Machine Translation at the European Commission and the Relation to Terminology Work
**Structure of the presentation**

- Usage Scenarios for Translation
- Technological Paradigms for MT
  - Statistical Machine Translation (SMT)
  - Rule-based Machine Translation (RBMT)
  - Hybrid MT
- MT@EC: Recent Developments and Perspectives
- Relation to Terminology Work
Requirements for MT depend on the way it is used

a) MT for assimilation „inbound“

b) MT for dissemination „outbound“

c) MT for direct communication

Speech recognition errors, specific style (chat) context dependence

Robustness
Coverage
Practically unlimited demand; but free web-based services reduce incentive to improve technology

Textual quality
Publishable quality can only be authored by humans; Translation Memories & CAT-Tools are almost mandatory for professional translators

Topic of many running and completed research projects (VerbMobil, TC Star, TransTac, …) US-Military uses systems for spoken MT, first applications for smartphones
Statistical Machine Translation: Theory

- Developed by F. Jelinek at IBM (1988-1995), based on „distorted channel“ Paradigm (successful for pattern- and speech recognition)

\[
\begin{align*}
\text{P(E)} & \rightarrow \text{E} \rightarrow \text{P(F|E)} \rightarrow \text{F} \\
\text{E*} = \text{argmax}_E \text{P(E|F)} &= \text{argmax}_E \text{P(E,F)} = \text{argmax}_E \text{P(E) * P(F|E)}
\end{align*}
\]

- Decoding: Given observation \text{F}, find most likely cause \text{E*}

\[
\begin{align*}
\text{P(E)}: \text{(Target) Language Model} & \quad \text{nGram-Models} \quad P(e_1 \ldots e_n) = \prod P(e_i|e_{i-2} e_{i-1}) \\
\text{P(F|E)}: \text{Translation Model} & \quad \text{Transfer of „phrases“} \quad P(F|E) = \prod P(f_i|e_i) * P(d_i) \\
\text{Search for E*: Decoding, MT} & \quad \text{Heuristic (beam) search}
\end{align*}
\]

- Three subproblems each has approximate solutions

- Models are trained with (parallel) corpora, correspondences (alignments) between languages are estimated via EM-Algorithm (GIZA++ by F.J.Och) search/decoding possible via Moses (Koehn e.a.)
Basic Architecture for Statistical MT

Translation Model (Adequacy)
- Parallel Corpus
- Alignment, Phrase Extraction
- Phrase Table

Language Model (Fluency)
- Monolingual Corpus
- Counting, Smoothing
- nGram-Model

Machine translation
- Source Text
- Target Text
- N-best Lists

Decoder
Examples of SMT Models

A selection of 23 out of 3897 ways to translate operations from EN to FR:

- operations || action || (0) || (0) || 0.00338724 0.0017156 0.00316685 0.0034059 2.718
- operations || actions || (0) || (0) || 0.0575431 0.0534003 0.0731052 0.0958526 2.718
- operations || activité || (0) || (0) || 0.0102038 0.0079204 0.00744879 0.0084917 2.718
- operations || activités || (0) || (0) || 0.019962 0.0194538 0.0366753 0.0451576 2.718
- operations || des actions || (0,1) || (0) || 0.0304499 0.0269505 0.00973472 0.00438066 2.718
- operations || des activités || (0,1) || (0) || 0.00877089 0.00997725 0.00246435 0.00206379 2.718
- operations || des opérations || (0,1) || (0) || 0.294821 0.281318 0.0406896 0.0238681 2.718
- operations || exploitation || (0) || (0) || 0.0437821 0.0365346 0.0208856 0.029298 2.718
- operations || fonctionnement || (0) || (0) || 0.0141471 0.01165 0.00919948 0.0099513 2.718
- operations || gestion || (0) || (0) || 0.00141338 0.0013098 0.00286578 0.0032669 2.718
- operations || intervention || (0) || (0) || 0.00561479 0.0026006 0.00110394 0.0013554 2.718
- operations || interventions || (0) || (0) || 0.0830237 0.0778631 0.0102142 0.0149096 2.718
- operations || les actions || (0,1) || (0) || 0.0339458 0.0271478 0.00931099 0.00712787 2.718
- operations || les activités || (0,1) || (0) || 0.00915348 0.0101746 0.00296613 0.00335805 2.718
- operations || les interventions || (0,1) || (0) || 0.0565693 0.0393793 0.00207406 0.00110872 2.718
- operations || les opérations || (0,1) || (0) || 0.413399 0.281515 0.0564235 0.0388363 2.718
- operations || manipulations || (0) || (0) || 0.0985325 0.183951 0.00104818 0.0034523 2.718
- operations || operations || (0) || (0) || 0.786026 0.557952 0.00200716 0.0023981 2.718
- operations || opération || (0) || (0) || 0.0245776 0.021675 0.00785022 0.0085959 2.718
- operations || opérationnel || (0) || (0) || 0.00656403 0.0069192 0.0012266 0.0013902 2.718
- operations || opérations effectuées || (0,1) || (0) || 0.110801 0.285316 0.00132696 0.00229301 2.718
- operations || opérations || (0) || (0) || 0.636821 0.562135 0.409237 0.522254 2.718
- operations || travaux || (0) || (0) || 0.00273044 0.0024213 0.00194025 0.0023517 2.718
SMT from a Translator’s perspective

