Example-based Machine Translation

Cristina Vertan

University of Hamburg • Informatics Department

Natural Language Systems Group

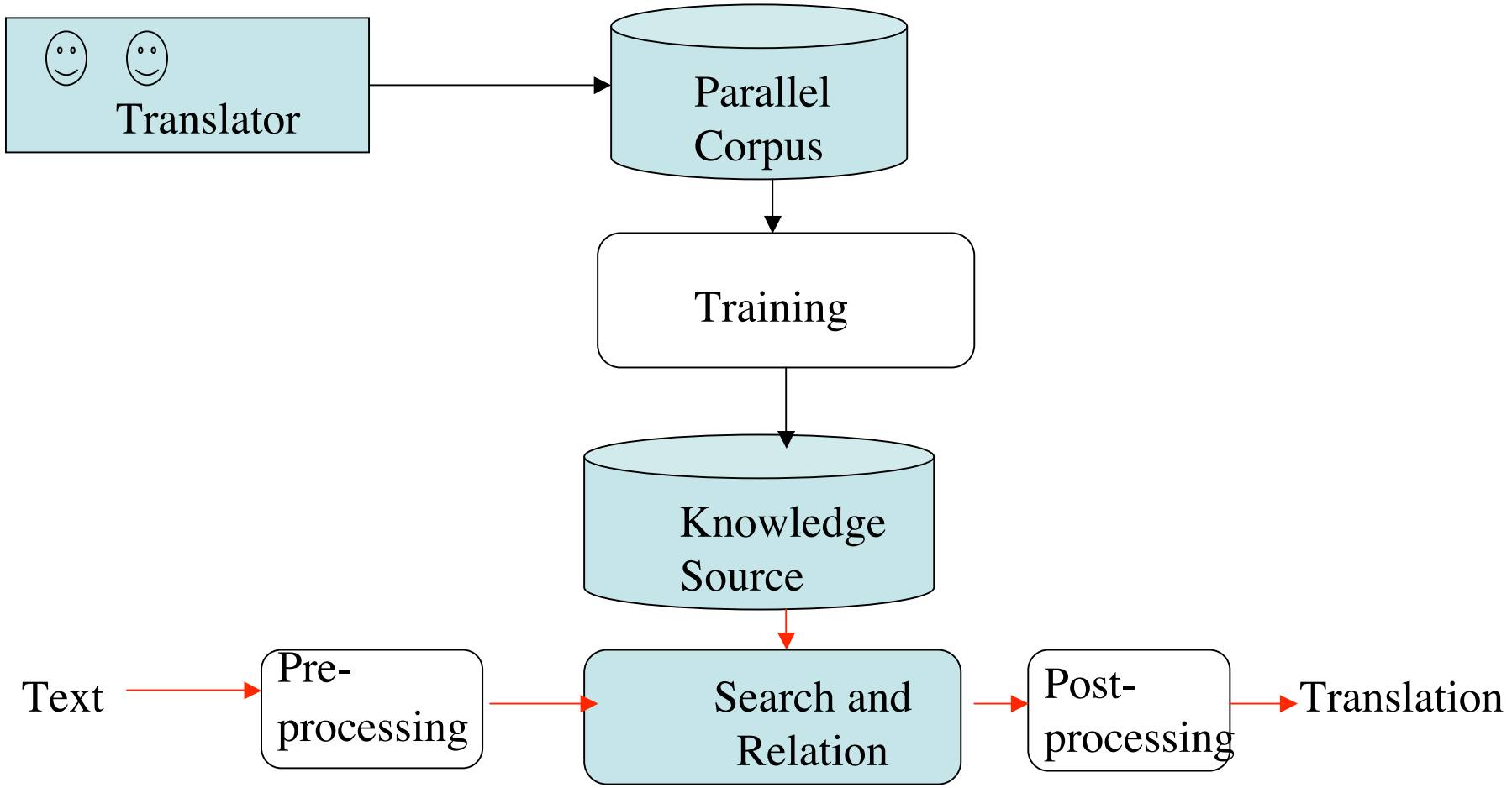
WWW: http://nats-www.informatik.uni-hamburg.de/~cri/

