Computational Linguistics (2001-2002 fall semester)

Computer Science Division, EECS, KAIST, Deajeon, Korea

Prof. Feng Zhiwei

Ch1 Development of Computational linguistics

1.1 Nature of computational linguistics
-- linguistic problem
-- linguistics formalism
-- computational formalism
-- computational implementation

1.2 Rudimentary stage of CL
-- Legent on Babel Tower
-- Digit-based dictionary (Descartes, Leibniz, Cave Beck, Kircher, Becher)
-- universal language: Interlingua
(Wilkins: <An Essay towards a Real Character and Philosophical Lan guage, 1668)>
-- Zifferngrammatik : term 'ein mechanisches Uebersetzen' (Rieger, <Univers al Language, Couturat, Leau, 1903>)
-- mechanical brain (Artsouni, 1933, Fr)
-- electromechanical machine (Troyansky, Ru) 3 steps

-- Weaver (Rockefeller Foundation) & Booth (British crystallographer)

In a memorandum written by Weaver in 1949 to the Rockerfeller F oundation which included the following two sentences.

> ``I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained i n the text."
> Universal Language (Interlingua):" The contents of source languag e and target language are the same".

-- This memorandum sparked a significant amount of interest and research, and by the early 1950s there was a large number of research groups working i n Europe and the USA, representing a significant financial investment (equivale nt to around1 000 000 US Dollars).

-- But, despite some success, and the fact that many research questions w

ere raised that remain important to this day, there was widespread disappointment on the part of funding authorities at the return on investment that this represented, and doubts about the possibility of automating translation in general, or at least in the current state of knowledge.

-- ALPAC Report (Automatic language Processing Advisory Committee)

. No reason to support to MT: there was no shortage of human translators, and that there was no immediate prospect of MT producing useful translation of general scientific texts. This report led to the virtual end of Government funding in the USA. Worse, it led to a general loss of morale in the field, as early hopes were perceived to be groundless.

. Semantic barrier

-- three systems in regular, if not extensive, use (one at the Wright Patterson USAF base, one at the Oak Ridge Laboratory of the US Atomic Energy Commission , and one the EURATOM Centre at Ispra in Italy).

-- The theoretical doubts were voiced most clearly by the philosopher Bar-Hillel in a 1959 report, where he argued that fully automatic, high quality , MT (FAHQMT) was impossible, not just at present, but in principle. The problem he raised was that of finding the right translation for pen in a context like the following:

> John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

The argument was

(i) here pen could only have the interpretation play-pen, not th

e alternative writing instrument interpretation,

(ii) this could be critical in deciding the correct translation for p en,

(iii) discovering this depends on general knowledge about the wo rld, and

(iv) there could be no way of building such knowledge into a co mputer.

Some of these points are well taken. Perhaps FAHQMT is impossib le. But this does not mean that any form of MT is impossible or usele ss.

-- the research should focus on more fundamental issues in the processing and understanding of human languages.

-- the useful MT is neither science fiction, nor merely a topic fo r scientific speculation. It is a daily reality in some places, and for som e purposes.

-- Term "Computational Linguistics" was proposed (ALPAC report, D. G. Hays, 1964)

-- Importance of linguistic study:

.The examples of automatic mis-translatioms: English-> Russian -> English

The spirit is willing, but the flesh is weak
=> The whiskey is alright, but the meat is rotten

Out of sight, out of mind
=> Invisible idiot

. The syntactic ambiguity in the source language:

Julia flew and crashed the air plane
Julia (flew and crashed the air plane)
(Julia flew) and (crashed the air plane)

Susan observed the yacht with telescope
Susan observed the man with a beard

Old men and women
(Old men) and women
Old (men and women)

dangerous cyanide and chlorine fumes
(dangerous cyanide) and (chlorine fumes)
dangerous (cyanide and chlorine) fumes

. lexical differences between source language and target language

The men killed the women. Three days later they were caught. (they = men)

The men killed the women. Three days later they were buried. (they = women)

English	German	French
know	wissen	savoir
	kennen	connaitre

The watch included two new recruits that night (watch = guard or clock)

. Syntactic differences between source and target

- German:

Auf dem Hof sahen wir einen kleinen Jungen, der einem Ferkel na chlief.

