Who am I?

- My name is Mārcis Pinnis
- I work as a researcher in Tilde
- My current research focuses on neural machine translation (NMT)
- You can find out more about the research we carry out at: https://www.tilde.com/research
<table>
<thead>
<tr>
<th>Wide range of services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise or cloud-based services</td>
</tr>
<tr>
<td>Secure and scalable solution</td>
</tr>
<tr>
<td>General and customer-tailored system development</td>
</tr>
<tr>
<td>Data storage and cleaning</td>
</tr>
<tr>
<td>Data pre-processing/post-processing</td>
</tr>
<tr>
<td>SMT and NMT system support for training and decoding</td>
</tr>
<tr>
<td>Document, Web site and text translation</td>
</tr>
<tr>
<td>Translation API for third party integration</td>
</tr>
<tr>
<td>Integrated with the Tilde Terminology platform</td>
</tr>
<tr>
<td>and more...</td>
</tr>
<tr>
<td>Popular CAT tool plugins</td>
</tr>
<tr>
<td>Trados SDL Studio</td>
</tr>
<tr>
<td>memoQ</td>
</tr>
<tr>
<td>GlobalSight</td>
</tr>
<tr>
<td>Memsource</td>
</tr>
<tr>
<td>MateCat</td>
</tr>
</tbody>
</table>
MT System Development
Typical Workflow for MT System Development

1. Data acquisition and management
2. Data pre-processing
3. MT system training
4. MT system deployment and integration
Data Pre-processing
Re-formatting and Parallel Data Extraction

• Parallel data is available in an abundance of different formats
  • To name but a few that do not cause MT developers headaches:
    • Translation Memory eXchange (TMX)
    • Different flavours of XML Localization Interchange File Format (XLIFF)
    • Simple text formats, e.g., Comma Separated Values (CSV), Tab Separated Values (TSV), Moses
    • Spreadsheet documents (XLS, XLSX, ODS, etc.)
    • TermBase eXchange (TBX)
  • To name but a few that customers often have, but require significant efforts to extract parallel data:
    • Portable Document Format (PDF)
    • HTML (i.e., the “our data is on the Web” format)
    • Word processing documents (DOC, DOCX, ODT, etc.)

• The only format supported by (almost) all MT development toolkits is Moses
• Online MT development platforms usually support also different TM formats
Data Filtering

- **Data does contain noise!**
- Translation memories also contain noise
- MT methods have varied levels of noise tolerance
- **Data has to be filtered**, especially for NMT

Here are two examples of noisy data extracted from “parallel” corpora:

<table>
<thead>
<tr>
<th>The date of application in Article 15 should be 1 January 2006.</th>
<th>Lēmuma 15. pantā piemērošanas datumam jābūt 2006.</th>
</tr>
</thead>
<tbody>
<tr>
<td>This Decision shall enter into force on 1 December 2008.</td>
<td>Šis lēmums stājas spēkā 2008.</td>
</tr>
</tbody>
</table>

(2) See Article 1 (4) of the proposed regulation.

<table>
<thead>
<tr>
<th>Germany</th>
<th>A. Allemagne</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>Allemagne</td>
</tr>
<tr>
<td>Germany</td>
<td>B. Allemagne</td>
</tr>
<tr>
<td>Germany</td>
<td>D. Allemagne</td>
</tr>
<tr>
<td>Germany</td>
<td>Danemark</td>
</tr>
<tr>
<td>Germany</td>
<td>Deutschland</td>
</tr>
<tr>
<td>Germany</td>
<td>Espagne</td>
</tr>
<tr>
<td>Germany</td>
<td>F. Allemagne</td>
</tr>
<tr>
<td>Germany</td>
<td>France</td>
</tr>
<tr>
<td>Germany</td>
<td>Géorgie</td>
</tr>
</tbody>
</table>
Is Filtering Really Necessary?

- Yes, of course it is!
- Non-filtered data may hinder training a high quality system
- Systematic noise or disproportionately high levels of noise can ruin NMT systems

WMT 2017 EN-LV Shared task on News Translation
(comparison between light filtering and advanced filtering)

WMT 2018 DE-EN Shared Task on Data Filtering (work in progress)
Data Cleaning and Normalisation

• Removal of content unnecessary for MT system training:
  • E.g., formatting tags, control symbols, CAT tool or customer specific placeholders, etc.

• Normalisation:
  • Harmonisation of punctuations
  • Decoding of escaped symbols
  • Separation of ligatures (a common issue in data extracted using OCR)
Methods for Data Filtering, Cleaning, and Normalisation

• Baseline methods for data filtering, cleaning, and normalisation:
  • clean-corpus-n.perl from the Moses project ([https://github.com/moses-smt/mosesdecoder](https://github.com/moses-smt/mosesdecoder))

• A more sophisticated filtering method:
  • [https://github.com/hainan-xv/zipporah](https://github.com/hainan-xv/zipporah)

• MT providers usually provide more advanced methods for data filtering, cleaning, and normalisation. E.g., you can learn about the methods developed by Tilde by following Tilde’s publications

• If you are interested in state-of-the-art methods for data filtering, follow the WMT shared task on data filtering ([http://www.statmt.org/wmt18/parallel-corpus-filtering.html](http://www.statmt.org/wmt18/parallel-corpus-filtering.html))
A typical workflow (during translation):
- Tag removal
- Sentence splitting
- Term and named entity recognition
- Non-translatable entity recognition
- Tokenisation
- Truecasing
- Part of speech (or morpho-syntactic) tagging and/or syntactic parsing.
- Word splitting (for NMT)
MT System Training
The current state-of-the-art in MT is NMT