E-Mail: vertan@informatik.uni-hamburg.de

General Principles -Corpus based MT

- The linguistic phenomena in both languages as well as the transfer rules are no longer linguistically described but derived automatically from a parallel corpus.
- First an aligned corpus is built
- Next step is a training phase, in which are calculated the connections between elements in the source language as well as in the target language (sometimes the results are called „knowledge sources“)).
- The translation is the result of 2 processes:
  - A search process (of elements in the source language)
  - A best-evaluated relation with a target expression
- There are 2 types of corpus-based MT systems
  - Example based MT - The translation of a source text is based of translation examples in the database
  - Statistical MT - the alignment information from the corpus is used for the training of a statistical translation model

Generic Architecture of a corpus-based MT-system



Aligned Corpus

- A parallel Corpus:
  - Is a collection of texts in at least 2 languages. It is extremely important that the content is the same for both texts.
  - Examples: Official Documents from EU , Newspapers in Countries with more than 1 official language
  - Contains markers (tags) for content-identical elements (Sentences, Paragraphs) in Texts
- The parallel aligned Corpus has to be adequate for the translation domain.
- When searching such corpora the main problem is, that 1 chunk in the source has more than 1 translation in the target language and the choice is made according to the context.

Parallel aligned Corpus - Example 1-

<DOC de-news-1996-10-02-1>
<H1>
Streit um baden-wuerttembergischen Nachtragshaushalt
</H1>
Die Oppositionsfraktionen im baden wuerttembergischen Landtag haben scharfe Kritik an der Finanzpolitik der CDU/FDP-Koalition geuebt. Bei der Vorlage des zweiten Nachtragshaushalts fuer das laufende Haushaltsjahr begruessten sie zwar die strikte Begrenzung der Neuverschuldung, gespart werde aber vor allem bei den Familien und damit am falschen Fleck.Finanzminister Mayer-Vorfelder verteidigte den eingeschlagenen Sparkurs. Der Nachtragshaushalt soll rund 1,1 Mrd. DM Deckungsluecke ausgleichen, die vor allem durch Steuerausfaelle im Haushalt klaffen. Rund 800 Mio. DM sollen durch Investitions-und Sachmittelkuerzungen erbracht werden. Einsparungen im Personalbereich werden mit 130 Mio. DM beziffert.

<DOC de-news-1996-10-02-2>
.....

<DOC de-news-1996-10-02-1>
<H1>
Baden-Wurttemberg supplementary budget dispute
</H1>
The opposition parties in Baden-Wurttemberg's Landtag have strongly criticized the financial policies of the governing CDU/FDP coalition. Upon presentation of the second supplementary budget for the current budget year, they approved of the strict limitation of new borrowing,but said that savings were going to be realized in the wrong place - on the backs of families. Finance Minister Mayer Vorfelder defended the budget, saying it would equal out a shortfall of 1.1 billion marks in state finances, which was caused primarily by tax losses. Cuts to investments and materials are expected to yield 800 million marks. Savings on personnel are estimated
at 130 million Marks.

<DOC de-news-1996-10-02-2>
.....

Paragraph-Alignment

Parallel aligned corpus - Example 2-

<DOC de-news-1996-10-02-1>
Streit um baden - wuerttembergischen Nachtragshaushalt

Die Oppositionsfraktionen im baden - wuerttembergischen Landtag haben scharfe Kritik an der Finanzpolitik der CDU / FDP - Koalition geuebt .

Bei der Vorlage des zweiten Nachtragshaushalts fuer das laufende Haushaltsjahr begruessten sie zwar die strikte Begrenzung der Neuverschuldung , gespart werde aber vor allem bei den Familien und damit am falschen Fleck .

Finanzminister Mayer - Vorfelder verteidigte den eingeschlagenen Sparkurs . Der Nachtragshaushalt soll rund 1,1 Mrd. DM Deckungsluecke ausgleichen , die vor allem durch Steuerausfaelle im Haushalt klaffen .
.....

<DOC de-news-1996-10-02-1>
Baden - Wurttemberg supplementary budget dispute

The opposition parties in Baden - Wurttemberg 's Landtag have strongly criticized the financial policies of the governing CDU / FDP coalition .

Upon presentation of the second supplementary budget for the current budget year , they approved of the strict limitation of new borrowing , but said that savings were going to be realized in the wrong place - - on the backs of families .

Finance Minister Mayer - Vorfelder defended the budget , saying it would equal out a shortfall of 1.1 billion marks in state finances , which was caused primarily by tax losses .

...

