




Experiments with Matching Algorithms in Example Based Machine Translation

Cristina Vertan, Vanesa Espin Martin
University of Hamburg
vertan@informatik.uni-hamburg.de

Contents

- General principles of Example based machine translations 
- Case Study 1: Corpus: technical documents
- Case Study 2: Corpus: newspapers & WWW texts
- Conclusions and further work

General Principles of 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

„A good translator is a lazy translator“

EBMT Sources: Theory of Translation

A new translation 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 stylistic 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.

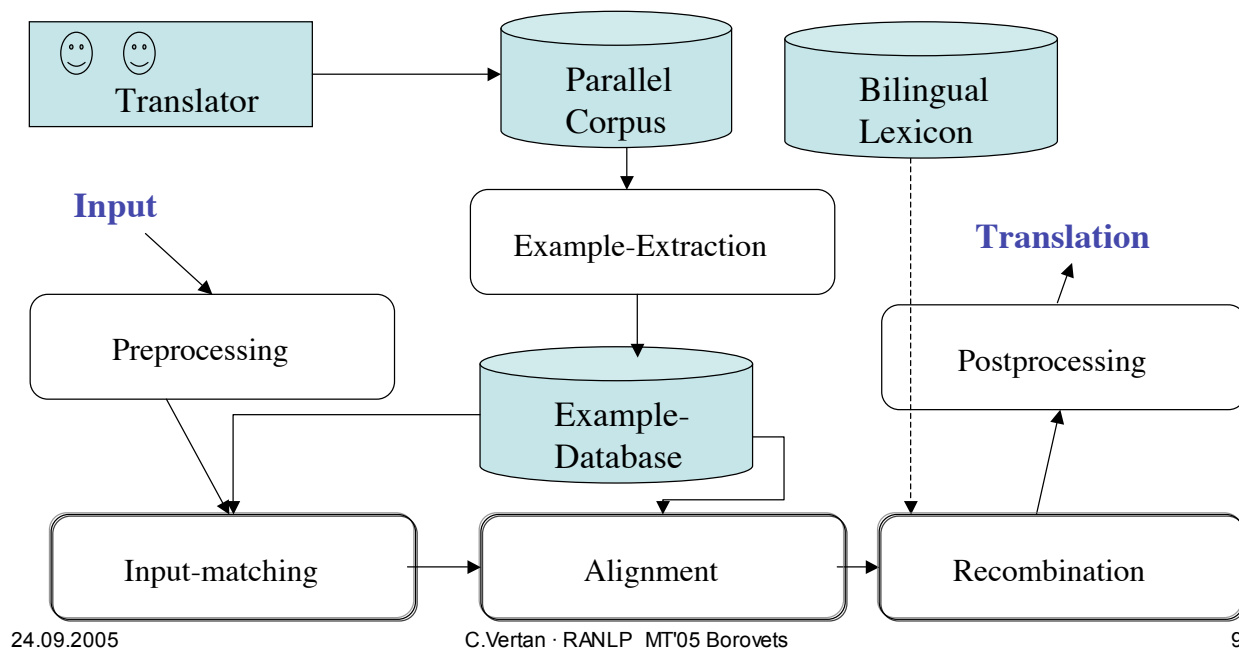
General Principles of EBMT

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

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 alignment phase, the corresponding parts in the target 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-occurrences (reflexiv, particle verbs, etc.)
- For a good syntactic coverage:
 - Structures 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 quantifiers, conjunctions, pronouns, question markers, etc.
- E.g. <QUANT> all uses (EN)
- <QUANT> tous usages (FR)


Length 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
- Usually 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*

Contents

- General principles of Example based machine translations
- Case Study 1: Corpus : technical documents 
- Case Study 2: Corpus: newspapers & WWW texts
- Conclusions and further work

