9 October 2002

ralf+gcc.cs.cmu.edu

Carnegie Mellon University

Ralf Brown

A Tutorial

Translation

Example-Based Machine
Overview

- CMU's Generalized EBM'T system
- Hands-On Exercise
- Sample Systems
- Partial Matching in EBM'T
- Relationship between EBM'T and other techniques
- Types of EBM'T
- What is EBM'T?
cal translation. Other corpus-based methods include translation memories and statistical-EWT is closely tied to case-based reasoning. EWT is sometimes called Memory-Based, Similarity-Based, etc. learn how to translate a collection of pre-translated texts as training material to automatically evolve explicitly encode translation rules. Rather than having EWT is one of a variety of corpus-based methods.

What is Example-Based Translation?
Gennifer Flowers is said to have had an affair with President Clinton for many years.

Yesterday, 200 delegates met behind closed doors to discuss the new tax code.
Example: IBM’s TM2 system.

Manual translation can be translated by the TM, leaving only the modifications to be re-translation or manually. The parts which remain unchanged previously-translation documents — the parts which remain unchanged. Translation memory is most useful when translating revised versions of translation from scratch. If done well, this is still much faster than generating a translation. More sophisticated translation memories retrieve the nearest match (if "close enough") and let the user fix up the retrieved version. More sophisticated translation memories retrieve the simplest version: if we are given one of the units in the corpus, retrieve a tool to aid a human translator.

Translation Memory is not, in itself, a translation system, but rather a
was actually matched. Example base corresponds to the portion of the source sentence that was not retrieved and which piece of the translated sentence it is the target to be translated, and combine the pieces later. For this to work, we need a way of determining whether the nearest matching portion of the input sentence matches portions of both source and target language. This can require considerable knowledge of both source and target language to the translation.

**Alternative: Find the nearest exact matches of portions of the input sentence in the corpus, and determine how to transfer any remaining differences.** Translation memory can be generalized: find the nearest matching sentence.
7-8 MIPS. The PDA running this pre-processing power. The PDA might sport 2 MB main memory and a 50 MB hard disk, and provide 2-4 MIPS of processing power.

PC might sport 2 MB main memory and a 50 MB hard disk, and provide 2-4 MIPS of processing power.

Historical Perspective: In 1984, a workstation or high-end desktop further, but necessary computational resources were not yet available.

The idea of storing large numbers of translation examples goes back to a guide for the translation of the input sentence. The given input sentence and an example sentence, which can be a guide for the translation of the input sentence. It is to find out the similarity of the most important function. What is now known as EBMFT was first proposed in 1981 by Makoto Nagao in a paper titled “Translation by Analogy.”

Origins of EBMFT
Types of ELM:

- Deep (parse tree-based)
- Shallow (morphological / part-of-speech analysis)
- Less shallow (lexical)
The red ball rolled

"der rote Ball rollte"

Language-independent rep

(after Vauquois)

"der rote Ball rollte"

The red ball rolled

<det><adj><N><V><det><adj><N><V>

NPV S NPV S

The red ball rolled

<det> <adj> S<det> <adj> S

NPV

The red ball rolled

<det> <adj> <det> <adj>

"der rote Ball rollte"
intensive interlingua approach to translation. This is also one of the major advantages of the much more knowledge-immediately possible between any pair of the languages in the corpus. Although most EBM systems are trained on bi-lingual corpora, it a
EBMT Resources

as a means of validating translations
as a source of parallel text

The World Wide Web is becoming an important resource for EBMT:
synthetic parser, dependency parser, etc.
thesaurus for computing semantic similarity
bilingual dictionary
parallel text

Types of data/knowledge required by EBMT systems:
CMU's Generalized EBiMT system

Hands-On Exercise

(break)

Sample Systems

Partial Matching in EBiMT

Relationship between EBiMT and other techniques

Types of EBiMT

What is EBiMT?
output language model

• word re-ordering probabilities

• translation probabilities

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

Unlike EBM, does not retain original examples once trained.