- The diagram depicts state-of-the-art architecture of recurrent neural networks (up to 2017)
- Convolutional neural networks
- Transformer networks (the current state-of-the-art in 2018)
Neural Machine Translation Toolkits

• Different from SMT where Moses is considered the go-to toolkit for do-it-yourself MT development, NMT has an abundance of toolkits
  • https://github.com/jonsafari/nmt-list

• The most promising (based on my subjective hands-on experience) toolkits are:
  • Marian (https://github.com/marian-nmt/marian)
    • Used by Tilde in production
  • Amazon Sockeye (https://github.com/awslabs/sockeye)
    • Used by Tilde in the WMT 2018 Shared Task on news translation
  • OpenNMT (http://opennmt.net/)
    • Has received a lot of attention by developers and may be easier for beginners, however, lacks behind state-of-the-art (in terms of quality)
Hardware Requirements and Time

• For training one NMT system:
  • A computer with an NVIDIA GPU (Titan X or GeForce GTX 1080 Ti)
  • Or access to cloud-computing services (e.g., Amazon EC2 P3.2xlarge nodes)
• Training time varies depending on:
  • The size of the training dataset
    • 4 million sentence systems can be trained as fast as within two days (using Amazon EC2 P3.8xlarge nodes; or up to 2 weeks on slower hardware)
    • 20 million sentence systems can be trained as fast as within four days (or up to 3 weeks on slower hardware)
  • Quality of the dataset (it will take longer to train a system on noisy data)
  • Language (and language characteristics)

• For translation:
  • A computer with an NVIDIA GPU (GeForce GTX 1080 Ti is more than enough)
Typical Training Scenarios (1)

- Customer has **enough in-domain data**
- In most cases, >500 thousand sentence pairs or more
- Train an in-domain NMT system on customer data only

**English-Russian IT Domain Systems**

[Graph showing BLEU scores with different curves for In-domain system, Broad domain system, and Domain adaptation.]
Typical Training Scenarios (2)

- Customer has a small in-domain corpus
- Typically <500 thousand sentence pairs
- Train a broad(er) domain system first
- Perform domain adaptation of the system using the customer’s data
Typical Training Scenarios (3)

- Customer has only monolingual data available
- Train a broad domain system (target-to-source)
- Back-translate the monolingual data
- Depending on the amount of the back-translated data:
  - If the data are sufficient (>500,000), combine the back-translated data with parallel data (1:1) and train a new system
  - If the data are limited (<500,000), adapt a pre-translated broad-domain (or close-domain) system using the back-translated data combined with parallel data (1:1)
MT System Deployment and Integration
What are the Options?

• MT providers offer a wide variety of integration possibilities
  • Popular CAT tool integration
  • Content management system integration
  • Desktop, Mobile, and Web application integration
  • Custom integration (upon customers’ requests)

• If you decide to build your own MT systems, you will have to handle the integration yourself as MT toolkits:
  • either do not provide any deployment functionality at all
  • or provide MT engine deployment as Web services, e.g.:
    • OpenNMT (http://opennmt.net/OpenNMT/tools/servers/)
    • Modern MT (https://github.com/ModernMT/MMT)
    • Marian https://marian-nmt.github.io/docs/cmd/marian-server
Integration Options provided by Tilde MT

- API for third party developers
- Translation Website and/or widget
- Mobile applications
- Desktop application
- CAT tool plugins
- Custom integration
EU Council Presidency Translator

Enabling multilingualism at the 2017-2018 EU Council Presidency
Custom NMT systems, developed specifically for the needs of the presidencies, offering higher quality translations on topics covered by the presidencies.

MT systems for the 24 official EU languages, enabling translation of full documents, preserving text formatting and style.
The motto of the Estonian presidency of the Council of the European Union is “Unity through balance”. We believe that Europe has enough common sense to find solutions to all challenges and to exploit all the capabilities that we face. We need to facilitate openness in both our economies and in society, while securing security and security. The task of Estonia as the presidency is to find a balance between the various views, traditions and interests of member states in order to achieve the best possible outcome for European citizens.
Machine translation results for documents help to understand the meaning of a source text, but do not equal translation by a human.
Links to Estonian media and information resources:

- **Delfi**
  - News site

- **ERR.ee**
  - News site

- **Æripäev**
  - News site

- **Eesti Ekspress**
  - Business news

- **Postimees**
  - Public e-services

- **EESTI.EE**

translate2017.eu
Plugin for Professional Translators
From September, 2017 to May, 2018 the EU Council Presidency Translator has processed:

- >11.5 million words
- >1.1 million sentences
- >0.3 million translation requests

~74 books (there are 155 thousand words on average in one Harry Potter book)
Usage of the custom NMT systems

Words

m6 m7 m8 m9 m10 m11 m12 m1 m2 m3 m4 m5 (half)

y2017

EN-ET  ET-EN  EN-BG  BG-EN

y2018
Types of translation requests

- Sentences:
  - Document translation: 52%
  - Other (text snippets, CAT): 31%
  - Website translation: 17%

- Words:
  - Document translation: 43%
  - Other (text snippets, CAT): 48%
  - Website translation: 10%

Legend:
- Blue: Document translation
- Orange: Other (text snippets, CAT)
- Grey: Website translation
Translation directions

• Most of the translations are performed by the systems translating from/to the language of the hosting country
• Statistics in terms of translated words:
Thank you!

Time for questions!