# Example-Based Machine Translation

# A Tutorial

Ralf Brown

Carnegie Mellon University

ralf+@cs.cmu.edu

9 October 2002

# Overview

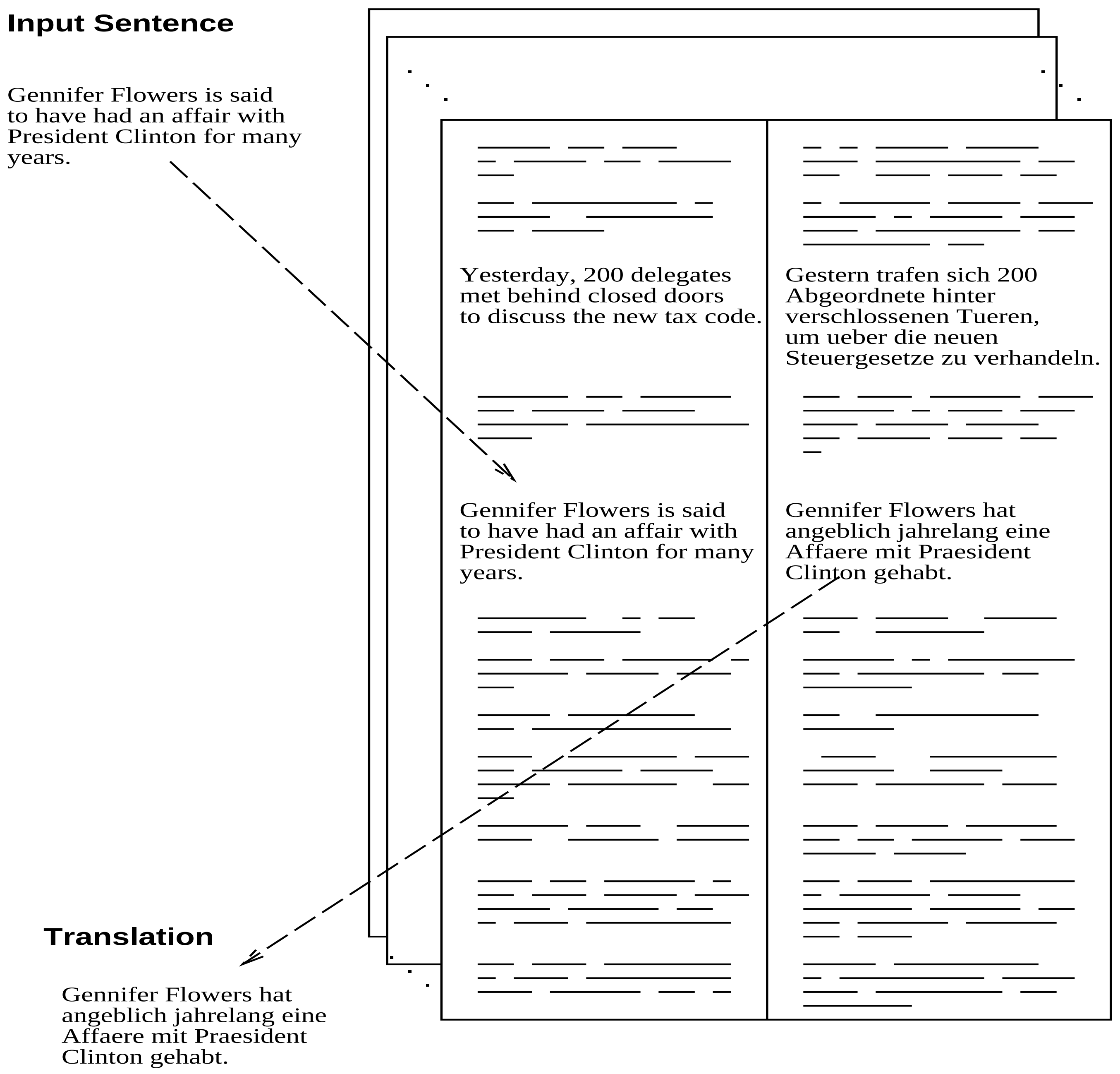
- What is EBMT?
- Types of EBMT
- Relationship between EBMT and other techniques
- Partial Matching in EBMT
- Sample Systems
- (break)
- Hands-On Exercise
- CMU's Generalized EBMT system

# 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, Similarity-Based, etc.
- EBMT is closely tied to Case-Based Reasoning

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



**Input Sentence**

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.

Gestern trafen sich 200
Abgeordnete hinter
verschlossenen Tueren,
um ueber die neuen
Steuergesetze zu verhandeln.

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

Gennifer Flowers hat
angeblich jahrelang eine
Affaere mit Praesident
Clinton gehabt.

**Translation**

Gennifer Flowers hat
angeblich jahrelang eine
Affaere mit Praesident
Clinton gehabt.

# Translation Memory

Translation Memory is not, in itself, a translation system, but rather a tool to aid a human translator.

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 1981 by Makoto Nagao in a paper titled “Translation by Analogy”.

