<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Asistent -- a machine translation system for Slovene, Serbian and Croatian</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Arcan, Mihael; Popovic, Maja; Buitelaar, Paul</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2016-09-29</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>University of Ljubljana</td>
</tr>
<tr>
<td><strong>Link to publisher's version</strong></td>
<td><a href="http://www.sdjt.si/wp/dogodki/konference/jdth-2016-english/">http://www.sdjt.si/wp/dogodki/konference/jdth-2016-english/</a></td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/14893">http://hdl.handle.net/10379/14893</a></td>
</tr>
</tbody>
</table>

Downloaded 2019-08-03T18:56:21Z

Some rights reserved. For more information, please see the item record link above.
Asistent
A Machine Translation System for Slovene, Serbian and Croatian

Mihael Arcan, Maja Popovic, Paul Buitelaar

Insight Center for Data Analytics, NUI Galway, Ireland
Humboldt University of Berlin, Germany
Asistent

ASISTENT (or assistant@en) was developed to translate text between English and the morphological complex south Slavic languages: Slovene, Serbian and Croatian.

Select translation table option/approach:
- phrase based
- factored
- hierarchical
- direct translation
- pivot translation

Please choose the translation direction:
- English → Slovene
- Slovene → English
- English → Croatian
- Croatian → English
- English → Serbian
- Serbian → English
- Slovene → Croatian
- Croatian → Slovene
- Serbian → Slovene
- Slovene → Serbian
- Croatian → Serbian
- Serbian → Croatian

This service is brought to you by
http://nlp.insight-centre.org/

http://server1.nlp.insight-centre.org/asistent/
## Motivation


### Machine Translation

<table>
<thead>
<tr>
<th>Excellent support</th>
<th>Good support</th>
<th>Moderate support</th>
<th>Fragmentary support</th>
<th>Weak/no support</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>French</td>
<td>Catalan</td>
<td>Basque</td>
<td>Croatian</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>Dutch</td>
<td>Bulgarian</td>
<td>Czech</td>
</tr>
<tr>
<td></td>
<td></td>
<td>German</td>
<td>Danish</td>
<td>Croatian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hungarian</td>
<td>Estonian</td>
<td>Finnish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Italian</td>
<td>Finnish</td>
<td>Galician</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polish</td>
<td>Greek</td>
<td>Greek</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Romanian</td>
<td>Icelandic</td>
<td>Greek</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Irish</td>
<td>Icelandic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Latvian</td>
<td>Irish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lithuanian</td>
<td>Latvian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Maltese</td>
<td>Lithuanian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Norwegian</td>
<td>Maltese</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Bokmål, Nynorsk)</td>
<td>Norwegian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Portuguese</td>
<td>Norwegian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slovak</td>
<td>Polish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slovene</td>
<td>Romanian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Serbian</td>
<td>Romanian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slovak</td>
<td>Russian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slovene</td>
<td>Russian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Swedish</td>
<td>Russian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Welsh</td>
<td>Russian</td>
</tr>
</tbody>
</table>
What Do We Need?

Translation Toolkit:
- Moses (text tokenisation, lower-casing)
- Giza++ (word alignment)
- KenLM (language model)

Parallel data
- Europarl
- DGT
- OpenSubTitles

[Image: http://server1.nlp.insight-centre.org/asistent/]

http://server1.nlp.insight-centre.org/asistent/
## Parallel Corpora Used for Asistent