Sentence-Alignment

2 Sentences in the DE-Corpus correspond to 1 Sentence in the EN-Corpus

Parallel aligned Corpus - Example 3 -

Die $_1$ Oppositionsfraktionen $_2$ im $_3$ baden - wuerttembergischen $_4$ Landtag $_5$ haben scharfe $_6$ Kritik $_7$ an der Finanzpolitik $_8$ der $_9$ CDU / FDP - $_{10}$ Koalition $_{11}$ geuebt $_{12}$.

The (1) opposition parties (2) in (3) Baden - Wurttemberg 's(4) Landtag(5) have strongly (6) criticized(7+12) the financial policies(8) of (9) the governing CDU / FDP (10) coalition(11) .

Fertitlity of a source word = the number of words in the target text

e.g. *fertility(Oppositionsfraktionen) = 2*

Distortion = Source and target words do not appear in the same place e.g. Koalition und coalition

Alignment-Methods

- Manual :
  - Extreme time consuming, because for real applications the corpus has to be really big.
  - Specialists with very good knowledge in both languages are needed
- Automatic with help of statistical procedures
  - E.g. length-based methods (number of words in the source and target text has to be close one to another)
  - Difficult to identify at the word-level, because for e.g. in:

Haben Scharfe Kritik geuebt ↔ *have strongly criticized*

The POS change and the semantic combinat

Lexicalization + Categorial divergence

„A good translator is a lazy translator“

EBMT Sources: Theory of Translation

A new transation may use as much material as possible from old translations (produced within the same domain, time, etc.).

Advantages of this approach:

- spares time

- ensures the terminological and stilistic consistency

Many human translations are revisions, improvements, changes of previous translations.

EBMT sources: cognition science

- Human translations are mostly not the result of deep linguistic analysis but more of an appropriate,
  - Division of the sentence in chunks followed by
  - Translation of the components as well as
  - Combination of these components.
- The translation of the components is done through analogy with previous existent translations.

EBMT source: MAHT

- Translators use often big databases with translation examples (Translator‘s workbenches /Translation memories).
- E.g. TRADOS - a TM-system for 12 European languages
- The system searches in the database all entries in the source language similar with the input and shows their translations
- The human translator identifies the pieces which he needs, and performs their recombination.

