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

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 **PCs**
      - (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


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.

<table>
<thead>
<tr>
<th>Year</th>
<th>EBMT</th>
<th>Computer</th>
<th>Cost/Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981</td>
<td>Birth</td>
<td>Mainframe</td>
<td>1</td>
</tr>
<tr>
<td>1989-</td>
<td>Small-scale</td>
<td>Workstation</td>
<td>100</td>
</tr>
<tr>
<td>2000-</td>
<td>Large-scale</td>
<td>PC</td>
<td>10,000</td>
</tr>
</tbody>
</table>
The Birth of EBMT (1)


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.
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.

\[ n\text{-lingual texts} \rightarrow n(n-1) \text{ MTs} \]
\[ (n=6000 \rightarrow 36 \text{ million MTs}) \]
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

1. Bilingual corpus
2. Thesaurus
3. Bilingual dictionary

Input sentence → Retrieve & Adapt → Output translation

Similarity metric
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.

<table>
<thead>
<tr>
<th>Japanese Phrase</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>youka no gogo</td>
<td>for (A) (B) of (A)</td>
</tr>
<tr>
<td>kaigi no sankaryou</td>
<td>for (A) (B) for (A)</td>
</tr>
<tr>
<td>kyouito no kaigi</td>
<td>in (A) (B) for (A) for the conference in Kyoto</td>
</tr>
<tr>
<td>issyuukan no kyuuuka</td>
<td>(A) s’ (B) one week’s holiday</td>
</tr>
<tr>
<td>mittsu no hoteru</td>
<td>(A) (B) three hotels</td>
</tr>
</tbody>
</table>
Bilingual Corpora (Types)

1. Comparable
   - Share the topic

2. Parallel
   - Translated
     - Documents in an international company
     - Canadian parliament proceedings
   - Aligned
     - Paragraph-Aligned
     - Sentence-Aligned
     - Word-Aligned
Bilingual Corpora (Sentence count)

- **Small-scale**
  - $10^1 \sim 10^3$
  - Many systems

- **Large-scale**
  - $10^4 \sim 10^5$
  - PanEBMT@CMU, D³@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)

<table>
<thead>
<tr>
<th>Level of MSCA</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3 (same class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>1</td>
<td>2/3</td>
<td>1/3</td>
<td>0</td>
</tr>
</tbody>
</table>

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

(Watanabe 92)
Matching

- Character-based
  - EDIT DISTANCE between character sequence
  - Eg. translation ~ translation

- 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

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

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

An alignment on the Rosetta Stone
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

I. Concepts and Features

II. Elements

III. Case studies
   1. Dp-match Driven transDucer (D^3)
   2. Hierarchical Phrase Alignment (HPA)
   3. HPA-based Translation (HPAT)

IV. Remarks
1. Translation using **DP-matching**

**D³** is an EBMT system.

**Input** いろ/が/気/に/入り/ません

**Example** デザイン/が/気/に/入り/ません
I do not like the **design**.

**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 D³

(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.

\[
\text{dist (input sentence, example source sentence)} < \delta \ (=1/3)
\]

**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

©ATR Spoken Language Translation Laboratories
Distance between **word** sequences

- Distance, \( \text{dist} \) is computed by **DP-maching**.
- **Semantic distance**, \( \text{SEMDIST} \) is incorporated.

\[
\text{dist} = \frac{I + D + 2 \sum \text{SEMDIST}}{L_{\text{input}} + L_{\text{example}}}
\]

**Semantic distance**

\[ \text{SEMDIST} = \frac{K}{N} \]

(Sumita 1991)

**Thesaurus**

- Word
  - apple
  - orange
  - carrot
  - potato
  - beef
  - chicken

- Hierarchical class
  - fruits
  - vegetables
  - meat

- Most specific common abstraction: ingredients

- TOP: food
Sample of \textit{dist} calculation

\begin{itemize}
  \item \textbf{Example source:} \textit{いろ}が気に入りません
  \item \textbf{Design source:} \textit{デザイン}が気に入りません
\end{itemize}

\begin{align*}
\text{SEMDIST} &= 1.0 \\
\text{Deletion} &= 0 \\
\text{Insertion} &= 0 \\
\text{Substitution} &= 1 \\
\text{dist} &= \frac{0 + 0 + 2 \times 1.0}{6 + 6} = 0.167
\end{align*}
Step (2) **Generate** Translation Patterns

\[
\begin{align*}
\text{INPUT: いろが気にいりません} \\
\text{EXAMPLE_1} \\
\text{SOURCE: デザインが気にいりません} \\
\text{TARGERT: I do not like the design} \\
\text{(1) } X=\text{色} \\
\text{(2) PATTERN_1} \\
\text{SOURCE: } X \text{が気にいりません} \\
\text{TARGERT: I do not like the } X
\end{align*}
\]
I do not like the design
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.

**PATTERN_1**

Xが気にいりません

*I do not like the X*

SUBSTITUTE

I do not like the **color**

**LOOK-UP**

X=色

X=color

**Bilingual dictionary**
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: 紅茶はありませんか。
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

<table>
<thead>
<tr>
<th>Coverage Type</th>
<th>Sentences (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXACT (0=dist)</td>
<td>46.4</td>
</tr>
<tr>
<td>DP (0&lt;dist≤1/3)</td>
<td>43.4</td>
</tr>
<tr>
<td>No output</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Covers about 90%
## Coverage vs. sentence length

<table>
<thead>
<tr>
<th></th>
<th>%</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXACT</td>
<td>46.4</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>DP</td>
<td>43.4</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>No output</td>
<td>10.2</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>ALL</td>
<td>100.0</td>
<td>1</td>
<td>30</td>
</tr>
</tbody>
</table>

Non-covered sentences are LONGER.
Quality

About 80% are good.