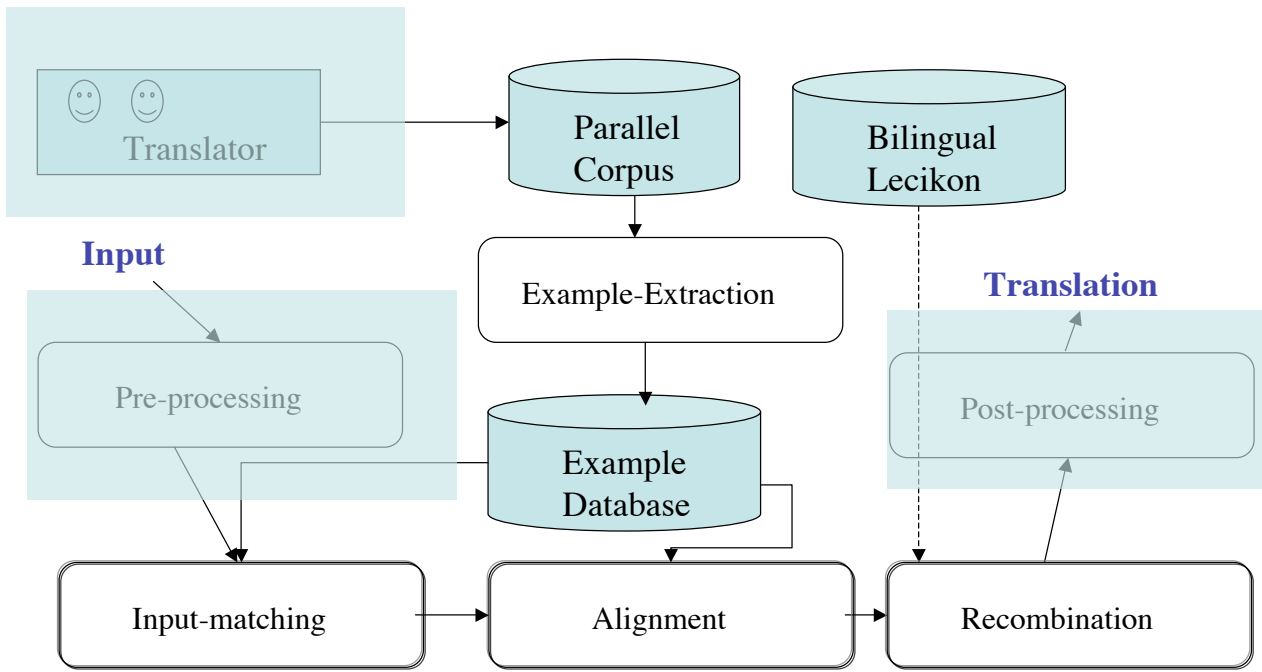
Case study 1: translation of specialised dialogue

- A German-English-German EBMT System
- Domain: car repairing
- Type of the input: simulated spoken language dialogues, under the scenario („Telefon service at a VW-Service-Center“):
 - *Ich habe Winterdiesel. Mein Auto springt trotzdem nicht an.*
 - *Ich verliere Bremsflüssigkeit. Was soll ich machen?*

Resources

- English and German Versions of the Volkswagen-manual (as parallel Corpus)
 - Simple sentences, quite often, 1 verb per sentence
- A conceptual structure of a German-English Lexicon
- Test-Corpus (examples of input sentences)