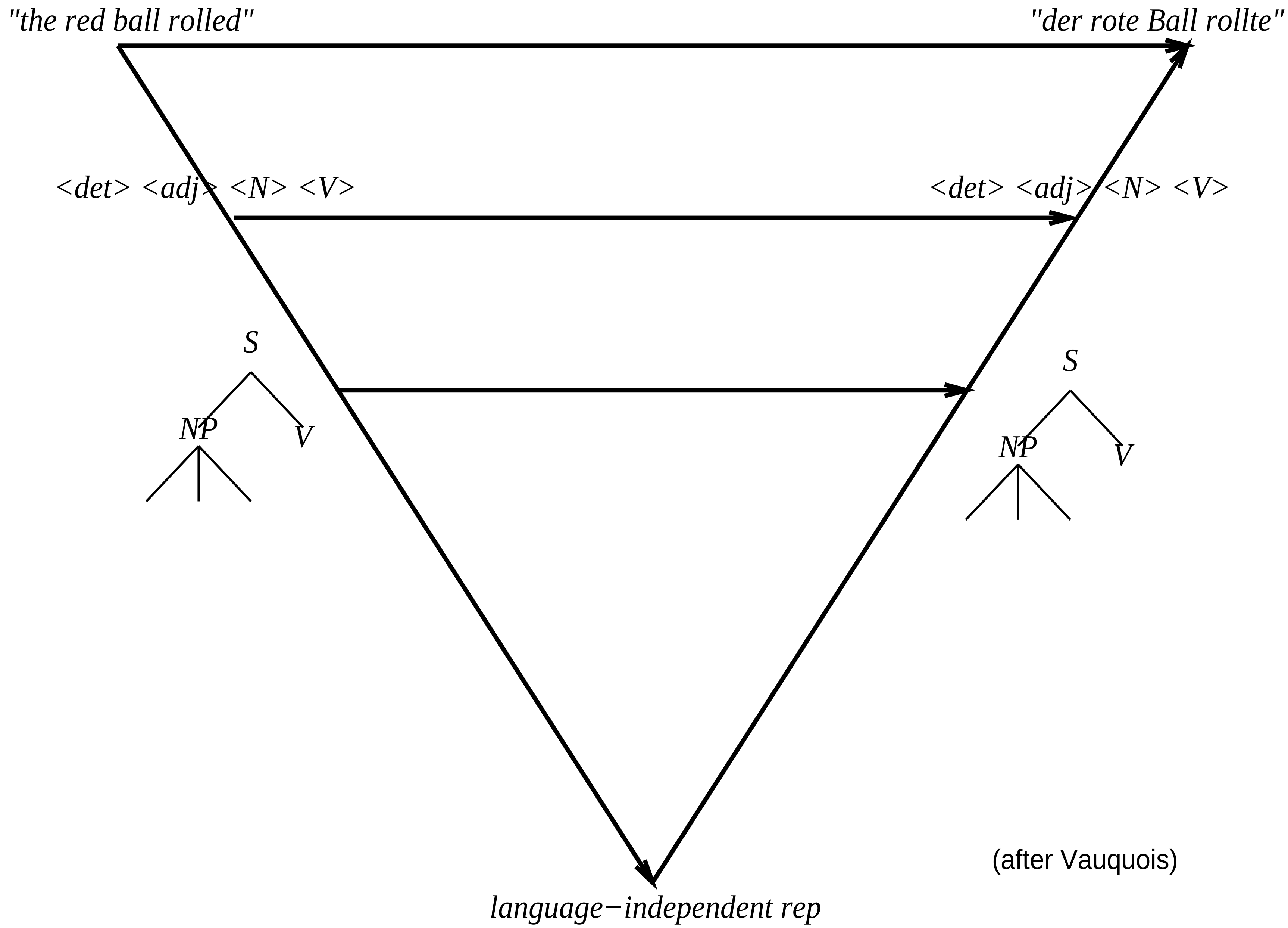
> *The most important function .. is to find out the similarity of the given input sentence and an example sentence, which can be a guide for the translation of the input sentence.*

The idea of storing large numbers of translation examples goes back further, but necessary computational resources were not yet available.

- *Historical Perspective:* in 1984, a workstation or high-end desktop PC might sport 2 MB main memory and a 50 MB hard disk, and provide 2-4 MIPS of processing power. The PDA running this presentation has 8 MB main memory, 128 MB secondary storage, and 7-8 MIPS.

# Types of EBMT

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



(after Vauquois)

# Multilingual EBMT

Although most EBMT systems are trained on bilingual corpora, if a multilingual corpus is available (or can be constructed), translation is immediately possible between any pair of the languages in the corpus. This is also one of the major advantages of the much more knowledge-intensive interlingua approach to translation.

# EBMT Resources

Types of data/knowledge required by EBMT systems:

- parallel text
- bilingual dictionary
- thesaurus for computing semantic similarity
- syntactic parser, dependency parser, etc.

The World Wide Web is becoming an important resource for EBMT:

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

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

# 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 has been combined with most other translation techniques:

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

Additionally, it can be used as a subroutine within a larger MT system.

# 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.

Philippe Langlais and Michel Simard are working on a hybrid EBMT/SMT system (paper to be presented Saturday morning)

# 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

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.

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

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

# Handling Partial Matches

Any EBMT system which permits partial matches against the training corpus needs a way of identifying corresponding segments between the two languages.

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

Additionally, there is the problem of *boundary friction*.

# Handling Partial Matches: Word Alignment

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.

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

Word alignments work best when there is a one-to-one correspondences between words.

# Word Alignment: Difficulties

- Many-to-one mappings can cause extraneous information to be included
  - cancer *patients were treated*
    *Krebspatienten wurden behandelt*
- How to deal with insertions?
  Translating into English from a language without determiners (e.g. Croatian) requires adding the correct determiner
- How to deal with word order variation?
  - *They were treated* yesterday
    *Sie wurden* gestern *behandelt*

# Boundary Friction

When translating based on multiple partial matches of the input, the resultant partial translations may not “fit together” properly:

- word level alignment may have included extraneous words or missed a necessary word
- one or more fragments may have the wrong case, number, etc.
- fragments may not show the correct agreement with each other
  - *His face was a / open book.*

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

# Overview of EBMT systems

- early systems: Sumita *et al*
- Veale & Way: Gaijin
- Michael Carl: EDGAR
- Brona Collins: ReVerb
- Guvenir & Cicekli: Generalized EBMT
- Sumita: $D^3$
- Imamura: HPA/HPAT
- CMU: G-EBMT

# Early EBMT Systems

Satoshi Sato and Makoto Nagao (1990)

- operated on dependency trees
- correspondence points between source- and target-language trees for an example provide the ability to replace portions of the sentence to match previously-unseen text
- hand-coded thesaurus for computing semantic distance to select among translation candidates

# Early EBMT Systems (2)

Eiichiro Sumita *et al* (1991,1993)

- translated only Japanese phrases of the form

  NOUN1 **no** NOUN2

- in most contexts, the English translation is

  NOUN1 **of** NOUN2

- system used a commercial thesaurus of everyday Japanese and calculated the semantic distance of the nouns, searching up the hierarchy for the most specific common abstraction

# System: Gaijin

(Veale & Way 1997)

German-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); however, phrases within the translation example can be templatized

# System: Gaijin (2)

Phrasal segmentation using Marker Hypothesis

- putative psycholinguistic constraint on grammatical structure
- states that natural languages are marked for grammar by a closed set of lexemes and morphemes
- Gaijin exploits such markers as signals for beginning and end of a phrasal segment
  - prepositions: in, out, on, with, ...
  - determiners: the, those, a, an, ...
  - quantifiers: all, some, many, ...
- markers not considered to start a new segment if previous/next segment would consist entirely of marker words

# System: Gaijin (3)

Segment Alignment

- possible segment correspondences between source and target are evaluated using segment length and word correspondence weights
- bonus for having leading marker of the same category type (e.g. “with” and “mit” )
- many-one segment mappings are (partially) handled by merging contiguous segments which all map to same segment in the other language
- non-contiguous mappings are considered unusable and will not be variablized

# System: Gaijin (4)

Templates

- all well-formed segment mappings are converted into variables, generating a template for the translation example
- infrequent marker words are removed from the variablized segment and retained in the template literally
- to simplify lookups, segment merging is represented in the target side only; when source segments need to be merged, the system uses a compound variable on the target side

# System: Gaijin (5)

## Template Example

```
E: Displays controls for coloring the extruded surfaces
G: Durch Klicken auf dieses Symbol lassen sich Optionen
  zum Kolorieren der extrudierten Flaechen anzeigen

{\bf Template}
E: {_ A} {prep B} {det C}
G: Durch Klicken auf {prep A} {prep B} {det C} anzeigen

{\bf Chunks}
A: Displays Controls
   dieses Symbol lassen sich Optionen

B: for coloring
   zum kolorieren

C: the extruded surfaces
   der extrudierten Flaechen
```