• like EBM, trained from parallel text.

EBMT.

EBMT and SMT
inducing grammar rules •
inducing morphology rules •
finding equivalence classes among words •
extracting bilingual terminology •
on

much recent work has focused

matricially learn translation rules. In fact, much include a component to auto-
are hybrid systems which do, but may include a manuallly-written rules (there

EMT and rule-based systems
It is conceivable to create an example-based interlingual system for a simple and task-based (capturing only the essential meaning) interlingual and then to generate a translation from the interlingual representation, using EBM translation technique to convert text into an interlingual-based interlingual system.
Additionally, it can be used as a subroutine within a larger MT system.

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

ETMT has been combined with most other translation techniques.

Hybridization
system (paper to be presented Saturday morning)

Philippe Langlais and Michel Simard are working on a hybrid EBM/MT

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

Other techniques developed for statistical MT can also be applied to

dictionaries is using statistical techniques on the training corpus. One obvious way to generate such a
cross-language correspondences. One obvious way to find

Many EBM systems require some form of bilingual dictionary to find

MT

Hybrids: EBM + Statistical
input for which EDGAR has no examples

- CAT2 translates linguistic structures and those portions of the
  EDGAR provides word and phrase translations

- tight integration

  CAT2 implements a semantic theory

  EDGAR uses morphological and syntactic information

  CAT2 rule-based system + EDGAR EBM system

added (Carl et al. 1999)

A number of rule-based systems have had data-driven components

Hybrids: EBM + CAT2 + rule-based
Integration uses string-based TM with EDBAR as fallback

• EBM'T has broadest coverage

• String-based TM is very precise, but has low coverage

• Experimented with a string-based translation memory, a lexeme-

memory

Hybrids: EBM'T + translation
and new text

trained NN then scores nearness of match between training examples

neural network learns salient terms from parallel corpus

EBMT using connectionist matching

(Ian McLean 1992)

Hybrids: EBM + neural nets
late the input translation, in which case another engine is given a chance to trans-

• Fall-over: one primary engine is used unless it fails to produce a
- tion, and the best one is selected by an external process
- after-the-fact selection: each engine generates a complete translation
- negotiation
- tight coupling: selecting at a substantial level or using inter-engine

Three main approaches to multi-engine combination:

- engine’s weaknesses.
- engine (so that one engine’s strengths can compensate for another en-
- gine) behind multi-engine approaches is to combine multiple methods (en-
- gine methods have strengths and weaknesses, the idea

Hybrids: multi-engine

combinations
• CMU's Generalized EMN system
• Hands-On Exercise (break)
• Sample Systems
• Partial Matching in EMN
• Relationship between EMN and other techniques
• Types of EMN?
• What is EMN?
Additionally, there is the problem of boundary friction. To each other.

For a deep EBM system, parse trees must have their nodes matched between the halves of a training example. For a shallow system, this takes the form of word-level alignments.

Handling Partial Matches:

Any EBM system which permits partial matches against the training corpus needs a way of identifying corresponding segments between the two languages.
Word alignments work best when there is a one-to-one correspondence between words.

The mapping may be a strict binary decision or a set of probabilistic weights.

A word-level alignment between two sentences specifies, for each source language word, which (if any) target language words are produced by that word when the sentence is translated.

Word Alignment

Handling Partial Matches:
Word Alignment Difficulties:

Why weren't they treated yesterday?

- They were treated yesterday.

How to deal with word order variations?

Croatian requires adding the correct determiner.

Translating into English from a language without determiners (e.g.):

Krebshauten weren't treated.

- Cancer patients weren't treated.

Many-to-one mappings can cause extraneous information to be in-

Sie wurden gestern behandelt
His face was a / open book.

