Basic Concepts and Technologies of Machine Translation

Walther v.Hahn, Cristina Vertan
University of Hamburg • Informatics Department
Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/~cri/
E-Mail: {vhahn,vertan}@informatik.uni-hamburg.de
Machine Interpretation

• New research and technology domain with applications in
  – Consecutive interpreting
  – Simultaneous interpretation
  – Dialogue interpretation
• Interesting because of the connection between
  – Signal level ⇔ Phonetics and
  – Text level ⇔ Linguistics
• High Relevance for cognitive linguistics
  – Interpreting strategy
  – Understanding
  – Time behaviour
  – Mapping of speaker- and language features
Lexical differences between languages

- One word in a source language can be replaced (translated) through more words or multi-word expressions in the target language.
- One word can be unambiguous in the target language, but not from the perspective of the source language.
- Ambiguity can be found: in one language and across languages.
- Lexical differences across languages have their source in
  - Difference between notions
  - Grammatical differences
  - Stylistic differences
Translation and understanding

Example: English ⇒ Spanish

*While driving down route 72, John swerved and hit a tree*

Problem:

In Spanish “hit” can be translated with:

1. *pear* to move with intention
2. *chocar* accidentally due to a moving object
3. *acertar* guessing
4. *golpear* (other possibilities) etc.

How can an MT system choose the right alternative?
It is hopeless without (at least lexical) semantics.
The MT-Triangle

- **Statistical Translation**
- **Direct Translation**
- **Shallow Transfer**
- **Deep Transfer**
- **Interlingua**
Transfer-System with 3 Languages

Spanish
Analysis

German
Analysis

English
Analysis

Spanish-German Transfer

German-Spanish Transfer

Spanish-English Transfer

English-Spanish Transfer

German-English Transfer

English-German Transfer

Spanish Generation

German Generation

English Generation
Interlingua-System with 3 Languages

- Spanish Analysis
- German Analysis
- English Analysis
- Spanish Generation
- German Generation
- English Generation
Interlingua- vs. Transfer-Systems

• Each module is independent from all other analysis and generation modules
• Target languages have no influence on the analysis process.
• For a new language only 2 new modules have to be added
• „back-translation“ possible (useful for system evaluation)
• Complicated representation even for languages belonging to the same family

• Language-dependent
• For each new language a high number of new modules must be implemented (for n languages: $n \times (n-1)$ modules)
• Straight-forward representation
• Local definition of similarities among languages.
Standard-Architecture of MT-Systems

Input (SL)

Preprocessing

Analysis

Transfer

Generation

Output (TL)

Postprocessing

Lexicon (SL)

Grammar (SL)

Discourse (SL)

Lexicon (TL)

Grammar (TL)

Discourse (TL)

Concept & Domain Knowledge
Direct System Architecture

Input (SL)

Preprocessing

Morphology analysis

Transfer

Output (TL)

Postprocessing

Morphology generation

Lexicon (SL)

Lexicon (TL)
Transfer-Systems Architecture

Input (SL)

Preprocessing

Morphology analysis

Syntax analysis

Semantic analysis

Transfer

Output (TL)

Postprocessing

Morphology generation

Syntax generation

Semantic generation

Lexicon (SL)

Grammar (SL)

Lexicon (TL)

Grammar (TL)
Interlingua-Systems Architecture

Input (SL)
- Preprocessing
  - Morphology analysis
    - Syntax analysis
      - Semantic analysis
        - Abstract Knowledge Representation
          - Concept & Domain Knowledge

Output (TL)
- Postprocessing
  - Morphology generation
    - Syntax generation
      - Semantic Discourse (TL)
        - Grammar (TL)
          - Discourse (SL)
            - Lexicon (SL)
              - Grammar (SL)
                - Discourse (SL)

Lexicon (SL)
Grammar (SL)
Discourse (SL)
Lexicon (TL)
Grammar (TL)
Discourse (TL)
MT-specific Pre-editing

• Checking source texts for foreseeable problems for the system and trying to eradicate them
• It can include:
  – Identification of names (proper nouns)
  – Marking of grammatical categories of homographs
  – Indication of embedded clauses
  – Bracketing of coordinate structures
  – Flagging or substitution of unknown words
  – Extreme form: Reformulation of the text using a “controlled language” and a corresponding editor
Pre-editing - Controlled Language

- Adaptation of source texts to the vocabulary such constructions which the system can translate
- The writers of texts for translation are restricted to
  - particular types of constructions
  - the use of terminology,
  - predefined meanings of every-day words
- E.g the sentence: *Loosen main motor and drive shaft and slide back until touching back plate* must be rewritten into:

  *Loosen the main motor. Loosen the drive shaft. Slide both parts until they touch the back plate.*
Post-Editing -1-

• Correction of the output from the MT-System to an agreed standard:
  – Minimal for assimilation purposes
  – Thoroughly for dissemination purposes
• E.g. Spanish ⇒ English output of an MT system:
  *En este estudio se buscará contestar dos preguntas fundamentales*
  In this study it will be sought to answer two fundamental questions
• The best post-edition may be:
  *This study will seek to answer two fundamental questions*
Post-Editing -2-

• Interactive post-editing:
  – The system alerts the editor of sentences or phrases which may be incorrectly translated (e.g. which contain an unresolved ambiguity, or a construction which could not be analysed)
  – It provides the option of correcting similar errors automatically throughout the text, once the editor has replaced a mistranslation

• Linguistically intelligent word processors:
  – Can spot some types of structural ambiguities
  – Can generate alternative structures
  – Change automatically gender agreement in a whole phrase
  – Insert automatically appropriate prepositions (e.g. if *discuss* is changed to *talk* then *about* is inserted before the direct object)
Evaluation of MT-Systems

• In contrast to other software there is no “best solution” by human translators, which can be compared with the output of the system
• I.e., for one input sentence there are many different correct translations
• Quality measurement of an MT System depends on its purposes and on the requirements of potential users.
• Possible participants in evaluation :
  – Researchers
  – Research sponsors
  – Purchasers
  – Translators
Evaluation strategies

Black Box vs. Glass Box

- MT system is seen as a black box, whose operation is treated purely in terms of its input-output behaviour
- Should not be conducted by the developers
- Tests: functionality, volume of data handled, recovery situations

- Components of the system are inspected as well as their effect in the system
- Relevant to researchers and developers
- Static analysis: checking the system without running it (automatic syntax and type checking by a compiler, manual inspection of the system, symbolic execution, data flow analysis)
- Dynamic glass box requires running the program (e.g. trying the program on many logical paths and ensuring that every logical branch is executed at least once).
Evaluation strategies

Test Suite vs. Test corpus

- Carefully constructed set of examples, each testing a particular linguistic or translation problem (e.g. different lexical and structural differences)
- Problem: it is assumed that the behaviour of a system can be projected from carefully constructed examples to real texts
- Test suite evaluations are difficult to compare

- An adequate corpus (for the domain of the system) is used as input
- Problem: it does not test systematically all possible sources of incorrect translations, but considers the most frequent constructions
- It is difficult to estimate the behaviour of the system for other types of text
Evaluation - Linguistic Quality measures

- **Intelligibility** - measures the fluency and grammaticality of the TL text, with concern for whether it faithfully conveys the meaning of the SL.
- **Accuracy** - indicates how the translated text preserves the content of the source text. (A highly intelligible sentence may not convey the meaning of the source text because of incorrect disambiguation.)
- **Error analysis**: e.g. count the number of words inserted, modified, deleted and moved by a post-editor. However, deciding what is an acceptable translation is subjective.
Evaluation - Software criteria

- Functionality - determines the degree to which it fulfills the stated or implied needs of a user
- Reliability - if the system maintains its level of performance under specified conditions and for a specified period of time
- Usability - indicates the effort needed to use the software by a stated or implied set of users
- Efficiency - relationship between the level of performance of the software and the amount of resources used to achieve that level of performance under specified conditions
- Maintainability - effort needed to make specified modifications to the software
- Portability - indicates the ability of the software to be transferred from one environment to another.
Wir treffen uns vor der Pizzeria Lorenzo. Ein italienisches Restaurant.

we meet in front of the seats and that way an italian restaurant

How long is the drive to Hanover?

wie lange fahren wir nach Hannover
Circa zwei Stunden. Wir sind dann um neun Uhr in Hannover.

Dialog act

great
bye
introduce
politeness formula
thank
deliberation
backchannel
init
defeat
close
commit
offer

Comparison

Cased based translation

about two hours then at nine o'clock , in Hannover.

Statistical translation

Translation Mismatch

Translation Soundness

Translation Quality

Syntactically Correct

Semantically Correct

Possible Misunderstandings

Next Turn

lost Information Elements

Translating Information Elements

Added Information Elements

File icon

Update

Dismiss
Different Approaches to MT

- Rule-based MT
- Knowledge-based MT
- Statistical-based MT
- Example-based MT

Other approaches to computer assisted translation
- Machine Aided Translation
- Translation Memories