# System: Gaijin (6)

Retrieving Examples

- examples indexed under both the phrasal chunks they contain and under the sequence of marker-word types
- prior example would be indexed under
  - “displays controls”
  - “for coloring”
  - “the extruded surfaces”
  - ?-prep-det

# System: Gaijin (7)

Adaptation

- *grafting*: replacing one phrasal segment with another from a different example
- *keyhole surgery*: replacing or morphologically fine-tuning individual words in a target segment
- Gaijin tries to minimize boundary friction during grafting by ensuring that the replacement is as compatible with the template position as possible
  - when multiple options are available, choose the one which shares the most words with the phrase that was in the original from which the template was formed

# 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)

English-German, Irish-English translation

- 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: ReVerb (2)

Knowledge Representation

- corpus is converted into a Case Base
- each sentence pair is stored as a case; cases refer to chunks, which may be replaced on adaptation
- individual word types have separate WORD objects indexing their occurrences in cases and chunks

A translation dictionary is generated from the word-to-word correspondences in the case base.

# System: ReVerb (3)

Template Creation

- examples are generalized where chunks can “safely” be replaced or otherwise adapted
- heuristic determination:
  - translation probability between SL and TL words in chunk
  - functional equivalence on either side of chunk
- for restricted domains, “careful” generalization is used, which merely masks the surface details of chunks and does not assume modularity between levels of linguistic description

Coverage vs. Accuracy tradeoff can be set at run-time by selecting a threshold of adaptibility.

# System: ReVerb (4)

Case Creation

- bitext alignment and linking of possibly-corresponding words using a bilingual dictionary; chunks will be aligned using linkage pattern
- case-based parsing to generate chunks
- chunk-boundary adjustments
  - fragmentation
  - extending chunk to include an additional word not otherwise covered
  - statistics used to increase the likelyhood of a good chunk boundary

Adaptation

- most reliable: replacing an entire chunk with another from the same context
- more work: glue together chunks from different examples

# System: ReVerb (5)

Case Retrieval and Adaptation

- retrieval metric is a combination of similarity score and adaptibility score
  - it is often better to retrieve an easily adaptable case that requires multiple adaptations than a poorly-adaptable case that requires only one change
- adaptation-safety knowledge – a quantification of the risk of choosing a particular case given that the source-language differences must be transferred across SL-TL links
  - related to the compositionality of the solution
- chunk-level adaptation dictionary
- keyhole adaptation within a chunk

# 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

# System: $D^3$

$D^3$: DP-match Driven transDucer

Eiichiro Sumita (2001)

- similarity metric includes edit and semantic distance
- generates translation patterns on the fly, selects most commonly used pattern
- adapts examples by substituting target words for variables
- 90% coverage of "travel conversation" sentences with 200K training examples, about 80% good quality

# System: HPA/HPAT

Kenji Imamura (2001) Hierarchical Phrase Alignment

- works by finding equivalent phrases from bilingual text
  - corresponding content words
  - same syntactic category
- parse failures cause problems; try to alleviate by combining partial trees

HPAT: HPA-based Translation

- generate transfer patterns from HPA-processed corpus
- parse source using source patterns, map to target patterns, then translate leaves of tree using a dictionary
- about 70% good quality translation of “travel” sentences using 125K training examples

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

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

# Hands-On Exercise

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

# Exercise Number 1: Translation Memory

Find close matches for the text

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

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

3. In diese Projekte seien seit 1991 insgesamt 5 Mio. Mark investiert worden.

# Exercise Number 2: Lexical EBMT

Find examples containing phrases from

1. Nach einem Bericht der ”Bild am Sonntag” plant Verkehrsminister Wissmann am kommenden Donnerstag die Abschibung bosnischer Fluechtlinge aus Frankfurt am Main.

# Exercise Number 3: Templatized EBMT

Find examples for a template matching

- Die Besprechungen wuerden an einem noch nicht angesagtem Ort stattfinden.

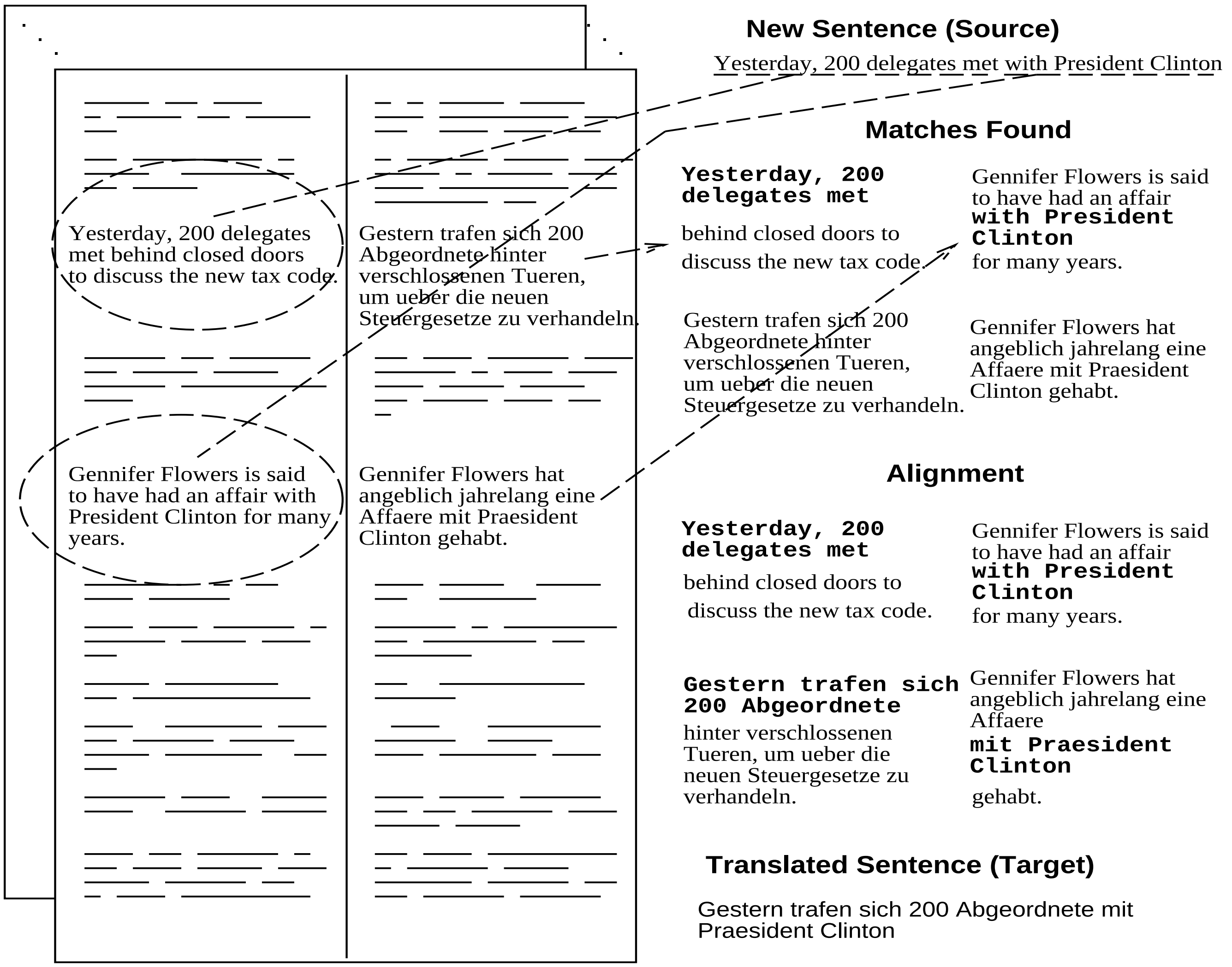
  $NP_1$ *wuerden an* $NP_2$ *stattfinden.*<br>
  $NP_1$ = *Die Besprechungen*<br>
  $NP_2$ = *einem noch nicht angesagtem Ort*

