MT Research in Japan

Hitoshi ISAHARA

National Institute of Information and Communications Technology

- R&D of EBMT
  - Japanese-Chinese Machine Translation System
  - Application
  - Linguistic resources
- Medical Examination Support System
- Translation of Aviation Manuals
- Asia-Pacific Association of Machine Translation（AAMT)

R&D of Japanese-Chinese Machine Translation System

- Five year national project from 2006 FY.
  - Budget: 2 million US$ per year.
- Our objectives are;

  Making scientific and technological information in China and Japan easily usable in both countries.
- Translation engines will be based on example-based machine translation.
- Syntactic information will be used in many parts of the translation engines.

Linguistic Resources

Dictionaries
(Providing glosses for specialist terminology
and general vocabulary, case frames,
semantic systems)

Corpora
(With syntactic annotations,
parallel corpora)

Expansion of Linguistic resources

Analysis engine

Translation engine

Scientific and technical papers
Japanese/Chinese

Scientific and technical papers
Chinese/Japanese

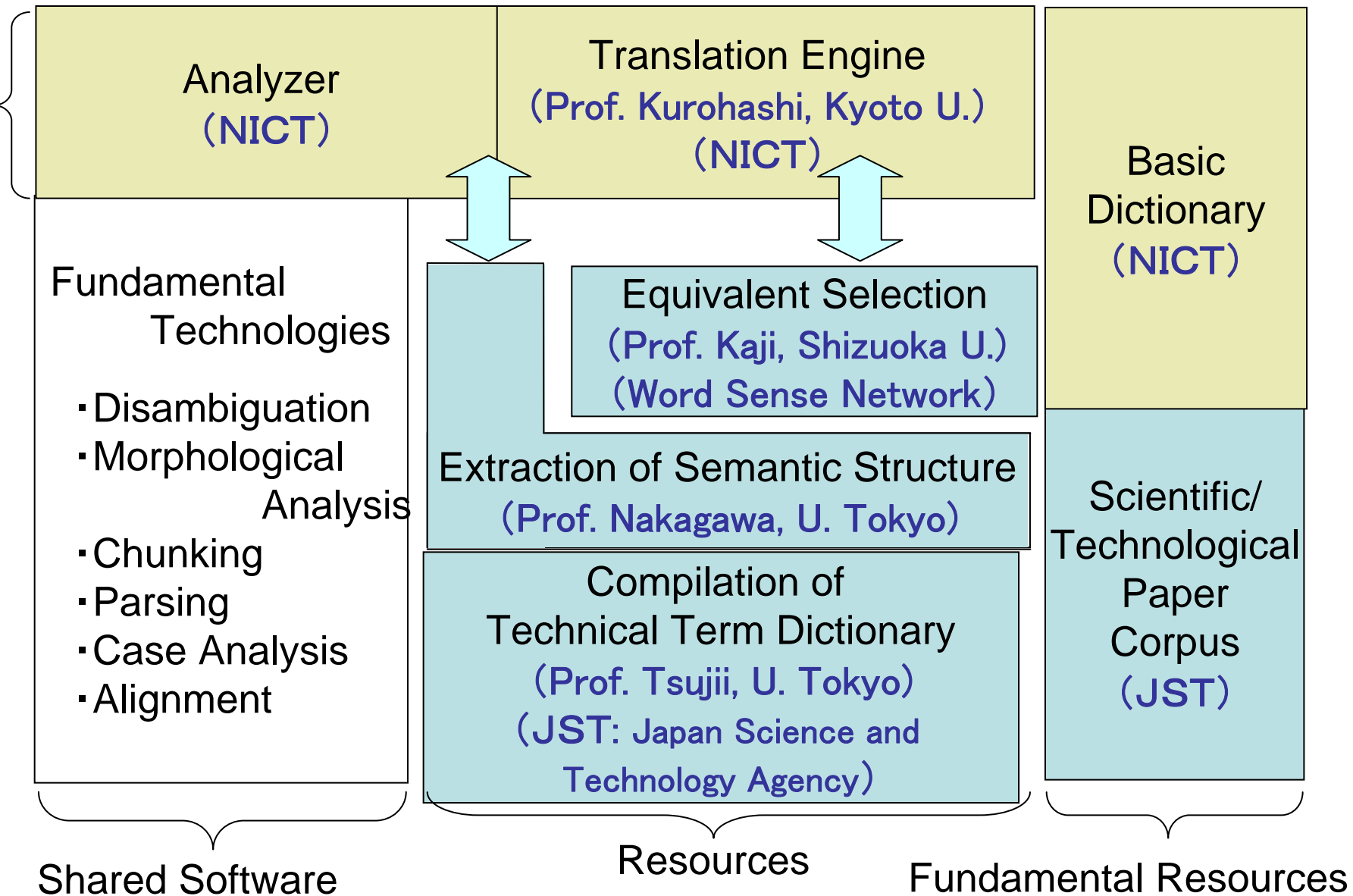**Example-based translation with deep consideration to syntactic structure**