- Fragments may not show the correct agreement with each other
- One or more fragments may have the wrong case, number, etc.
- A necessary word
- Word level alignment may have included extraneous words or missed
- Resilient partial translations may not "fit together" properly

When translating based on multiple partial matches of the input, the

Boundary Friction
CMU's Generalized EMT system

Hands-On Exercise

(break)

Sample Systems

Partial Matching in EMT

Relationship between EMT and other techniques

Types of EMT

What is EMT?
Overview of EBM

• CMU: G-EBMT
• Imamurah: HPA/HPAT
• Sumita: D3
• Guvener & Cicelkii: Generalized EBM
• Brona Colliums: Reverb
• Michael Carli: EDGAR
• Veale & Wav: Galiin
• Early systems: Sumita et al.
among translation candidates. 
- Hand-coded thesaurus for computing semantic distance to select.
- To match previously-unseen text for an example provide the ability to replace portions of the sentence.
- Correspondence points between source- and target-language trees.
- Operated on dependency trees.

Satoshi Sato and Makoto Nagao (1990)

Early EM algorithms
For the most specific common abstraction
leveled the semantic distance of the nouns, searching up the hierarchy
system used a commercial thesaurus of everyday Japanese and calcu-

**noun1 of noun2**

in most contexts, the English translation is

**noun1 no noun2**
translated only Japanese phrases of the form

Eiichiro Sumita et al. (1991, 1993)

Early EMT Systems (2)
can be templated partial matching); however, phrases within the translation example

matching performed at the level of complete tag sequences (no speech tags

transliteration examples converted into templates consisting of part-of-

part-of-speech tagging in both languages

German-English translation

System: Gaijin
segment would consist entirely of marker words

markers not considered to start a new segment if previous/next

... - quantifiers: all, some, many,'

... - determiners: the, those, a, an,'

... - prepositions: in, out, on, with,'

phrasal segment

Garlin exploits such markers as signals for beginning and end of a

set of lexemes and morphemes

states that natural languages are marked for grammar by a closed

putative psycholinguistic constraint on grammatical structure

phrasal segmentation using Marker Hypothesis

System: Garlin (2)
non-contiguous mappings are considered unusable and will not be

visible

tiguous segments which all map to same segment in the other lan-

many-one segment mappings are (partially) handled by merging con-

„with“ and „mit“

possible segment correspondences between source and target are

Segment Alignment

System: Gašicín (3)
uses a compound variable on the target side.

1. When source segments need to be merged, the system
   side only.
   • Simplify lookups. Segment merging is represented in the target
   and retained in the template literally.

2. Interchange marker words are removed from the varialized segment.

3. Erasing a template for the translation example:
   • all well-formed segment mappings are converted into variables.

System: Caliin (4)
K: Durch Klettan�t \{prep A\} det C \{det B\} \{prep B\} zum Klotorieren ändern ändern

\textbf{Example Template}

\textbf{System: Gaßen (5)}
– ʔ-prep-det
– „the extruded surfaces”
– „for coloring”
– „displays controls”

Prior example would be indexed under
under the sequence of marker-word types

Examples indexed under both the phrasal chunks they contain and
Retrieving Examples

System: Gaijin (6)
which the template was formed. The most words with the phrase that was in the original from
when multiple options are available, choose the one which shares
possible
that the replacement is as compatible with the template position as
Galilin tries to minimize boundary friction during grafting by ensuring
words in a target segment
Keypole surgery: replacing or morphologically fine-tuning individual
ent example
grafting: replacing one phrasal segment with another from a different
Adaptation System: Galilin (7)
target language

matched chunks from case base are re-specialized and refined in the
multiple levels of generalization

induces translation templates from analyzed reference translations

applies morphological analysis to both languages

Michael Carlet et al., University of Saarbrücken

System: EDGAR
retrieved examples are adapted to fit the text to be translated.

Retrieval criterion is combination of similarity and adaptability.

That feature lists shallow processing than original Nagao/Sato approach, using
tation training examples are abstracted to syntactic dependency repre-

English-German, Irish-English translation

(Brona Collins 1996, 1999)

System: Reverb
denotes in the case base. A translation dictionary is generated from the word-to-word correspon-

dences in cases and chunks. 