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

# CMU’s Generalized EBMT System

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

# EBMT Paradigm



**New Sentence (Source)**

Yesterday, 200 delegates met with President Clinton

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

Gestern trafen sich 200 Abgeordnete hinter verschlossenen Tueren, um ueber die neuen Steuergesetze zu verhandeln.

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

Gennifer Flowers hat angeblich jahrelang eine Affaere mit Praesident Clinton gehabt.

**Matches Found**

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

Gennifer Flowers is said to have had an affair **with President Clinton** for many years.

Gestern trafen sich 200 Abgeordnete hinter verschlossenen Tueren, um ueber die neuen Steuergesetze zu verhandeln.

Gennifer Flowers hat angeblich jahrelang eine Affaere mit Praesident Clinton gehabt.

**Alignment**

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

Gennifer Flowers is said to have had an affair **with President Clinton** for many years.

**Gestern trafen sich 200 Abgeordnete** hinter verschlossenen Tueren, um ueber die neuen Steuergesetze zu verhandeln.

Gennifer Flowers hat angeblich jahrelang eine Affaere **mit Praesident Clinton** gehabt.

**Translated Sentence (Target)**

Gestern trafen sich 200 Abgeordnete mit Praesident Clinton

Maximal−Length match of source substrings and concatenation of intra−sentence aligned text

# 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

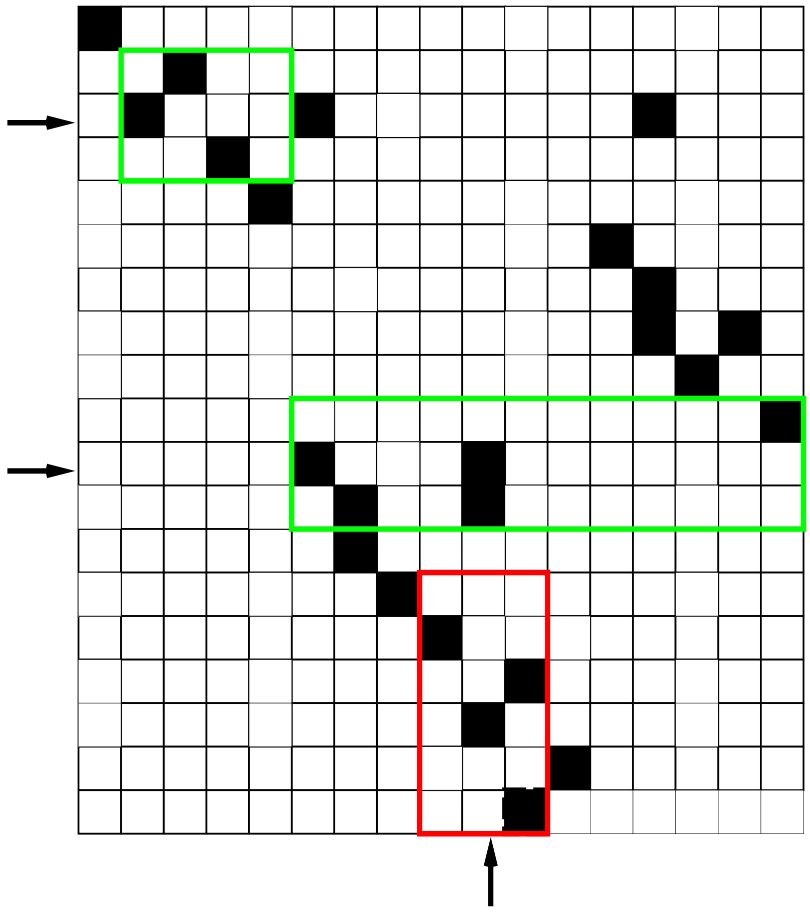
When the system partially matches a training example, the hard part is determining which portion of the translation corresponds to the matched text.

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

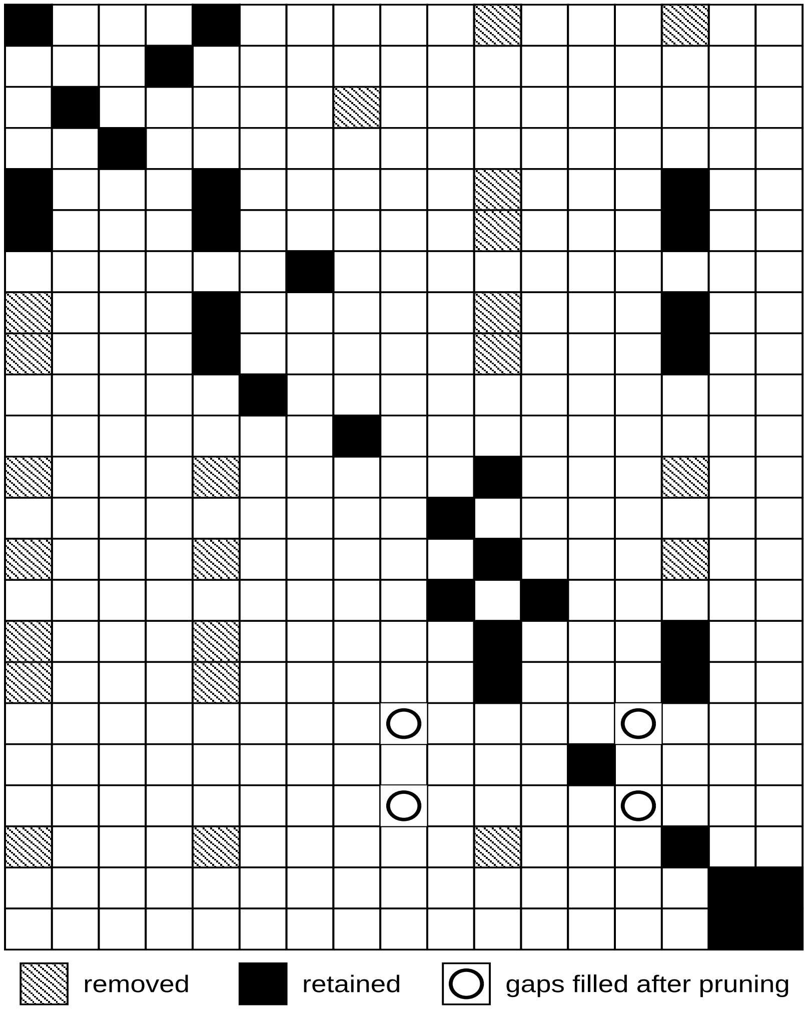
- common location in sentence
- difference in length
- words known to translate as empty string

