From MT@EC to eTranslation in CEF

Overview for the DGT QT21 Workshop

Andreas Eisele
Machine Translation Engines Team
Directorate-General for Translation R3.2

Luxembourg, 15.3.2017
From MT@EC to eTranslation in CEF

• What are MT@EC and eTranslation?
• Our users and some key requirements
• Architecture: From pure Statistical to Hybrid MT
• How to improve the translation quality
  • Quality improvements for end users
  • Improvements for translators
• Chances and risks of Neural MT
• Our Questions for Researchers from QT21
What are MT@EC and eTranslation?

**MT@EC:**

- Since **2010**, replacing an obsolescent rule-based solution (ECMT)
- Aimed at officials in EU and member states' administrations, both *translators* and *end users*
- Covering all **24 EU languages** in all combinations (78 LPs directly)
- Based on open source Statistical MT technology (*Moses*), co-funded by EU Framework Programmes for research and innovation
- Developed by **DGT R3.2**, using co-funding by ISA programme
- Real-life trial with DGT translators since 2011, released in **June 2013**
- Has been used to translate **tens of millions of pages** (up to 0.5M pages on a single day!)
- Usage and feed-back from *translators* are one of the main sources of inspiration for **improvements** of *MT quality*
First released: 26 June 2013

Languages: all directions between 24 EU languages
78 direct + domain specific engines

Technology: statistical machine translation
using the open source software Moses co-funded by the
EU Framework Programmes for research and innovation
Delivery:  - web user interface (human to machine)  
- web services (machine to machine)

Security:  host EC data centre + access: EC Authentication Service (ECAS) + transfer: sTESTA (private network of EU public administrations) or https

Special features:
- Source document format/formatting maintained [not for pdf]
- Specific output formats for translation: tmx and xlfiff
- Indication of quality for language pairs

Especially for the web client:
- Responsive web design
- User interface in 24 languages
- Translation can also be returned by email
- Can translate multiple documents to multiple languages
- Feedback mechanism etc.
Who can use MT@EC today?

→ **European institutions and bodies:**
  - Commission
  - Parliament
  - Council
  - Court of Justice
  - Court of Auditors
  - Economic and Social Committee
  - Committee of the Regions
  - European Central Bank
  - European Investment Bank, etc.

→ **Public administrations in the EU countries, Norway and Iceland**

→ **Online services** funded or supported by the EU
Good morning to everyone!

Buongiorno a tutti!
How to submit a translation request via the MTatEC webservice

InboundConnectorPublicSimpleProxyService

The service is accessible via the following URL on the stesta network: https://mtatecservice.ec.testa.eu/MtateOsbConnector/inboundConnectorPublicSimpleProxyService?WSDL

The entry point of the service conforms to SOAP protocol.

The structure of the XML message to send a request is the following:

```
<soapenv:Envelope xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/
xmlns:mtat="http://mtatec.dgt.ec.europa.eu/"
   <soapenv:Header/>
   <soapenv:Body>
     <mtat:asktranslation>
       <arg0>
         <documentToTranslate>ftp://user:password@my-ftp-server/my-document-to-translate-path</documentToTranslate>
         <priority>5</priority>
         <externalReference>clientRequestId</externalReference>
       </arg0>
       <sourceLanguage>DE</sourceLanguage>
       <targetLanguage>EN,FR</targetLanguage>
       <targetTranslationPath>email:toto@ec.europa.eu</targetTranslationPath>
       <username>toto</username>
       <applicationName>CLIENT_APPLICATION_NAME</applicationName>
       <domains>all</domains>
       <requestType>doc</requestType>
       <institution>eu.europa.ec</institution>
       <departmentNumber>DGT.R.3.002</departmentNumber>
     </mtat:asktranslation>
   </soapenv:Body>
</soapenv:Envelope>
```
MT@EC project 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

MT action lines

3. Service
2. Engines
1. Data
What next?

from MT@EC...

to CEF eTranslation
What are MT@EC and eTranslation?

**CEF = Connecting Europe Facility**
Union financial assistance to *trans-European networks* in order to support projects of common interest in the sectors of transport, *telecommunications* and energy infrastructures and to exploit potential synergies between those sectors. Resources are to be made available under the multiannual financial framework for the years 2014-2020.