Delivering a practical machine translation system within a new framework

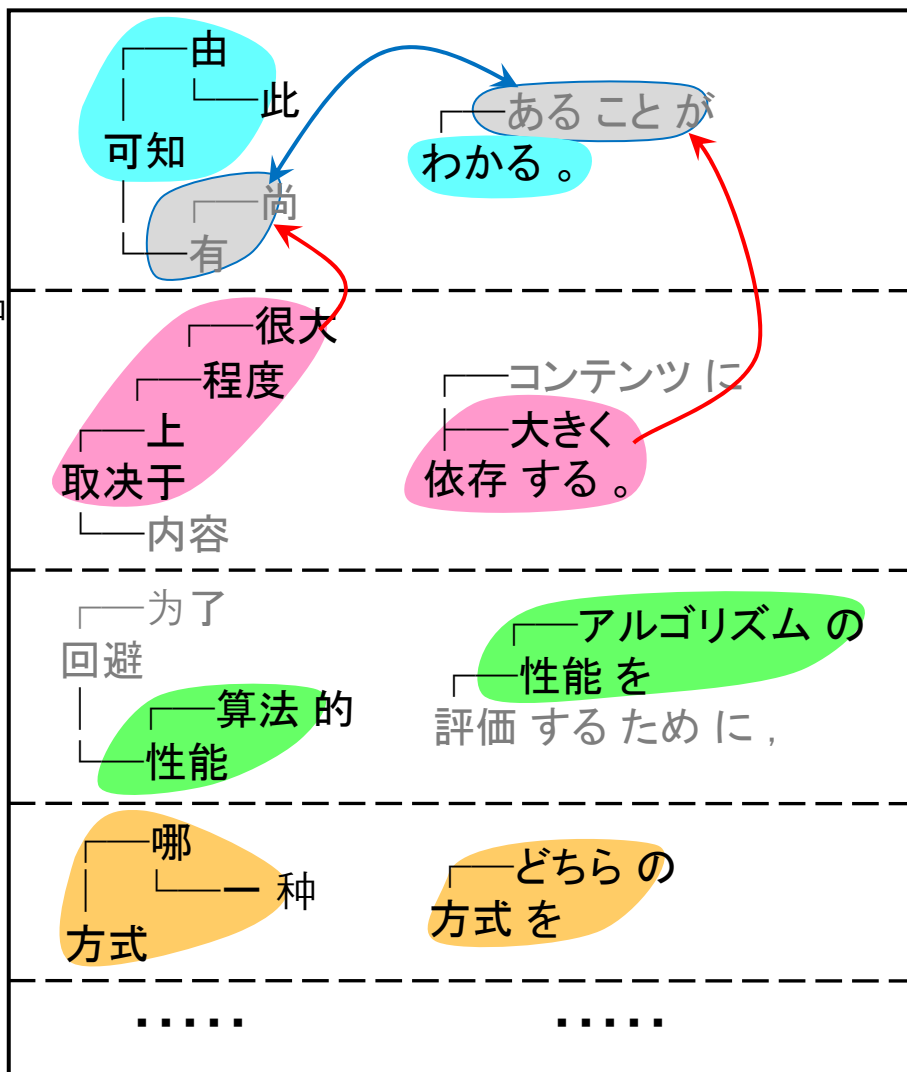# Collaboration between System development and Dictionary development

