CSE6339 3.0 Introduction to Computational Linguistics
Instructor: Nick Cercone – 3050 LAS – nick@cse.yorku.ca
Mondays, Wednesdays 10:00-11:20 – North Ross 836A
Winter Semester, 2014

Introduction to the course

Some NLP Applications

◂ Magic?

Some NLP Applications

Natural language interfaces to databases

Natural language interfaces to search engines

Generate and Repair Machine Translation (GRMT)

Natural Language Interfaces - SystemX

In the 1990's Simon Fraser University researchers were engaged in a long-term project entitled Assessing Information with Ordinary Language which was realized in several versions of SystemX. Initial SystemX NL interface prototypes were modularly designed utilizing proven technologies, e.g., augmented transition network grammars. SystemX served as an umbrella project for new ideas and technologies, as a testbed for various techniques espoused by students and for experimenting with incompletely specified theories, HPSG's.

Early SystemX

SystemX handled quantifiers, which were problematic because SQL was not able to express queries with quantification in an obvious fashion. Thus, Has every cmpt major taken at least 3 math courses?, combined the problem of quantification with that of data organization and calculation.

```
Ctr    select a.student#
       from student a
          where a.major = 'cmpt'
             and not exists (select e.student#
                from course b, class c, offering d,
                enroll e where b.dept = 'math'
                          and b.cname = d.cname
                          and c.offer# = d.offer#
                          and a.student# = e.student#
                          and c.class# = e.class#
                          and 883 > d.semester
                group by e.student#
                having 3 <= count(distinct d.cname))
```

Translate as "The answer is *no* if there is a student who is a cmpt major and it is not the case that the student is a member of the set of students who have taken at least 3 math courses."

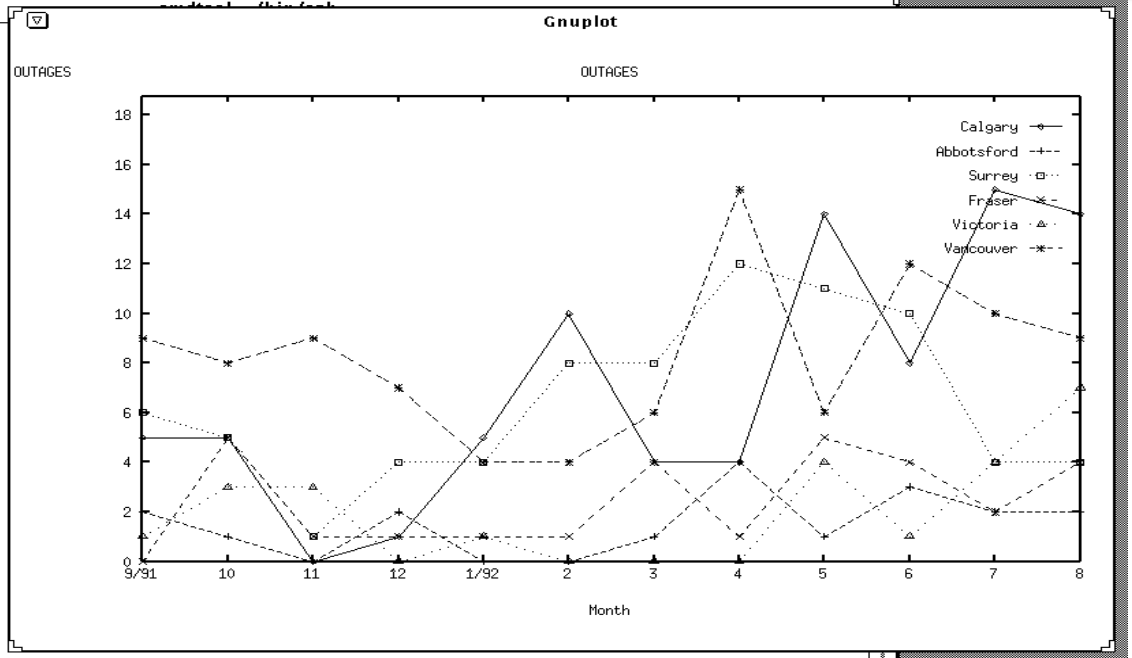
At Roger's Cablesystems Ltd., the vice president for customer service enters the following into his computer, Give me the Western region outage log for June. Within seconds SystemX presents him with a neatly formatted table (or graph) of the data retrieved from Rogers' relational database. He could have said, What's the outage log for the Western region for June?, or Tell me the June regional outage log for the West. or Find the Western outages for June., etc.

SystemX can determine that, whichever phrase he uses, he means the same thing. Such flexibility in parsing is nontrivial.

An Example

SystemX is able to display responses to requests for trends in statistical data graphically. The user has the choice of inputting his trend request using English, using menus (in the case of "canned" trends) or using a combination of English and menu responses. "Canned" trends display data that is predictably desired on a reasonably frequent basis, accessed for a minimum of keystrokes. "Canned" trends are those available through the first eight menu items (below).

```
1 New query
2 Display SQL
3 Save Last Response
4 Print Saved Reponses
5 Display a Trend
6 Print Last Trend
7 Make a Comment
8 Stop
WHAT NEXT? > 5
Display a Trend


Trend Menu

1 # of C/S Representatives
2 Service Call Ratios
3 System Reliability
4 Subscribers per Employee
5 Subscribers per Km of Plant
6 Maintenance Calls
7 Repair Calls per TSR FTE
8 Repair Calls per TSR Hours
9 Specify an Ad Hoc Trend
10 Cancel Trend Request
WHICH TREND? > 9
Specify an Ad Hoc Trend

Enter phrase describing data to be graphed.
(Return to cancel): the unscheduled outages

Parsing Completed.

Do you wish the current, 12 month trend for the Western divisions? (Y or N): y
```



Natural language interfaces to internet search engines - NLAISE and MATISE

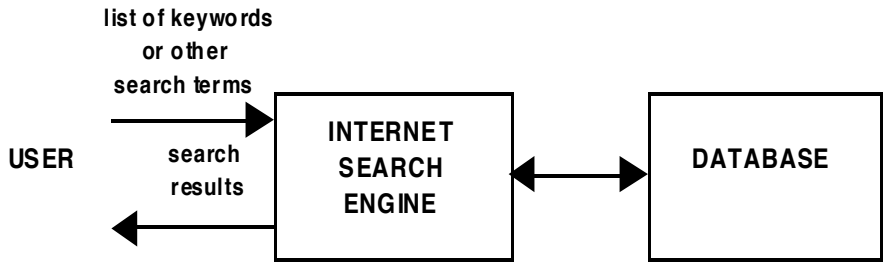
Despite the many search engines available, searching for a relevant site remains difficult. One major reason for this difficulty is that search engines do not analyze queries semantically; in contrast, most search engines perform keyword matching.

How can use of NL semantics improve internet searching?