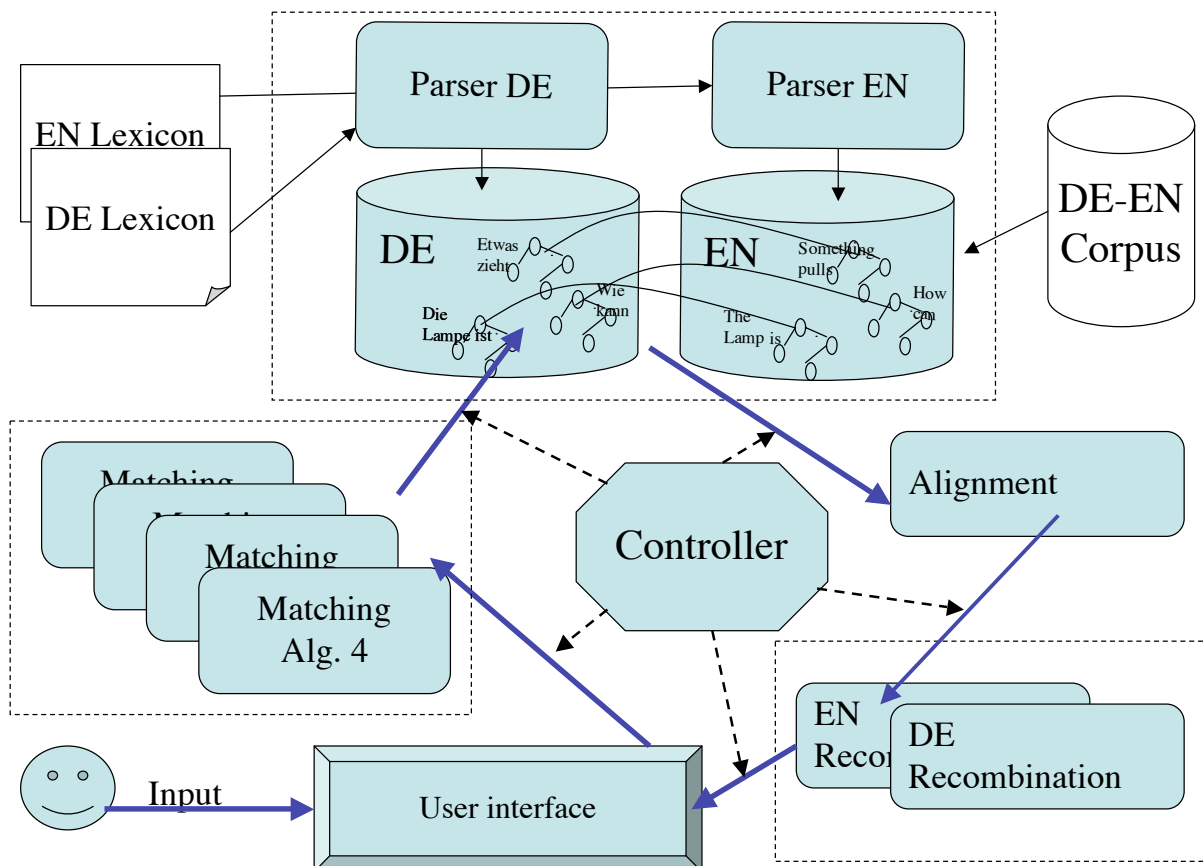
Architecture of the Autoreparatur HILfsystEm (AHILE)



