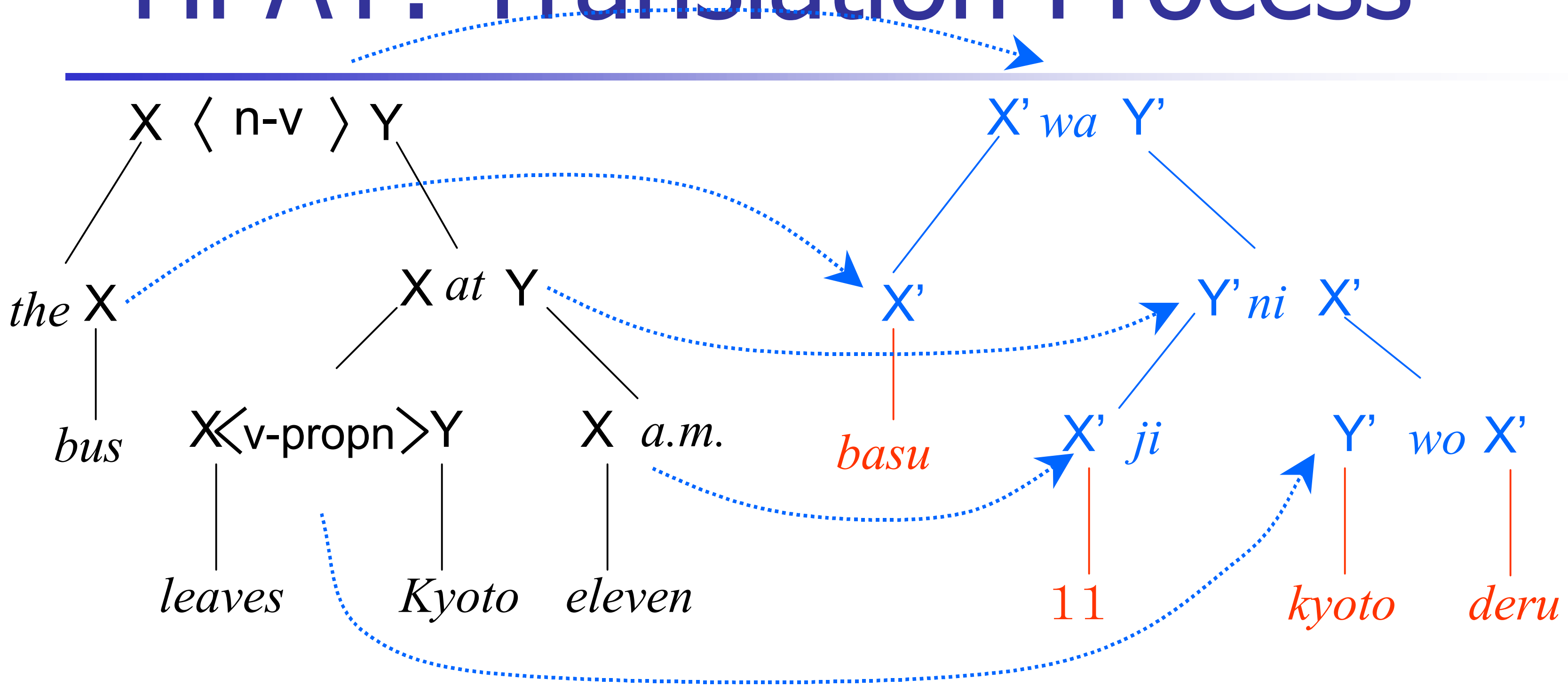
Mapping of source and Target patterns

Conditions of mapping from corpus

HPAT: Pattern Generation



HPAT: Translation Process



(1) Parse source language using source patterns.

(2) Map source patterns to target patterns.

(3) Translate leaves by referring to a dictionary.

Experiments: Settings

- A collection of phrases for overseas tourists.