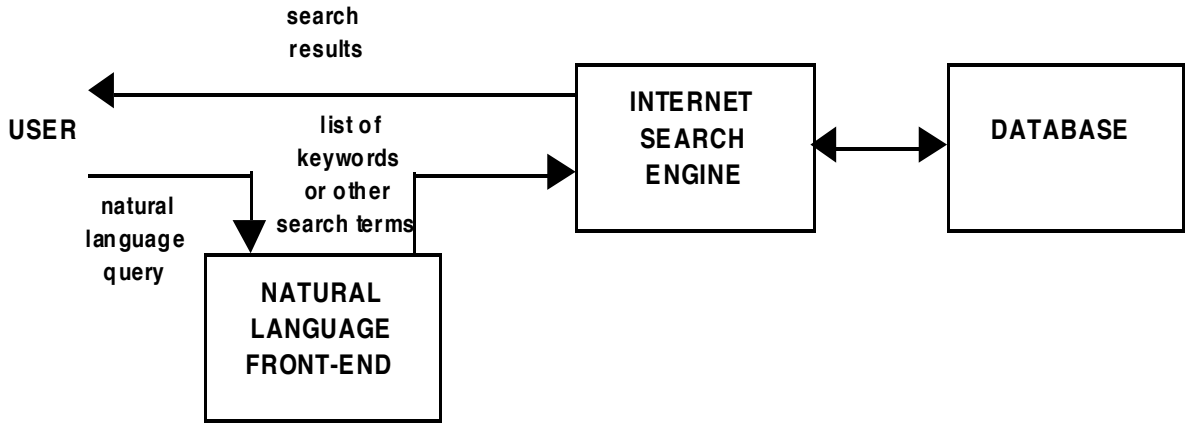
Keywords	Infoseek	Yahoo	Lycos
top, computer science	4,606,219	7	17,374
+top, computer science	78,414	3071	17,374
top, +computer science	76,989	1064	17,374
+top, +computer science	2,132	7	17,374
top, computer science department	4,606,545	2	28,403
+top, computer science department	78,386	3071	28,403
top, +computer science department	76,991	101	28,403
+top, +computer science department	2,131	2	28,403
top, computer science program	4,606,672	1 (RHIT)	23,925
+top, computer science program	78,415	3071	23,925
top, +computer science program	76,991	1(N. Ga.)	23,925
+top, +computer science program	2,132	162	23,925
rank & computer science program	7,570,132	16460	29,205
rank & computer science department	4,604,421	133,690	28,694
rank and computer science department	23,267,722	133,690	28,694
rank | computer science department	4,117	133,690	28,694
rank or computer science department	6,070,374	133,690	28,694

Natural Language	# hits
Which is the best computer science department?	22,497,252
Which is the best computer science program?	22,497,257
Which is the top computer science department?	22,497,276
Which is the top computer science program?	22,497,290
the best computer science department?	22,497,339
the best computer science program?	22,497,350
the top computer science department?	22,497,299
the top computer science program?	22,497,305
best computer science department?	4,931,348
best computer science program?	7,895,146
top computer science department?	4,681,660
top computer science program?	7,653,131
computer science department?	4,667,918
computer science program?	7,638,548
computer science?	4,665,750
computing science?	2,588,408

The figure shows the representation of existing search engines compared with the NL front-ends.