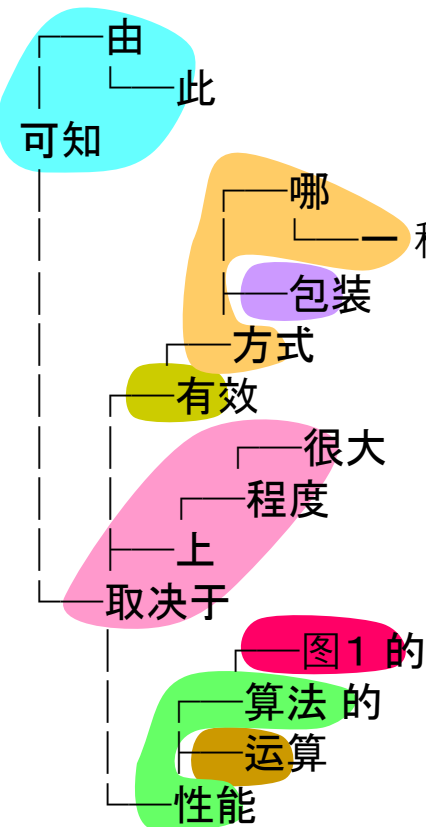**Machine Translation System（Example-based MT）**

Analyzer （NICT）	Translation Engine （Prof. Kurohashi, Kyoto U.） （NICT）	Basic Dictionary （NICT）

Fundamental Technologies

・Disambiguation
・Morphological Analysis
・Chunking
・Parsing
・Case Analysis
・Alignment

Equivalent Selection
（Prof. Kaji, Shizuoka U.）
（Word Sense Network）

Extraction of Semantic Structure
（Prof. Nakagawa, U. Tokyo）

Compilation of
Technical Term Dictionary
（Prof. Tsujii, U. Tokyo）
（JST: Japan Science and Technology Agency）

Scientific/
Technological
Paper
Corpus
（JST）

Shared Software

Resources

Fundamental Resources

Example-based translation with deep consideration to syntactic structure

入力文：由此可知哪一种包装方式有效很大程度上取决于图1的算法的运算性能

用例データベース

入力依存構造木

| ─由
| └─此
可知

| ─哪
| └─一 种
| └─包装
─方式
─有効
| ─很大
| ─程度
| └─上
─取决于
| ─图1 的
| ─算法 的
| ─运算
└─性能

─由
└─此
可知
└─尚
有

わかる 。

─あること が

─很大
─程度
─上
取决于
└─内容

コンテンツ に
─大きく
依存 する 。

─为 了
回避
| ─算法 的
└─性能

─アルゴリズム の
─性能 を
評価 する ため に，

─哪
└─一 种
方式

─どちら の
方式 を

・・・・・

・・・・・

出力依存構造木

─どちら の
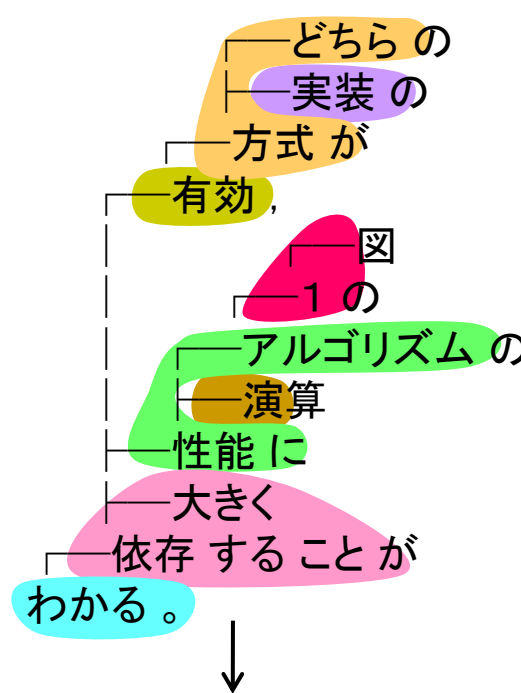─実装 の
─方式 が
─有効，
| ─図
─1 の
─アルゴリズム の
─演算
─性能 に
─大きく
─依存 する こと が
わかる 。

↓

出力文：
どちらの実装の方式が
有効，図1のアルゴリズ
ムの演算性能に大きく
依存することがわかる。

Mission Statement

- Goal for the 3rd year

  – Evaluate the Japanese-Chinese machine translation prototype system for specific target domains.