24.09.2005

C.Vertan · RANLP_MT'05 Borovets

17



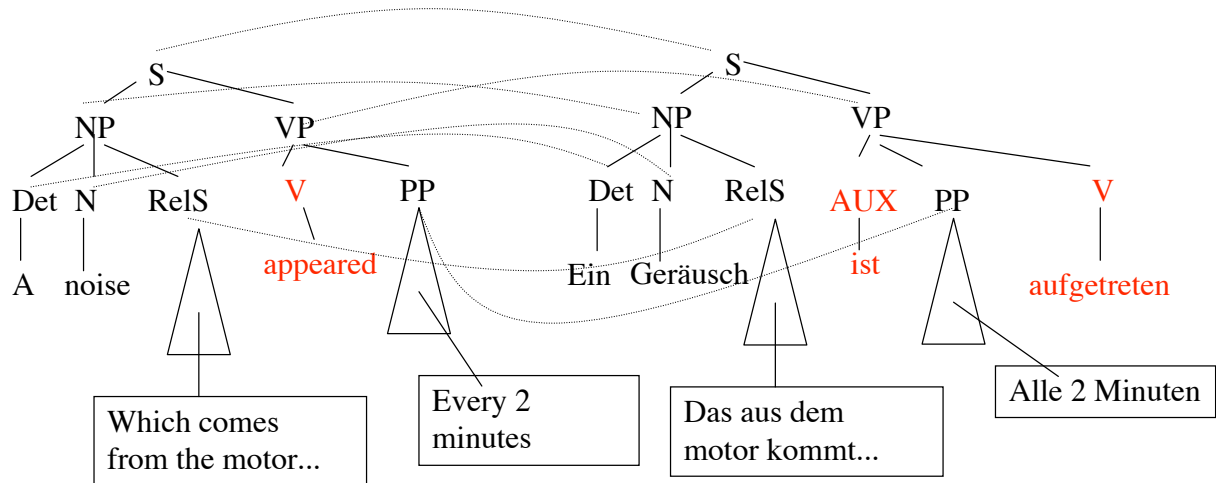
24.09.2005

C.Vertan · RANLP_MT'05 Borovets

18

EBMT with linguistic knowledge

- The translation patterns are not words, but syntactical structures in both languages with corresponding links



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 and statistical-based methods (e.g. big probability for multi-word lexemes).
- Matching approaches:
 - Edit distance
 - Angle of similarity
 - Semantic similarity

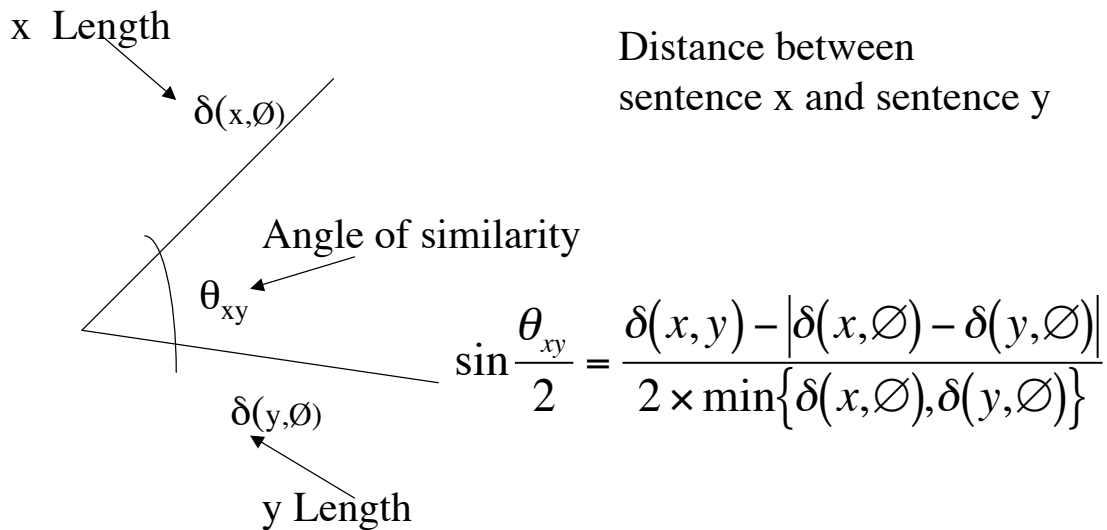
Word-based matching - 1 -

- A “thesaurus“ is used additionally; it is a kind of semantical network where distances between words express semantic similarity.
- For example for the following entries in the DB:
 - Der *Abstand* zwischen den Kontrollen soll 2 Jahre nicht überschreiten
↔ The *interval* between 2 general checks should not exceed 2 years.
 - Der normale *Abstand* zwischen den Nebelleuchten ist x cm.
↔ The normal *distance* between fog-lights is x cm.
- The input : *Wo finde ich den Abstand zwischen den Rädern?*
 - *Räder* in the thesaurus is closer to *Nebelleuchten*, therefore *Abstand* is translated by *distance*,
although the edit distance between *Räder* and *Kontrolle* is smaller than the edit distance between *Räder* and *Nebelleuchte*.