Dem Jungen folgte ein grosser Hund.

-- English:

In the yard we saw a small boy running after a piglet.

A large dog followed the boy.

The boy was followed by a large dog.

Is means that the computational linguistics is absolutely necessary.

1.3 Growing stage of CL

Natural Language Understanding (NLU):

First Generation of NLU)

-- Special Format system:

. SAD-SAM (R. Lindsay, 1963)

kinship relation database

. STUDENT (D. Bobrow, 1968)

- to resolve the application questions in mathematics exercise:

eg. If the number of customers Tom gets is twice the square of 20 per cent of the number of advertisements he runs, and the number of advertisements he runs is 45, what is the number of customers Tom gets?

Answer: "The number of customers Tom gets is 162"

- some basic patterns:
(what are * and *)
(what is *)
(How many *1 is *)
(How many * do * have)
(How many * does * have)
(find *)
(* is multiplied by *)
(* is divised by *)
(* (*1/verb) * as many * as (*1/verb) *)

- REMEMBER Module (common knowledge):
feet is the plural of foot
one half always means 0.5
Succesful candidates sometimes means students who passed the ad missions test
distance equals speed times time
one foot equals 12 inches

. BASEBALL (B. Green, 60s)

Q: How many games did the Yankees play in July?
format expression:

TEAM	=YANKEES
MONTH	=JULY
GAMES	=?

-- Text-based system

. PROTOSYNTHEX-I system (R.F. Simmons, J.F. Burger, R.E. Long, 1966)

-- Finite logical system

. SIR (Semantic Information Research, B. Raphael, 1968)

28 pattern:

* is *

* is part of *

there are * on *

Is * * ?

How many * does * have?

What is the * of * ?

* = noun or noun phrase

Man-Machine dialogue:

Person: Every boy is a person.

Machine: I understand.

P: A finger is part of a hand.

M: I understand.

P: There are two hands on each person.

M: I understand.

P: How many finger does John have?

M: The above sentence is ambiguous but I assume (has) means (h as a parts). I don't know whether finger is part of John.

P: John is a boy.

M: I understand.

P: How many fingers does John have?

M: The above sentence is ambiguous but I assume (has) means (h as as parts). How many fingers per hand?

P: Every hand has 5 fingers.

M: The above sentence is ambiguous but I assume (has) means (h as as parts). The answer is 10.

Above dialogue needs following common knowledge:

1. Every boy is a person.
2. The finger is a part of hand.
3. There are two hands on each person.
4. Every hand has 5 fingers.

. DEDUCOM (J.R. Slagle, 1965)

. DEACON (F.B. Thompson, 1966)

. CONVERSE (C. Kelleg, 1968)

-- General deduction system

. Some girls are pretty

Every girl is pretty

.QA2 and QA3 systems

Second generation system

-- LUNAR (W. Woods, 1972): to help geologist to study the rock samples from Appolo-11.

-- SHRDLU (T. Winograd, 1972)
block world:
the pyramide is on the table can be expressed:
ON (PYRAMIDE TABLE)
MICRO-PLANNER:

```
THGOAL(ON    ?X    ?Y)
    (OR(ON-TOP     ?X    ?Y)
        (AND(CLEAR-TOP ?X)
                (CLEAR-TOP  ?Y)
                (PUT-ON  ?X     ?Y))))
```

-- MARGIE (Meanning Analysis, Response Generation and Inference on English, R. Schank, 1975): CD (Conceptual Dependency) expression

. Paraphrase:

"John eats the ice cream with a spoon" can be paraphrased as:

John INGESTs the ice cream by TRANSing the ice cream on a spoon to the mouth, by TRANSing the spoon to the ice cream, by GRASPing the spoon, by MOVing his hand to the spoon, by MOVing his hand muscles.

"INGEST, TRANS, GRASP, MOV" are basic actions.

. Inference:

from "John hits Mary", the system can deduce out following sentences:

John was angry with Mary.
Mary might hit John back.
Mary might get hurt.