- Goal for the 5th year

  – Improve the Chinese analysis performance, and complete demonstration experiments on the Japanese-Chinese and Chinese-Japanese machine translation prototype system

Performance

	Training Corpus	Parsing	EBMT		RBMT		SMT（Moses）	
			BLEU4	Adequacy (3&more)	BLEU4	Adequacy (3&more)	BLEU4	Adequacy (3&more)
J→E	1million	90%（E）	18.4%	70.4%	11.6%	61.2%	20.9%	57.0%
E→J		90%（J）	21.9%	—	10.3%	—	21.4%	—
J→C	0.6 million	73%（C）	11.2%	—	8.9%	—	18.1%	—
C→J		90%（J）	11.4%	—	6.9%	—	22.1%	—

↳ Target：80%

Resources and Analyzers

- Parallel corpus
  - 1 million sentence pairs
- Dictionary
  - Technical term dictionary
  - EDR Japanese-Chinese dictionary
- Chinese analyzer
  - Combination of methods
  - Use of huge raw corpus
  - Morphological analysis: 94%
  - Syntactic analysis: 89%

Application

- Games of the XXIX Olympiad Beijing 2008
  - Fluency, Adequacy: more than 60%
  - Speed

- Thailand-Japan distance education
  - Thailand
    - Thai Computational Linguistics Laboratory of NICT
    - High school in Mae Hong Son in Thailand
  - Japan
    - Keihanna Laboratories, Kyoto
    - High school in Uji, Kyoto

Machine Translation for Official Multilingual Service at Beijing Olympics

NICT
www.nict.go.jp

Beijing Tourist Information Service System
Using Multilingual Tourist Information
Database (at 2008 JOC Japan House)

NICT

Beijing Organizing
Committee of Olympic
Games (BOCOG)

Olympic spectator
service information

Translator
(110)

CAPINFO (Multilingual Information Database)

Multilingual Machine
Translation System

Multilingual Database
(Olympics and Beijing)

Translation
Aid

Translation（E & F）

Web
(SOHO)

Mobile Phone
（China Mobile）

Information Booth
5-60 booths with
English translator,
hand language translator
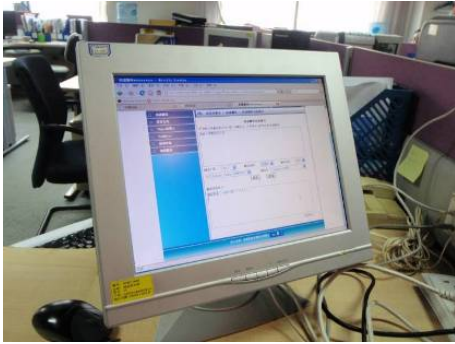and Beijing guide.

**Multilingual
Guide via
Telephone**
44 languages

Beijing Olympics Official Multilingual Spectator Service System

Tourist Information Services at Beijing Olympics

Machine Translation System for spectator service information of Beijing Olympics
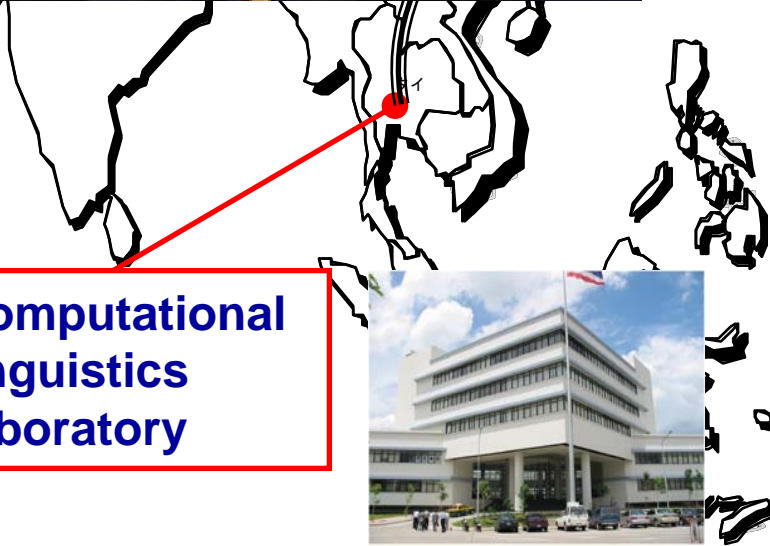

Tourist Service at a hotel in Beijing

Machine Translation System used in Information Booths

Thai Computational Linguistics Laboratory

Keihanna Research Labotatories

Thai Computational Linguistics Laboratory

Linguistic Resources by NICT

- Corpus
  - NICT Multilingual Corpus (CJE)
  - Large-scale JC Parallel Corpus
  - News Paper Article Parallel Corpus (JE)
  - Patent Parallel Corpus (JE)
  - Corpus of Spontaneous Japanese
  - NICT JLE Corpus