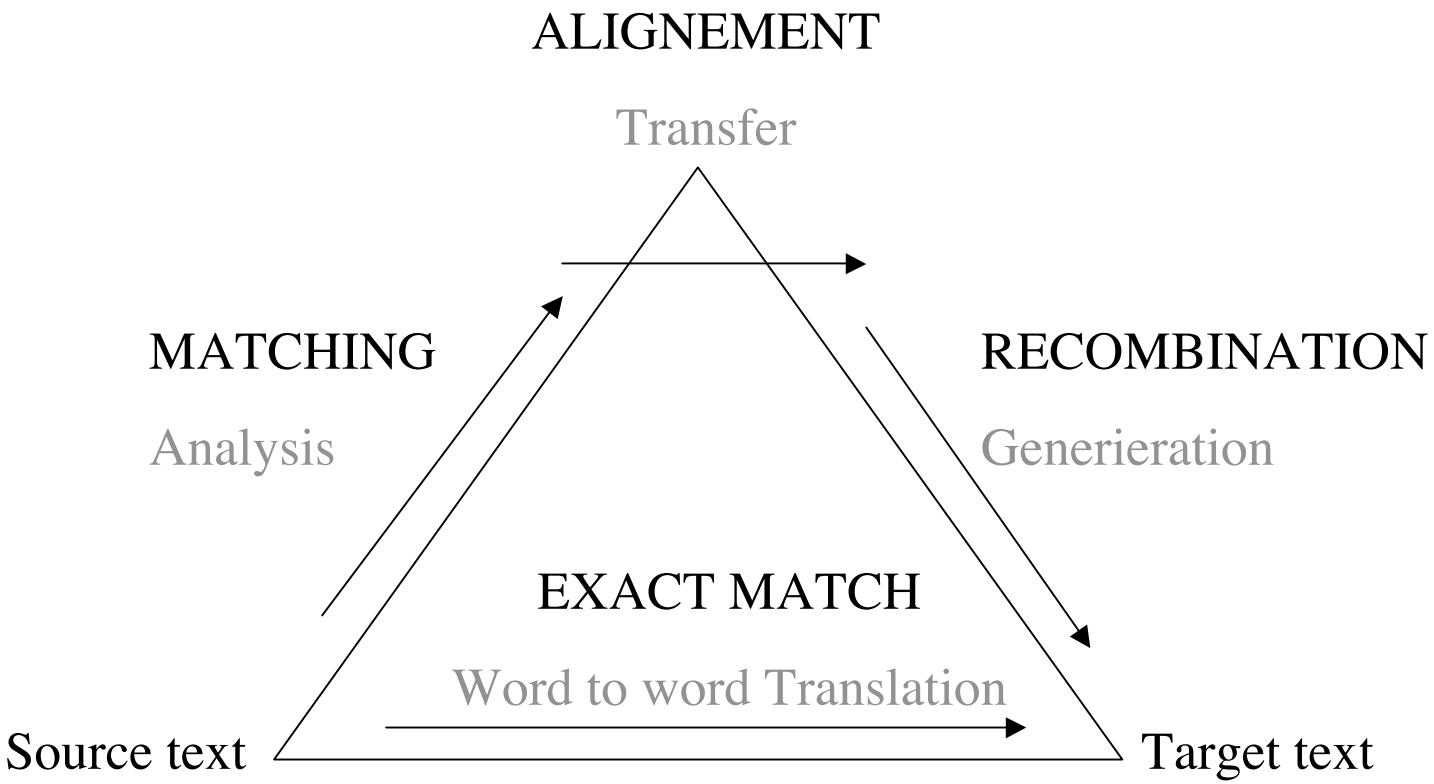
General Principles of EBMT

- A parallel Corpus is used
- Part of the input text are compared with source chunks in the corpus
- The translation of the retrieved parts are put together and form the translation.

Or

- The most similar sentences to the input in the SL corpus are retrieved (a distance is defined)
- The corresponding translations are combined to form an output