**eTranslation will:**
- build on the *existing* MT@EC service *but* not be limited to it
- put emphasis on *secure, quality, customisable* MT for pan-European online services (DSIs) *but* not be limited to them
- be a *multilingualism enabler, not only MT*

**Platform is being built to serve**
CEF Digital Service Infrastructures (DSI), other public online services, public bodies in the EU Member States, and European institutions/bodies.
eTranslation will be looking for:

- even better **language technologies**
  both **open source** and **commercial** products
- even more quality **language resources**
  both **open** and **licenced** data
- even better **ways to enable multilingualism**
  both from **research** and **market**
  (and appropriate infrastructure to **scale up**)
## Customisation for online services

**connecting to the CEF.AT platform (starting from MT@EC)**

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europeana</td>
<td>The digital European Library, common, multilingual access point to digital resources of European heritage.</td>
</tr>
<tr>
<td>ODP*</td>
<td>The pan-European <strong>Open Data Portal</strong> for accessing open data infrastructures distributed over a EU and MS data repositories.</td>
</tr>
<tr>
<td>EESSI</td>
<td>The <strong>Electronic Exchange of Social Security Information</strong>, a platform between 32 countries (EU+EFTA).</td>
</tr>
<tr>
<td>ODR*</td>
<td>The <strong>Online Dispute Resolution</strong> platform for resolution of online contractual disputes between consumers and traders, linking all national ADR entities.</td>
</tr>
<tr>
<td>e-justice*</td>
<td>A portal which is a single point of access to law, enabling EU judicial cooperation.</td>
</tr>
<tr>
<td>SaferInternet</td>
<td>Services to make Internet a trusted environment for children.</td>
</tr>
</tbody>
</table>

*already connected to MT@EC*
Building MT Engines for CEF
Major Milestones

3 July 2017
- Fast text snippets on the cloud (webservice)

November 2017
- Document translation (webservice and web page)

Domain adaptation
- Ongoing and on demand
**eTranslation Milestones 2017**

- Pilot Cloud implementation: 12/2016
- eTranslation API for snippet translation on cloud
- Move of document conversion server to Cloud: 07/2017
- Engine factory: multiple engine on 1 server: 09/2017
- Engine factory: multiple engines on multiple servers: 10/2017
- Smart Engines Repository: 11/2017
- Language Resources Repository (ELRC2)
- First NMT engines in production: Q2/2018
- Cloud secure according to DIGIT Analysis (provisional)
- Continuous collaboration with DSIs (eTranslation System Integration, Specific Linguistic Data...)

<table>
<thead>
<tr>
<th>eTranslation Milestones</th>
<th>Planned Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Engine Factory phase 1 - build 1 engine on 1 server</td>
<td>09/2016</td>
</tr>
<tr>
<td>2 Pilot Cloud implementation</td>
<td>12/2016</td>
</tr>
<tr>
<td>3 Engine Factory: engine repository (storage and installation)</td>
<td>06/2017</td>
</tr>
<tr>
<td>4 Fastest engines possible with loss of no more than 10% BLEU score</td>
<td>Q2 2017</td>
</tr>
<tr>
<td>5 Interfacing with ELRC</td>
<td>Q2 2017</td>
</tr>
<tr>
<td>6 Separate workflows enabling snippet translation on cloud (at client’s choice) (webservice for technical users such as IMI, N-Lex)</td>
<td>03/07/2017</td>
</tr>
<tr>
<td>7 Move of document conversion server to cloud platform</td>
<td>09/2017</td>
</tr>
<tr>
<td>8 Document translation on cloud (potential eTranslation 1.0 launch)</td>
<td>15/11/2017</td>
</tr>
<tr>
<td>9 NMT engines for HU and DE in production for DGT</td>
<td>12/2017</td>
</tr>
<tr>
<td>10 Engine factory: building of multiple engines on 1 server</td>
<td>12/2017</td>
</tr>
<tr>
<td>11 Engine factory: building of multiple engines on multiple servers</td>
<td>Q2/2018</td>
</tr>
<tr>
<td>12 Render cloud secure to (Limited High) according to DIGIT analysis (not yet delivered)</td>
<td>Q2 2018 (provisional)</td>
</tr>
</tbody>
</table>
# DSIs using eTranslation