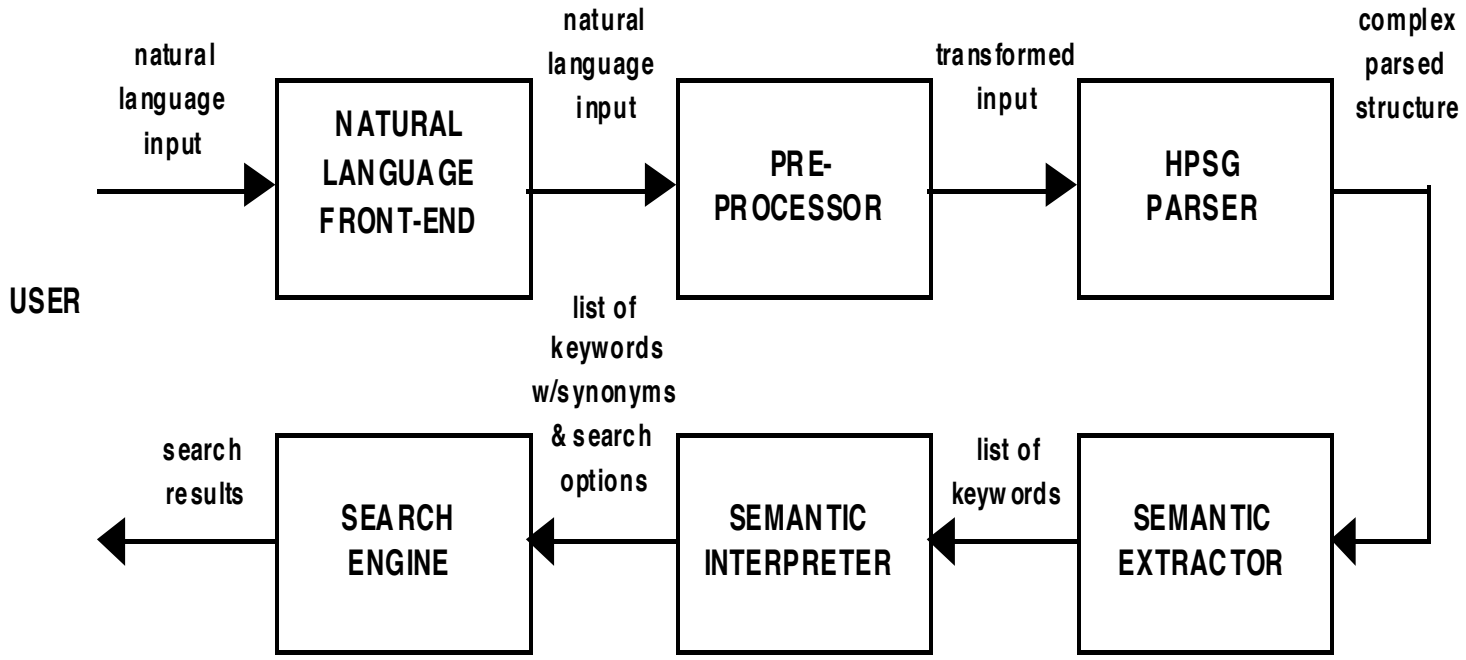


EXISTING SEARCH ENGINES

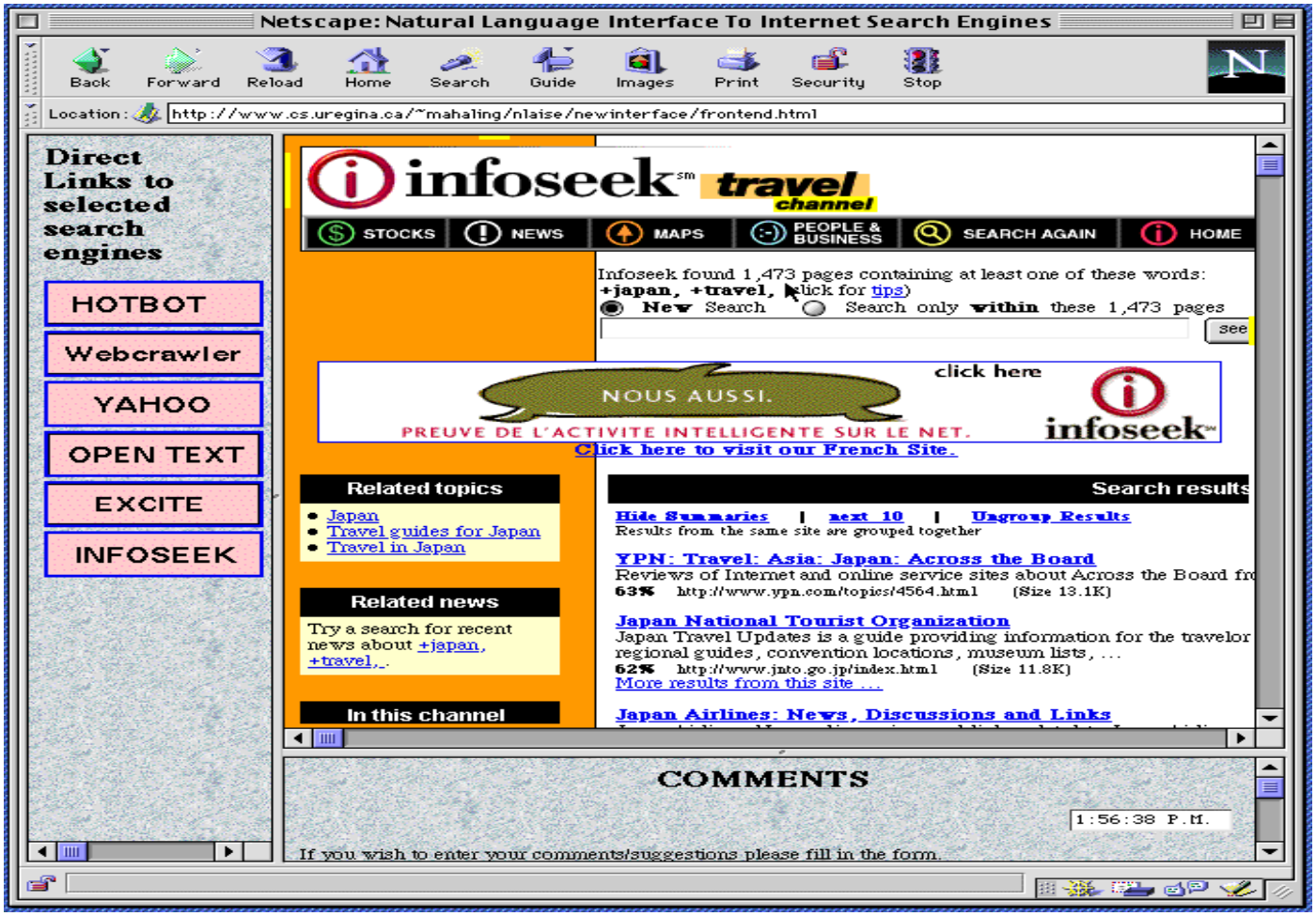


SEARCH ENGINES WITH NATURAL LANGUAGE "FRONT-END"

NLAISE allows users to choose the search engine best suited for their search and enter the query in English. The NL query is analyzed both syntactically and semantically in order to select the most appropriate keywords describing sought information. Keywords are interpreted to provide more meaningful search terms by using keyword synonyms in conjunction with Boolean operators supported by specific search engines.


USER
natural language input
NATURAL LANGUAGE FRONT-END
natural language input
PRE-PROCESSOR
transformed input
HPSG PARSER
complex parsed structure
SEMANTIC EXTRACTOR
list of keywords
SEMANTIC INTERPRETER
list of keywords w/synonyms & search options
SEARCH ENGINE
search results

Shown below is output from asking NLAISE to parse the phrase I want to schedule a trip to Japan and generate appropriate keywords for search engine examination. NLAISE was also requested to use Infoseek. Inspection of the 1,473 web pages returned verified that 80% were relevant. Note the choice of keywords "Japan" and "travel" which indicates the level of sophistication of NLAISE's semantic interpretation of the original input phrase.


Netscape: Natural Language Interface To Internet Search Engines
Back
Forward
Reload
Home
Search
Guide
Images
Print
Security
Stop
Location : http://www.cs.uregina.ca/~mahaling/nlaise/newinterface/frontend.html
Direct Links to selected search engines
HOTBOT
Webcrawler
YAHOO
OPEN TEXT
EXCITE
INFOSEEK
infoseek travel channel
STOCKS
NEWS
MAPS
PEOPLE & BUSINESS
SEARCH AGAIN
HOME
Infoseek found 1,473 pages containing at least one of these words:
+japan, +travel, (click for tips)
New Search
Search only within these 1,473 pages
see
click here
NOUS AUSSI.
PREUVE DE L'ACTIVITE INTELLIGENTE SUR LE NET.
infoseek
Click here to visit our French Site.
Related topics
Japan
Travel guides for Japan
Travel in Japan
Related news
Try a search for recent news about +japan, +travel,..
In this channel
Search results
Hide Summaries | next 10 | Ungroup Results
Results from the same site are grouped together
YPN: Travel: Asia: Japan: Across the Board
Reviews of Internet and online service sites about Across the Board fr
63% http://www.ypn.com/topics/4564.html (Size 13.1K)
Japan National Tourist Organization
Japan Travel Updates is a guide providing information for the travelor
regional guides, convention locations, museum lists, ...
62% http://www.jnto.go.jp/index.html (Size 11.8K)
More results from this site ...
Japan Airlines: News, Discussions and Links
COMMENTS
1:56:38 P.M.
If you wish to enter your comments/suggestions please fill in the form.

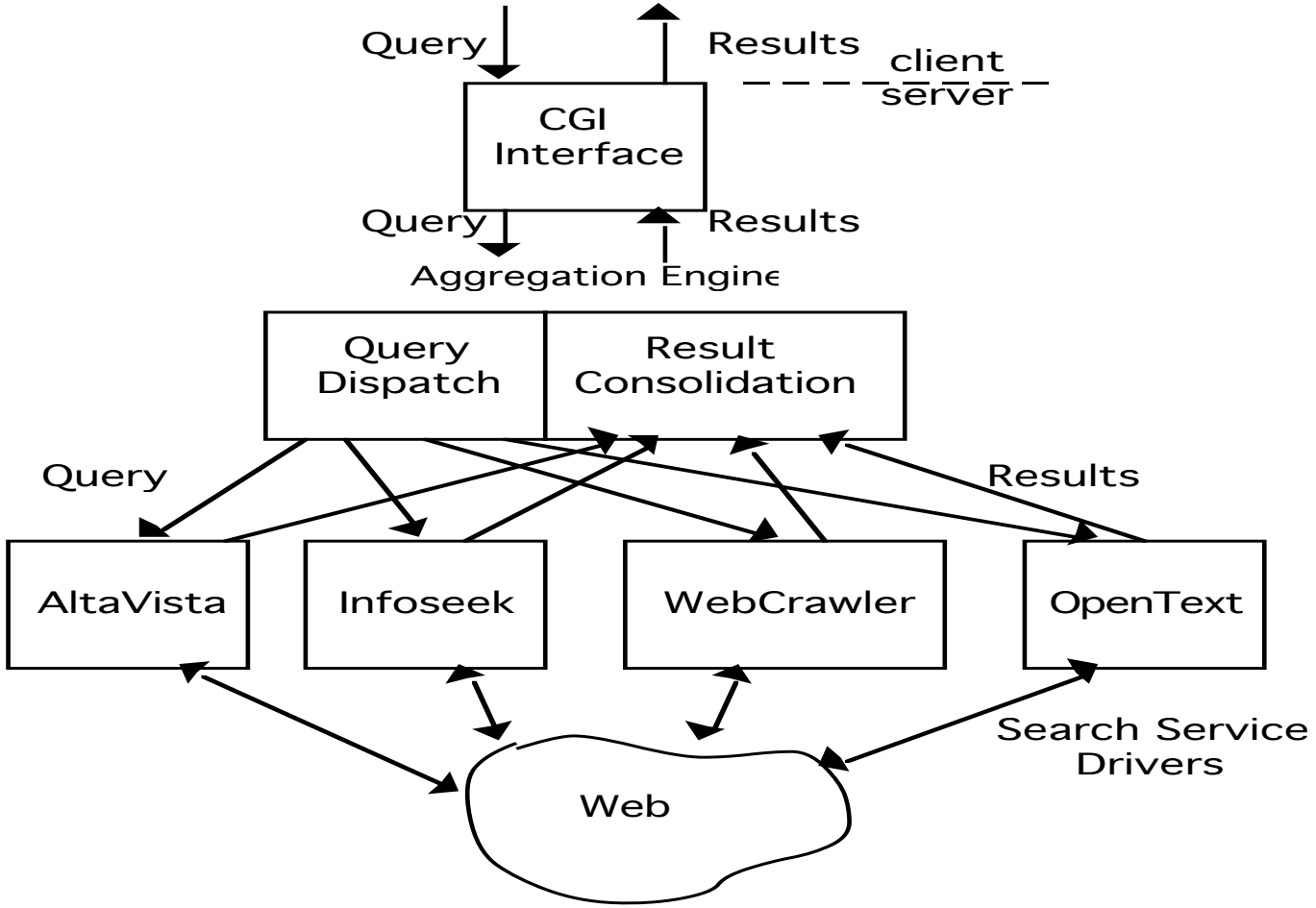
EMATISE extended NLAISE in 3 user-oriented ways:

(1) enhanced semantic interpretation eliminating much ambiguity over multiple domains;

(2) sent out term expanded queries to multiple search engines in parallel, reranked results and returned a single relevant high precision list for the user; and

(3) a higher level of abstraction above conventional search services presented the user with a single, central and natural search interface with which to interact. For example:

Ematise's modular design, depicted in the next slide, consists of a CGI interface, aggregation engine, and search service drivers. The CGI interface passes a user's query option in a logical format, search service neural from Web client, to the meta search engine server. The logical query is passed to the aggregation engine, responsible for concurrently dispatching the query to selected search services, obtaining initial results from each service, eliminating duplicate results, re-ranking and consolidating the results, and finally creating HTML pages from the results, to be properly displayed back at the Web client.

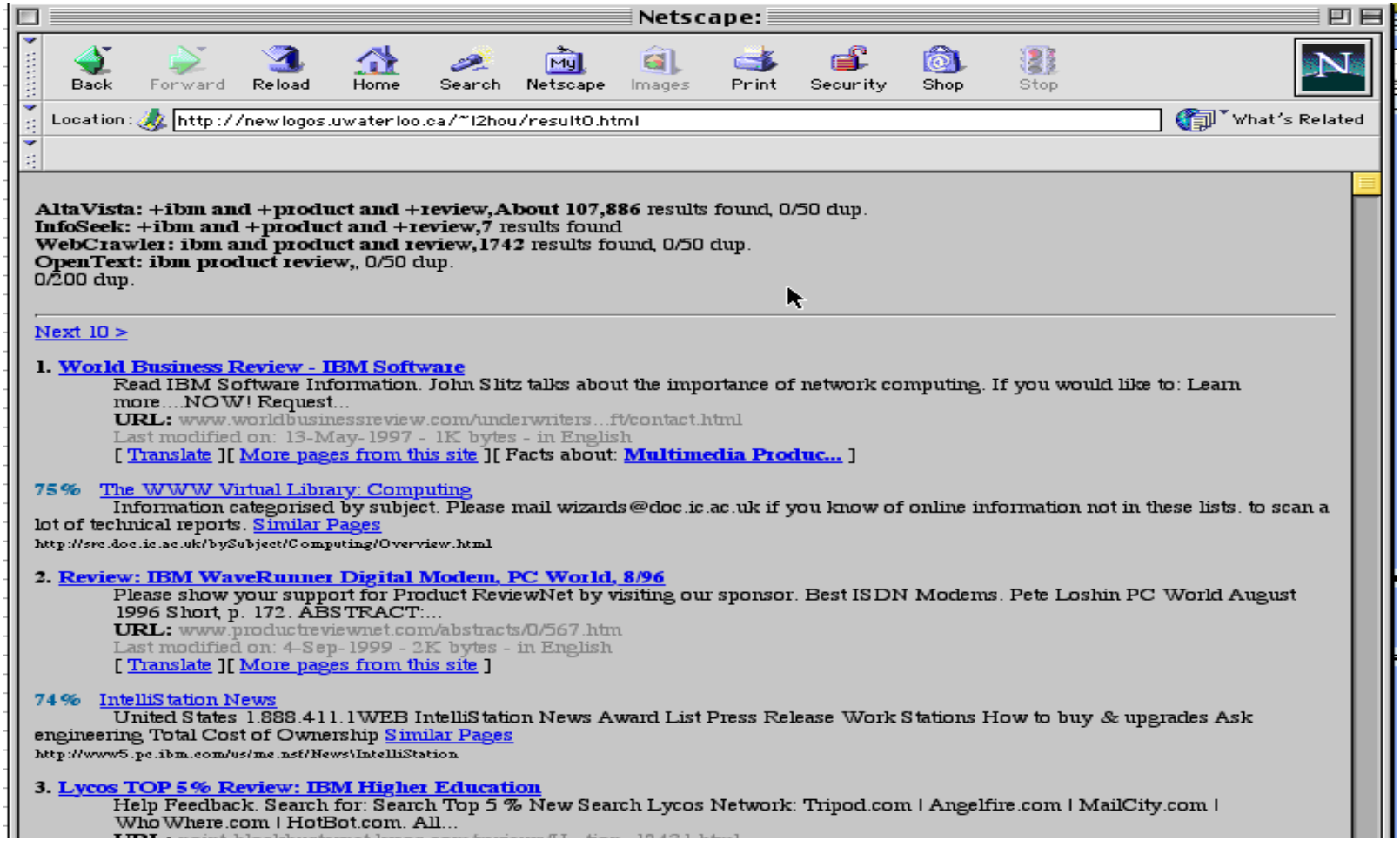

Query
Results
client
server
CGI
Interface
Query
Results
Aggregation Engine
Query
Dispatch
Result
Consolidation
Query
Results
AltaVista
Infoseek
WebCrawler
OpenText
Search Service
Drivers
Web

Ematise provides a layer of abstraction above traditional services and incorporates several desirable features.

As the Web grows and changes, search services become volatile. Interfaces of existing search services change often due to enhancements which impact query input and output format. Also a number of services are retired or replaced. Ematise's modular design, especially the search service driver classes, provides a wrapper around this service specific information, effectively encapsulately them, and allows for services to be added, modified, and removed easily.

Ematise has a meta search engine that does not require large databases or large amounts of memory. Since both the server and client side of the meta search engine are implemented in Java, they are easily portable to different platforms without the extra effort of changing the code.