- SMT can be seen as a generalisation of Translation Memory to sub-segmental level
- The phrases are text snippets taken from real-world translations (i.e. as good as what you entered)
- Re-combination of those phrases in new contexts may lead to significant problems:
  - Alignment errors ➔ spurious/lost meaning
  - Ignorance of morphology
  - Grammatical errors
  - Wrong disambiguation
- SMT will not recover implicit information from source text nor handle structural mismatches

current research prototypes include some linguistics & show significant improvements
Architectures for Rule-Based (RB) MT

The „Vauquois-Triangle“ (Vauquois, 1976)
The past: ECMT

- Single technological solution ("one-size-fits-all")
- Developed between 1975 and 1998
- 28 language pairs available (ten languages)
- Suspended since December 2010

The future?

- Hands-on workshop at DGT on Apertium (May 2011)
- Open-source solution, backed by a strong developer community, originally focused on regional languages
- Lexicons for many European languages being developed
- Could provide building blocks for hybrid solution...
Strengths and Weaknesses of MT Paradigms

(RBMT: translate pro ↔ SMT: Koehn 2005, examples from EuroParl)

EN: *I wish the negotiators continued success with their work in this important area.*

RBMT: *Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzten.*

*continued*: Verb instead of adjective

SMT: *Ich wünsche der Verhandlungsführer fortgesetzte Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.*

*three wrong inflectional endings*
## Strengths and Weaknesses of MT Paradigms

<table>
<thead>
<tr>
<th>English</th>
<th>RBMT: translate pro</th>
<th>SMT: Koehn 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>We seem sometimes to have lost sight of this fact.</strong></td>
<td>Wir scheinen manchmal Anblick dieser Tatsache verloren zu haben.</td>
<td>Manchmal scheinen wir aus den Augen verloren haben, diese Tatsache.</td>
</tr>
<tr>
<td><strong>The leaders of Europe have not formulated a clear vision.</strong></td>
<td>Die Leiter von Europa haben keine klare Vision formuliert.</td>
<td>Die Führung Europas nicht formuliert eine klare Vision.</td>
</tr>
<tr>
<td><strong>I would like to close with a procedural motion.</strong></td>
<td>Ich möchte mit einer verfahrenstechnischen Bewegung schließen.</td>
<td>Ich möchte abschließend eine Frage zur Geschäftsordnung ε.</td>
</tr>
</tbody>
</table>
Problems with Reliability of Lexicon Acquisition

[November 2007, corrected in the meantime]

See translationparty.com for more hilarious examples
In the early 90s, SMT and RBMT were seen in sharp contrast. But advantages and disadvantages are complementary.

Search for integrated (hybrid) methods is now seen as natural extension for both approaches.