. Inference + Paraphrase:

"John killed mary by choking her" can be paraphrased as:

John strangle mary.
John choked mary and she died because she was unable to breathe.

-- SAM (Script Applier mechanism, R. Schank & R. Abelson, 1975):

. script:

- Persons: customer, server, casher.
- Things: restaurant, table, menu, food, check, tip, payment.
- Events:

1. Customer goes to restaurant.
2. Customer goes to table.
3. Server brings menu.
4. Customer orders food.
5. Server brings food.
6. Customer eats food.
7. Server brings check.
8. Customer leaves tip for server.
9. Customer gives payment to casher.
10. Customer leaves restaurant.

"John went to restaurant. He sat down. He got mad. He left."
is paraphrased as:

"John was hungry. He decided to go to a restaurant. He went to one. He sat down in a chair. a waiter did not go to the table. John became upset. He decided he was going to leave the restaurant. He left it."

-- PAM (Plan Applier Mechanism, R. Wilinsky, 1978): plan, plan box, goal.

Machine Tramslation

-- MT research became the preserve of groups funded by the Mormon Church, who had an interest in bible translation (the work that was done at Brigham Young University in Provo, Utah ultimately led to the WEIDNER and ALPS systems, two notable early commercial systems)

-- A handful of groups in Canada (notably the TAUM group in Montreal, who developed the METEO system), the USSR (notably the groups led by Mel′cuk, and Apresian) , and Europe (notably the GETA group in Grenoble, probably the single most influential group of this peri

od, and the SUSY group in Saarbrueken).

-- A small fraction of the funding and effort that had been devoted to MT was put into more fundamental research on Computational Linguistics, and Artificial Intelligence, and some of this work took MT as a long term objective, even in the USA

-- the late 1970s that MT research underwent something of a renaissance. There were several signs of this renaissance.

. The Commission of the European Communities (CEC) purchased the English-French version of the SYSTRAN system, a greatly improved descendent of the earliest systems developed at Georgetown University (in Washington, DC), a Russian-English system whose development had continued throughout the lean years after ALPAC , and which had been used by both the USAF and NASA . The CEC also commissioned the development of a French-English version , and Italian-English version.

. At about the same time, there was a rapid expansion of MT activity in Japan, and the CEC also began to set up what was to become the EUROTRA project, building on the work of the GETA and SUSY groups. This was perhaps the largest, and certainly among the most ambitious research and development projects in Natural Language Processing. The aim was to produce a `pre-industrial' MT system of advanced design (what we call a Linguistic Knowledge system) for the E

C languages.

. Also in the late 1970s the Pan American Health Organization (PAHO) began development of a Spanish–English MT system (SPANAM) , the United States Air Force funded work on the METAL system at the Linguistics Research Center, at the University of Texas in Austin, and the results of work at the TAUM group led to the installation of the METEO system. For the most part, the history of the 1980s in MT is the history of these initiatives, and the exploitation of results in neighbouring disciplines.

. As regards the practical and commercial application of MT systems. The systems that were on the market in the late 1970s have had their ups and downs, but for commercial and marketing reasons, rather than scientific or technical reasons, and a number of the research projects which were started in the 1970s and 1980s have led to working, commercially available systems. This should mean that MT is firmly established, both as an area of legitimate research, and a useful application of technology.

. But researching and developing MT systems is a difficult task both technically, and in terms of management, organization and infrastructure, and it is an expensive task, in terms of time, personnel, and money. From a technical point of view, there are still fundamental problems to address.