Language	English	Japanese
Sentence#	125,579	
Total Word#	721,848	774,711
Vocabulary#	9,945	14,494
Equivalent Phrase#	**404,664**	

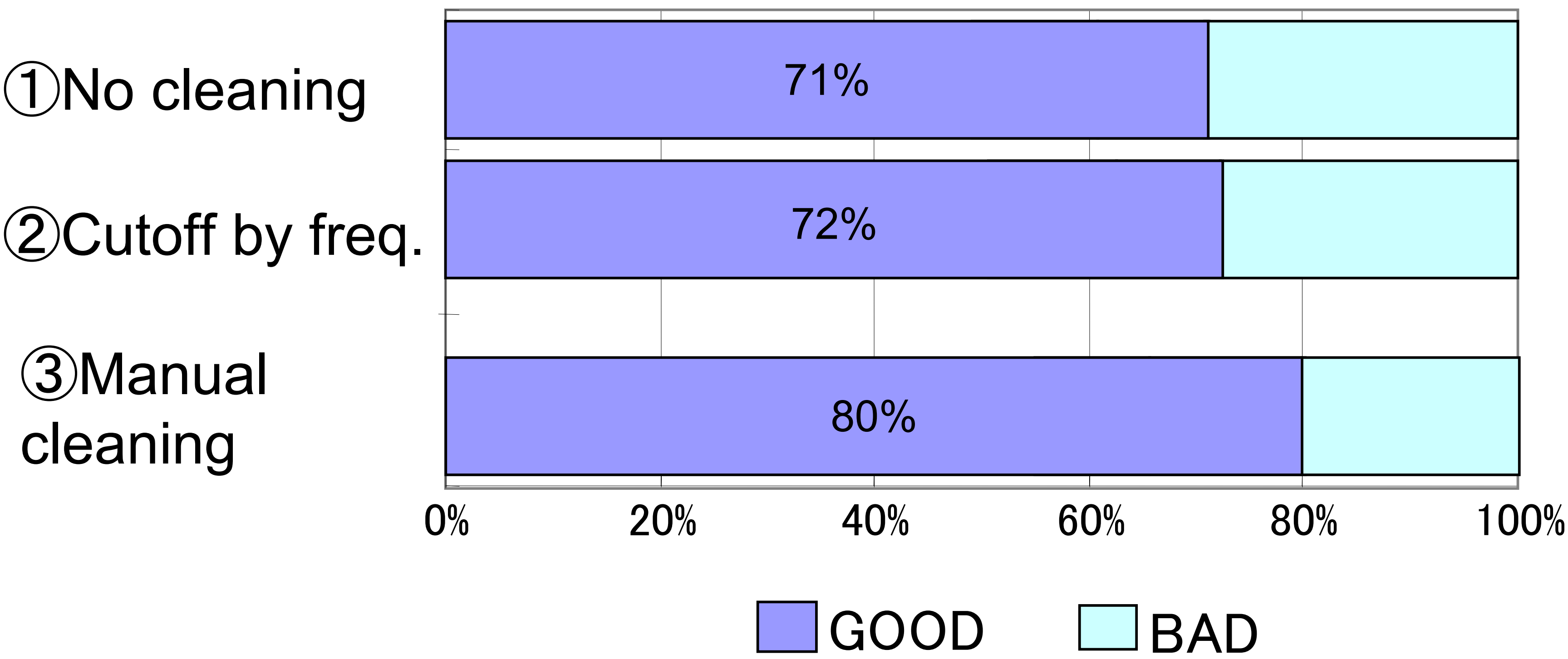
Results (1) Transfer Pattern Number

Cleaning Method	Pattern	Transfer Pattern#
①No cleaning	All	56,910
②Cutoff by freq.	More than 2 times	5,478
③Manual cleaning	Manually selected	635

1/10

1/10

Results (2) Translation Quality



Wrap-up of **HPAT**

- HPAT **automatically acquires transfer patterns** from a bilingual corpus by using HPA.
- Translation system based on the patterns achieved about **70%** accuracy.
- The upper-bound of the translation accuracy (**80%**) is estimated by selecting the subset of patterns by hand.
- We are working on **automatic selection of transfer patterns**.