<table>
<thead>
<tr>
<th></th>
<th>RBMT</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax, Morphology</td>
<td>++</td>
<td>--</td>
</tr>
<tr>
<td>Structural Semantics</td>
<td>+</td>
<td>--</td>
</tr>
<tr>
<td>Lexical Semantics</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Lexical Adaptivity</td>
<td>--</td>
<td>+</td>
</tr>
<tr>
<td>Lexical Reliability</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>
Hybrid MT Architectures (from EuroMatrix/Plus)

Possible ways to combine SMT with RBMT

1) Syntactic selection
- Source Text → SMT-engine(s) → Hypotheses → Selection → Target Text

2) Stochastic selection
- Source Text → Rule-based MT engines → Hypotheses → Selection → Target Text

3) SMT feeds rule-based MT
- Source Text → Rule-based MT engine → Target Text

4) SMT has the last word
- Parallel Corpus → Alignment, Phrase Extraction → Phrase-Table → Dyn. PT → SMT Decoder → Target Text

5) SMT corrects RBMT output
- Source Text → Rule-based MT engine → Error-Model → MT + Ref. T → Hypotheses → SMT Decoder → Target Text

6) Rule-based transfer architecture interleaved with stochastic ranking
- Source Text → Rule-based analysis → Stochastic ranking → Rule-based transfer → Stochastic ranking → Rule-based generation → Stochastic ranking → Target Text
Technological approach for MT@EC

- Started in June 2010 to implement an action plan
- Start with SMT as baseline technology
- Integrate linguistic knowledge as needed
- For morphologically simple/structurally similar LPs, baseline technology may be “good enough”
- For more challenging languages, techniques and tools from market and research will be incorporated
- Collaboration with DGT’s LDs will be crucial
MT@EC architecture

outline

Users and Services

DISPATCHER
managing MT requests

MT engines
by language, subject...

MT data
language resources specific for each MT engine

Language resources
built around Euramis

Customised interfaces

ENGINES HUB

DATA HUB

MT action lines

3. Service

2. Engines

1. Data
MT@EC overall planning

If all goes well…

➔ 2011 MT engines available to DGT staff to use as a CAT tool ("benchmark" engines, 
  quality enhancement via feed-back loop)

➔ 2012 Beta versions of the MT@EC service for selected test users outside DGT (comparison of engines)

➔ 2013 Operational MT@EC service for Commission, other EU institutions, and public administrations
Purpose
- Collect first round of feed-back about main issues in the current baseline engines
- Identify engines that could already be useful as they are now
- Limit the effort for the translators involved

Approach
- Let translators compare several hundred MT results with reference translations, showing differences in color
- Ask via web interface whether editing effort appears acceptable (“useful”) or not (“useless”)
Highlighting of differences between translations:

- Words of both translations are shown in black if the same word and both neighbours appear in the other translation as well.
- Words are shown in blue if the same word appears in the other translation, but at least one of its neighbours differs.
- Words that do not show up in the other translation (omissions, insertions, different lexical choice) are shown in red.
- If common parts of unmatched words are identified, they are displayed in violet.

**SRC (3g6558):** the date, time and location of the inspection, and
**DE REF:** das Datum, die Uhrzeit und den Inspektionsort und
**DE MT:** Datum, Uhrzeit und Ort der Inspektion sowie

†useful†useless †irrelevant
Maturity Check: Summary of Results

61 translators from 21 language departments provided more than 16000 individual judgments

DGT's SMT maturity check outcome as a (useful useless) sentences ratio + morphology
Translators were also asked (via a wiki page) to rank main types of observed errors. The following error types were ranked highest across all languages:

- Words or sub-sentences misplaced
- Word prefixes/infixes/suffixes wrong
- Terms usage inconsistent within the text
- Words/stems/vocabulary wrong
- Words missing
- Congruence wrong
How MT relates to Terminology Work

- **MT should respect existing terminology**
  - In case of doubt, “official” terms should be preferred over alternative wordings
  - SMT models can be tuned to respect such preferences

- **Training corpora contain inconsistent terminology**
  - Causes inconsistencies in MT results, unless properly handled
  - Systematic detection of such cases will improve MT quality

- **Training SMT from translation memories can identify new terminology as used in practice**
  - Frequent terms not in IATE can be identified and manually validated
  - This can speed up the development of IATE for new languages
    - Experiments with RO LD ongoing
    - first results: 2275 out of 2415 manually checked RO terms were good
      - precision of 94%
Thank you