Word-based Matching: "Angle of similarity" - 2 -

- A trigonometrical distance is computed.
- The distance between 2 sentences corresponds to a difference function δ .
- This difference function works similar as the string-based matching (the number of operations is calculated)
- The operations are weighted, e.g. the insertion of a comma has a smaller weight than the absence of an adjective.
- The weights are defined according to the system and the translation domain

Word-based Matching: "Angle of similarity" - 3 -



Word-based Matching - "Angle of similarity" Example

1. *Lesen Sie Seite 3 im Kapitel "Benzin"*
 2. *Lesen Sie Seite 3 im Kapitel "Benzin" und Seite 5 in Kapitel "Länderspezifische Bemerkungen"*
 3. *Lesen Sie Seite 4 im Kapitel "Bremsen".*
- String-based matching gives a closer similarity between sentence 1 and sentence 3 because they differ only by 1 word.

However: Sentence 2 is actually a better choice as sentence 1 is contained entirely. This choice is made by the "angle distance".

Database search (Alignment) -1-

- In the ideal case an identical entry in the database is found.
- Usually it is necessary to cut necessary parts off the example translations
- The simpler the database structure (no POS-alignment, no additional tags) the more difficult is this step.

Database-search (Alignment) -2-

- There are statistical procedures for automatic alignment. They are based on statistical models of the source and target language
- The easiest way: the syntactical structure of the examples (in both languages) as well as relations among them are saved in the DB
- By the relations, parts of the structures can be identified.

Composition of the Output (Recombination)

- It is very difficult without relying on grammatical structures
- When tree structures are available, the process reduces to unification of trees
- The unification is done according to a recombination grammar.

Conclusions and further work

- For strong inflected languages it is reasonable to work with lemmas for measuring the matching between input and examples in the translation DB
- Multi-word Terminological expressions have to be recorded in a special lexicon (eventually also in lemmatized form)
- In the recombination phase the choice of the article form must be done by statistical computations in the corpus, or this step is part of post- processing (especially in the case of strong inflected languages)
- The example DB has to be improved with spoken language expressions, as the manual style does not offer the required syntactic variety.

Contents

- General principles of Example based machine translations
- Case Study 1: Corpus : technical documents
- Case Study 2: Corpus: newspapers & WWW texts ←
- Conclusions and further work

Case Study 2: Translation of touristic information

- Languages : English, Spanish
- Corpus:
 - WWW: Texts about tourism extracted from <http://www.spain.info>; Languages Spanish – English (10.000 words)
 - Articles from the LDC „UN Corpus“

Examples Data Base (I)

Complete Sentences Alignment

“Some areas are covered in mountains while others are full of plains”



“En algunas zonas abundan las montañas, mientras que en otras predominan las llanuras”

- Tagging: First Tagging.
 - Language: SGML
 - <sent + “alignment”>
 - <verb>; <N>; <PN + “type”> ...

Task Environment

- JAVA platform -> XML
 - API JAXP
 - SAX -> For Reading XML texts. Sequential access.
 - DOM -> For Reading and manipulating XML texts. Tree Structure access.
 - Corpus Portion
 - Sentences
 - Spanish -> 241
 - English -> 223
 - Words per sentence (average)
 - Spanish -> 19
 - English -> 20

MATCHING methods

- Edit distance: Character by character. Number of Insertions, Deletions and Substitutions to achieve the target.
- String matching: Exact match between two character sequences.
- Word by word (with the help of a Thesaurus): Compare sentences word by word. Introduction of semantics.

Edit Distance (I)

CORPUS: “Design is **the key player in the Spanish fashion industry**, which in recent years has gone from strength to strength abroad, thanks to the work of creators such as Jesús del Pozo, Adolfo Domínguez, Paco Rabanne, Pedro del Hierro, and the increasing presence of Spanish models on international catwalks”

INPUT: “**The key player in the Spanish fashion industry**”

DISTANCE?