Comparison with Menezes's Approach

	HPAT	Menezes's
Corpus	•Phrases for overseas tourists	•Help documents
PA	•Phrase structure •General rules	•Logical Form •Heuristic rules
Translator	•Constituent boundary anchor •Semantic distance based	•Content word anchor •Frequency based

Outline

I. Concepts and Features
II. Elements
III. Case studies
IV. Remarks

Comparison of **E**BMT and **S**BMT

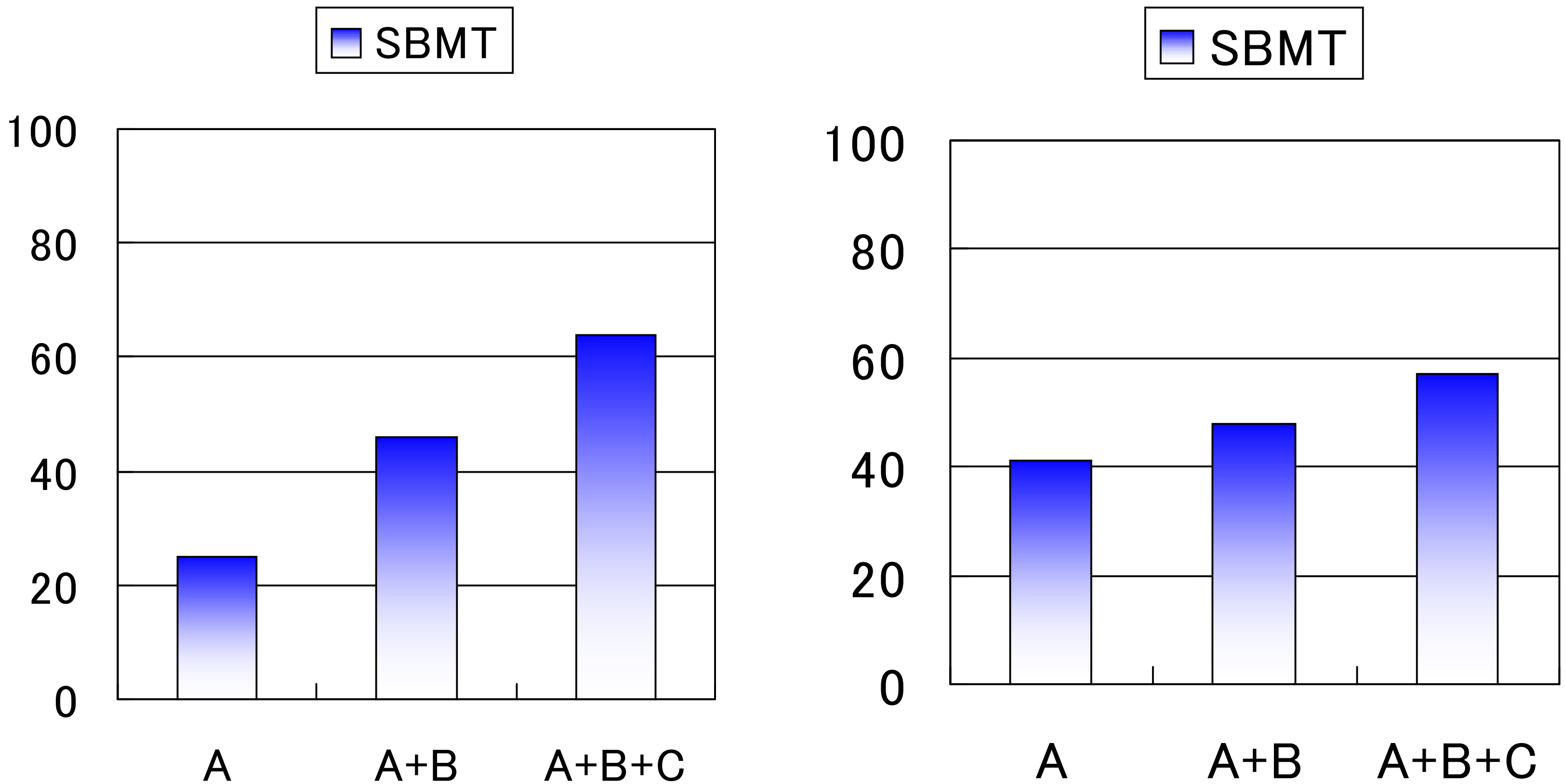
EBMT has been applied mainly to **Japanese and English**.

