Fifth Biennial Conference of the
Association for Machine Translation in the Americas

Tutorial Notes

Example – Based Machine Translation

Ralf Brown
Carnegie Mellon University
Language Technologies Institute

October 9, 2002
Tiburon Lodge
Tiburon, California

Example-Based Machine Translation

A Tutorial

Ralf Brown

Carnegie Mellon University

ralf+@cs.cmu.edu

9 October 2002

Overview

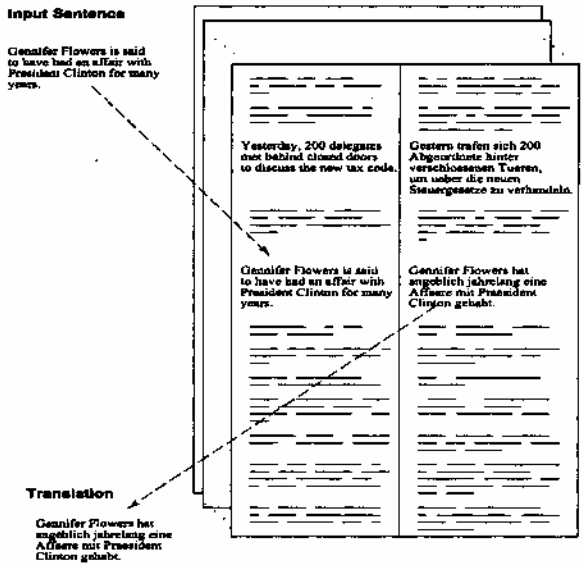
- What is EBMT?
- Types of EBMT
- Relationship between EBMT and other techniques
- Sample Systems
- (break)
- Hands-On Exercise
- CMU's Generalized EBMT system

What is Example-Based Machine Translation?

EBMT is one of a variety of *corpus-based methods*. Rather than having someone explicitly encode translation rules, corpus-based methods use a collection of pre-translated texts as training material to automatically learn how to translate.

- EBMT is sometimes called Memory-Based Translation
- EBMT is closely tied to Case-Based Reasoning

Other corpus-based methods include translation memories and statistical translation.



Translation Memory

Simplest version: if we are given one of the units in the corpus, retrieve its translation. More sophisticated translation memories retrieve the nearest match (if "close enough") and let the user fix up the retrieved translation. If done well, this is still much faster than generating a translation from scratch.

Translation memory is most useful when translating revised versions of previously-translated documents – the parts which remain unchanged can be translated by the TM, leaving only the modifications to be re-translated manually.

Example: IBM's TM2 system.

Example-Based Machine Translation

Translation memory can be generalized: find the nearest matching sentence in the corpus, and determine how to transfer any remaining differences to the translation.

Drawback: this can require considerable knowledge of **both** source and target language.

Alternative: Find the largest exact matches of portions of the input to be translated, and combine the pieces later. For this to work, we need a way of determining which piece of the translated sentence in the example base corresponds to the portion of the source sentence that was actually matched.

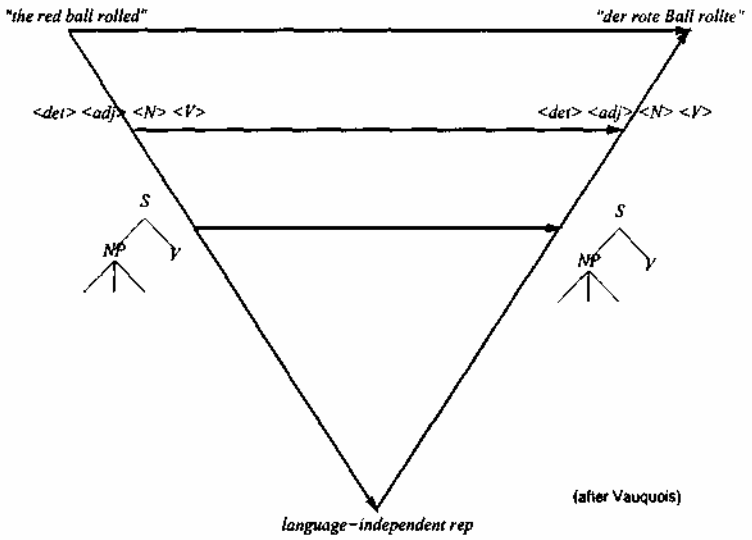
Origins of EBMT

What is now known at EBMT was first proposed in 1984 by Makoto Nagao. The idea of storing large numbers of translation examples goes back much further, but necessary computational resources were not yet available.

