ADDRESSING NMT CHALLENGES IN TRAINING WITH LIMITED DATA

RIHARDS KRIŠLAUKS
rihards.krislauks@tilde.lv
AI BASED LANGUAGE TECHNOLOGY

NEURAL MACHINE TRANSLATION

NATURAL LANGUAGE PROCESSING

SPEECH PROCESSING

VIRTUAL ASSISTANTS
TILDE SUCCESS FACTORS

- RESEARCH & DEVELOPMENT & DATA CURRATION
- FAST INNOVATION CYCLE FROM CUTTING-EDGE RESEARCH TO SCALABLE APPLICATIONS
- EFFICIENT PAN-EUROPEAN COLLABORATION
Popular CAT tool plugins

Rich functionality for state-of-the-art system development

Wide range of services

• Enterprise or cloud-based services
• Secure and scalable solution
• General and customer-tailored system development

• Data storage and cleaning
• Data pre-processing/post-processing
• SMT and NMT system support for training and decoding
• Document, Web site and text translation
• Translation API for third party integration
• Integrated with the Tilde Terminology platform and more...

• Trados SDL Studio
• memoQ
• GlobalSight
• Memsource
• MateCat
CHALLENGES DUE TO LIMITED DATA.
MT requires lots of data

• Good to have 1-4M parallel sentences. More is even better.
• SMT is as data-hungry as NMT.
• However, SMT has a lot of options to deal with it:
  • Using language models trained on monolingual data;
  • Domain adaptation – using out-of-domain data;
  • Term glossaries;
  • SMT is better at remembering rare events.
A typical SMT training setup

- 50-100k in-domain sentence pairs
- 3-7M out-of-domain sentence pairs
- 5-20M monolingual sentences (target language)
- 1-2k in-domain sentence pairs for evaluation
- A few hundred terms

SMT involves a bunch of different components trained separately on each of the above types of data
How about NMT?

• Can be trained using only parallel sentences (without doing anything weird)
• The best case scenario – we have some 3-7M parallel in-domain sentences at our disposal
• What if we don’t? (And usually, we don’t)
Working around limited in-domain data

Two options:

• Use out-of-domain data
• Use monolingual data
Using out-of-domain data

1) Mix out-of-domain data with in-domain data during training
   • Lots of options on how to do the “mixing”

2) Pre-train the NMT system on out-of-domain data and tune it on in-domain data afterwards
Using monolingual data

• Produce more in-domain data via back-translating monolingual data from target language to source language:
  \[ \text{monolingual in-domain} \rightarrow \text{parallel in-domain} \]
• Afterwards use any of the methods in the previous slide
To summarize

• Can an NMT system be trained on only a corpora with 5k sentences?
  • No, at least not with current methods.
• What can be done?
  • Identify out-of-domain data. Preferably it is similar.
  • Identify monolingual data in the same domain. It’s usually a lot simpler to acquire.
  • Can’t just throw it all together – the training process must be structured accordingly. Depends on the size of gathered data.
CASE STUDY 1.

Tilde’s WMT 2017 NMT systems
WMT 2017

• WMT is an annual workshop/conference featuring multiple machine translation tasks.

• News translation task – English-Latvian and Latvian-English. What is given?
  • 4.5M parallel out-of-domain sentences (general domain);
  • 38M/370M monolingual in-domain sentences (news);
  • 2k sentence in-domain development set;
  • No parallel in-domain data.
Data cleaning allows to train the same or better quality NMT systems with less data.

For SMT systems, data filters tend to lower MT system quality. For NMT, the opposite is true.

