Machine Translation at the European Commission

META forum, Budapest
28 June 2011
- EU Official languages: 23
- EC procedural languages: 3 (EN, FR, DE)
- DGT: 1750 linguists and 600 support
- Where: in Brussels, Luxembourg and in local offices in Member States
Machine Translation at the EC

The past: ECMT

- Rule-based machine translation
- Developed between 1975 and 1998
- 28 language pairs available (ten languages)
- Since 2006 only linguistic maintenance work on a couple of systems
- Suspended in 12/2010

The future: MT@EC

- 05/2010 Commission Task Force confirmed need for new MT for the Commission
- 06/2010 Action plan approved by management
- 09/2010 Work started for MT@EC
Task Force outcome (May 2010)

Conclusions

- MT@EC is necessary for the Commission
  (trust, confidentiality, continuity)
- Data-driven systems: a major technological breakthrough
- Any solution should be flexible and sustainable and ensure technological independence

First steps towards MT@EC

- Collection of user requirements
- Elaboration of an “architecture” (outline)
- Proposal for organisational and financial arrangements
Machine Translation Service
Outline of the proposed architecture

Users and Services
Customised interfaces

ENGINES HUB

MT engines
by language, subject...

MT data
language resources
specific for each MT engine

Language resources
built around Euramis

DATA HUB

DISPATCHER
managing MT requests

DATA MODELLING

USER FEEDBACK
MT Action plan

- Adopted in June 2010 by DGT
- Implementation started immediately
- Work along three action lines:
  1. data
  2. engines
  3. service
MT@EC architecture outline

Users and Services

Customised interfaces

MT engines by language, subject...

MT data language resources specific for each MT engine

Language resources built around Euramis

ENGINES HUB

MT engines by language, subject...

MT data language resources specific for each MT engine

DATA HUB

CUSTOMISED INTERFACES

3. Service

2. Engines

1. Data

MT action lines

DISPATCHER managing MT requests

DATA MODELLING

USER FEEDBACK
Action line 1: MT data

Objective: Infrastructure for the data required for the MT engines and the operation of the MT@EC service

Challenge: be ready for optimising all kinds of data for MT

Started with: internal DGT translation memories

Tasks: extract data, and put in place automatic procedures for cleaning, filter and processing them for MT

Now:

✓ Initial processing for internal datasets defined

Next:

✓ Work with other datasets

✓ Implementation of automatic procedures and process in the context of a database
Action line 2: MT engines

Objective: Set up MT engines and develop the necessary knowhow and processes for linguistic support in DGT.

Challenge: Compare alternative systems (both commercial and non-commercial) in terms of quality of output, price (total cost of ownership), feasibility, language coverage etc.
Action line 2: MT engines

**Started with:** open source tools

- **Basis:** SMT system *Moses* to establish internal *benchmarks*.
  - **tasks:** set up SMT engines and develop user interfaces and tools for capturing feedback in order to improve them.

- In parallel looking also into open source rule-based tool like Apertium (Luxembourg workshop) – technology watch
Action line 2: MT engines

- Now:
  Engines built: 50 (EN->X, X->EN, Other ECMT language pairs)

- Use:
  ✓ Limited custom access to engines since March
  ✓ Maturity check: prioritise
  ✓ “Real-life trial”
Action line 2: MT engines

Next:
- Improve SMT language pairs to use as benchmarks
  - Through user feedback
  ✓ Linguistic and translators’ perspective: started in DGT
  ✓ End users outside DGT: bilateral arrangements
  - Through targeted linguistic interventions pre-translation/translation (calls for tenders - demonstration)
- Prepare for comparisons (target end of the year)
Action line 3: MT service

**Objective:** Infrastructure for operating the MT@EC service

**Challenge:** flexible and sustainable implementation and governance of MT service

**Started with:** proof of concept

  **tasks:** design and implement “proof of concept”:
  - to analyse the scope of the service and primary scenarios of behaviour (that will drive the system's functionality)
  - to provide estimates for next phase

**Now:**

- Architecture confirmed (SOA)
- Elaboration of prototype service starting in June

**Next:**

- Testing
Service

- June 2011: start elaboration
- End 2011: tests with prototype (tbc)
- June 2012: start development
- 2013 (2nd half): operational baseline MT@EC service

Engines

- End 2011: benchmark/baseline versions
- Continuous actions to improve performance:
  - quality of MT output (though use and targeted linguistic interventions)
  - speed (through engineering interventions)
Open to the market

- Language technology watch (continuous)
- Linguistic interventions - demonstration projects in 2011
- Comparison of baseline engines to market offerings - 2012
  check:
  http://ec.europa.eu/dgs/translation/workwithus/calls/open/index_en.htm

... and to research

- Using Moses
- A major institutional user of MT
- Involvement in projects (e.g. Multilingual web)
- Conferences for EU institutions staff (e.g. EM+ workshop)
- Provider of language resources...
A closing word on data

- by September: update DGT-Translation Memory ("Acquis")
- Where: most probably on JRC web site [http://langtech.jrc.it/DGT-TM.html](http://langtech.jrc.it/DGT-TM.html)
- Now: 2,187,504 source segments
  16,883,981 target segments
- To add: 2,415,739 source segments
  38,168,330 target segments
  (corresponding to all 2004-2010 EU legislation)
Thank you