Improving the method:

- **Moving in the sentence -> not enough**
- **Reduce the alignment size to sub-sentences. -> new size??**

Edit Distance (II)

- New Alignment: 1-1.
 - Spanish: 507 sentences, 9 words/sentence
 - English: 503 sentences, 9 words/sentence

CORPUS: “Design is **the key player in the Spanish fashion industry**”

INPUT : “**the key player in the Spanish fashion industry**”

- Drastically dependent of the size of the sentences:

CORPUS: “It is in the south of Spain”

-> is its ED lower than the Threshold??

INPUT : “**Design is the key**” ->REDUCE THE CORPUS AGAIN?

RECURSION Technique for searching sentences in the Corpus →
SUCCESS with “Design is the Key”

ED=0

String Matching (I)

- Retrieving of **exact** matches of the input sequence in the corpus

CORPUS: “The autonomous region of Andalusia lies in **the south of Spain**”

INPUT: “**the south of Spain**”

The algorithm success without changes

String Matching (II)

CORPUS:

“The autonomous region of Andalusia lies in **the south of Spain**”

“The Basque Country **is full of plains**”

INPUT: “**the south of Spain is full of plains**”

- The algorithm fails: no exact matches of the input sentence.
- Improved by RECURSION in the **input** (minimum: fragments of two words)

Comparing the methods (I)

- Computational Cost:
 - ED: High Cost
 - SM: Low Cost
- Locating input fragments:
 - ED: Worse results (threshold??)
 - SM: Better results
- Sentences size:
 - ED: Strongly dependent
 - SM: Weak dependence

Comparing the methods (II)

- Similarity:
 - ED: Retrieving of sentences that are *close* to the input
 - SM: Only exact matches of fragments

CORPUS: “Cultural landscape of Aranjuez”

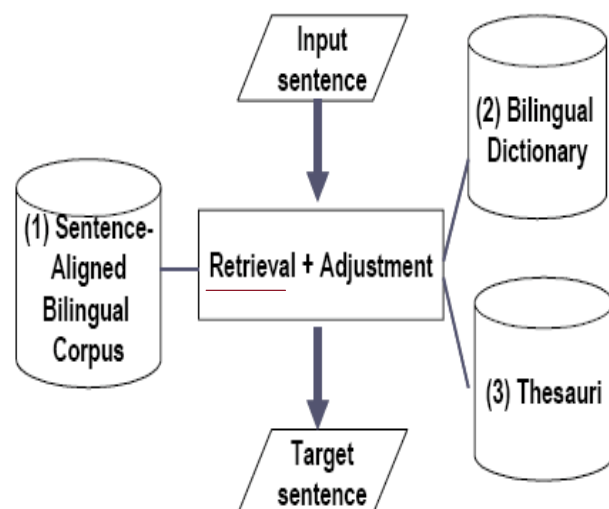
INPUT: “Cultural landscapes of **Seville**”

ED: Success

SM: Fail

Word by Word

- Retrieval of the most similar translation pair.
- Based in the words of the sentences
- Resources
 - Bilingual Corpus
 - Thesaurus
 - Bilingual Dictionary



Retrieval

- By measuring the *distance* between the word sequences of the input and example sentences of the bilingual corpus, retrieving the examples with the minimum distance (smaller than a threshold)