Initially we will serve 6 DSIs; the list may grow during the runtime of CEF

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europeana</td>
<td>The digital European Library, common, multilingual access point to digital resources of European heritage.</td>
</tr>
<tr>
<td>ODP</td>
<td>The pan-European <strong>Open Data Portal</strong> for accessing open data infrastructures distributed over a EU and MS data repositories.</td>
</tr>
<tr>
<td>EESSI</td>
<td>The <strong>Electronic Exchange of Social Security Information</strong>, a platform between 32 countries (EU+EFTA).</td>
</tr>
<tr>
<td>ODR</td>
<td>The <strong>Online Dispute Resolution</strong> platform for resolution of online contractual disputes between consumers and traders, linking all national Alternative Dispute Resolution (ADR) entities.</td>
</tr>
<tr>
<td>e-justice</td>
<td>A portal which is a single point of access to law, enabling EU judicial cooperation.</td>
</tr>
<tr>
<td>SaferInternet</td>
<td>Services to make Internet a trusted environment for children.</td>
</tr>
</tbody>
</table>
Main Usage Scenarios of MT

Requirements depend on the way MT is being used

a) MT for end users
   (assimilation, inbound)

b) MT for translators
   (dissemination, outbound)

c) MT for direct communication

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

Textual quality, terminology, accuracy
Publishable quality can only be authored by humans; MT needs to be embedded into CAT Tools

Ill-formed input, recognition errors, specific style (chat), context dependence
MT as a module in larger information systems covering specific scenarios
Main Requirements

For end users (scenario a), EC and other administrations
- Provide MT as a (simple and robust) service
- Optimise quality for understandability (gisting)
- Deal with many domains, document types, formats, ...
- Scale to huge volumes

For outbound translation (scenario b), DGT & similar institutions
- Provide MT as a tool within a CAT workflow
- Develop new ways to incorporate feed-back from translators
  - explicit feed-back on MT quality
  - implicit feed-back via TM
  - improvements requiring language-specific knowledge
  - towards linguistically informed/hybrid approaches
- Optimise quality for post-editing
Basic Architecture for Statistical MT

Translation Model \textit{(Adequacy)}
- Parallel Corpus
- Alignment, Phrase Extraction
- Phrase Table

Target Language Model \textit{(Fluency)}
- Monolingual Corpus
  - Counting, Smoothing
- Language Model