Individually word types have separate WORD objects indexing their occurrences in cases, and each sentence pair is stored as a case; cases refer to chunks, which may be replaced on adaptation. 

Knowledge Representation

System: ReVerb (2)
The threshold of adaptability: Coverage vs. Accuracy tradeoff can be set at run-time by selecting a balance between levels of linguistic description that do not assume modularity for restricted domains. "Careful" generalization is needed, which merely masks the surface details of chunks and does not assume modularity:

- Functional equivalence on either side of chunk
- Translation probability between SL and TL words in chunk
- Heuristic determination: otherwise adapted examples are generalized where chunks can "safely" be replaced or

Template Creation

System: Reverb (3)
more work: glue together chunks from different examples
context
most reliable: re-parsing an entire chunk with another from the same
Adaptation

any
statistics used to increase the likelihood of a good chunk bound-

extend chunks to include additional word not otherwise cov-

Fragmentation
chunk-boundary adjustments

case-based parsing to generate chunks

a bilingual dictionary; chunks will be aligned using linkage pattern

bilingual alignment and linking of possibly-corresponding words using

Case Creation

System: Reverb (4)
Keyhole adaptation within a chunk
chunk-level adaptation dictionary
- Related to the compositionality of the solution
- Must be transferred across SL-TL links

In a particular case given that the source-language differences must
adaptation-safety knowledge – a quantification of the risk ofchoosing
only one change
multiple adaptations than a poorly-adaptable case that requires
- It is often better to retrieve an easily adaptable case that requires
score

Retrieval metric is a combination of similarity score and adaptibility
Case Retrieval and Adaptation

System: Reverb (5)
similar sentence pairs should correspond

• Generalization based on the heuristic that differences in mostly
  word stems and morphemes by co-indexed variables
• Training examples are abstracted into templates by replacing certain

(1996-)

System: Guverir & Cicelli
Mary+DAT X'T+ACC VER+PAST+1SG
I give+PAST the X'S to Mary

System: Güvenir & Çicekli (2)

Template: (Çicekli & Güvenir, 1996)
examples, about 80% good quality
90% coverage of “travel conversation” sentences with 200K training
adapts examples by substituting target words for variables
used pattern generates translation patterns on the fly, selects most commonly
similarity metric includes edit and semantic distance
Eiichiro Sumita (2001)
D3: DP-matched DF driven transducer

System: D3
125K training examples
about 70% good quality translation of "travel" sentences using
translate leaves of tree using a dictionary
then parse source using source patterns, map to target patterns, then
generate transfer patterns from HPA-processed corpus
HPAT: HPA-based Translation

trees
parse failures cause problems; try to alleviate by combining partial

same syntactic category – corresponding content words
works by finding equivalent phrases from bilingual text
Keni! Imamura (2001) Hierarchical Phrase Alignment

**System:** HPAT/HPAT
• CMU's Generalized EBM T system
• Hands-on Exercise
  (break)
• Sample Systems
• Partial Matching in EBM T
• Relationship between EBM T and other techniques
• Types of EBM T
• What is EBM T?
• CMU's Generalized EMT system
• Hands-On Exercise
  (break)
• Sample Systems
• Partial Matching in EMT
• Relationship between EMT and other techniques
• Types of EMT
• What is EMT?
Hands-On Exercise

**Emulate an EMT system**
**Emulate a translation memory**
**Distribute bilingual corpus to tutorial participants**

2. Das Ergebnis der Steuererhebung wird am Donnerstag offiziell bekanntgegeben.

1. Die Arbeitslosenquote sank von 10,7 auf 9,3 Prozent.

Find close matches for the text.

Translation Memory

Exercise Number 1:
Fliechtlinie aus Frankfurt am Main.

Voraussichtliche Reduzierung der Abschreibungsbasis

Nach meinen Berichten der plant Verkehrsminister

Find examples containing phrases from

lexicon EBM

Exercise Number 2:
Find examples for a template matching...
What is EBMT?