- Lexicon
  - EDR dictionary
    - Word dictionary: Japanese and English
    - Bilingual dictianary: J-E, E-J and J-C
    - Concept dictionary
    - Co-occurrence dictionary: J and E
  - Japanese WordNet
    - http://nlpwww.nict.go.jp/wn-ja/index.en.html
    - Asian Wordnet

Natural Language Processing Portal Site

[**English** | Japanese]

by Knowledge Information Processing Technologies Committee
JEITA(Japan Electronics and Information Technology Industries Association)
| About this site | Linking to this site | nlp_portal at nlp.kuee.kyoto-u.ac.jp | 1 5 5 3 0

Conferences (Domestic and International)

To see the Japanse version, click here.

Product News (Links to news release pagess of HLT related products in Japan)

To see the Japanse version, click here.

Related Topics

- Language Resources
  - Catalogue of LRs in Asia (in Japanese)
  - Catalogue of LRs in Japan (in Japanese) (in English)
- Projects, research organizations, academic associations
  - Language Resources initiatives around the world (in Japanese) (in English)
  - Related academic associations, research organizations in Japan, etc. (in Japanese)
- Glossary
- Natural Language Processing portal sites around the world
  - Language Technology World
  - HLT Central

Windows Media | Windows | Picasa ウェブ アルバム … | 5月9日（土）の番組表 …

その他のブックマーク

SHACHI
LANGUAGE RESOURCES METADATA DB

▶Home ▶Sitemap ▶Contact ▶Japanese

search...

▶ Language Resources Catalog

Catalog List

- Catalog List
- Catalog List(Asia Version)

Catalog Search

- Catalog Search
- Catalog Search(Asia Version)

SHACHI

Home

About SHACHI

Lanuguage Resources Catalog

Publications

News

Information

Contact

Sitemap

Language Resource Search

SHACHI LANGUAGE RESOURCE SEARCH for Global

Medical Examination Support System

- There are around 40,000 Japanese residents in Thailand, and more than one million Japanese tourists and business travelers visit Thailand every year.

- As a result, serious language problems arise in various situations, especially in hospitals.
  - When Japanese workers in Thailand, who generally cannot speak Thai, want to be asked detailed questions in Japanese about their health by a local doctor, even in Bangkok there are few doctors who can speak Japanese.

- We wish to develop a medical examination support system using NLP technology which enables local doctors and foreign patients to communicate using their own mother tongue.

- The basic idea and prototype have already been developed as the MedSLT System (http://www.issco.unige.ch/en/research/projects/medslt/)
- Multilingual spoken language translation system tailored for various domains of medicine.
  - Currently the system supports speech recognition and translation for the headache, chest pain and abdominal pain domains.
  - Supported languages are English, French, Japanese, Spanish, Catalan and Arabic.

- We will use a trilingual (J-T-E) list of illustrative sentences in the medical domain as a resource for this R&D,
- develop a Thai system by compiling Thai grammar and a dictionary, and
- extend the system to wider domains.

- Our plan includes on-the-spot inspection of the system at hospitals in Thailand.

Business Translation

- Manuals
  - Manuals are revised frequently and there may be similar manuals for similar products, with some parts of the manuals being identical.
  - Intuitively, manuals for similar products, or manuals for different versions of the same product, are likely to resemble each other.
  - Therefore, even with only a small amount of training data, a corpus-based MT system can output useful translations.

- Aviation Manuals
  - Aircraft manufacturers produce series of aircraft, but the manuals for each series are written independently.
  - Airline companies whose official language is other than English must first write aviation manuals in their own language, by editing contents based on procedures defined in their operation policy and the optional systems they select, for each type of aircraft in their fleet, and then translate them into English for overseas use.