Source Text \rightarrow \text{Decoder} \rightarrow \text{Target Text}

N-best Lists
Sample from a Phrase Table: EN ➞ DE

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
<th>Score 4</th>
<th>Score 5</th>
<th>Score 6</th>
<th>Score 7</th>
<th>Score 8</th>
<th>Score 9</th>
<th>Score 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>competition</td>
<td>wettbewerb</td>
<td>0.792718</td>
<td>0.824898</td>
<td>0.471616</td>
<td>0.386947</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>84532</td>
<td>142086</td>
<td>67010</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbs</td>
<td>0.820595</td>
<td>0.871534</td>
<td>0.119012</td>
<td>0.098244</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>20607</td>
<td>142086</td>
<td>16910</td>
</tr>
<tr>
<td>competition</td>
<td>den wettbewerb</td>
<td>0.649377</td>
<td>0.824898</td>
<td>0.061237</td>
<td>0.010077</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>13399</td>
<td>142086</td>
<td>8701</td>
</tr>
<tr>
<td>competition</td>
<td>des wettbewerbs</td>
<td>0.309026</td>
<td>0.871534</td>
<td>0.022071</td>
<td>0.001277</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>10148</td>
<td>142086</td>
<td>3136</td>
</tr>
<tr>
<td>competition</td>
<td>auswahlverfahren</td>
<td>0.196723</td>
<td>0.150756</td>
<td>0.021212</td>
<td>0.021199</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>15321</td>
<td>142086</td>
<td>3014</td>
</tr>
<tr>
<td>competition</td>
<td>der wettbewerb</td>
<td>0.578915</td>
<td>0.824898</td>
<td>0.019981</td>
<td>0.032399</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>4904</td>
<td>142086</td>
<td>2839</td>
</tr>
<tr>
<td>competition</td>
<td>auswahlverfahrens</td>
<td>0.196723</td>
<td>0.150756</td>
<td>0.021212</td>
<td>0.021199</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>5989</td>
<td>142086</td>
<td>2365</td>
</tr>
<tr>
<td>competition</td>
<td>konkurrenz</td>
<td>0.575426</td>
<td>0.544313</td>
<td>0.013316</td>
<td>0.013055</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>3288</td>
<td>142086</td>
<td>1892</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsrechtlichen</td>
<td>0.528922</td>
<td>0.587997</td>
<td>0.008044</td>
<td>0.009912</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>2161</td>
<td>142086</td>
<td>1143</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsbedingungen</td>
<td>0.137052</td>
<td>0.149663</td>
<td>0.007453</td>
<td>0.020569</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>7727</td>
<td>142086</td>
<td>1059</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbs-</td>
<td>0.448318</td>
<td>0.580939</td>
<td>0.005159</td>
<td>0.004711</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>1635</td>
<td>142086</td>
<td>733</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbspolitik</td>
<td>0.156701</td>
<td>0.509962</td>
<td>0.002301</td>
<td>2.00395e-06</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>4365</td>
<td>142086</td>
<td>684</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerb zu</td>
<td>0.489235</td>
<td>0.824898</td>
<td>0.004638</td>
<td>0.010709</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>1347</td>
<td>142086</td>
<td>659</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsrechtliche</td>
<td>0.314149</td>
<td>0.454091</td>
<td>0.002766</td>
<td>0.001325</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>1251</td>
<td>142086</td>
<td>393</td>
</tr>
<tr>
<td>competition</td>
<td>den wettbewerb</td>
<td>0.531609</td>
<td>0.824898</td>
<td>0.004711</td>
<td>0.010709</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>696</td>
<td>142086</td>
<td>370</td>
</tr>
<tr>
<td>competition</td>
<td>competition</td>
<td>0.939394</td>
<td>0.787368</td>
<td>0.002399</td>
<td>0.002125</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>363</td>
<td>142086</td>
<td>341</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbspolitik zuständige</td>
<td>0.751724</td>
<td>0.509962</td>
<td>0.002301</td>
<td>2.00395e-06</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>435</td>
<td>142086</td>
<td>341</td>
</tr>
<tr>
<td>competition</td>
<td>für wettbewerbspolitik</td>
<td>0.603383</td>
<td>0.509962</td>
<td>0.002259</td>
<td>0.000419</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>532</td>
<td>142086</td>
<td>321</td>
</tr>
<tr>
<td>competition</td>
<td>konkurrenzen</td>
<td>0.115158</td>
<td>0.113537</td>
<td>0.002209</td>
<td>0.002216</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>2718</td>
<td>142086</td>
<td>313</td>
</tr>
<tr>
<td>competition</td>
<td>für wettbewerbspolitik zuständige</td>
<td>0.71659</td>
<td>0.509962</td>
<td>0.002188</td>
<td>2.86217e-08</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>343</td>
<td>142086</td>
<td>321</td>
</tr>
<tr>
<td>competition</td>
<td>dem wettbewerb</td>
<td>0.531609</td>
<td>0.824898</td>
<td>0.004711</td>
<td>0.010709</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>1721</td>
<td>142086</td>
<td>295</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsrechts</td>
<td>0.10087</td>
<td>0.490932</td>
<td>0.002041</td>
<td>0.015385</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>2875</td>
<td>142086</td>
<td>290</td>
</tr>
<tr>
<td>competition</td>
<td>führen</td>
<td>0.00311379</td>
<td>0.0028466</td>
<td>0.001900</td>
<td>0.002574</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>86711</td>
<td>142086</td>
<td>270</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsfähigkeit</td>
<td>0.00331525</td>
<td>0.003633</td>
<td>0.001879</td>
<td>0.001591</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>80537</td>
<td>142086</td>
<td>267</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsrecht</td>
<td>0.0941485</td>
<td>0.483305</td>
<td>0.001766</td>
<td>0.021142</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>2666</td>
<td>142086</td>
<td>251</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerb ,</td>
<td>0.0736518</td>
<td>0.824898</td>
<td>0.001682</td>
<td>0.050882</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>3245</td>
<td>142086</td>
<td>239</td>
</tr>
<tr>
<td>competition</td>
<td>führt</td>
<td>0.00342564</td>
<td>0.0027133</td>
<td>0.001506</td>
<td>0.001710</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>62470</td>
<td>142086</td>
<td>214</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerb herrscht</td>
<td>0.291549</td>
<td>0.426895</td>
<td>0.001456</td>
<td>0.000545</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>710</td>
<td>142086</td>
<td>207</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerbsdruck</td>
<td>0.0640314</td>
<td>0.05135</td>
<td>0.001379</td>
<td>0.017619</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>3061</td>
<td>142086</td>
<td>196</td>
</tr>
<tr>
<td>competition</td>
<td>wettbewerb in</td>
<td>0.0863615</td>
<td>0.824898</td>
<td>0.001287</td>
<td>0.011242</td>
<td>2.718</td>
<td>0</td>
<td>0</td>
<td>2119</td>
<td>142086</td>
<td>183</td>
</tr>
</tbody>
</table>