*http://opus.lingfil.uu.se/index.php*

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>En-Sl</th>
<th>En-Hr</th>
<th>En-Sr</th>
<th>Sl-Hr</th>
<th>Sl-Sr</th>
<th>Hr-Sr</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGT</td>
<td>legal</td>
<td>1.8M</td>
<td>196K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECB</td>
<td>finance</td>
<td>79K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMEA</td>
<td>medical</td>
<td>253K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europarl</td>
<td>EP proceedings</td>
<td>599K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gnome</td>
<td>IT</td>
<td>998</td>
<td>5K</td>
<td>126</td>
<td>4K</td>
<td>600K</td>
<td>300K</td>
</tr>
<tr>
<td>hrenWaC</td>
<td>Web Corpus</td>
<td>86K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JRC-Acquis</td>
<td>legal</td>
<td>29K</td>
<td>38K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KDE</td>
<td>IT</td>
<td>58K</td>
<td>32K</td>
<td>85K</td>
<td>49k</td>
<td>33.2k</td>
<td></td>
</tr>
<tr>
<td>LangCourse</td>
<td>education</td>
<td></td>
<td></td>
<td>3K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHP</td>
<td>IT</td>
<td>1K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OpenSubtitles</td>
<td>subtitles</td>
<td>10.1M</td>
<td>16.3M</td>
<td>20.5M</td>
<td>6.1M</td>
<td>13.3M</td>
<td>22.3M</td>
</tr>
<tr>
<td>SETimes</td>
<td>Web Corpus</td>
<td>198K</td>
<td>209K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tatoeba</td>
<td>education</td>
<td>777</td>
<td>633</td>
<td></td>
<td></td>
<td></td>
<td>200K</td>
</tr>
<tr>
<td>TED</td>
<td>education</td>
<td>13K</td>
<td>76K</td>
<td>1K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ubuntu</td>
<td>IT</td>
<td>8K</td>
<td></td>
<td>557</td>
<td>86K</td>
<td>51K</td>
<td></td>
</tr>
</tbody>
</table>
Concatenated Data for Assistant Training

Millions

- En-Sl
- En-Hr
- En-Sr
- Sl-Hr
- Sl-Sr
- Hr-Sr

- Vocabulary L1
- Vocabulary L2
- Parallel sentences

Millions

- En-Sl
- En-Hr
- En-Sr
- Sl-Hr
- Sl-Sr
- Hr-Sr
Concatenated Data for Assistant Training

- **En-Sl**, **En-Hr**, **En-Sr**
- **Sl-Hr**, **Sl-Sr**, **Hr-Sr**

- Vocabulary L1
- Vocabulary L2
- Parallel sentences

![Bar Chart](chart.png)
Concatenated Data for Assistant Training

- **Millions**
  - En-SI
  - En-Hr
  - En-Sr
  - Sl-Hr
  - Sl-Sr
  - Hr-Sr

- **vocabulary L1**
- **vocabulary L2**
- **parallel sentences**

- **Millions**
  - En-SI
  - En-Hr
  - En-Sr
  - Sl-Hr
  - Sl-Sr
  - Hr-Sr

- **Insight**
Data Preparation

Data cleanup

- Falsely encoded special characters removed/fixed
- Cyrillic $\rightarrow$ Latin (Serbian)
- removing special symbols, as "#", "%" and "@"
- Slovak bi-text was removed from the Tatoeba corpus

Corpora filtering based on the sentence length proportions

- too short and too long sentences were not included into the training set (average length +/- standard deviation)

Evaluation / Development set creation

- avoiding too short/long sentences (10 and 40 words / 5 to 15 words)
Data Preparation (example English-Slovene)

- **parallel sentences**
  - non-preprocessed: 15.5
  - preprocessed: 13

- **entries in TM**
  - non-preprocessed: 230
  - preprocessed: 610

- **unique source words in TM**
  - non-preprocessed: 590
  - preprocessed: 545
Data Preparation (example English-Slovene)

- **Parallel sentences**
  - Non-preprocessed: 14
  - Preprocessed: 13

- **Entries in TM**
  - Non-preprocessed: 225
  - Preprocessed: 610

- **Unique source words in TM**
  - Non-preprocessed: 540
  - Preprocessed: 610
Data Preparation

BLEU

- non-preprocessed
- preprocessed

English->Slovene: +0.41
Slovene->English: +1.15
Automatic Evaluation, In-Domain (English-Slavic)

BLEU

Google  Asistent

En->Sl  SL->En  En->Hr  Hr->En  En->Sr  Sr->En

http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz
Automatic Evaluation, In-Domain (Slavic-Slavic)

BLEU

Google  Asistent

http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz
Pivot Translation

... can enable a bridge between languages, when existing parallel corpora are under-resourced ...