The figure below shows EMATISE results after a simple translation of the sentence "I want to visit the homepage of IBM product review" into search engine neutral search terms, term expanded by the drivers for particular search engines. Figure 6 illustrates the results of this query after the aggregation engine assembles the results.


Netscape:
Back
Forward
Reload
Home
Search
Netscape
Images
Print
Security
Shop
Stop
Location: http://newlogos.uwaterloo.ca/~l2hou/result0.html
What's Related
AltaVista: +ibm and +product and +review,About 107,886 results found, 0/50 dup.
InfoSeek: +ibm and +product and +review,7 results found
WebCrawler: ibm and product and review,1742 results found, 0/50 dup.
OpenText: ibm product review,, 0/50 dup.
0/200 dup.
Next 10 >
1. World Business Review - IBM Software
Read IBM Software Information. John Slitz talks about the importance of network computing. If you would like to: Learn more....NOW! Request...
URL: www.worldbusinessreview.com/underwriters...ft/contact.html
Last modified on: 13-May-1997 - 1K bytes - in English
[Translate][More pages from this site][Facts about: Multimedia Produc...]
75% The WWW Virtual Library: Computing
Information categorised by subject. Please mail wizards@doc.ic.ac.uk if you know of online information not in these lists. to scan a lot of technical reports. Similar Pages
http://src.doc.ic.ac.uk/bySubject/Computing/Overview.html
2. Review: IBM WaveRunner Digital Modem, PC World, 8/96
Please show your support for Product ReviewNet by visiting our sponsor. Best ISDN Modems. Pete Loshin PC World August 1996 Short, p. 172. ABSTRACT:...
URL: www.productreviewnet.com/abstracts/0/567.htm
Last modified on: 4-Sep-1999 - 2K bytes - in English
[Translate][More pages from this site]
74% IntelliStation News
United States 1.888.411.1WEB IntelliStation News Award List Press Release Work Stations How to buy & upgrades Ask engineering Total Cost of Ownership Similar Pages
http://www5.pc.ibm.com/us/me.nsf/News\IntelliStation
3. Lycos TOP 5% Review: IBM Higher Education
Help Feedback. Search for: Search Top 5 % New Search Lycos Network: Tripod.com | Angelfire.com | MailCity.com | WhoWhere.com | HotBot.com. All...

Machine Translation

Generate and Repair Machine Translation (GRMT)

Imagine picking up the phone in Toronto, dialing your Japanese friend in Tokyo. You speak English; she hears Japanese. Fortunately it is 2020 and your English is automatically translated into Japanese in the time it takes to transfer your words. Impossible you say??!!!

Machine translation has fascinated and frustrated researchers for over 50 years. Recent success in statistical, nonlinguistic and hybrid systems provides hope that we will not be confined to traditional direct, transfer and intralingual approaches. We provide an approach following from CS methodology: generate and repair machine translation. (GRMT).

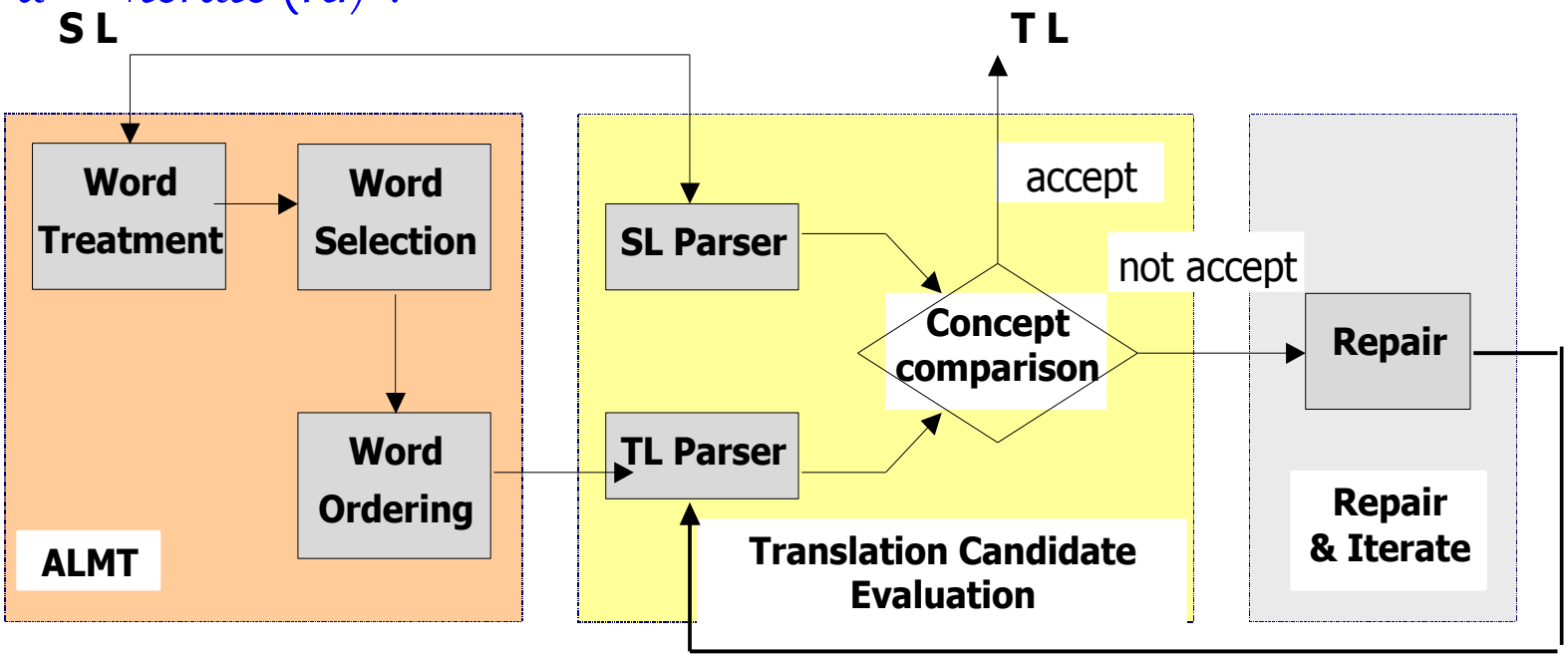
Comparison of Three Traditional Approaches

Attributes	Direct MT	Transfer MT	Interlingual MT
Factors of Accuracy	dictionary and mapping rule	transfer rule	concept representation
Levels of Linguistic Analysis	word	meaning	concept
Intermediate Representation	no Representation	language dependent representation	language independent represenation
Modularlity	depends on system desigr	depends on system design	analysis and generation modules independent
Multilingual System (add the nth language to the (n-1) languages system	needs 2(n-1) mapping rules	needs (n-1) analysis, (n-1) generation and 2(n-1) transfer	needs 1 analysis and 1 generation
Extendibility (in terms of integrating new language to system)	needs mapping rules	needs analysis, generatior and transfer	needs analysis and/or generation