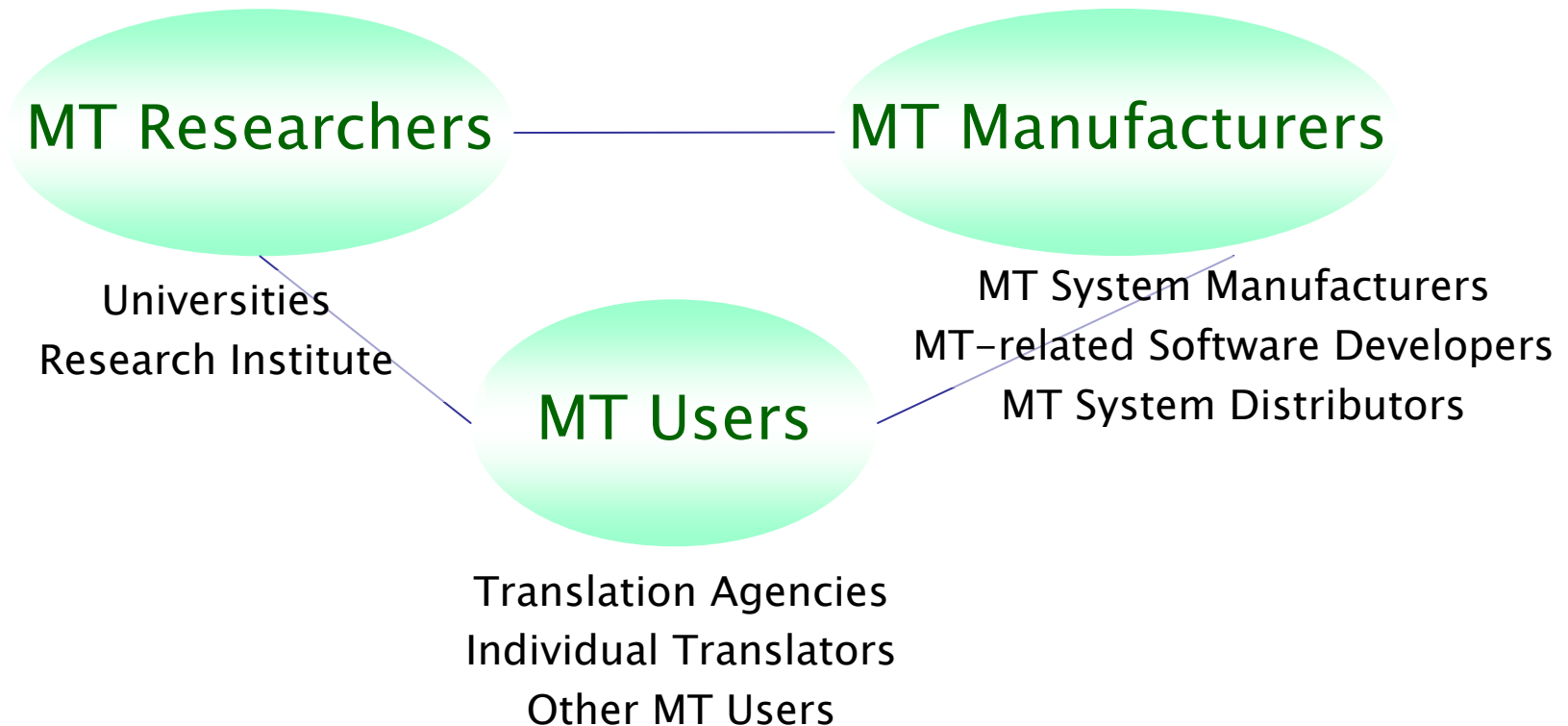- Although manuals are usually created separately for each aircraft, they may be very similar in style and contents.
- Furthermore, a series of aircraft sold by an aircraft manufacturer is likely to be similar in operation, and so the manuals will also be similar.
- However, airline companies have traditionally translated each manual from scratch.
- Machine translation technology could be usefully tailored to such companies.

Summary of Experiments

- Generally speaking, corpus-based MT requires huge amounts of text data, e.g. 1 million parallel sentences.

- However, though the numbers of sentences used in our experiments were rather small, the case-insensitive BLEU scores ranged from 45.6 to 60.3.

- These scores are very high given the small size of the training data, and so SMT is very promising for this task.

Asia-Pacific Association of Machine Translation（AAMT)

- International Association of Machine Translation

MT Researchers ——————— MT Manufacturers

Universities
Research Institute

MT Users

MT System Manufacturers
MT-related Software Developers
MT System Distributors

Translation Agencies
Individual Translators
Other MT Users

Activities

- Committee for Seeking Future Direction of MT
- AAMT/Japio Patent Translation Workshop
- AAMT Journal
- AAMT Nagao Award
  - Given in recognition of individuals or groups who have contributed to R&D leading to or facilitating commercialization of MT systems
- China Chapter and Korea Chapter

Asia-Pacific Association for Machine Translation

AAMT

Korea
AAMT

The association is comprised of three entities: researchers, manufacturers, and users of machine translation systems. The association endeavors to develop machine translation technologies to expand the scope of effective global communications. For this purpose the association is engaged in machine translation system development, improvement, education, and publicity.