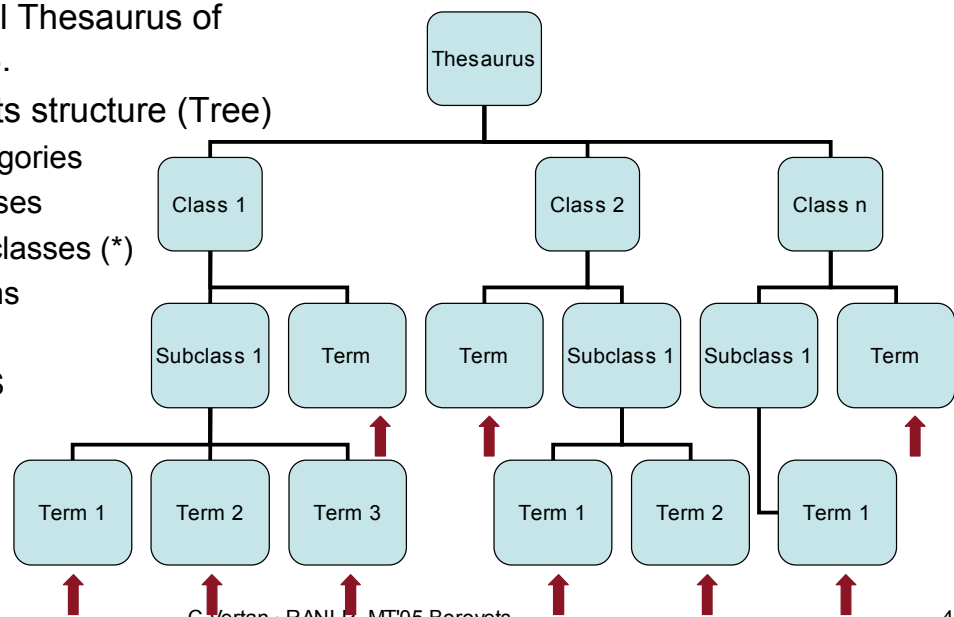
$$dist = \frac{I + D + 2 \sum semdist}{L_{input} + L_{example}}$$

$$semdist = \frac{K}{N} \quad \longrightarrow \quad \text{Semantic Distance}$$

Thesaurus Construction (I)

- Bilingual Thesaurus of NOUNS.
- Elements structure (Tree)
 - Categories
 - Classes
 - Subclasses (*)
 - Terms

NOUNS



Thesaurus Construction (II)

- Spanish Culture
 - Entertainment
 - Fashion
 - Sports
 - Religion
 - Dietary Habits
 - Mediterranean Diet
 - Typical Food
 - Tapas
 - Art
 - Monuments
 - Mosque
 - Museum
 - Monastery
 - ...
- Spanish Geography
 - Territories (“map”)
 - Autonomous Region
 - City
 - Province
 - Town
 - ...
 - Geographical Quirks (“geo”)
 - Mount
 - Mountain
 - Mountain Range
 - River
 - Ocean
 - ...
 - Cardinal Points

24.09.2005

C.Vertan · RANLP_MT'05 Borovets

43

Thesaurus Construction (III)

• Preferred Terms / Non Preferred Terms

• scopeNote

• Synonyms

• BT, NT, RT

• UF, USE

• Translation

Word Net
Structure



XML Format

24.09.2005

C.Vertan · RANLP_MT'05 Borovets

44

Bilingual Dictionary (I)

- Full-Form lexicon Dictionary
 - XML format
 - Inflected forms (verbs + nouns)
 - Lemma of the entry
 - Concept of the entry (very important for Spanish verbs)
 - Verbal time
 - Genre
 - Number
 - ...

Bilingual Dictionary (II)

- Automatically build extracting from the corpus
 - Attributes addition to the corpus
 - `<verb lex = "eat">have eaten</verb>`
 - `<n lex = "mountain">mountains</n>`
 - `<pn lex = "geo">Sacratif</pn>`
- Consistency with the Thesaurus
 - The lemmas of the nouns must appear with the same lemma in the Thesaurus

Measuring the Distance. Getting Ready

- Compile the corpus (Only necessary when changes in the corpus are made)
 - Lemmas extraction -> full-form-lexicon automatic construction
 - Substitution of each word in the corpus sentences by its lemma.
 - Note: JAVA DOM ->high cost in time and memory
- Input
 - Substitution of each word in the input by its lemma (searching in the full-form-lexicon)