**Direct Translation:**
- source language → target language

**Pivot Translation:**
- source language → pivot language → target language

Diagram:
- English → Serbian
- Serbian → Croatian
- Croatian → Slovene
Pivot Translation

- TM (En→Sr) + LM (Sr)
- TM (Hr→Sl) + LM (Sl)
- TM (Sr→Sl) + LM (Sr)
- TM (Hr→Sl) + LM (Sl)

Languages:
- English
- Serbian
- Croatian
- Slovene

Insight
Evaluation for Pivot Translation, In-Domain (English-Slavic)

<table>
<thead>
<tr>
<th>Language Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-&gt;SI</td>
</tr>
<tr>
<td>SI-&gt;En</td>
</tr>
<tr>
<td>En-&gt;Hr</td>
</tr>
<tr>
<td>Hr-&gt;En</td>
</tr>
<tr>
<td>En-&gt;Sr</td>
</tr>
<tr>
<td>Sr-&gt;En</td>
</tr>
</tbody>
</table>

**BLEU**

- Asistent (direct)
- Asistent (pivot)
Evaluation for Pivot Translation, In-Domain (Slavic-Slavic)

BLEU

Asistent (direct)  Asistent (pivot)

- Sl->Sr
- Sr->Sl
- Hr->Sr
- Sr->Hr
- Sl->Hr
- Hr->Sl
Automatic Evaluation, Out-Domain (Massive Open Online Courses MOOCs)

BLEU

<table>
<thead>
<tr>
<th>Translation</th>
<th>Google</th>
<th>Asistent (direct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Croatian-&gt;English</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>English-&gt;Croatian</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>Serbian-&gt;English</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>English-&gt;Serbian</td>
<td>18</td>
<td>15</td>
</tr>
</tbody>
</table>
Automatic Evaluation, Out-Domain
(Massive Open Online Courses MOOCs)

BLEU

Google  Asistent (direct)  Asistent (pivot)

Croatian->English  English->Croatian  Serbian->English  English->Serbian
ASISTENT (or assistant@en) was developed to translate text between English and the morphological complex south Slavic languages: Slovene, Serbian and Croatian.

Select translation table option/approach:
- phrase based
- factored
- hierarchical
- direct translation
- pivot translation

Please choose the translation direction:
- English → Slovene
- Slovene → English
- English → Croatian
- Croatian → English
- English → Serbian
- Serbian → English
- Slovene → Croatian
- Croatian → Slovene
- Serbian → Slovene
- Slovene → Serbian
- Croatian → Serbian
- Serbian → Croatian

This service is brought to you by
http://nlp.insight-centre.org/
Filtered vs Original Translation Models

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Asistent</th>
<th>Original TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-&gt;Sl</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>Sl-&gt;En</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>En-&gt;Sr</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Sr-&gt;En</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>En-&gt;Hr</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Hr-&gt;En</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Sl-&gt;Sr</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Sr-&gt;Sl</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Sl-&gt;Hr</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Hr-&gt;Sl</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Sr-&gt;Hr</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Hr-&gt;Sr</td>
<td>20</td>
<td>30</td>
</tr>
</tbody>
</table>
Accusations of witchcraft are also common in other African countries.
Accusations of witchcraft are also common in other African countries.
Future Work

Comparison between different SMT methods
- hierarchical models
- factored models (added linguistic information)
- neural machine translation

Manual Evaluation
- pivot translations
- phrase-based vs hierarchical models

Making Asistent/TM (more) accessible
Asistent - A Machine Translation System for Slovene, Serbian and Croatian

Mihael Arcan

mihael.arcan@insight-centre.org

Thank You
Hierarchical Models Evaluation (BLEU)

- PB-SMT
- Hierarchical