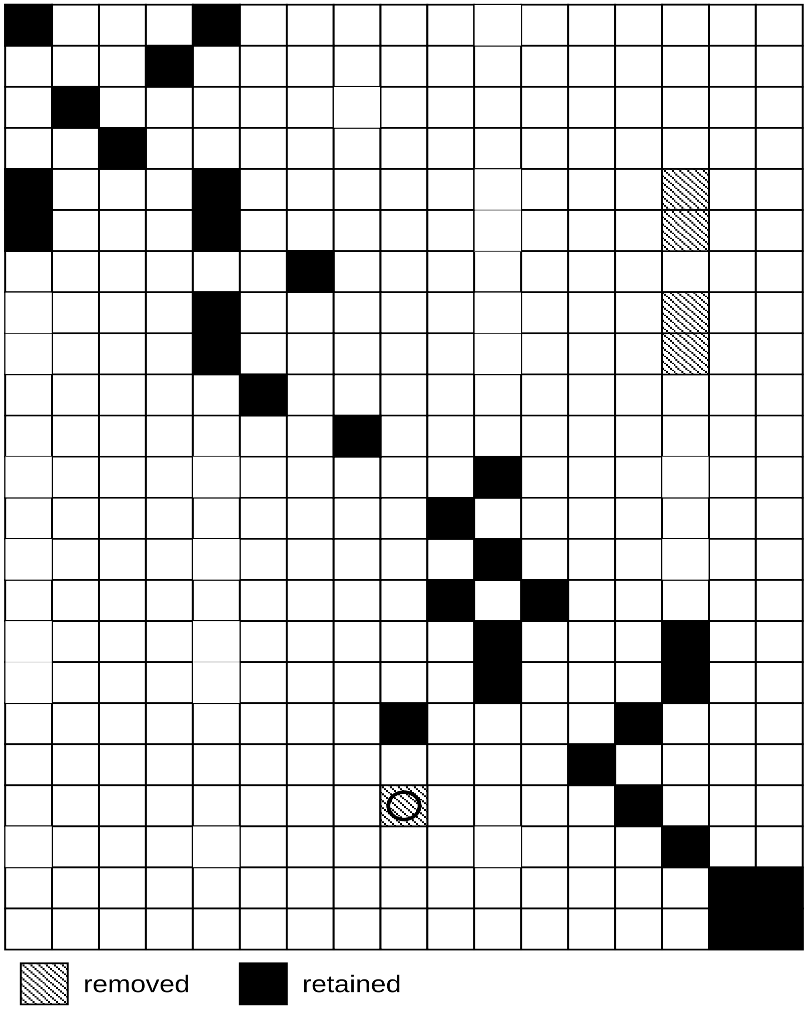
to find the best-scoring substring of the translation.