Translation Examples by a Commercial System

English	Translation to French	Translation back to English
1. Hatchery officials are having to teach the fish to like worms.	1. Les fonctionnaires de l'incubateur doivent apprendre le poisson pour aimer des vers.	1. Civil servants of the incubator must learn fish for aimerdes toward.
2. It does not matter if you are born in a duck yard.	2. Il n'importe pas si vous naissez dans un jardin du canard.	2. He/it doesn't import if you are born in a garden of the duck.
3. Only a life lived for others is a life worth while.	3. Seulement une vie pour les autres vaut pendant que.	3. Only a life for other is worth while.
4. I never think of the future.	4. Je ne pense jamais du futur.	4. I never think the future.
5. You can take a fish to school but you can not make them think.	5. Vous pouvez prendre un poisson pour scolariser, mais vous ne pouvez pas les faire penser.	5. You can take a fish to school, but you don't can pasles to make think.
6. It's no go.	6. Il n'est pas aucun entrain.	6. Him estpas no liveliness.
7. Never mind.	7. Ne faites jamais attention.	7. Don't make attention never
8. BOISE, Idaho (AP)- Trout in Idaho are not just swimming in schools- they are going to school.	8. BOISE, Idaho (AP) - Truite dans Idaho ne nage pas dans les écoles juste - theyare aller scolariser.	8. BOISE, Idaho (AP) - Trout in Idaho doesn't swim rightly in schools-they are to be going to school.

GRMT is composed of 3 phases: "Analysis Lite Machine Translation (ALMT)", "Translation Candidate Interpretation (TCI)" and "Repair and Iterate (RI)".



The 3 Phases

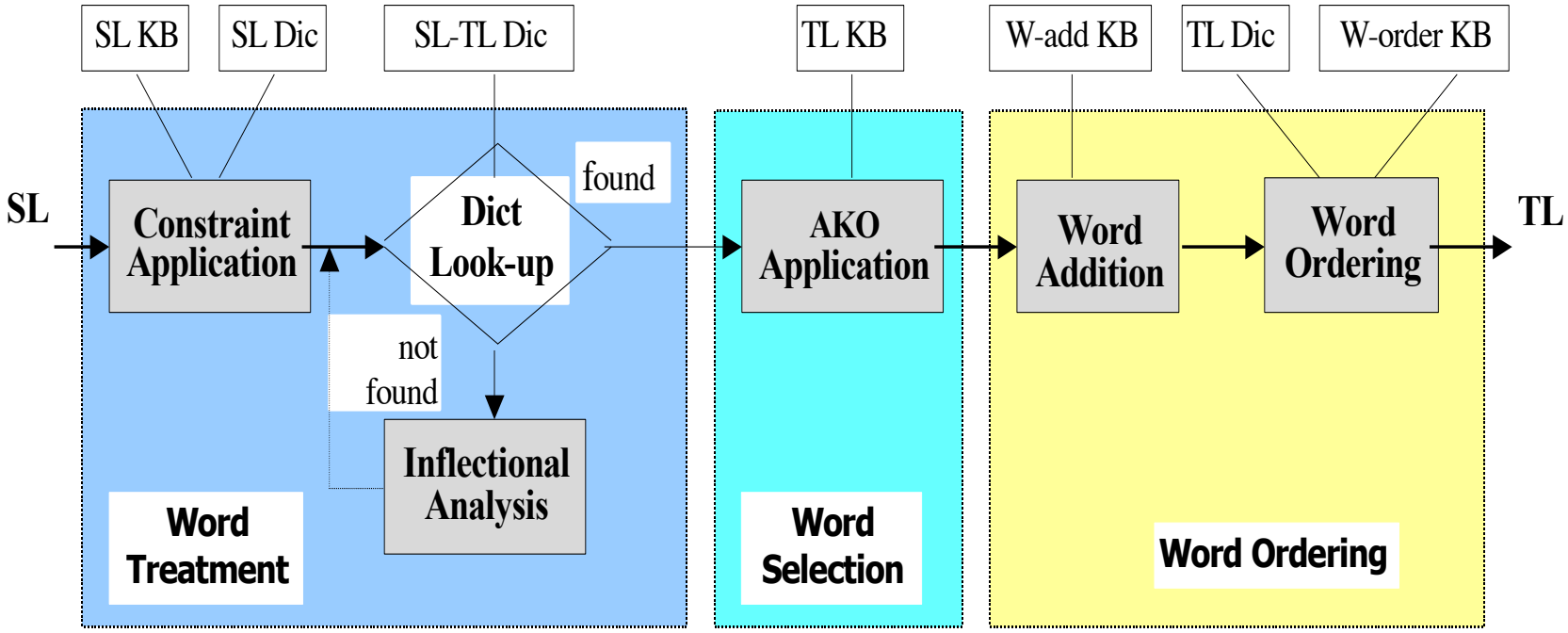
ALMT generates translation candidates (TC) by considering syntactic and semantic differences between language pairs without any sophisticated analysis. This ensures the TC is generated quickly and efficiently.

Next, the system interprets the TC to see if it retains the meaning of the SL. If so, that TC will be considered a translation. If not, that TC will be repaired based on the diagnosis indicated in the TC1 phase.

The repaired TC will be re-interpreted to determine if it still has a different meaning from the SL. These two processes iterate until the TC conveys the same meaning as the SL. The TCI and RI stages ensure the accuracy of the translation result.

ALMT Architecture

ALMT was designed around two simple notions: first, the more accurately ALMT generates a TC, the less work is required in the latter phases; and second, generating the TC must be done quickly. Therefore, ALMT generates a TC by considering the difference between language pairs in terms of syntax and semantics without performing any sophisticated analysis. ALMT performs its task by judiciously selecting a few simple heuristics, constraints, and semantic principles, to apply when appropriate, into a simple direct framework for translation.


SL KB
SL Dic
SL-TL Dic
TL KB
W-add KB
TL Dic
W-order KB
SL
Constraint Application
Dict Look-up
found
not found
Inflectional Analysis
AKO Application
Word Addition
Word Ordering
TL
Word Treatment
Word Selection
Word Ordering

Example of a Correctly Generated ALMT TC

An old woman lived in the cottage, with a fat black cat and a plump brown hen.

TC: ผู้หญิง แก่ คนหนึ่ง ได้ อยู่ ใน กระท่อม หลัง นั้น กับ แมว สี ดำ อ้วน ตัว หนึ่ง และ ไก่ สีน้ำตาล อวบ ตัว หนึ่ง

CT: ผู้หญิง แก่ คนหนึ่ง ได้ อยู่ ใน กระท่อม หลัง นั้น กับ แมว สี ดำ อ้วน ตัว หนึ่ง และ ไก่ สีน้ำตาล อวบ ตัว หนึ่ง

(phûujiˇŋ- *woman*) (kɛ`ɛ- *old*) (khon- clas) (ny`ŋ- *an*) (dâj- *past*) (ju`u- *live*) (naj- *in*) (krathɔˆm- *cottage*) (laˇŋ- clas) (nán- *the*) (kàb- *with*) (mɛɛw- *cat*) (siˇidam- *black*) (?ûan- *fat*) (tua- clas) (ny`ŋ- a) (lɛ`- *and*) (kàj- *hen*) (siˇinámtaan- *brown*) (?ùab- *plumb*) (tua- clas) (ny`ŋ- a)

In the example, some words have more than one meaning e.g., old, in, live, with … The appropriate meaning of old and in, can be selected by considering the semantic relationship between words. However, appropriate words for live and with cannot be selected in the same manner because there is no explicit relationship between these words and words in their proximity. Therefore, the first meaning appearing on the list of meanings for each word is selected. The word **ได้** (dâj) is added to clarify the past tense (lived). The classifiers **คน** (khon), **หลัง** (laùŋ) and **ตัว** (tua) are also added according to Thai grammar. The indefinite determiners a & an in this expression correspond to the word **หนึ่ง** (nyˋŋ) in Thai indicate the need for classifiers for the words **ผู้หญิง** (phûujiùŋ- woman), **แมว** (mɛɛw- cat) and **ไก่** (kàj- hen) respectively.