Nagao referred to his method as translation by analogy.

Types of EBMT

- lexical (shallow)
- morphological / part-of-speech analysis (less shallow)
- parse tree-based (deep)



EBMT and SMT

Statistical MT, as another corpus-based method, is closely related to EBMT.

- like EBMT, trained from parallel text
- unlike EBMT, does not retain original examples once trained

A trained statistical MT system essentially consists of one or more mathematical models:

- translation probabilities
- word re-ordering probabilities
- output language model

EBMT and rule-based systems

A purely corpus-based system doesn't use manually-written rules (there are hybrid systems which do), but may include a component to automatically learn translation rules. In fact, much recent work has focused on

- extracting bilingual terminology
- finding equivalence classes among words
- inducing morphology rules
- inducing grammar rules

EBMT and interlingual systems

An interlingual translation system tries to analyze its input all the way to a language-independent representation of the underlying meaning, before generating a translation in the other language. Interlinguas vary from extremely detailed (trying to capture every last nuance) to fairly simple and task-based (capturing only the essential meaning).

It is conceivable to create an example-based interlingual system for a task-based interlingua, using EBMT techniques to convert text into an interlingual representation, and then to generate a translation from the interlingua.

Hybridization

- EBMT + rule-based
- EBMT + translation memory
- EBMT + statistical
- EBMT + neural nets
- multi-engine

Hybrids: EBMT + Statistical MT

Many EBMT systems require some form of bilingual dictionary to find cross-language correspondences. One obvious way to generate such a dictionary is using statistical techniques on the training corpus.

Other techniques developed for statistical MT can also be applied to EBMT, such as word-level alignments.

Hybrids: EBMT + rule-based

A number of rule-based systems have had data-driven components added (Carl et al 1999)

CAT2 rule-based system + EDGAR EBMT system

- EDGAR uses morphological and syntactic information
- CAT2 implements a semantic theory
- tight integration
  - EDGAR provides word and phrase translations

  - CAT2 translates linguistic structures and those portions of the input for which EDGAR has no examples

Hybrids: EBMT + translation memory

(Michael Carl and Silvia Hansen 1999)

- experimented with a string-based translation memory, a lexeme-based translation memory, and the EDGAR EBMT system
- string-based TM is very precise, but has low coverage
- EBMT has broadest coverage
- integration uses string-based TM with EDGAR as fallback

Hybrids: EBMT + neural nets

(Ian McLean 1992)

EBMT using connectionist matching

- neural network learns salient terms from parallel corpus
- trained NN then scores nearness of match between training examples and new text

Hybrids: multi-engine combinations (1)

Since all translation methods have strengths and weaknesses, the idea behind multi-engine approaches is to combine multiple methods (engines) so that one engine's strengths can compensate for another engine's weaknesses.

Hybrids: multi-engine combinations (2)

Three main approaches to multi-engine combination:

- tight coupling: selecting at a subsentential level or using inter-engine negotiation
- after-the-fact selection: each engine generates a complete translation, and the best one is selected by an external process
- fail-over: one primary engine is used unless it fails to produce a translation, in which case another engine is given a chance to translate the input

Overview of EBMT systems

- Veale & Way: Gaijin
- Michael Carl: EDGAR
- Brona Collins: ReVerb
- Guvenir & Cicekli: Generalized EBMT
- ...
- CMU: G-EBMT

System: Gaijin

(Veale & Way 1997)

Japanese-English translation

- part-of-speech tagging in both languages
- translation examples converted into templates consisting of part-of-speech tags
- matching performed at the level of complete tag sequences (no partial matching)