24.09.2005

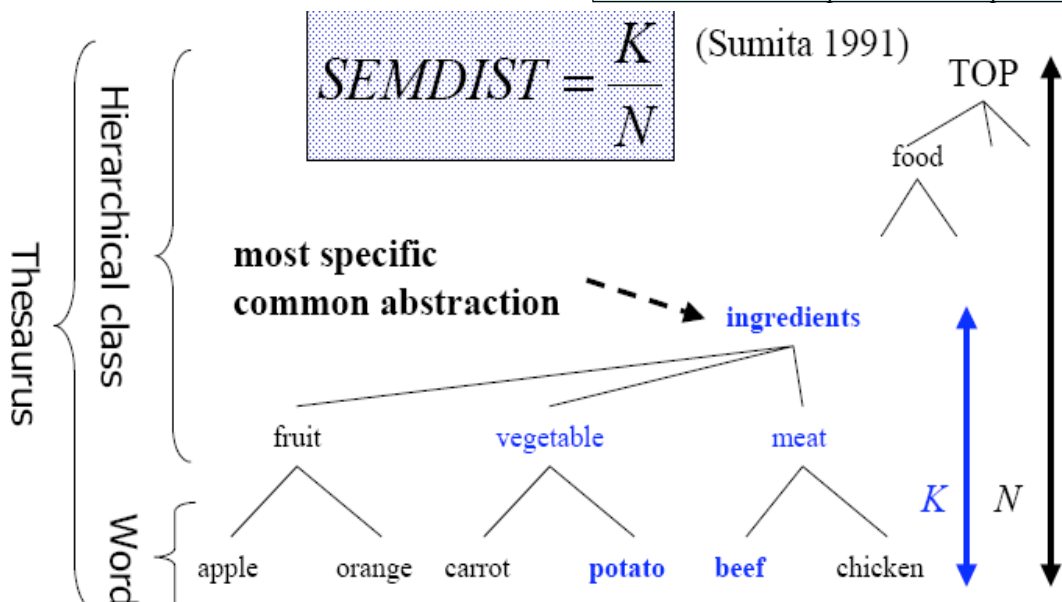
C.Vertan · RANLP_MT'05 Borovets

47

Measuring the Distance (II)

$$dist = \frac{I + D + 2 \sum semdist}{L_{input} + L_{example}}$$

- Distance




24.09.2005

48

Measuring the Distance (II)

- Semantic Distance
 - If two words are in the same subclass -> Semantic Distance = 0. Maximal Similarity.
 - Sea – Mountain -> SD = 0
 - If they are in different categories -> Semantic Distance = 1. Completely Dissimilar.
 - Sea – Museum -> SD = 1

Measuring the Distance. Sample

- Initial sentence manipulation (lexicon):
 - INPUT: "I have seen the Alhambra of Granada"
- 
"see the monum of map"
- CORPUS : "You will see the Mosque of Cordoba"


"see the monum of map"

**0 insertions 0 deletions 0 substitutions
dist = 0**

Measuring the Distance. Sample

- Initial sentence manipulation (lexicon):
 - INPUT: “The autonomous region of Andalusia lies in the south of Spain”



- “The region of map lie in the cardinal point of map”**
- CORPUS : “The gulf of Almeria lies in the east of Andalusia”



“The gulf of map lie in the cardinal point of map”

**0 insertions 0 deletions 1 substitutions
semdist (region, gulf) = 0.5
dist = (0+0+2*0.5) / (11+11)**

24.09.2005

51

Conclusions

- Humans don't translate character by character.
- The semantic, introduced in the last method, is a very important improvement and is close to humans behavior.
- The more detailed, clear, complete and concise the Thesaurus is, the more the good results in retrieval
- Domain in which we are working, is strong related with the input of the users

24.09.2005

C.Vertan · RANLP_MT'05 Borovets

52

Further Work

- Thesaurus improvement
 - Disambiguation
 - More relationships
 - More words
- Full-Form-Lexicon extension
 - Adjectives ?
 - More concepts (more tagging)
- Extending the Corpus

THANK YOU !