SBMT has been applied mainly to pairs of **European languages**.

We applied **S**BMT and **E**BMT to **the same Japanese and English corpus.**

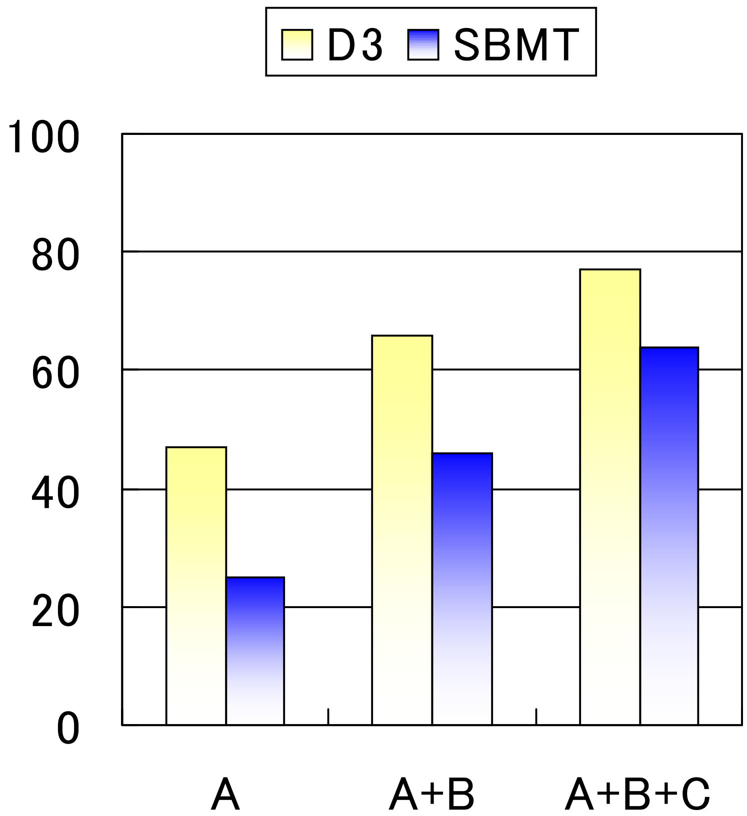
Taro Watanabe, Kenji Imamura, Eiichiro Sumita, 2002, *Statistical MT Based on Hierarchical Phrase Alignment*, Proc. of TMI.