Translation pyramid for EBMT

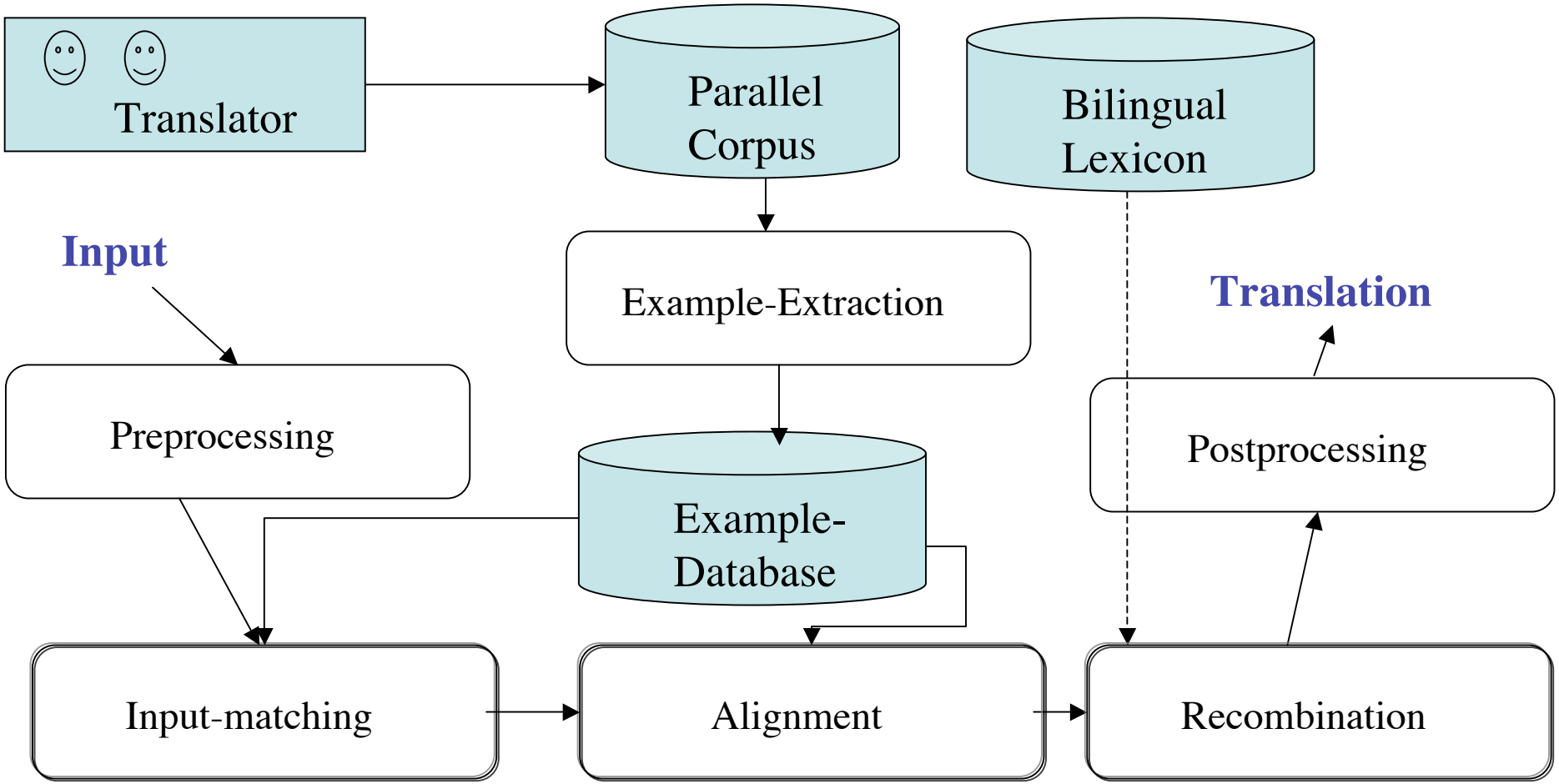


@A. Way

Functionality of an EBMT-System

- Relevant examples from a parallel corpus are extracted and saved in a database
- The input is compared with entries in the database(matching-phase).
  - Either the system looks for the identity of (parts of the) input with the database entries or
  - a distance between the input and the database entries is computed, and the database entry with the minimal distance to the input is chosen.
- Further on, in the alignement phase, the corresponding parts in the taget language are retrieved (this is trivial when the whole identical input is found in the DB)
- The corresponding chunks in the target language are recombined and build the output

Architecture of an EBMT-System



Relevant Examples?

- For a good lexical coverage:
  - a lot of domain relevant words
  - As much as possible with co-occurences (reflexiv, particle verbs, etc.)
- For a good syntactic coverage:
  - Sturcutres containing main and relative clauses
  - Active and passive voice sentences
  - questions
  - Sentences with embedded structures, e,g attribute sentences, conjunction sentences

Corpus-Tagging for EBMT

- It is possible to mark in the corpus words or morphemes, which delimit a clear co-text: like quantifers, conjunctions, pronouns, question markers, etc.
- E.g. <QUANT> all uses
- <QUANT> tous usages

Lenght and Size of Examples

- The *size* of the example database varies between some hundreds and 800.000 sentences.
- The bigger the database, the better the system works
- There is no ideal *length* for the examples:
  - The longer the examples, the lower the chance for a match
  - The shorter the example the bigger the chance to have some ambiguities
- Usualy the standard *unit* for the examples is a sentence

EBMT - Example

- Input: *Ungeeigneter Kraftstoff kann zu Motorschäden führen*

- the translation database contains:

- *Starke Motorbelastung kann zu Motorschäden führen - High engine loading can cause engine damage*

-*Ungeeigneter Kraftstoff darf nicht benutzt werden.- Unsuitable fuel must not be used*

- Following chunks are identified

- *kann zu Motorschäden führen - can cause engine damage.*

-*Ungeeigneter Kraftstoff - Unsuitable fuel*

•The translation is then:

– *Unsuitable fuel can cause engine damage*

Input for Matching

- The problem is to find out, which parts of the input can be retrieved in the database
- This is done through a combination of string-based, statistical-based methods (e.g. big probability for multi-word lexemes), and help of additional linguisitc knowledge.
- String-based matching approaches:
  - Edit distance
  - Angle of similarity
  - Semantic similarity

String-based Matching

- The similarity is measured between the input string and each string in the database. Following distances are used:
  - "longest common sequence"
  - "Edit distance": how many operations (Insert, Delete, Replacement) are necessary to transform the input string into an entry in the Database
- These methods can be implemented easier through greedy algorithm, or dynamic programming

Tomorrow

- How to improve EBMT through:
  - Linguistic knowledge
  - Rule-based approaches
- How rule-based MT works ?

Architecture of the System -Version 1