# 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.

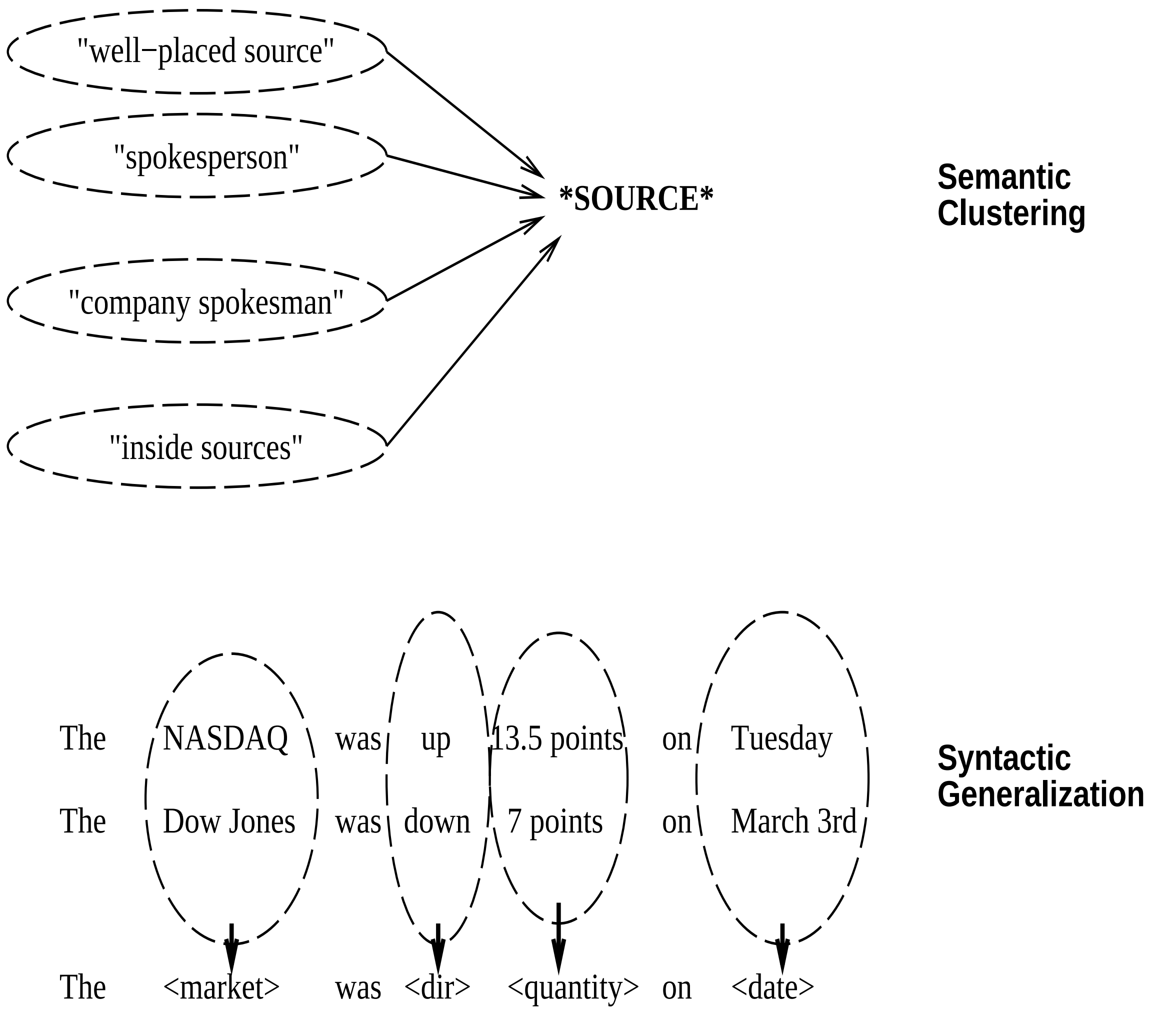
# Sample Dictionary

```
(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


"well−placed source"
"spokesperson"
"company spokesman"
"inside sources"
*SOURCE*
Semantic Clustering
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>
Syntactic Generalization

# 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

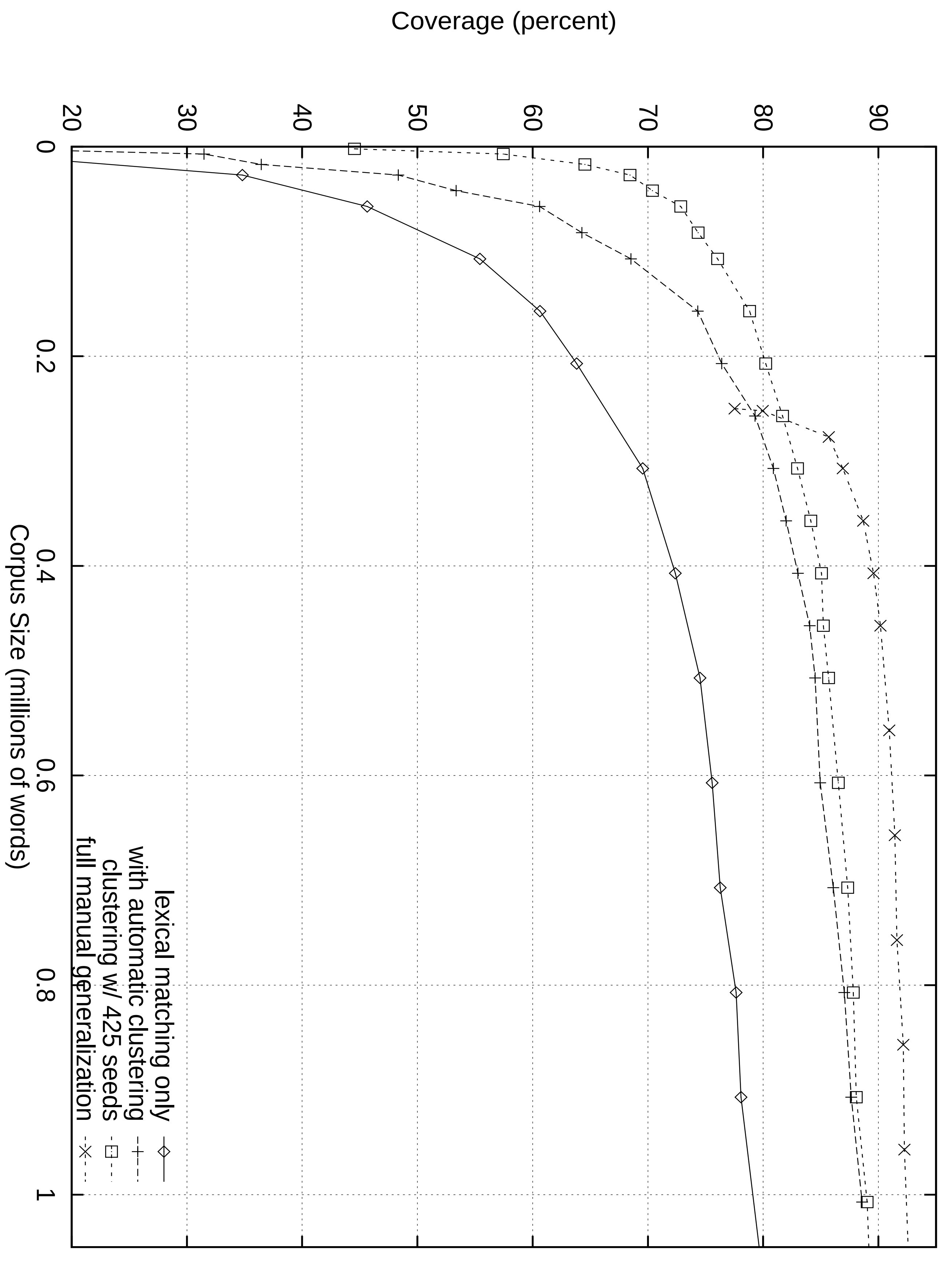
These bilingual word pairs also serve to provide a rough separation of a word into its senses.

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 | | |


Coverage (percent)
20
30
40
50
60
70
80
90
Corpus Size (millions of words)
0
0.2
0.4
0.6
0.8
1
lexical matching only
with automatic clustering
clustering w/ 425 seeds
full manual generalization

# 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 .

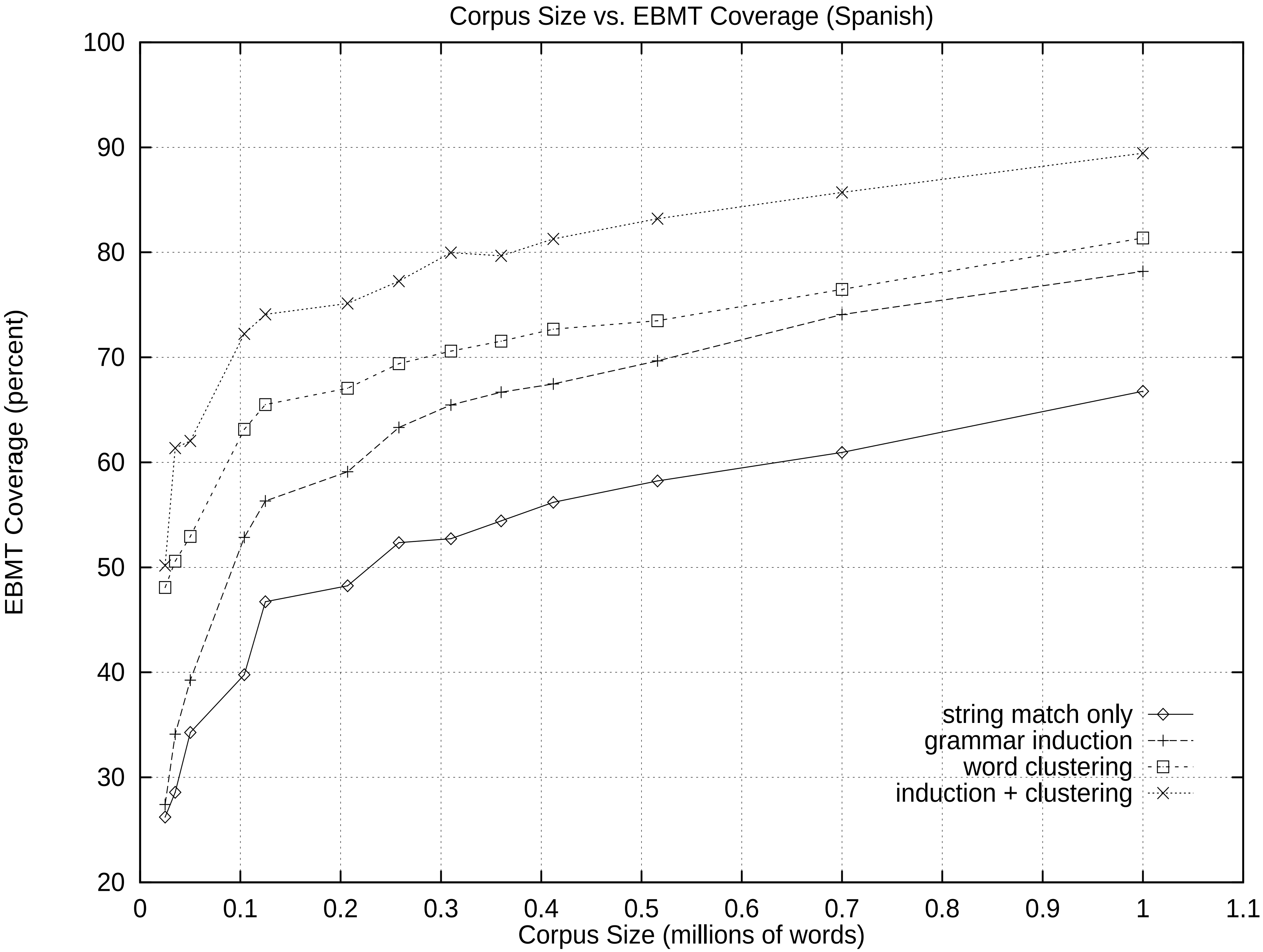
we are watching <CL_0> supplies .
nous regardons les approvisionnements en <CL_0> .

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"


Corpus Size vs. EBMT Coverage (Spanish)
EBMT Coverage (percent)
100
90
80
70
60
50
40
30
20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
Corpus Size (millions of words)
string match only
grammar induction
word clustering
induction + clustering

# 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.

# Word Decompounding (2)

Cognate Scoring

- a form of Longest Common Substring
- simplest form counts number of common in-sequence characters, allowing letters to be skipped:
  - **do**c**umente**d  
    **do**k**ument**i**e**rt
