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

Add text to be translated and specify the translation direction:

- English Example
- Slovene Example
- Croatian Example
- Serbian Example

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></td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>Dutch</td>
<td>Hungarian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>German</td>
<td>Italian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hungarian</td>
<td>Polish</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Italian</td>
<td>Romanian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basque</td>
<td>Croatian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bulgarian</td>
<td>Czech</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Croatian</td>
<td>Danish</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Estonian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Finnish</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Galician</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Greek</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Icelandic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Irish</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Latvian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lithuanian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Maltese</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Norwegian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Bokmål, Nynorsk)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Portuguese</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Serbian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slovak</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slovene</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Swedish</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Welsh</td>
<td></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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>En-SI</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></td>
<td>557</td>
<td>86K</td>
<td>51K</td>
</tr>
</tbody>
</table>

*http://opus.lingfil.uu.se/index.php*
Concatenated Data for Assistant Training

- **Millions of Vocabulary**
  - L1: Blue bars
  - L2: Red bars
  - Parallel Sentences: Green bars

- **Languages**
  - En-Sl
  - En-Hr
  - En-Sr
  - Sl-Hr
  - Sl-Sr
  - Hr-Sr

- **Counts**
  - En-Sl: 0.6
  - En-Hr: 0.4
  - En-Sr: 1.2
  - Sl-Hr: 0.4
  - Sl-Sr: 1.2
  - Hr-Sr: 1.6

- **Counts**
  - En-Si (millions): 10
  - En-Hr (millions): 15
  - En-Sr (millions): 25
  - Sl-Hr (millions): 5
  - Sl-Sr (millions): 15
  - Hr-Sr (millions): 20
Concatenated Data for Assistant Training

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

Millions

- vocabulary L1
- vocabulary L2
- parallel sentences

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

Millions
Concatenated Data for Assistant Training

- **En-SI**, **En-Hr**, **En-Sr**: Vocabulary L1
- **Sl-Hr**, **Sl-Sr**, **Hr-Sr**: Vocabulary L2
- **Parallel sentences**

**Bar Graphs**:
- Millions

Counts:
- En-US: 1.6
- En-HE: 1.2
- En-UA: 1.4
- Sl-C: 0.8
- Sl-SL: 0.6
- Sl-CR: 0.4
- Hr-SR: 0.2

**Comparative Analysis**:
- Vocabulary L1 and L2
- Parallel sentence counts for different language pairs
Concatenated Data for Assistant Training

- **En-Sl**: 0.6
- **En-Hr**: 0.8
- **En-Sr**: 1.6
- **Sl-Hr**: 0.4
- **Sl-Sr**: 1.2
- **Hr-Sr**: 1.4

- **vocabulary L1**: 10
- **vocabulary L2**: 15
- **parallel sentences**: 20

**Legend**:
- Blue: vocabulary L1
- Red: vocabulary L2
- Green: parallel sentences
Data Preparation

Data cleanup
- Falsely encoded special characters removed/fixed
- Cyrillic → 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 to 16
  - preprocessed: 13 to 13.5

- entries in TM:
  - non-preprocessed: 200 to 205
  - preprocessed: 230 to 235

- unique source words in TM:
  - non-preprocessed: 520 to 530
  - preprocessed: 580 to 590
Data Preparation (example English-Slovene)

- Parallel sentences
- Entries in TM
- Unique source words in TM

Comparison of non-preprocessed vs. preprocessed data.
Data Preparation

- English -> Slovene: +0.41
- Slovene -> English: +1.15

BLEU scores for non-preprocessed vs preprocessed data.
Automatic Evaluation, In-Domain (English-Slavic)

BLEU

Google  Asistent

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

BLEU

Google  Asistent

Sl->Sr  Sr->Sl  Sl->Hr  Hr->Sl  Sr->Hr  Hr->Sr

http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz
can enable a bridge between languages, when existing parallel corpora are under-resourced ...

Direct Translation: source language $\rightarrow$ target language

Pivot Translation: source language $\rightarrow$ pivot language $\rightarrow$ target language

English $\rightarrow$ Serbian $\rightarrow$ Slovene $\rightarrow$ Croatian
Pivot Translation

English

TM (En→Sr) + LM (Sr)

TM (En→Hr) + LM (Hr)

Serbian

Croatian

TM (Sr→Sl) + LM (Sl)

TM (Hr→Sl) + LM (Sl)

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

- **Google**
- **Asistent (direct)**

- **Croatian→English**
- **English→Croatian**
- **Serbian→English**
- **English→Serbian**
Automatic Evaluation, Out-Domain (Massive Open Online Courses MOOCs)

BLEU

Google  Asistent (direct)  Asistent (pivot)

Croatian->English  English->Croatian  Serbian->English  English->Serbian
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
- Slovene → Serbian
- 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/
Filtered vs Original Translation Models

BLEU

Asistent  original TM

En->Sl  Sl->En  En->Sr  Sr->En  En->Hr  Hr->En  Sl->Sr  Sr->Sl  Sl->Hr  Hr->Sl  Sr->Hr  Hr->Sr
Asistent’s API request

Translation request (json)

Provided translation by Asistent

```json
{
    "nbest": "5",
    "translation_direction": "en_sl",
    "method": "phrase_based",
    "approach": "direct",
    "text2translate": [
        {
            "source": "Accusations of witchcraft are also common in other African countries."
        }
    ]
}
```

http://server1.nlp.insight-centre.org/asistent/rest_service.html
Asistent’s API request

{
"time":"6 wallclock secs ( 0.02 usr 0.01 sys + 5.16 cusr 0.42 csys = 5.61 CPU)",
"translation_direction":"en_sl",
"nbest":"3",
"method":"phrase_based",
"approach":"direct",
"text2translate":[
{
"source":"Accusations of witchcraft are also common in other African countries.",
"possible_translations":{
"obtožbe so pogosti tudi v čarovništva, druge afriške države.":"-9.741",
"obtožbe čarovništva so pogosti tudi v drugih afriških državah.":"-9.644",
"obtožbe o čarovništvu so pogosti tudi v drugih afriških državah.":"-9.706"
},
"best":"obtožbe čarovništva so pogosti tudi v drugih afriških državah."
}
],
}
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

BLEU