7941 ENTRIES, covering 142086 occurrences!

Total size of this Phrase Table: 502 M entries
Sample from a Phrase Table (EN → DE, II)

competition ||||, des bundeskartellamts ||| 0.0880911 0.0314685 3.71991e-06 8.74674e-08 2.718 ||| 0-2 ||| 6 142086 1
competition ||||, derweil der wettbewerb ||| 0.528547 0.824898 3.71991e-06 1.27812e-09 2.718 ||| 0-3 ||| 1 142086 1
competition ||||, der der konkurrenz ||| 0.528547 0.544313 3.71991e-06 1.20352e-05 2.718 ||| 0-3 ||| 1 142086 1
competition ||||, der den wettbewerbsregeln ||| 0.528547 0.443897 3.71991e-06 1.10122e-05 2.718 ||| 0-3 ||| 1 142086 1
competition ||||, den wettbewerb zu beschränken ||| 0.0101644 0.824898 3.71991e-06 1.34967e-09 2.718 ||| 0-2 ||| 52 142086 1
competition ||||, den wettbewerb in ||| 0.0097258 0.824898 3.71991e-06 3.85031e-05 2.718 ||| 0-2 ||| 53 142086 1
competition ||||, den wettbewerb in der ||| 0.040456 0.824898 3.71991e-06 3.22387e-06 2.718 ||| 0-2 ||| 12 142086 1
competition ||||, den wettbewerb auf ||| 0.00960994 0.824898 3.71991e-06 1.55564e-05 2.718 ||| 0-2 ||| 55 142086 1
competition ||||, den wettbewerb auf dem ||| 0.0176182 0.824898 3.71991e-06 1.04796e-07 2.718 ||| 0-2 ||| 30 142086 1
competition ||||, dass wettbewerbsprobleme ||| 0.528547 0.544313 3.71991e-06 0.2718 ||| 0-6 ||| 1 142086 1
competition ||||, dass es sie ||| 0.00788876 3.4e-06 3.71991e-06 1.35692e-10 2.718 ||| 0-2 ||| 67 142086 1
competition ||||, dass einem wettbewerb ||| 0.528547 0.824898 3.71991e-06 7.98821e-07 2.718 ||| 0-3 ||| 1 142086 1
competition ||||, dass ein wettbewerb ||| 0.105709 0.824898 3.71991e-06 1.90358e-06 2.718 ||| 0-3 ||| 5 142086 1
competition ||||, dass die übernahme den wettbewerb ||| 0.528547 0.824898 3.71991e-06 3.30452e-11 2.718 ||| 0-5 ||| 1 142086 1
competition ||||, dass die früher ||| 0.0755067 0.0002227 3.71991e-06 1.1282e-09 2.718 ||| 0-3 ||| 7 142086 1
competition ||||, dass das konkurrenzen ||| 0.528547 0.113537 3.71991e-06 1.43608e-08 2.718 ||| 0-3 ||| 1 142086 1
competition ||||, dass aufgrund ||| 0.00030676 8.2e-06 3.71991e-06 1.67179e-08 2.718 ||| 0-2 ||| 1723 142086 1
competition ||||, das heisst die teilnahme ||| 0.528547 0.0019528 3.71991e-06 2.38986e-14 2.718 ||| 0-4 ||| 1 142086 1
competition ||||, das dem wettbewerb ||| 0.132137 0.824898 3.71991e-06 2.27147e-06 2.718 ||| 0-3 ||| 4 142086 1
competition ||||, damit der wettbewerb ||| 0.031091 0.824898 3.71991e-06 2.46123e-06 2.718 ||| 0-3 ||| 17 142086 1
competition ||||, behindere den wettbewerb ||| 0.528547 0.824898 3.71991e-06 7.95114e-10 2.718 ||| 0-3 ||| 1 142086 1
competition ||||, aus denen der wettbewerb ||| 0.528547 0.824898 3.71991e-06 3.06829e-08 2.718 ||| 0-4 ||| 1 142086 1
competition ||||, also um zu vermeiden ||| 0.264273 0.00062696 3.71991e-06 3.39252e-13 2.718 ||| 0-4 ||| 2 142086 1
competition ||||, wettbewerbs ||| 0.105709 0.871534 3.71991e-06 0.00688575 2.718 ||| 0-1 ||| 5 142086 1
competition ||||, die konkurrenz ||| 0.176182 0.544313 3.71991e-06 6.17479e-06 2.718 ||| 0-2 ||| 3 142086 1
competition ||||, den wettbewerb ||| 0.0587274 0.824898 3.71991e-06 7.60326e-05 2.718 ||| 0-2 ||| 9 142086 1
competition ||||, darstellen ||| 0.00352364 0.0007677 3.71991e-06 1.47395e-06 2.718 ||| 0-1 ||| 150 142086 1
competition ||||, besteht ||| 0.000597228 0.001879 3.71991e-06 1.5931e-06 2.718 ||| 0-1 ||| 885 142086 1
competition ||||, (vgl. ||| 1.62126e-05 5.93e-05 3.71991e-06 9.22169e-08 2.718 ||| 0-1 ||| 1 32601 142086 1
competition ||||, (pro auswahlverfahren ||| 0.176182 0.150756 3.71991e-06 3.74981e-09 2.718 ||| 0-2 ||| 3 142086 1
competition ||||, (pro auswahlverfahren ||| 0.528547 0.150756 3.71991e-06 2.62817e-11 2.718 ||| 0-2 ||| 1 142086 1
competition ||||, (europe competition ||| 0.04404566 0.787368 3.71991e-06 3.61082e-10 2.718 ||| 0-2 ||| 12 142086 1
competition ||||, (derzeit immer intensiver werdenden ) wettbewerbs ||| 0.528547 0.871534 3.71991e-06 4.44457e-25 2.718 ||| 0-6 |||
competition ||||, (d. h. wettbewerb ||| 0.528547 0.824898 3.71991e-06 4.18807e-07 2.718 ||| 0-2 ||| 1 142086 1
competition ||||, (aeu) darstellen ||| 0.105709 0.0007677 3.71991e-06 1.51821e-13 2.718 ||| 0-3 ||| 5 142086 1
Where do wrong phrase table entries come from?