The word **ผู้หญิง** (phûujiùŋ- woman) belongs to the class Female (1-1-1-1-1-2), a subclass of Human (1-1-1-1). A noun that belongs to the class 1-1-1-1 is compatible with a classifier with the WordAsso number 2-4-2-1-1-1 based on classifier relations. The classifier **คน** (khon) with 2-4-2-1-1-1 is selected for **ผู้หญิง** (phûujiùŋ-woman). The words **แมว** (mɛɛw-cat) and **ไก่** (kàj-hen) belong respectively to the classes mammal (1-1-1-2-1-1) and fowl (1-1-1-2-1-2-2), subclasses of animal (1-1-1-2). Since a noun with WordAsso number 1-1-1-2 relates to a classifier with 2-4-2-1-1-2 according to classifier relations, the classifier **ตัว** (tua) with 2-4-2-1-1-2 is selected for **แมว** (mɛɛw-cat) and **ไก่** (kàj-hen).

The definite determiner the corresponds to the word **นั้น** (nán) in Thai and indicates the need for classifiers for **กระท่อม** (krathɔ̂m- cottage). The word **กระท่อม** belongs to the class Housing (1-1-2-1-1-2-2) which is compatible with a classifier with the WordAsso number 2-4-2-1-2-10 based on classifier relations. Therefore, the classifier **หลัง** (laˇŋ) with the WordAsso 2-4-2-1-2-10 is selected for the word **กระท่อม** (krathɔ̂m- cottage).

The selected words in each noun phrase, an old woman, the cottage, a fat black cat and a plump brown hen are rearranged into Thai grammatical order as illustrated. The generated TC for the example is exactly the same as the Correct Translation (CT).

The required GRMT knowledge bases include:

Constraints, Dictionaries, Grammar & Lexicon

Constraints

SL constraints are the characteristics of the SL which are different from those of the TL. They are used to simplify the structure of the SL and to narrow the scope of possible TL words that correspond to each SL word. TL constraints are the characteristics of the TL that are different from those of the SL. They are required, not only to retain the meaning of SL but also to make them grammatically correct.

Dictionaries

GRMT uses 3 types of dictionaries, the SL dictionary, TL dictionary and a bilingual dictionary. Entries in the SL and TL dictionaries can be single word, some inflected and derived forms which cannot be easily handled by rules. Compound words are also included. Each entry has morphological, syntactic and semantic information. The Thai dictionary entry contains word form and word subcategory. The English dictionary contains the category used in the inflectional analysis step. The Bilingual dictionary contains the English entry and all corresponding Thai words and AKO number for each Thai word, e.g., the word "dream" in English has 3 Thai words which express differences in meaning and usage.

All Thai words which correspond to each English entry are ordered based on usage frequency. The first meaning is selected once constraint and AKO fail.

The SL dictionary is used by ALMT in the constraint application and inflectional analysis steps. The SL and the TL are put into correspondence via the SL-TL dictionary. The SL-TL dictionary contains the SL entry and its all possible corresponding words in the TL.

Grammars and lexicons of both SL and TL are required in the TCE analysis process. They are developed principally based on HPSGs.

Experiments of ALMT (English to Thai) indicate that TCs can be generated with relative accuracy. The table below shows all the steps in an earlier example of applying ALMT.

English	Constraint Application output	Word Treatment output	Selected Word in Thai	Word Addition	Word ordering
An	An	Ë?Öè§	Ë?Öè§		?ÙéË-Ô§ (phûuji¡ŋ)
old	old	á¡è (kɛ̀ɛ), à¡èÔ (kàw)	á¡è (kɛ̀ɛ)	€? (khon)	á¡è (kɛ̀ɛ) €? (khon)
woman	woman	?ÙéË-Ô§¡ŋ)	?ÙéË-Ô§¡ŋ)	ä´(ɛ́aj)	Ë?Öè§ ä´(ɛ́aj)
lived	live	ÍÂÙÈ (jùu), ÁÔª (mii), ´ÔÃ§ªÔÇÔ (damroŋchiiwíd)	ÍÂÙÈ (jùu)		ÍÂÙÈ (jùu)
in	in	ã? (naj), à¢éÒÁÔ (khâwmaa), à¢éÒ (khâwpaj), ÍÂèÔ§ (nay)	ã? (naj)		ã? (naj)
the	the	?Ñé? (nán)	?Ñé? (nán)	ËÅÑ§ (lɛ́ŋ)	¡ÃD · èÍÁ (krathˆm) ËÅÑ§ (lɛ́ŋ)
cottage	cottage	¡ÃD · èÍÁ (krathˆm)	¡ÃD · èÍÁ (krathˆm)		?Ñé? (nán)
with	with	¡Ñº (kàb), ´éÇ (dûaj), «Ô§Ãy (sˆŋ)	¡Ñº (kàb)		¡Ñº (kàb)
a	a	Ë?Öè§	Ë?Öè§	µÑÇ (tua)	áÁÇ (mɛɛw) ÊÓ´Ó (si¡ídam)
fat	fat	ÍéÇ (?ûan), ä¢ÁÑ? (khâwjman)	ÍéÇ (?ûan)		ÍéÇ (?ûan)
black	black	´Ó (dam), ÊÓ (sˇdam)	ÊÓ (sˇdam)		µÑÇ (tua)
cat	cat	áÁÇ (mɛɛw), ?ÙéË-Ô§äÁè´Õ (phûuji¡ŋmâjdìi)	áÁÇ (mɛɛw)		Ë?Öè§
and	and	áÅÐ (lɛ́`)	áÅÐ (lɛ́`)		áÅÐ (lɛ́`)
a	a	Ë?Öè§	Ë?Öè§	µÑÇ (tua)	ä¡è (kàj) ÊÓ?éÓµ (si¡inámtaan)
plumb	plumb	ÍÇº (?ùab)	ÍÇº (?ùab)		ÍÇº (?ùab)
brown	brown	ÊÕ?éÓµÔÅ (si¡inámtaan), à¡ÃÔÅ (kaafée), €Å (fée)	ÊÕ?éÓµÔÅ (si¡inámtaan)		µÑÇ (tua)
hen	hen	ä¡è (kàj)	ä¡è (kàj)		Ë?Öè§

Translation Candidate Evaluation (TCE)

The second phase of GRMT, TCE, analyzes the generated translation candidate to determine whether the TC retains the meaning of the source language. TCE analyzes both the SL and the TC in parallel, then compares the parses semantically alone, since there are syntactic level differences between languages. If the semantic results are the same, the TC will be deemed an appropriate translation. If the semantic results are different, the TC will be repaired in the third phase, Repair and Iterate. TCE comprises two modules: the analyzer and semantic comparison.