System: EDGAR

Michael Carl *et al* @ University of Saarbrücken

- applies morphological analysis to both languages
- induces translation templates from analyzed reference translations
- multiple levels of generalization
- matched chunks from case base are re-specialized and refined in the target language

System: ReVerb

(Brona Collins 1996, 1999)

- explicitly uses Case-Based Reasoning
- training examples are abstracted to syntactic dependency representation
  - shallower processing than original Nagao/Sato approach, using flat feature lists
- retrieval criterion is combination of similarity and adaptability
- retrieved examples are adapted to fit the text to be translated

System: Guvenir & Cicekli

(1996-)

- training examples are abstracted into templates by replacing certain word stems and morphemes by co-indexed variables
- generalization based on the heuristic that differences in mostly-similar sentence pairs should correspond

System: Guvenir & Cicekli (2)

Sample of differences and similarities:

I give+PAST the book to Mary
Mary+DAT kitap+ACC ver+PAST+1SG

I give+PAST the pencil to Mary
Mary+DAT kurgun kalem+ACC ver+PAST+1SG

(Cicekli & Güvenir, 1996)

Template:

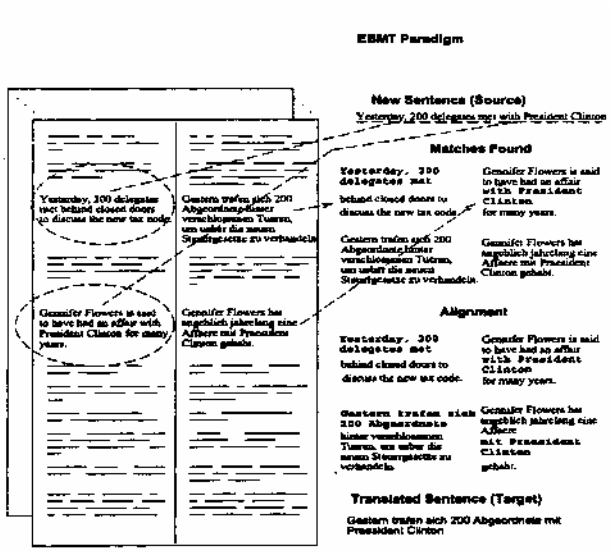
I give+PAST the X^S to Mary
Mary+DAT X^T+ACC ver+PAST+1SG

Hands-On Exercise

- distribute bilingual corpus to tutorial participants
- emulate a translation memory
- emulate an EBMT system

CMU's Generalized EBMT System

- simple lexical match
- inexact matching
- generalizing into templates
  - manually
  - automatically (machine learning)
- multi-engine



G-EBMT: Lexical Matching

Shallow processing:

- string match of surface forms
  - **Advantage**: little or no need for linguistic knowledge
  - **Disadvantage**: requires large amounts of training text
- convert text into templates, then use string match of templates
  - **Advantage**: requires less training text
  - **Problem**: how to produce good-quality general templates?

G-EBMT: Inexact Matching

We can get many more (and longer) matches against the corpus if we can make a match where not all words are matched.

A recent addition to the system is allowance for a one-word gap in the middle of a match, *provided there is a reasonably unambiguous translation known* for that word. Reasonably unambiguous means that the word either

- has only one translation listed in the dictionary
- has its most-common translation occurring more than twice as frequently as the next translation

This fuzzy matching proved helpful on limited training data, but did not improve quality when more data was available.

G-EBMT: Word-Level Alignment

When the system partially matches a training example, the hard part is determining which portion of the translation corresponds to the matched text.