<table>
<thead>
<tr>
<th>Rank</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>41.4</td>
</tr>
<tr>
<td>B</td>
<td>25.2</td>
</tr>
<tr>
<td>C</td>
<td>11.8</td>
</tr>
<tr>
<td>D</td>
<td>10.8</td>
</tr>
<tr>
<td>No output</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Quality vs. $\text{dist}$

Outputs reliability values and performs cooperatively with users.
Relationship between length and $\text{dist}$

The longer the input is, the larger the proportion of distant examples is.
Less frequent errors - *collocation*

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

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?" 適い/あり/ます

Translation 1. Yes, we do.
Translation 2. Yes, we have a shuttle bus.
D³ 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

D³ uses **DP-matching**, featuring **semantic distance** between words.

D³ demonstrates **good quality** and **short turnaround** in a travel conversation such as these in a phrase-book.

D³ shows that **distance provides reliability**.
Future work in $\mathbb{D}^3$

- **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.

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)**

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

**English:**
*I have just arrived in New York.*

**Japanese:**
*NewYork ni tsui ta bakari desu ga*

---

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

Tagger

Word Alignment

Tagger

Parser

HPA

Search for Equivalent Phrase Sequences

Japanese Sentence

Tagger

Parser

Equivalent phrases
Problem common to previous works

- Previous works of phrase alignment:
  - Between dissimilar language families
    - Kaji et al. (1992)
    - 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.
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.
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.

# equivalent phrases = 1

# equivalent phrases = 3
Experimental Settings

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

<table>
<thead>
<tr>
<th></th>
<th>rule#</th>
<th>coverage</th>
<th>accuracy</th>
<th>ambiguity</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>284</td>
<td>67%</td>
<td>44%</td>
<td>4.18</td>
</tr>
<tr>
<td>Japanese</td>
<td>256</td>
<td>67%</td>
<td>52%</td>
<td>1.97</td>
</tr>
</tbody>
</table>

- 300 bilingual sentences used for evaluation.
**HPA outperformed previous works**

<table>
<thead>
<tr>
<th>Equivalent Phrase#</th>
<th>correct</th>
<th>Context-dependent</th>
<th>wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPA</td>
<td>1,676</td>
<td>86.2%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Previous work</td>
<td>726</td>
<td>86.5%</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

Compared with previous work, the proposed method extracted **twice as many equivalent phrases** with almost no deterioration in accuracy.
3) **HPAT** (HPA based Translation)

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

---

## HPAT: Transfer Pattern

<table>
<thead>
<tr>
<th>Syn. Cat.</th>
<th>Source Pattern</th>
<th>Target Pattern</th>
<th>Source Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP</td>
<td>$X_{VP} \text{ at } Y_{NP}$</td>
<td>$Y' \text{ de } X'$</td>
<td><em>(present, conference)</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Y' \text{ ni } X'$</td>
<td><em>(stay, hotel)</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Y' \text{ wo } X'$</td>
<td><em>(arrive, p.m.)</em></td>
</tr>
<tr>
<td>NP</td>
<td>$X_{NP} \text{ at } Y_{NP}$</td>
<td>$Y' \text{ no } X'$</td>
<td><em>(man, front desk)</em></td>
</tr>
</tbody>
</table>

### Mapping of source and Target patterns

### Conditions of mapping from corpus
The bus leaves at eleven a.m.
(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.

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence#</td>
<td>125,579</td>
<td></td>
</tr>
<tr>
<td>Total Word#</td>
<td>721,848</td>
<td>774,711</td>
</tr>
<tr>
<td>Vocabulary#</td>
<td>9,945</td>
<td>14,494</td>
</tr>
<tr>
<td>Equivalent Phrase#</td>
<td>404,664</td>
<td></td>
</tr>
</tbody>
</table>
## Results (1) Transfer Pattern Number

<table>
<thead>
<tr>
<th>Cleaning Method</th>
<th>Pattern</th>
<th>Transfer Pattern#</th>
</tr>
</thead>
<tbody>
<tr>
<td>① No cleaning</td>
<td>All</td>
<td>56,910</td>
</tr>
<tr>
<td>② Cutoff by freq.</td>
<td>More than 2 times</td>
<td>5,478</td>
</tr>
<tr>
<td>③ Manual cleaning</td>
<td>Manually selected</td>
<td>635</td>
</tr>
</tbody>
</table>

© ATR Spoken Language Translation Laboratories 81
Results (2) Translation Quality

① No cleaning: 71%
② Cutoff by freq.: 72%
③ Manual cleaning: 80%

GOOD | BAD
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

<table>
<thead>
<tr>
<th></th>
<th>HPAT</th>
<th>Menezes’s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus</strong></td>
<td>• Phrases for overseas tourists</td>
<td>• Help documents</td>
</tr>
<tr>
<td><strong>PA</strong></td>
<td>• Phrase structure</td>
<td>• Logical Form</td>
</tr>
<tr>
<td></td>
<td>• General rules</td>
<td>• Heuristic rules</td>
</tr>
<tr>
<td><strong>Translator</strong></td>
<td>• Constituent boundary anchor</td>
<td>• Content word anchor</td>
</tr>
<tr>
<td></td>
<td>• Semantic distance based</td>
<td>• Frequency based</td>
</tr>
</tbody>
</table>
Outline

I. Concepts and Features
II. Elements
III. Case studies
IV. Remarks
Comparison of EBMT and SBMT

EBMT has been applied mainly to Japanese and English.

SBMT has been applied mainly to pairs of European languages.

We applied SBMT and EBMT to the same Japanese and English corpus.
SBMT works in E-to-J and J-to-E

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
- 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!