Types of EBMT

Relationship between EBMT and other techniques

Sample Systems

Partial Matching in EBMT

Hands-On Exercise

CMU's Generalized EBMT system
System

CMU's Generalized EBM

• multi-engine

  - automatically (machine learning)

  - manually

• generalizing into templates

• inexact matching

• simple lexical match
Gennifer Flowers is said to have had an affair with President Clinton for many years.

Yesterday, 200 delegates met behind closed doors to discuss the new tax code.

Translational Sentence (Target)

Gennifer Flowers hat angeblich jahrelang eine Affäre mit President Clinton gehabt.

Alignment

Matches Found

New Sentence (Source)
Problem: how to produce good-quality general
  Advantage: requires less training text
  Disadvantage: requires large amounts of training text
  Advantage: little or no need for linguistic knowledge

Shallow processing:

G-EBMT: Lexical Matching
Improve quality when more data was available.

This fuzzy matching proved helpful on limited training data, but did not
guarantee as the next translation
has its most-common translation occurring more than twice as fre-

• has only one translation listed in the dictionary
the word either

unambiguously means that translation known for that word. Reasonably unambiguously means that the middle of a match, provided there is a reasonably unambiguous

A recent addition to the system is allowance for a one-word gap in

can make a match where not all words are matched.

G-EMT: Inexact Matching
to find the best-scoring substring of the translation.

- words known to translate as empty string
- difference in length
- common location in sentence

functions such as dictionary. It then uses the translations along with heuristic scanning
to perform word-level alignment, the EBM system needs a bilingual
text.

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

**Word-Level Alignment**

**G-EBMT:**
Pruning Correspondences
times the two words co-occurred. Where \( \text{C is the number of } \)
\( P(W_i|W_j \wedge \text{C}) \) and \( P(W_j|W_i \wedge \text{C}) \)

The threshold is based on Mutual Conditional Probability:

Errors by sacrificing some words – we can get a larger vocabulary at the cost of more errors, or reduce

Statistical dictionaries can be tuned: there is a size/accuracy tradeoff.

and output any remaining entities as probable mutual translations. We can extract bilingual dictionaries such as the one required for word-

**Term-Substitution Dictionary**
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))
(COMPLEMENTARTY (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))
The NASDAQ was up 13.5 points on Tuesday.

The Dow Jones was down 7 points on March 3rd.

The market was dir<quantity> on date<date>.

*SOURCE*

well-placed source

company spokesman

spokesperson

significant source
Generalization

G-EMT: Manual
<table>
<thead>
<tr>
<th>Semantic:</th>
<th>Syntactic:</th>
</tr>
</thead>
<tbody>
<tr>
<td>shapes</td>
<td>They may be changed slightly. They may be used interchangeably.</td>
</tr>
<tr>
<td>colors</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
<tr>
<td>first-person verbs</td>
<td></td>
</tr>
<tr>
<td>names of cities</td>
<td></td>
</tr>
<tr>
<td>plural adjectives</td>
<td></td>
</tr>
<tr>
<td>days of the week</td>
<td></td>
</tr>
<tr>
<td>masculine nouns</td>
<td></td>
</tr>
<tr>
<td>numbers</td>
<td></td>
</tr>
<tr>
<td>or syntactically</td>
<td></td>
</tr>
</tbody>
</table>
When translating, perform the same substitutions, but remember the appropriate translation for each occurrence. Match the resulting term against the indexed corpus, and substitute the remembered transliterations into the translated template.

25 players met in London yesterday.

Given a set of equivalence classes, replace each occurrence in the train-

Equivalence classes (1)

G-EBMT Generalization:

<number> players met in <city>.

<time> players met in <city>.
the proper translation for the abstracted words.

Even though the example on the previous slide would not have matched

```ruby
>12 players met in Paris last Tuesday.

>numbert players met in city time.
```

Thus, we try matching not only the surface form, but also the template.

**Equivalence Classes (2)**

G-EBMT Generalization:
12 Spieler trafen sich Letzen Diesttag in Paris.

