Motivation

Course Overview

What are the problems?

MT history

Machine Translation

Overview & Introduction

Jörg Tiedemann

ejorg.tiedemann@lingfil.uu.se

Department of Linguistics and Philology

Uppsala University

Why Machine Translation?

Translation is expensive

On-line demand for translation (on-the-fly)

Globalization, growing export, lots of language pairs

Political issues (EU, support of minority languages...)

Tourism, movies, news

... 

MT combines various aspects of computational linguistics

Why is translation hard?

Languages are different on many levels:

- Lexicon, syntax, semantics
- Source language ambiguity
- Cross-lingual divergences
- Target language variation
- Pragmatics, style, culture, background

Computers have big problems with ambiguity.
Translation by humans

A human translator needs...

▶ to understand the source language
▶ to know how to speak the target language (well!)
▶ knowledge about the relationship between source and target language
▶ knowledge about the topic of the text to be translated
▶ knowledge about culture, values, traditions and expectation av speakers of source and target language

Computers have big problems with all of these issues

What can we do?

▶ Balance between quality and input restrictions

Automatic translation depending on task

<table>
<thead>
<tr>
<th>General purpose</th>
<th>Semi-automatic</th>
<th>Sublanguage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rough translation</td>
<td>Post-editing</td>
<td>Domain specific</td>
</tr>
<tr>
<td>Browsing quality</td>
<td>Editing quality</td>
<td>Publishing quality</td>
</tr>
<tr>
<td>Fully automatic</td>
<td>Computer-aided (human) translation</td>
<td>Fully automatic</td>
</tr>
<tr>
<td>(gisting)</td>
<td>(CAT)</td>
<td>(FAHQMFT)</td>
</tr>
<tr>
<td>On-line service</td>
<td>Localization, ...</td>
<td>Product manuals, limited domains</td>
</tr>
</tbody>
</table>
General Course Related Information

Course: 5LN426 (7.5hp)

- introductory course
- presentation of main approaches (lectures)
- focus on practical sessions
- test & use some existing systems and tools
- develop & present projects

Examination:

- pass obligatory lab assignments (G)
- translation project (graded)
  - define & develop your own MT project
  - present the project in class (2 seminars at the end)
  - describe project in a report

Literature:

- Philipp Koehn: Statistical Machine Translation
  Cambridge University Press
- on-line material (scientific papers, slides, ...)

For the labs:

- you need an account on the STP machines ...

Website:

- http://stp.lingfil.uu.se/~joerg/mt11
  Additional information might appear at any time during the course.

What exactly is MT?

- MT = automatic translation from one language (source language) to another (target language) using computers
- MT ≠ translation memories and bilingual dictionaries
- MT - often: sentence-by-sentence translation
- MT often refers to translation of written text (not speech)
- semi-automatic: CAT = computer-aided translation
Why is MT hard?

Processing of languages is hard:
- analysis, information extraction, message understanding
- natural language generation, discourse fluency
- ambiguity resolution, synonymy & redundancy
- non-compositionality, idioms ...
- domain detection & adaptation

Languages are different on many levels:
- lexicon, syntax, pragmatics
- style, culture, background

Why is it hard to accept MT?

- unrealistic expectations
- bad translations in available systems
- unexpected (not humanlike) errors
- problems of workflow integration

→ changing now due to recent developments (!?)

What are the problems?

- source language ambiguity
- target language variation
- cross-lingual divergences

Source language ambiguity

- poäng
  - point, points, credit, ...
- var
  - was, were (verb)
  - each (pron)
  - where (adv)
  - every (adj)
- anta
  - suppose
  - admit (anta någon till ...)

→ Ambiguity is usually solved in context
Lexical ambiguities across languages

Language divergences and mismatches

Variation in Target language

Divergences

Variation in Target language

Redundancy of natural languages

Types (Habash, 2002)

1. categorical divergences:
   tener celos (N) (lit. “to have jealousy”) ↔ to be jealous (A)
2. conflation: ir flotando (lit. “to go floating”) ↔ to float
3. structural divergence: entrar en N (lit. “to enter in N”) ↔ to enter N
4. head swapping: entrar corriendo (lit. “to enter running”) ↔ to run in
5. thematic divergence: me gustan uvas (lit. “to-me they-please grapes”) ↔ I like grapes
“Linguist’s view” on MT

What linguistic knowledge is needed to translate?