without presenting ||| zonder
without presenting the ||| zonder de
without presenting the goods ||| zonder de goederen
without presenting the goods and ||| zonder de goederen noch
without presenting the goods and the ||| zonder de goederen noch de
without presenting the goods and the corresponding

presenting the ||| de
presenting the goods ||| de goederen
presenting the goods and ||| de goederen noch
presenting the goods and the ||| de goederen noch de
presenting the goods and the corresponding

the office of departure ||| het kantoor van vertrek aan te brengen
office of departure ||| kantoor van vertrek aan te brengen
of departure ||| van vertrek aan te brengen
departure ||| vertrek aan te brengen
. ||| aan te brengen .
. ||| te brengen .
. ||| brengen .
More examples of misalignment
Observations about typical errors

Typical errors depend mainly on target language (TL)

- Morphologically simple TL: Statistical models work reasonably well
- Strongly inflected TL: Word endings are often wrong
- Differences in order between SL and TL → more alignment errors → spurious deletions & insertion

Some frequent errors can be fixed with simple means

- Certain types of expressions can be treated with rules
- Normalising punctuation helps a lot

Errors caused by different word order can be reduced

- Re-ordering before alignment reduces alignment errors
- Effect on final translation quality positive, but no breakthrough
- Increases complexity of software infrastructure → not yet in use
Enhanced Architecture for Hybrid (Rule-Based + Statistical) MT

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

Target Language Model (Fluency)
- Monolingual Corpus
- Language Model
- Preprocessing, Counting, Smoothing

Source Text → Linguistic Preprocessing → Decoder → Linguistic Postprocessing → Target Text
Ways to improve translation quality

• Collect feed-back on frequent errors, refine rule-based modules
  ongoing work based on "Language Weeks"
• Integrate linguistic tools in pre- and post-processing/PT pruning
  part-of-speech models, morphology, parsing, reordering, ...
• Distinguish relevance of training data
  domain awareness, recency, data quality indicators, ...
• More data
  typically improves coverage, but may hurt disambiguation and
  domain-awareness if not done well
• Different types of data
  Lexicons, Terminologies, Ontologies, ...
How to better serve the needs of end users