Example

This figure illustrates the parse of the SL "The ugly duckling hides his head under his wing" and the parse of its TC. The TC is generated by ALMT. The parses are in ALE representation. Both SL and TC are licensed by our grammars. Their syntax are shown in the box.

```
STRING:
0 the 1 ugly 2 duckling 3 hides 4 his 5 head
6 under 7 his 8 wing 9
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
                        NUCLEUS beauty
                              WORDASSO 13-8-4
                              INSTANCE [4]
                              SURFACE ugly
. . . .

SYNSEM synsem
    LOC loc
      CAT cat
        COMPS e_list
        HEAD verb
          AGR agr
            GENN gend
            NUM num
            PER per
          AUX minus
          INV minus
          LANG eng
          MOD none
          NEG boolean
          PRD boolean
          VFORM fin
        MARKING unmarked
        SPR list_synsem
        SUBJ e_list
      CONT psoa
        NUCLEUS hide2
              WORDASSO 2-10-1
              HID1 [1]
              HID_PLACE [3]
              HIDER1 [4]
              SURFACE hide
        QUANTS e_list
      CONX conx
. . .
```

```
STRING:
0 ÅÙìà»ç´ 1 ¢Õéäëêãè 2 µÑÇ 3 ¹Ñé 4 «èÎ 5 ËÑÇ
6 ¢íŞà¢Õ7 ãµé 8 »Õì 9 ¢íŞà¢Ð
CATEGORY:

phrase
QSTORE ne_set_quant
. . .
                        NUCLEUS beauty
                              WORDASSO 13-8-4
                              INSTANCE [4]
                              SURFACE ¢Õéäëêãè
. . .

SYNSEM synsem
    LOC loc
      CAT cat
        COMPS e_list
        HEAD verb
          AGR agr
            GENN gend
            NUM num
            PER per
          AUX minus
          INV minus
          LANG thai
          MOD none
          NEG boolean
          PRD boolean
          VFORM bse
        MARKING unmarked
        SPR ne_list_synsem
          HD synsem
            LOC loc
              CAT cat
                COMPS list_synsem
                . . .
                MARKING comp
                SPR list_synsem
                SUBJ list_synsem
. . .
        SUBJ e_list
      CONT psoa
        NUCLEUS hide2
              WORDASSO 2-10-1
              HID1 [1]
              HID_PLACE [3]
              HIDER1 [4]
              SURFACE «èÎ
        QUANTS e_list
      CONX conx
. . .
```

Semantic Extraction extracts the semantic information of the SL and the TC from their parses so that we can compare the meaning of the TC with that of the SL. The semantic representation of an expression in GRMT is based on the HPSG representation provided. The features CONT, CONX and QSTORE hold the semantics of the object. The semantic representation described in CONT value of *hide* corresponds to the situation in which (*the hider*) hides (*the hid*) in (or under, from, ...) (*the hid_place*).

The features SURFACE and WORDASSO indicate the word form and WordAsso number of the lexical entry. Features SURFACE and WORDASSO are used in the Repair and Iterate phase. Feature QSTORE is a storage for the quantifiers. Feature CONX contains linguistic information that bears on certain context-dependent aspects of semantic interpretation.

Translation Candidate Evaluation – Semantic Comparison

In comparing the semantics between the SL and its TC, the values of the features CONT, QSTORE and CONX are considered. If the values of these features of the parsed SL are the same as those of the parsed TC, TCE concludes that the TC does not require repair. If any of these features are different, TCE will provide the information of the parsed SL which differs from that of TC parse. This information will be used in the next phase, Repair and Iterate.

Repair and Iterate (RI)

RI performs the repair process if required and returns the repaired TC to the TCE phase. TCE re-analyzes the repaired TC to determine if a different meaning from the source language remains.

RI examines the result of TCE. The TCE output is the TC with the semantic information of the SL which differs from that of the TC. With this information, RI is able to detect the part of the TC that causes the mis-translation. The mis-translated part will be replaced with a more appropriate translation.

RI searches the Word Treatment output for a more appropriate translation based on the information provided by TCE. The new TC is put through Word Ordering to revise its syntax. Once revision is complete, the repaired TC is returned to TCE.

If the CONT or QSTORE value of the SL is different from that of the TC, the SL CONT value: SURFACE and WORDASSO features are passed to RI. SURFACE indicates the surface form of the word which causes the mistranslation.

WORDASSO specifies the proper meaning of the word in question in terms of WordAsso number. To repair CONT or QSTORE of the TC, RI re-selects the corresponding word in the TL for the word specified in SURFACE. The re-selection is done by searching *Word Treatment output* for the corresponding word which has the same WordAsso number as specified in WORDASSO.

In the case that the CONX value of the SL differs from that of the TC, the SL CONX value: the BEARER and NAME features are passed to RI. BEARER specifies the index which associates with the certain name specified in the NAME value. RI repairs the TC by associating the right names to the right indices based on the information provided by TCE.

The CONT values of the SL and the TC parses of Example 5 are different as illustrated in Figure 9. The CONT value of the SL indicates that the word *like* is mis-translated (SURFACE: like) and its proper meaning in this expression is *to regard with pleasure or fondness* which is classified into the class of 2-6-2-4 (WORDASSO: 2-6-2-4). Therefore, RI begins the repair process by re-selecting the translation of the word *like*. RI searches the *WordTtreatment output* Table for the translation of *like* which has WordAsso number 2-6-2-4 and thus, the translation **ชอบ (chɔˆɔ),** is selected

Key Features of GRMT

Simplicity Each step performed by GRMT is straightforward to carry out.

Modularity GRMT's translation process is separated into three modules: ALMT, TCE and RI. Each module is comprised of sub-modules for easy modification and maintenance.

Extendibility GRMT is intended to be easily extendible to any other language. Since each component is separated not only in the translation process components but also in the knowledge bases, each component can be extended easily to a larger domain.

Multilinguality depends on the modularity and extendibility. GRMT is highly modular and extendible in two major ways. The treatment of SL and TL are independent. Required SL and TL knowledge bases are developed separately, hence it is easy to add new languages. For example, SL-constraints (e.g., plurality, continuous tense, etc.) required to translate English into Thai can be applied to translate English into Chinese and Japanese since these languages share those characteristics.

French and Spanish share the same syntactic features with English, features which differ from those of Thai, Chinese and Japanese, then GRMT requires six analyzers and two sets of constraints to perform the translation between the two language families. Transfer MT requires 6 SL analyzers, 6 TL generations and 18 sets of transfer rules.

Summary

NLU by computer is an enviable goal and many uses of this technology have already been put to the test.

Has research progressed to the point where it will actually be possible to begin to build the "ideal" NL system? As knowledge representation ideas become more precisely formulated, such an evolution is happening.

Despite these developments, the ability to incorporate knowledge is still a major source of difficulty confronting the designer of the ideal NL system. Much knowledge representation is not explicitly aimed at NLU, and less yet at the problem of integrating knowledge into the interpretation processes. Much work is highly theoretical. Finally, knowledge representation is a vast area of inquiry.

It appears that the ideal NL system is still some way off, at least in its full splendor. Nevertheless, the indicators are all very positive.

Concluding Remarks

On Problems

Our choicest plans have fallen through,
our airiest castles tumbled over,
because of lines we neatly drew
and later neatly stumbled over.