The translated template:

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

Equivalence Classes (3)

G-EMT Generalization:
A paired production-rule grammar to be created.

Pattern Replacement

G-EMT Generalization:
recursive replacement

G-EBMT generalization:
Generalize

G-EMTL: Learning How to

- Word decompounding
- Grammar induction
- Single-word equivalence classes via clustering

Three different learning mechanisms have been implemented to date:

- 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.
Information into the clustering.

**Solution**: use the approach of Baracchina and Villar to inject bilingual

of bilingual pairs.

**Problem**: this yields only a monolingual clustering, but we need a set

identifiable.

**Approach**: create a pseudo-document for each word, containing all

clustering.

we can use standard document-clustering techniques to perform word

sum of the words in the immediate neighborhoods of its occurrences,

**Observation**: if the context in which a word appears is defined as the

**Single-Word Equivalences**
3. Treat those word pairs as indivisible tokens in further processing.

2. Whenever there is a unique correspondence indicated by the bi-text.

1. Use a bilingual dictionary to create a rough bi-text mapping between the source-language and target-language halves of a sentence pair.

Into Monolingual Clustering

Interacting Bilingual Information
For example, word into its sense.

These bilingual word pairs also serve to provide a rough separation of a
<table>
<thead>
<tr>
<th>Subject</th>
<th>Liberal</th>
<th>Liberal</th>
<th>Democratic</th>
<th>Democratic</th>
<th>Tory</th>
<th>Conservative</th>
<th>Conservative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>EVIDENCE</td>
<td>EVIDENCE</td>
<td>CLEARLY</td>
<td>CLEARLY</td>
<td>FACON</td>
<td>CONSERVATIVE</td>
<td>CONSERVATIVE</td>
</tr>
<tr>
<td>Clarity</td>
<td>Reality</td>
<td>Surely</td>
<td>Certain</td>
<td>Certainly</td>
<td>Only</td>
<td>Probably</td>
<td>PROBLEMAVEL</td>
</tr>
<tr>
<td>Event</td>
<td>VRAIMENT</td>
<td>SURREMENT</td>
<td>VRAIMENT</td>
<td>SURREMENT</td>
<td>REN</td>
<td>ONE</td>
<td>PROBABLEMENT</td>
</tr>
<tr>
<td></td>
<td>CLOTHING</td>
<td>RETRAITES</td>
<td>SURREMENT</td>
<td>CERTAIN</td>
<td>TING</td>
<td>ONLY</td>
<td>PROBABLEMENT</td>
</tr>
<tr>
<td></td>
<td>VETERENS</td>
<td>RETRAITES</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>NEVER</td>
<td>TING</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>CLOTHING</td>
<td>RETRAITES</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>NEVER</td>
<td>TING</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>VETERENS</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>CLOTHING</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>VETERENS</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>CLOTHING</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>VETERENS</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>CLOTHING</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
<tr>
<td></td>
<td>VETERENS</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAIN</td>
<td>CERTAINMENT</td>
</tr>
</tbody>
</table>
are $S_1$ and $S_2$ if appropriate.
The various instantiations of $D$ are added to an equivalence class, as

... $S_1$ $D$ $S_2$

The initial implementation only searches for the pattern
to find constituents that can be used interchangeably.
Thus, we can search a corpus for patterns of similarity and dissimilarity.

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

Observation: Similar sentences in a corpus tend to differ by concrete

Grammar Induction
Grammar Induction (2)
Grammar Induction (3)

And apply it, removing resulting duplicates:

Make an equivalence class:

Find differences:
Grammar Induction (4)
Corpus Size vs. EBMT Coverage (Spanish)

- String match only
- Grammar induction
- Word clustering
- Induction + clustering

Corpus Size (millions of words)

EBMT Coverage (percent)
By looking at the examples in a parallel corpus, compounds of English terms, which provides the possibility of learning how to split cognate terms, particularly in technical domains, there may be a large percentage of.