1. Phonological knowledge (speech-to-speech)
2. Morphological knowledge (inflection)
3. Syntactical knowledge (grammar)
4. Semantic knowledge (meaning)
5. Pragmatic knowledge (fluency, style, politeness, ...)

“Engineer’s view” on MT

MT as decoding

"I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information coded in the text"

(Warren Weavers (1947/49), Rockefeller Foundation)

(... and computers are good in encoding/decoding)

What do we need for MT?

1. more training data
2. statistical models
3. machine learning and efficient computing
4. more data! more data! more data!

Fred Jelinek (IBM speech group) in the 80s:

“Every time I fire a linguist the performance goes up”

(... or something similar)
Some words about the history of MT

▶ early days of computers: translation was thought to be one of the first tasks to be solved by computers (1930’s & ’40s)
▶ MT = first task for language technology
→ development of NLP/computational linguistics
▶ MT has a long history with many ups & downs

A brief history of MT (based on Hutchins)

1. Precursors and pioneers, 1933–1945
   ▶ interest in cryptography, information theory, coding
   ▶ early computers, political/military interests

   ▶ Georgetown experiment (rus-eng) → big success
   ▶ three (classical) MT approaches (direct MT, interlingua-based, transfer-based)
   ▶ polarization: empiricists and theoretical (computational) linguists
   ▶ world-wide establishment of MT research groups
   ▶ problems: insufficient computer power, formal theories under development, no NLP tools/resources
   ▶ ALPAC report 1966: no need for MT (it’s not going to work in near future)

Misconceptions about MT

▶ "MT is a waste of time because you will never make a machine that can translate Shakespeare."
▶ "There was/is an MT system which translated The spirit is willing but the flesh is weak into the Russian equivalent of The vodka is good, but the steak is lousy. → MT is useless".

D.J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys and Louisa Sadler

A brief history of MT (based on Hutchins)

▶ The quiet decade, 1967–1976
   ▶ domain-specific translation (controlled language)
   ▶ mainly interlingua approaches but no real progress

▶ Operational and commercial systems, 1976–1989
   ▶ Systran, METAL
   ▶ computer-aided translation (especially in Japan)

▶ Revival of MT research, 1976–1989
▶ Data-driven MT, 1990–
   ▶ SMT, EBMT, synchronous grammars, hybrid models
Where are we now?

- Today: more automatic translations than manual
- ca 50 million translation requests every day on the Internet (van der Meer, 2008)
- Google: provides on-line translation for 63 languages

→ Tasks: higher quality, other domains, other language pairs

Two main paradigms

- approaches inspired by linguistic theory
- data-driven techniques (re-using existing translations)

→ Let’s look at the classical rule-based systems first

The Vauquois triangle (translation triangle)

→ some intermediate symbolic representation

(from Jurafsky & Martin, 2008: Speech and Language Processing)
Direct Translation

- morphological analysis
- word-by-word using large bilingual dictionaries (may include phrases)
- local reordering using simple rules
- morphological generation

Example (from Jurafsky & Martin)

Input: Mary didn’t slap the green witch
After 1: Morphology Mary DO-PAST not slap the green witch
After 2: Lexical Transfer Maria PAST no dar una bofetada a la verde bruja
After 3: Local reordering Maria no dar PAST una bofetada a la bruja verde
After 4: Morphology Maria no dió una bofetada a la bruja verde

Transfer-based MT

3 steps:
1. source language analysis
2. transfer (source to target structures)
3. target language generation

Transfer rules:
- lexical transfer
- syntactic transfer
- semantic transfer

Interlingua-based MT

Basic idea: use meaning representation
- semantic analysis of source language input
- (formal) representation of meaning using interlingua → no language-pair specific transfer necessary!
- target language generation from meaning representation

Nice idea, but ...
- Where are the tools?
- What is a good way of representing meaning?
- Many problems with ambiguity and language divergences
Interlingua-based MT

Example interlingual representation: *Mary did not slap the green witch*

![Interlingual representation diagram]

Summary

- long, bumpy history
- rule-based systems vs. data-driven systems
- linguistic theory vs. examples & statistics
- large demand for MT world-wide
- MT quality depending on task

→ MT = hot topic again

Next lecture:

- more on rule-based systems
- domain-specific approaches
- evaluation of MT output (manual & automatic)

First lab session:

- evaluation of MT output
- a little experiment with understandability