SBMT **works** in E-to-J and J-to-E

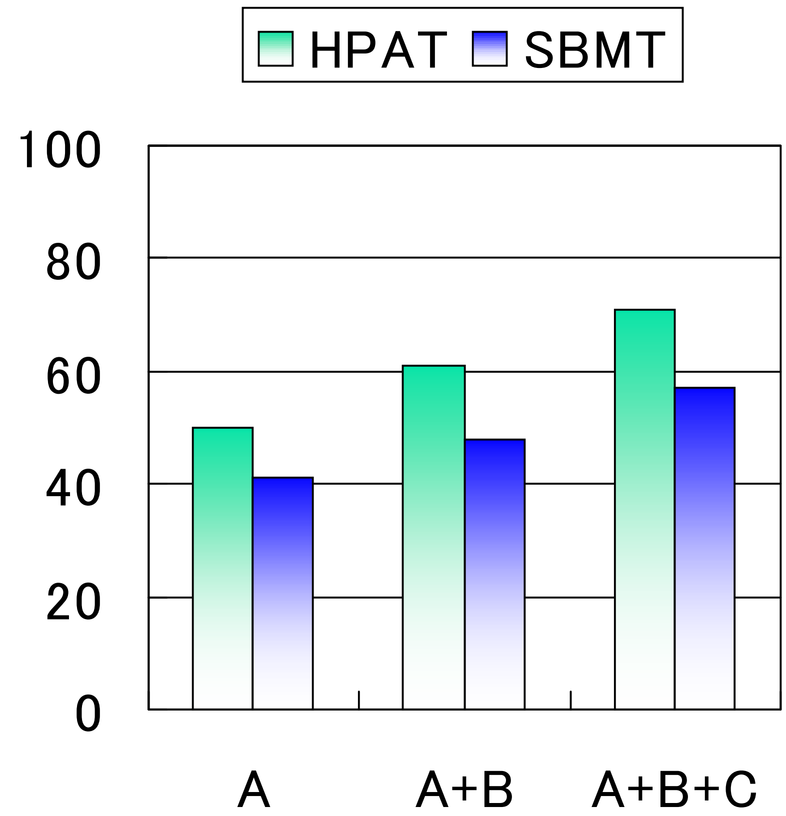


(Japanese to English) *(English to Japanese)*

EBMT surpasses SBMT
(as of October 2001)



(Japanese to English)



(English to Japanese)

Differences of EBMT and SBMT in **Japanese and English** translation

- **Unit**
  - EBMT (**sentence, phrase**) > SBMT (**word**)
- **Quality**
  - EBMT (good) > SBMT (poor)
- **Coverage**
  - EBMT (narrow) < SBMT (broad)
- **Robustness**
  - EBMT (less robust) < SBMT (robust)
- **Speed**
  - EBMT (fast) > SBMT (slow)

Outcome

- **Word-based SBMT**, a revival of the direct method of the ’50s, is suitable for pairs of **European languages** but not for **Japanese and English**.
- This is because **word-based SBMT** cannot capture the major differences between **Japanese and English**.
- Several organizations (Yamada 2001, Alshawi 2000) including ATR, are pursuing **syntax-based SBMT**.

- **Which is suitable for Japanese and English, syntax-based SBMT or EBMT?**

Corpus-related problems (1)

- EBMT is no longer a dream and **exhibits high quality** for a restricted domain such as travel conversation.
- EBMT will **grow rapidly** with SBMT.
- **Common underlying technology** such as phrase alignment will **support** two strategies of CBMT.

- A **common weak point is** that a **sentence-aligned large-scale corpus is not always available**.

Corpus-related problems (2)

- Corpus building
  - We do not have a way to estimate the **size** of the corpus needed for a domain.
  - We often do not have a **sentence-aligned** corpus or even a paragraph-aligned corpus.
  - We do not have a way to clean a **noisy** corpus.

Corpus-related problems (3)

- To realize broad-coverage and high-quality system:
  - We must exploit **heterogeneous corpora** of different types, cleaning levels, and other characteristics.

Other problems of EBMT

- Thesaurus
  - What is the best hierarchy?
  - How can we obtain a good thesaurus?
  - Can we cover specialized terms and proper nouns?
- What is the best definition of semantic distance?

Conclusions

- EBMT and SBMT are **attacking** problems.
  1. **Knowledge Building**
  2. **Translation Quality**
  3. **Quality Evaluation**.
- EBMT and SBMT are **solving** these problems.

- **Who** will **win** this interesting **race**?

Comments and questions

- Please e-mail to: eiichiro.sumita@atr.co.jp
- Thanks for coming!