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer patients</td>
<td>Krebspatienten</td>
</tr>
<tr>
<td>aortic isthmus stenosis</td>
<td>Aortenisthmussstenose</td>
</tr>
</tbody>
</table>

causes a mismatch between languages. Some languages readily form compound words, unlike English, which...
Word Decomposition (2)

- Letter pairings, such as C with K.
- Generalization: allow varying weight for related but non-identical documents.

- Customary
- documented

- allowing letters to be skipped:

- simplest form counts number of common in-sequence characters.
- a form of longest common substring

- Cognate Scoring
also use a dictionary translation of one or both to find non-cognate word pairs need not be composed of the original words; one can

For selected words, use the word pair that gave the highest score.

Select words for which some pair has a cognate score above threshold.

Select words with highest score similarity with words in the compoundind language.

Concatenate adjacent words in the non-compounding language and

Finding Candidate Compounds

Word Decomposing (3)
Word Decomposition (4)
### Results

<table>
<thead>
<tr>
<th>Decomponunded (Feed-S-S)</th>
<th>Decomponunded (dict-S)</th>
<th>Decomponunded (base)</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>75.61%</td>
<td>75.14%</td>
<td>74.60%</td>
</tr>
<tr>
<td>2.963</td>
<td>81.74%</td>
<td>81.21%</td>
<td>78.17%</td>
</tr>
<tr>
<td>2.943</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.992</td>
<td>80.81%</td>
<td>80.71%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>71.31%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matches (Words) (Words)</th>
<th>Coverage (Avg Len)</th>
<th>German-English EBM Fit Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.8%</td>
<td>0.150, 726</td>
<td>943, 290</td>
</tr>
<tr>
<td>6.6%</td>
<td>64.306, 227</td>
<td>644, 224</td>
</tr>
<tr>
<td>7.4%</td>
<td>82.91, 147</td>
<td>828, 147</td>
</tr>
<tr>
<td>6.8%</td>
<td>109.559, 604</td>
<td>482, 604</td>
</tr>
<tr>
<td>7.4%</td>
<td>128.847, 231</td>
<td>665, 231</td>
</tr>
<tr>
<td>4.6%</td>
<td>100.540, 415</td>
<td>521</td>
</tr>
<tr>
<td>1.0%</td>
<td>96.960, 383</td>
<td>0, 120</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Types Tokens Error Rate</th>
<th>Decomponunded Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run</td>
<td></td>
</tr>
</tbody>
</table>

---
no need to worry about selecting from among alternative translations

• no need to worry about combining the partial translations translation is correct

• doesn't try to generate translations unless reasonably certain

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

Use in Multi-Engine MT

G-EBMT:
lideres rusos firman pacto de paz civil politicos 

leaders Russians compact of peace of civilian leaders 

Putin's political leaders signed a compact of peace for civilian leaders.
C-EBMT

Applications of CML's
Speech-to-Speech translation (next slide)
AVENU.E: translation for endangered languages
data
Milli-RADD: Rapidly-Adaptable Data-Driven Translation (restricted)
Large amounts of data available
Mega-RADD: Rapidly-Adaptable Data-Driven Translation

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

Text Translation
Tongues: Follow-up for US Army Chaplain School: English-Croatian, with field-test using native Croatians speakers in Zagreb.

Tongues (2000-2001)

From available data, initial work on English-Korean; later built English-Spanish Creole, Haitian-Haitian & Haitian-English Speech translation on a laptop: English-Croatian, English-Haitian

Speech-to-Speech Translation

methods.

We have performed experiments using EBM and other corpus-based

• translate the document collection – likely to be impractical

for-word dictionary works best

• translate the query – suffers from lack of context; statistical word-

When using MT, can either

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

Cross-Language Retrieval
by a commercial MT system.

English yielded better results than the provided translations generated initially experiments using EBM to translate Chinese news stories into

- and do so across multiple languages!
- more about an event discussed by a specific!ed story
- the onset of a new event, or
- Find news stories of interest, either

**Topic Tracking**