To perform word-level alignment, the EBMT system needs a bilingual dictionary. It then uses the translations along with heuristic scoring functions such as

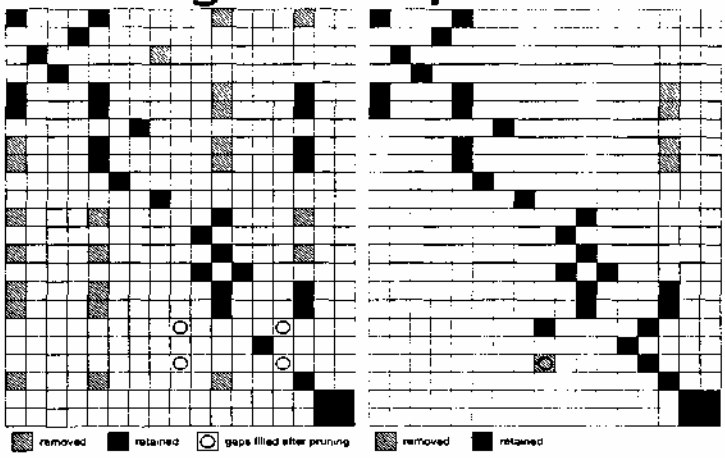
- common location in sentence
- difference in length

- words known to translate as empty string

to find the best-scoring substring of the translation.

G-EBMT: Correspondence Table

Pruning Correspondences



Term-Substitution Dictionary

We can extract bilingual dictionaries such as the one required for word-level alignment using fairly simple statistical techniques. One method: build a large table of co-occurrences, filter it using a threshold function, and output any remaining entries as probable mutual translations.

Statistical dictionaries can be tuned: there is a size/accuracy tradeoff – we can get a larger vocabulary at the cost of more errors, or reduce errors by sacrificing some words

The threshold is based on Mutual Conditional Probability: $P(W_s|W_t) \geq thr(C)$ and $P(W_t|W_s) \geq thr(C)$ where C is the number of times the two words co-occurred.

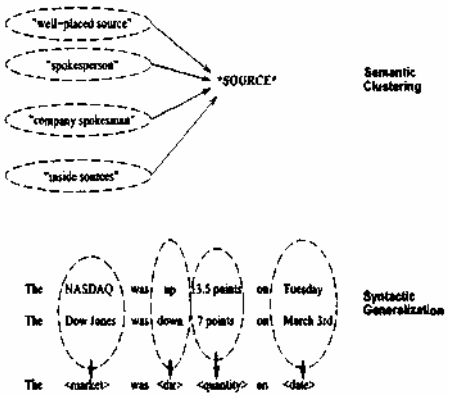
Sample Dictionary

(ABI (ABI 4)(BEVERAGE 2)(AMALGAMATED 2))
(ALMAHDI (AL-SADIQ 1)(AL-MAHDI 1))
(ARABSAT (ARABSAT 6))
(BIOTOPES (BIOTOPOS 2))
(BLEACH (LEJÍA 1))
(COMPLEMENTARITY (COMPLEMENTARIEDAD 77))
(D-1 (D-1 91)(D-2 43))
(DEEPEN (PROFUNDIZAR 17))
(DYNAMICS (DINÁMICA 77))
(EBW (HAZ 6)(ELECTRONES 6)(SOLDADORA 6))
(ESCOBAR (ESCOBAR 30))
(EXTRACONTINENTAL (EXTRACONTINENTALES 1))
(GEOSYSTEMS (GEOEX-1986 1)(GEOSISTEMAS 1))
(HU (HU 2)(XIAODI 2))
(KG (KILOGRAMOS 16)(KG 10))
(MILITARY-IDEOLOGICAL (MILITAR-IDEOLÓGICA 1))
(MONASTERY (MONASTERIO 2))
(NON-NUCLEAR-WEAPON (POSEEDORES 78))
(ORCI (OIRI 8))
(PASHTU (PASHTU 1)(BRISTISH 1))

(PYAT (PYAT 1)(CABARET 1)(PERODIN 1)(VILLARD 1))
(RAVANDI (RAVANDI 1)(KATCHOUI 1))
(REDISCOVER (REENCONTRAR 1))
(SCENES (ESCENAS 5))
(SECRECY (SECRETO 53))
(SHANKANGA (SHANKANGA 1))
(TECNOLÓGICO (UCMM 1))
(XXVI (XXVI 86))
(|1506TH| (|1506A| 8))

G-EBMT: Generalization

G-EBMT Augmentation



G-EBMT: Manual Generalization

- equivalence classes
- pattern replacement
- recursive replacement

Sample Equivalence Classes

Equivalence classes are sets of words/phrases which can be used interchangeably. They may be

semantic:	or syntactic:
numbers	masculine nouns
days of the week	plural adjectives
names of cities	first-person verbs
colors	etc.
shapes	
etc.	