-- diferenciation of grammar from algorithm (mechanism, Yingve)
-- <Framework for syntactic Transaltion>(V. Yingve, 1957)
. First step: expression of text structure of source language by digital code s
. Second step: transfer from structure of source language to the structure of target language,
. Third step: output of text of target language

-- CETA (B Vauquois)
. six steps of MT:
- Morphological Analysis of source language
- Syntactic Analysis of source language
- Lexical transfer from source language to target language
- Structure transfer from source langue to target language
- Syntactic generation of target language
. Morphological generation of target language

ARIANE - 78:
- ATEF: non-deterministic finite state transducer
- ROBRA: tree to tree transducer
- TRANSF: lexical transfer
- SYGMOR: deterministic tree-string transducer

-- Y.A. Wilks: preference semantics, semantic analysis

-- TAUM-METEO system (Montreal University, 1976)

-- Commercial MT system in Japan:
ATLAS-I, ATLAS-II (Fujitsu)
HICATS (Hitachi)
PIVOT (NEC)
MELTRAN (Mitsubishi)
TAURAS (Toshiba)
BRAVICE PAK 11/73 (BRAVICE International)

-- TITUS -IV (France)
-- SYSTRAN www.systransoft.com

Example 1:
English (Source): I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has be en coded in some strange symbols. All I need to do is strip off the code i n order to retrieve the information contained in the text.
French (Target): J'ai un texte devant moi ce qui est ecrit dans le Russe mais j e vais feindre qu'on lui ecrit vraiment en anglais et qu'il a ete code dans q

uelques symboles etranges. Tout que je dois faire doit decoller le code afi n de rechercher l'information contenue dans le texte.

Example 2:

English(SL): But researching and developing MT systems is a difficult t ask both technically, and in terms of management, organization and infrastructure, and it is an expensive task, in terms of time, person nel, and money. From a technical point of view, there are still funda mental problems to address.

French(TL): Mais recherchant et se developpants la TA des systemes est une tache difficile toutes les deux techniquement, et en termes de gestion, organisation et infrastructure, et c'est une tache chere, en termes de temps, personnel, et argent. D'un point de vue techn ique, il restent des problemes fondamentaux a adresser.

-- LOGOS -III
-- WEIDNER
-- METAL (Siemens and Texas University)
-- EUROTRA
-- Mu system
-- DLT
-- Speech Translation system
.Speech Trans (CMU, 1989)
. JANUS (CMU, 1992)
. SL-TRANS (ART, Japan, 1989)
-- Verbmobil project (1993-2001, Germany)
-- C-STAR (Consortium for Speech Translation Advanced Research, 1991

-- New theory of computational linguistics:
Lexcical Functional Grammar (LFG, R.M. Kaplan and J. Bresnan, 198 3)
Unification Grammar (UG, Martin Kay, 1983)

GPSG (Generalized Phrase Structure Grammarm G. Gazdar, E. Klein, I. Sag, G. Pullum, 1985)

Head-driving Phrase structure Grammar (HPSG, C. Pollard, 1985)

1.4 Prosperity stage of CL

COLING'90 (Helsinki, 1990, theory, method and tools for large-scale authen tic text processing)

TMI-92 (Conference for theory and method of machine translation, Montr eal, 1992, June): rationalism and empiricism in MT

New epoch of Machine Translation (J. Hutchins, MT Summit IV, 1993-Jul y, Japan):: corpus-based approach

-- Noisy channel theory

Source language -> noisy channel --> Target language

-- Hidden Markov Model (HMM)

-- Example-based MT (Makoto Nagao)

. MBT1 MBT2 (Kyoto University, Japan):
decomposition, transfer, composition

. PANGLOSS (Multi-engine MT, CMU)

. ETOC (ATR, Japan)

1.5 Practical tasks of computational linguistics

-- Indexing and retrieval in textual databases
-- Machine Translation (MT)
-- Natural Language Understanding (NLU)
-- Automatic text production
-- Automatic text checking
-- Automatic content analysis
-- Automatic tutoring (electronic text book)
-- Automatic dialog and information system

1.6 Difference between traditional linguistics and computational linguistics

-- Difference between natural language and artificial language

1. full ambiguity <--> without ambiguity
2. complax structure <--> relatively simple structure
3. description of meaning is difficult <--> meaning can be controlled by

people

4. multi to multi correspondence between meaning and structure <--> on e to one correspondence between meaning and structure

-- Difference between computational linguistics and traditional linguistics

1. general language phenomena <--> special linguistic phenomena
2. more practical <--> more theoretical
3. first analysis then understanding <--> first understanding then analysis
4. cross-branch research of language <--> pure research of language