Japan

Select language:　**English**　Japanese　Chinese　Korean　Thai

- Message from the President
- Organization
- Publication –AAMT Journal
- Standardization
- New Membership
- Contact Us
- MT Systems
  –MT Software –MT Services –TM tools

- AAMT Nagao Award
- Activities
- Supporting Organizations
- Related Organizations
- Articles

Events

Regional Information

Japan	China	Korea
Thailand	Malaysia	Australia
Myanmar	Laos	Hong Kong
Taiwan	Vietnam	India
Singapore	Indonesia	

TOPICS details

08 May　　New : MT Summit XII 3rd Workshop on Patent Translation
13 Apl.　　update : "MT Systems", "AAMT Journal", "Supporting Organizations", "Organization"

China

Malaysia

MT System

AAMT
Asia-Pacific Association for
Machine Translation

| MT Systems | MT Services | TM tools |

MT Software

- In this part, we list MT softwares which are sold in Japan or translate from/to Asian Languages.
- We use abbreviated name as software names.
- **AAMT is NOT responsible for any damage caused by use of the following sites or software.**
- No fixed order is observed.

Last update : 13 Apl. 2009

Company Name		Languages			OS		
–	Product Name	EJ	JE	others	W: Windows M: Mac OS		
● Toshiba Solutions (http://hon-yaku.toshiba-sol.co.jp/)							
–	The HONYAKU Premium	EJ	JE	–	W	–	–
–	The HONYAKU Business	EJ	JE	–	W	–	–
–	The HONYAKU Server	EJ	JE	C	W	–	UNIX
● TOSHIBA (http://www.toshiba.co.jp/index.htm)							
	LalaVoice HONYAKU		JE		W		
● NEC (http://www.nec.co.jp/middle/meshplus/)							
–	CROSSROAD for Enterprise	EJ	JE	CK	W	–	Linux
–	CROSSROAD	EJ	JE	–	W	–	–
–	Tabitsu	EJ	JE	–	W	–	–
–	Eibun Meibun Making	–	JE	–	W	–	–
–	PIVOT	EJ	JE	–	–	–	UNIX

sysmeta (http://www.transemall.co.kr/shop/shopdetail.html?brandcode=001008000012&search=&sort=price)

–	TransWiz	–	–	JK	W	–	–

Kingsoft (http://www.iciba.com/)

–	Kingsoft FastAIT	EJ	JE	JCK	W	–	–

Transtar (http://www.transtar.com.cn/)

–	Transtar Parsonal	–	–	EJC	W	–	–
–	Transtar Enterplize	–	–	EJC	W	–	–
–	Transtar C/J	–	–	CJ	W	–	–

Huajian (http://www.hjtek.com/)

–	Huajian Translation	–	–	EC	W	–	–

Inventec (Shanghai) Co., Ltd. (http://www.dreye.com.cn/)

–	Dr. eye	–	–	EC	–	–	Pocket PC

Otek International Inc. (http://www.ippc.biz/softasp/detail_allproducts.asp?index=02321)

–	TransPen	–	–	ECJ	–	–	–
–	TransWhiz Pro	–	–	ECJ	W	–	–
–	TransWhiz standard	–	–	ECJ	W	–	–
–	Secretary beats	–	–	ECJ	W	–	–

Kebutuhan Sistem (http://www.cdpenerjemah.cjb.net/)

–	TransTools	–	–	EIn	W	–	–

Axel Blume (http://www.ablume.com/)

–	AbcThai	–	–	EGT	W	–	–

Padideh Co. (http://www.irtik.com/archives/?id=1883)

–	Padideh Translator	–	–	E,Fa	W	–	–

Mabnasoft (http://mabnasoft.com/english/parstrans/index.htm)

–	PTRAN	–	–	E,Pe	W	–	–

Babylon (http://www.babylon.com/)

–	Babylon	EJ	JE	EFGIPSDC	W	–	–

Thank you!