G-EBMT Generalization: Equivalence Classes (1)

Given a set of equivalence classes, replace each occurrence in the training text by the class name, and index the resulting templates.

25 players met in London yesterday.
25 Spieler trafen sich gestern in London.

<number> players met in <city> <time>.
<number> Spieler trafen sich <time> in <city>.

When translating, perform the same substitutions, but remember the appropriate translation for each occurrence. Match the resulting template against the indexed corpus, and substitute the remembered translations into the translated template.

G-EBMT Generalization: Equivalence Classes (2)

Thus, we try matching not only the surface form, but also the templatized version of the input against the example base:

12 players met in Paris last Tuesday.
<number> players met in <city> <time>.

Even though the example on the previous slide would not have matched directly, the template is identical and therefore we have a successful match. We also know (from the definition of each equivalence class) the proper translation for the abstracted words.

G-EBMT Generalization: Equivalence Classes (3)

The final step is to substitute the proper word translations back into the translated template:

12 players met in Paris last Tuesday.
<number> players met in <city> <time>.
<number> Spieler trafen sich <time> in <city>.
<number> = 12, <city> = Paris, <time> = letzten Dienstag
12 Spieler trafen sich letzen Dienstag in Paris.

G-EBMT Generalization: Pattern Replacement

Members of an equivalence class need not be literal strings, which allows a paired production-rule grammar to be created.

<N-M>:

English	French
accessory	accessoire
book	livre
costume	accoutrement
subscription	abonnement

<NP-M>:

English	French
the <N-m>	le <N-m>
<poss-m> <N-m>	<poss-m> <N-m>
the <number> <N-m>	le <number> <N-m>
the <adj-m> <N-m>	le <N-m> <adj-m>
the <adj-m>1 <adj-m>2 <N-m>	le <N-m> <adj-m>2 <adj-m> 1
the <ordinal> <N-m>	le <N-m> <ordinal>
the <national-m> <N-m>	le <N-m> <national-m>

G-EBMT Generalization: Recursive Replacement

For historical reasons, the G-EBMT system has two separate but related mechanisms for specifying equivalence classes and rewriting rules. One is context-independent (applied unconditionally), while the other is used only if at least one adjacent word matches in some training example.

The two sets of rewriting rules are applied alternately until no more replacements are possible.

G-EBMT: Learning How to Generalize

While generalization is highly effective, creating all the rules manually is considerable work. Much recent development has focused on learning equivalence classes and rewriting rules automatically from the corpus.

Three different learning mechanisms have been implemented to date:

- single-word equivalence classes via clustering
- grammar induction
- word decompounding

Single-Word Equivalences

Observation: if the context in which a word appears is defined as the sum of the words in the immediate neighborhoods of its occurrences, we can use standard document-clustering techniques to perform word clustering.

Approach: create a pseudo-document for each word, containing all the words surrounding its occurrences; make the word the document identifier.

Problem: this yields only a monolingual clustering, but we need a set of bilingual pairs.

Solution: use the approach of Barrachina and Vilar to inject bilingual information into the clustering.

Injecting Bilingual Information into Monolingual Clustering

1. use a bilingual dictionary to create a rough bi-text mapping between the source-language and target-language halves of a sentence pair.
2. whenever there is a unique correspondence indicated by the bi-text mapping, generate a bilingual word pair consisting of the word and its translation.
3. treat those word pairs as indivisible tokens in further processing.

Bilingual Information

These bilingual word pairs also serve to provide a rough separation of a word into its senses.