- generalization: allow varying weight for related but non-identical letter pairings, such as $C$ with $K$.

# Word Decompounding (3)

Finding Candidate Compounds

- concatenate adjacent words in the non-compounding language and score similarity with words in the compounding language
- select words for which *some* pair has a cognate score above threshold
- for selected words, use the word pair that gave the highest score
- word pairs need not be composed of the original words; one can also use a dictionary translation of one or both to find non-cognate compounds

# Word Decompounding (4)

Finding Split Point

- find leftmost position that maximimize similarity with first word of pair, and rightmost position that maximizes similarity with second word
- if those two positions coincide, split there
- if we have a gap,
  - split after hyphen, if present in gap
  - prefer location in gap that produces known words
- if we have an overlap
  - split after hyphen, if present in overlap
  - if one word of the pair exactly matches prefix or suffix of compound, split there
  - try dropping a letter and split if new boundaries coincide

# Decompounding: Results

Decompounder Performance

| Run | Types | Tokens | Error Rate |
|---|---|---|---|
| Baseline | 66,960 | 383,120 | 1.0% |
| Dict-Single | 100,540 | 415,521 | 4.6% |
| Dict-Full | 128,847 | 665,231 | 7.4% |
| Feedback-S-S | 109,559 | 482,604 | 6.8% |
| Feedback-S-F | 143,151 | 828,147 | 7.4% |
| Feedback-F-S | 116,306 | 644,224 | 6.6% |
| Feedback-F-F | 150,726 | 943,290 | 11.8% |

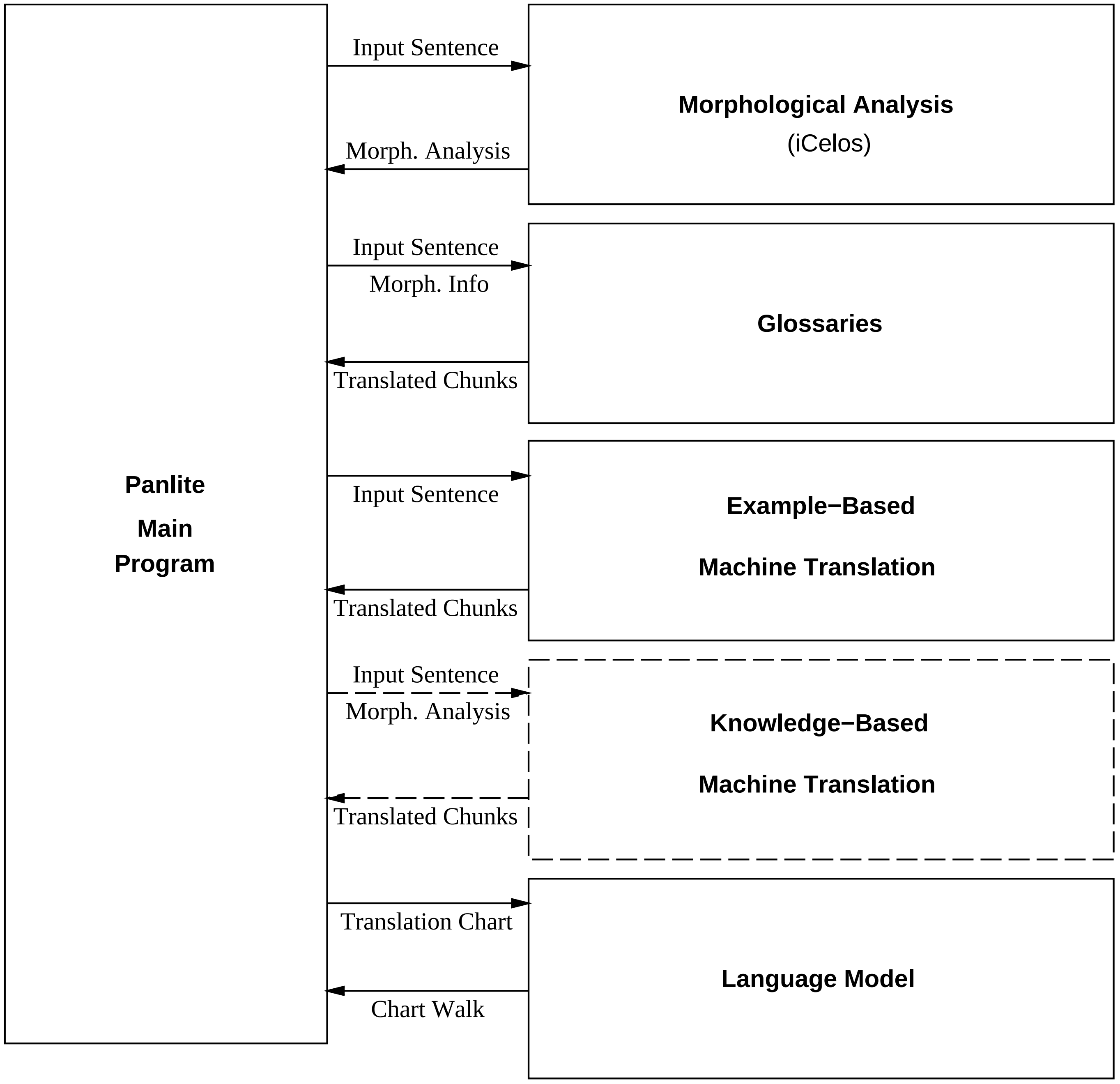