• Build models optimised for different use cases (domain adaptation), e.g. specific MT engines for DSIs that have sufficient training data
• Improve scalability
  • Better capacity and response times via cloud computing
  • Offer choice between speed and accuracy
• Better coverage of general-purpose vocabulary
• Better robustness when dealing with low-quality input
• Linguistic improvements will also help end users
How to better serve the needs of users

• Ongoing: **Implement improvements** identified during "language weeks": simple issues have been resolved
• Work on **domain adaptation** will help translators and end-users (Euramis covers many different domains)
• **Learn from** stream of **corrections** (implicit feed-back) done by translators using the system
• Explore new paradigm of **Neural MT**, especially for complex target languages
From a recent presentation by Philipp Koehn:
Neural Machine Translation (NMT)

What is NMT?
- Large Artificial Neural Networks are trained on existing translations to optimise MT performance
- NMT training induces hidden representations for textual data that can find and exploit hidden generalisations
- Radical departure from the phrase-based SMT approaches, where subcomponents are separately optimized

Why is it important?
- Can generate more fluent and grammatical translations and improves performance on data not seen in training
- Relevant for the highly inflected languages (e.g. Finno-Ugric, Baltic) where SMT performance so far was not yet convincing
- Recent highly promising results in comparative evaluations (WMT 2016) cause a dramatic shift in research focus towards NMT. SMT may become obsolete within few years
CEF and NMT: Activities and Key Challenges

What has been done within CEF?

- Members of the MT Engines team have started to explore NMT since late 2015
- We can reproduce recent research results and started to compare NMT-based engines against SMT baseline. First results look very promising, and translators in DGT's HU LD wait eagerly for NMT.
- We are about to explore NMT as a way to support domain adaptation, one of CEF's key requirements for eTranslation

What are the key challenges?

- NMT is a hot research topic; using it in a production setting is therefore still a challenge (rapid changes, speed, and scalability)
- Among other things, we need to ensure that gains in fluency do not deteriorate the adequacy of translations
- NMT requires specific hardware (GPUs for number crunching with big matrices). Currently we can only use cloud computing infrastructure available in CEF
However, expectations need to be managed carefully:
Next steps

• **Changes due to embedding of MT project into CEF**
  • New types of big users with specific needs for domain adaptation
  • Make effective use of Cloud computing infrastructure

• **Follow-up work on feed-back from users (Language Weeks)**
  • First changes of the MT system based on LWs were included in the 10th ... 12th generations of MT engines, more work is ongoing

• **Improve embedding in translators' work flow**
  • CAT tool integration, analyse changes made in post editing

• **Explore Neural MT**
  • Cloud infrastructure with GPUs, security aspects
  • Learn how to use existing tools from research community
  • Learn how to manage the fluency/accuracy tradeoff

• **Integrate more and more diverse training data**
Questions?

andreas.eisele@ec.europa.eu
DGT-MT@ec.europa.eu
Our Questions to QT21 Researchers

- Operations and Scalability
- NMT, PBSMT, pre- and post-processing
- Domain adaptation
- Low resources
- Morphology
Operations and Scalability

- MT@EC's way of deploying Moses is based on load distribution across multiple (single-threaded) Moses processes. Is there potential for speed-up using multithreading? Could another version of Moses help us to achieve the short response time some of our users wish us to deliver? Could a different layout of modules improve the overall responsivity?
NMT, PBSMT, and pre-/post-processing

- What modules could be shared across different approaches, how could a modular architecture look like?
- How can we obtain better fluency as provided in NMT, but still keep adequacy under control?
- How could hybrid combinations of different elements look like?
Domain adaptation

• How to obtain the best combined engine from (small) domain-specific and (big) generic training data? How to identify domain-specific subsets in large mixed-domain training sets?
• How to incorporate non-parallel (comparable) data for domain-adapted MT?
• For NMT: How to deal with domain-specific data that is not covered well by the original word embeddings? Can NMT engines be adapted to a new domain anyhow?
Low resources

• How can multi-source NMT be trained? How to obtain $N^*(N-1)$ translation directions with $O(N)$ effort?

• How can the main advantages of NMT and SMT be combined? How can we detect and manage inaccuracies in NMT?
Morphology

- To what extent does BPE replace explicit knowledge about word morphology? Where are the limits?
- How does rule-based morphological pre- and post-processing fit into the larger picture?