For example,

E: bank	G: Bank	financial institution
E: bank	G: Ufer	river-bank

Sample Clusters

HISTOIRE	HISTORY	HOMMES	POLITICIANS
ÉCONOMIE	ECONOMY	PRISONNIERS	PRISONERS
CERTAINEMENT	CERTAINLY	AVEUGLES	BLIND
CERTAINEMENT	SURELY	CHAUSSURES	SHOES
CERTES	SURELY	CONSTRUCTEURS	BUILDERS
JAMAIS	NEVER	PENSIONNÉS	PENSIONERS
PAS	NOT	RETRAITÉS	PENSIONERS
PEUT-ÊTRE	MAY	VÊTEMENTS	CLOTHING
PROBABLEMENT	PROBABLY		
QUE	ONLY		
RIEN	NOTHING		
SÛREMENT	CERTAINLY		
SÛREMENT	SURELY		
VRAIMENT	REALLY		
CONSERVATEUR	CONSERVATIVE	FAÇON	EVENT
CONSERVATEUR	TORY	ÉVIDENCE	CLEARLY
DÉMOCRATIQUE	DEMOCRATIC	ÉVIDENCE	OBVIOUSLY
DÉMOCRATIQUE	NDP		
LIBÉRAL	LIBERAL		

Grammar Induction

Observation: similar sentences in a corpus tend to differ by concrete constituents.

The team met *at the airport*.
The team met *in town*.

Thus, we can search a corpus for patterns of similarity and dissimilarity to find constituents that can be used interchangeably.

The initial implementation only searches for the pattern

$$S_1\ D\ S_2$$

The various instantiations of D are added to an equivalence class, as are S_1 and S_2 if appropriate.

Grammar Induction (2)

Sort sentences:

we are watching agricultural chemicals .
nous regardons les produits chimiques agricoles .

we are watching energy supplies .
nous regardons les approvisionnements en énergie .

we are watching equipment supplies .
nous regardons les approvisionnements en matériel .

we are watching fertilizer supplies .
nous regardons les approvisionnements en engrais .

we are watching steel production .
nous regardons la production de acier .

Sorted by reverse word order:

we are watching agricultural chemicals .
nous regardons les produits chimiques agricoles .

we are watching steel production .
nous regardons la production de acier .

we are watching energy supplies .
nous regardons les approvisionnements en énergie .

we are watching equipment supplies .
nous regardons les approvisionnements en matériel .

we are watching fertilizer supplies .
nous regardons les approvisionnements en engrais .

Grammar Induction (3)

Find differences:

we are watching **energy** supplies .
nous regardons les approvisionnements en **énergie** .

we are watching **equipment** supplies .
nous regardons les approvisionnements en **matériel** .

we are watching **fertilizer** supplies .
nous regardons les approvisionnements en **engrais** .

Make an equivalence class:

<CL_0>:
"energy" = "énergie"
"equipment" = "matériel"
"fertilizer" = "engrais"

And apply it, removing resulting duplicates:

we are watching agricultural chemicals .
nous regardons les produits chimiques agricoles .

we are watching <CL_0> supplies .
nous regardons les approvisionnements en <CL_0> .

we are watching steel production .
nous regardons la production de acier .

Grammar Induction (4)

Repeat process to get:
<CL_2>:
"<CL_0> supplies" = "les approvisionnements en <CL_0>"
"agricultural chemicals" = "les produits chimiques agricoles"

Word Decompounding

Some languages readily form compound words, unlike English, which causes a mismatch between languages:

German: Aortenisthmusstenose
English: aortic isthmus stenosis

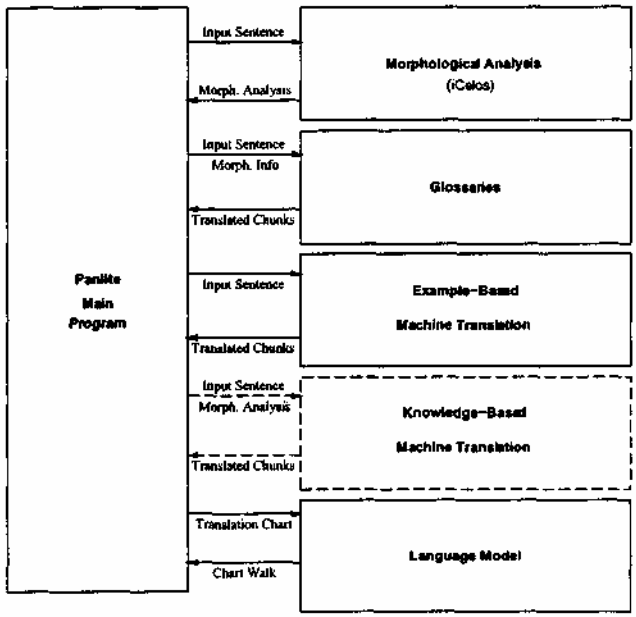
German: Krebspatienten
English: cancer patients