German-English EBMT Performance

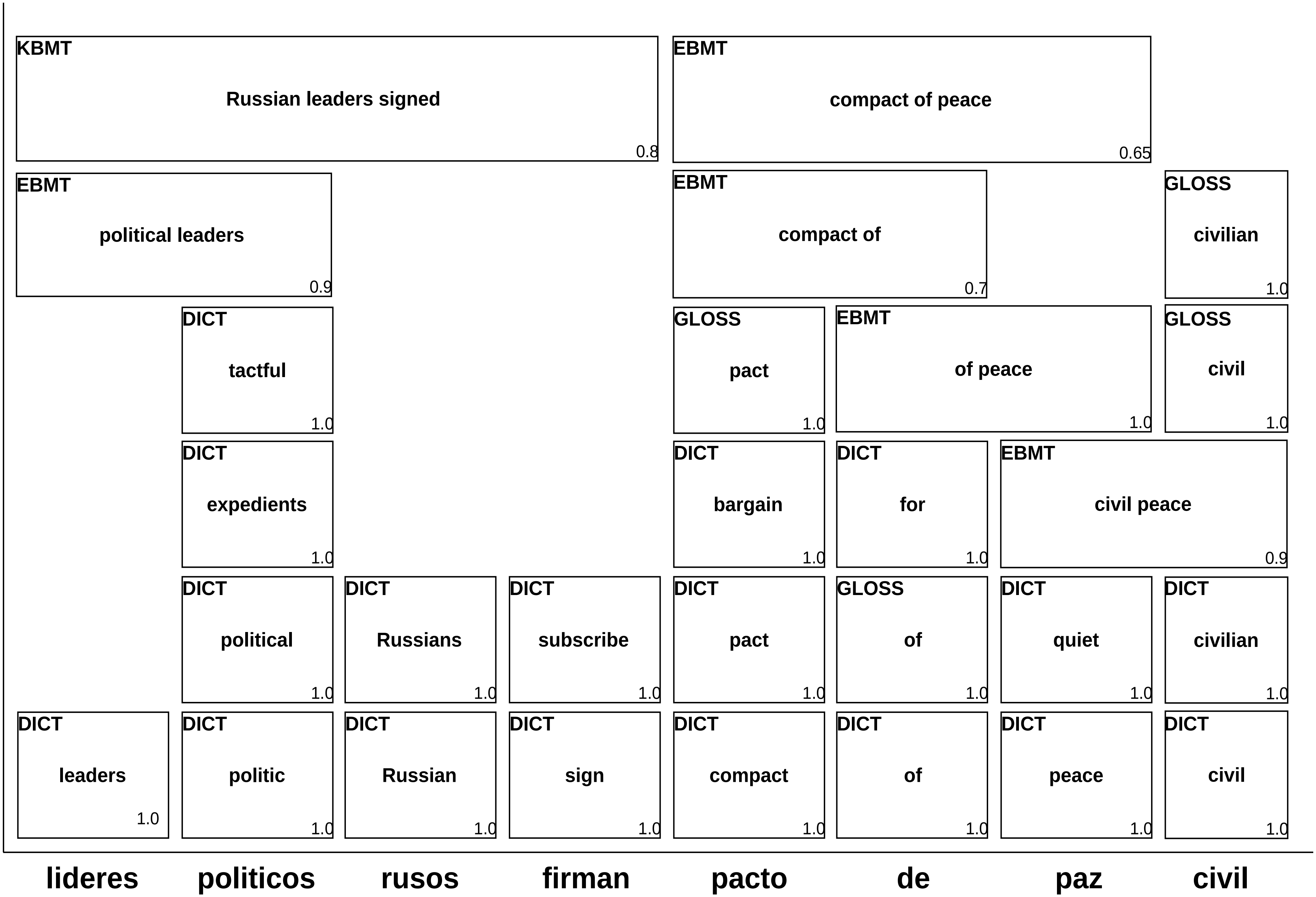
|  | Corpus | Coverage | Avg Len |
|---|---|---|---|
| Run | Matches | (words) | (words) |
| Baseline | 78.17% | 71.31% | 2.876 |
| Decompounded (base) | 80.81% | 74.40% | 2.992 |
| Decompounded (dict-S) | 81.21% | 75.14% | 2.943 |
| Decompounded (feed-S-S) | 81.74% | 75.61% | 2.963 |

# 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


Panlite
Main
Program
Input Sentence
Morph. Analysis
Morphological Analysis
(iCelos)
Input Sentence
Morph. Info
Translated Chunks
Glossaries
Input Sentence
Translated Chunks
Example−Based
Machine Translation
Input Sentence
Morph. Analysis
Translated Chunks
Knowledge−Based
Machine Translation
Translation Chart
Chart Walk
Language Model



KBMT
Russian leaders signed
0.8
EBMT
compact of peace
0.65
EBMT
political leaders
0.9
EBMT
compact of
0.7
GLOSS
civilian
1.0
DICT
tactful
1.0
GLOSS
pact
1.0
EBMT
of peace
1.0
GLOSS
civil
1.0
DICT
expedients
1.0
DICT
bargain
1.0
DICT
for
1.0
EBMT
civil peace
0.9
DICT
political
1.0
DICT
Russians
1.0
DICT
subscribe
1.0
DICT
pact
1.0
GLOSS
of
1.0
DICT
quiet
1.0
DICT
civilian
1.0
DICT
leaders
1.0
DICT
politic
1.0
DICT
Russian
1.0
DICT
sign
1.0
DICT
compact
1.0
DICT
of
1.0
DICT
peace
1.0
DICT
civil
1.0
lideres
politicos
rusos
firman
pacto
de
paz
civil

# 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.

# URLs

My Papers: `http://www.cs.cmu.edu/~ralf/papers.html`

Gaijin: `http://www.compapp.dcu.ie/~tonyv/papers/gaijin.html`

ReVerb: `http://citeseer.nj.nec.com/collins98examplebased.html`

ResearchIndex: `http://www.researchindex.com/`