The graph shows scores after the generic filters (from Moses SMT) were applied.
<table>
<thead>
<tr>
<th><strong>Corrupt data filters</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Corruption through encoding</td>
</tr>
<tr>
<td>• Corruption through optical character recognition</td>
</tr>
<tr>
<td>• Corruption through (external) data processing workflows</td>
</tr>
<tr>
<td>• Non-textual data filtering</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Non-parallel data filters</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Incorrect language detection</td>
</tr>
<tr>
<td>• Low content overlap filtering</td>
</tr>
<tr>
<td>• Length ratio filtering</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Data normalisation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Escaped symbol decoding</td>
</tr>
<tr>
<td>• Control symbol removal/replacement</td>
</tr>
<tr>
<td>• Tag removal</td>
</tr>
<tr>
<td>• Punctuation normalisation</td>
</tr>
</tbody>
</table>
Not all back-translated data are correct, but the models are able to learn from such data because:

1) the target is correct,
2) the data is from our required domain,
3) our original training data is from a broader domain.
CASE STUDY 2.

Creating custom systems with clients’ data
**DOMAIN ADAPTATION**

- Start out with a large general domain data-set.
- After the model has converged, continue training on a smaller domain-specific data set.
MT IN CAT TOOLS.
MT workflow

- User
- Document
- CAT tool
- MT plugin
- Translation suggestions
- MT system provider
- Translation suggestion
Plugins are CAT tool specific

- A dedicated plugin must be developed to support a specific CAT tool.
- Different ways to provide the plugin to the end-user:
  - Either already shipped with the CAT tool (e.g. Memsource, MemoQ, MateCat);
  - Or installed on the user’s computer (SDL Trados Studio).
- All provide same basic functionality – MT suggestions or pre-translation.
- Advanced functionality might differ – tag support, MT system selection capabilities.
Role of an MT system provider

- A dedicated infrastructure is necessary for MT system hosting:
  - Either in the cloud,
  - Or on premise.
- Provides interface for MT system training and/or trains MT systems and provides them to the user.
- Provides an API endpoint used by the MT plugins in CAT tools.
- Other features – data-set repository, user management, document translation in the web, etc.
NMT-specific limitations

• To enable tag support (essential for formatted documents):
  • NMT model architecture must support alignment extraction,
  • NMT systems must be trained to deal with unknown elements in text.

• Using terminology in NMT is an open research problem:
  • It is best to use a dedicated Terminology plugin to assist the translator.
New Project
Translation Memory and Automated Translation
Select translation memories and automated translation servers for the language pairs selected in the project.

All Language Pairs

Translation Memory and Automated Translation
Create a Translation Memory or select an existing one or connect to Machine Translation or other resources.

Language Resources
Estonian (Estonia) - English

PLUGIN SCREENSHOTS
To start translation, please choose a language pair and a translation system.

Group: EU Presidency

From: Estonian
To: English

EU Presidency NMT system (ET-EN) (Running)
Productivity gains with MT

Productivity increase

<table>
<thead>
<tr>
<th>Translator</th>
<th>Translated tokens/hour</th>
<th>Post-edited tokens/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57%</td>
<td>207%</td>
</tr>
<tr>
<td>2</td>
<td>1500</td>
<td>3000</td>
</tr>
<tr>
<td>3</td>
<td>1500</td>
<td>2250</td>
</tr>
<tr>
<td>4</td>
<td>1500</td>
<td>2250</td>
</tr>
</tbody>
</table>

Average: 149%
### SMT vs. Neural MT (human evaluation)

<table>
<thead>
<tr>
<th></th>
<th>(U) en-lv</th>
<th>(U) lv-en</th>
<th>(C) en-lv</th>
<th>(C) lv-en</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15.9%</td>
<td>13.9%</td>
<td>4.9%</td>
<td>7.7%</td>
</tr>
<tr>
<td></td>
<td>32.7%</td>
<td>17.1%</td>
<td>34.8%</td>
<td>31.9%</td>
</tr>
<tr>
<td></td>
<td>34.2%</td>
<td>50.9%</td>
<td>44.8%</td>
<td>44.2%</td>
</tr>
</tbody>
</table>

- **SMT**: 0%
- **Neither or both**: 20%
- **NMT**: 80%
- **100%**: 60%

Human evaluation: NMT is better.
The latest scientific publications on NMT by Tilde are:


Thank you!