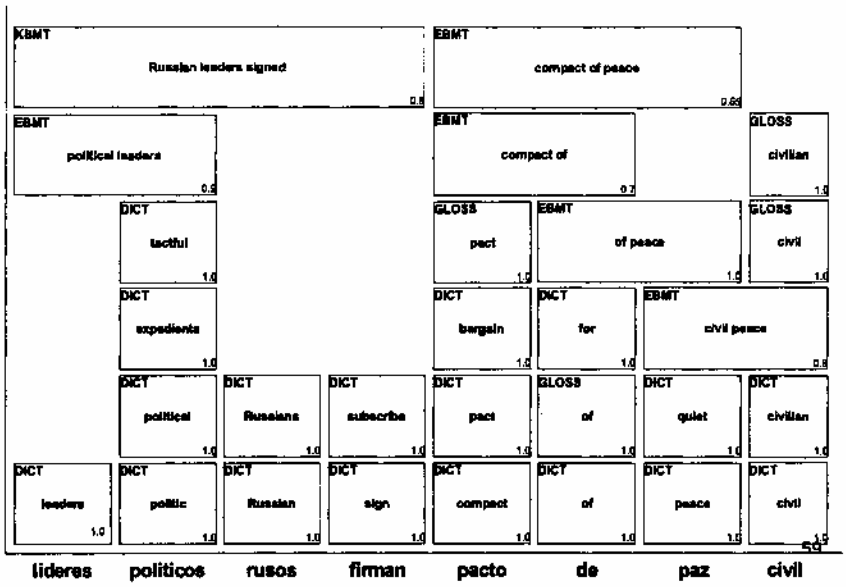
Particularly in technical domains, there may be a large percentage of cognate terms, which provides the possibility of learning how to split compounds by looking at the examples in a parallel corpus.

G-EBMT: Use in Multi-Engine MT

The G-EBMT engine was built from the ground up for use in a fine-grained multi-engine system.

- doesn't try to generate translations unless reasonably certain the translation is correct
- no need to worry about combining the partial translations
- no need to worry about selecting from among alternative translations





Applications of CMU's G-EBMT

- text translation
- speech translation systems
- cross-language information retrieval
- topic tracking

Text Translation

No current project specifically for developing EBMT, but it is used (and has been used) in numerous other projects at CMU:

- Pangloss (1995-1996)
- Mega-RADD: Rapidly-Adaptable Data-Driven translation (large amounts of data available)
- Milli-RADD: Rapidly-Adaptable Data-Driven translation (restricted data)
- AVENUE: translation for endangered languages
- Speech-to-Speech translation (next slide)

Speech-to-Speech Translation

- **DIPLOMAT (1996-1999)**
  Speech translation on a laptop: English-Croatian, English-Haitian Creole, initial work on English-Korean; later built English-Spanish from available data.
- **TONGUES (2000-2001)**
  Follow-on for US Army Chaplain School: English-Croatian, with field-test using naive native Croatian speakers in Zagreb.

Cross-Language Retrieval

Given a query in one language, find relevant documents in another.

When using MT, can either

- translate the query – suffers from lack of context; statistical word-for-word dictionary works best
- translate the document collection – likely to be impractical

We have performed experiments using EBMT and other corpus-based methods.

Current CLIR project: MUCHMORE (2000-2003)

Topic Tracking

Find news stories of interest, either

- the onset of a new event, or
- more about an event discussed by a specified story

– and do so across multiple languages!

Initial experiments using EBMT to translate Chinese news stories into English yielded better results than the provided translations generated by a commercial MT system.