Sisay Adugna
Haramaya University
Ethiopia
sisayie@gmail.com

Andreas Eisele
DFKI GmbH
Germany
eisele@dfki.de
• **Afaan Oromoo (ISO Language Code: om)**
  - 17 million people's mother tongue – MS Encarta
  - 24,395,000 people's Official working language-CSA
  - Spoken also in Kenya and Somalia

• **English (ISO Language Code: en)**
  - Lingua franca of online information.
  - 71% of all web pages – www.oclc.org
1. to test how far we can go with the available limited parallel corpus for the English – Oromo language pair and the applicability of existing Statistical Machine Translation (SMT) systems on this language pair.

2. to analyze the output of the system with the objective of identifying the challenges that need to be tackled.
Data

- Documents include the Constitution of FDRE (Federal Democratic Republic of Ethiopia),
- Proclamations of the Council of Oromia Regional State,
- Universal Declaration of Human Right and Kenyan Refugee Act
- Religious and medical documents

Source

- Council of Oromia Regional State (Caffee Oromiyaa)
- WWW
Size and organization

- 20K Sentence pairs (EN, OM) or (300,000 words) for TM
- 62K Sentences (OM) or (1,024,156 words) for LM
- 90% for training and 10% for testing
- Preprocessing: PERL and python scripts
- Language Modeling: SRILM
- Alignment: GIZA++
- Phrase-based Translation Modeling: Moses
- Decoding: Moses
- Postprocessing: PERL scripts
- Evaluation: PERL Script
- Demonstration: Python Scripts
Sentence aligner mistake in tokenization

- Due to appostrophe called \textit{hudhaa(\textbackslash`)} in Oromo
- Wrong tokenization bal'ina $\rightarrow$ bal ‘ ina
- Results in wrong alignment
- mis-alligned sentences pairs were found to cause lower BLUE score of 5.06%
- Example of wrongly aligned sentence pair

<table>
<thead>
<tr>
<th>Sentence 34</th>
<th>BLEU: 0.3672 (0.5455/0.40/3.33/0.25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>PART TWO Payment of Rural Land Use Payment And Income Tax.</td>
</tr>
<tr>
<td>Ref 0</td>
<td>2) Dhaabbinee invastarootni lafa baadiyya seeraan kenna meli hundumarratti kaffaltii itti fayyadama lafa baadiyya raawwachuq gabu.</td>
</tr>
<tr>
<td>Output 0</td>
<td>KUTAA LAMA kaffaltii kaffaltii itti fayyadama lafa baadiyyaasi gibrigalii.</td>
</tr>
</tbody>
</table>

- Correcting the sentence pairs manually improved BLUE score to 17.74%
Result and Discussion ...

- Result after improving the alignment

- Average BLEU Score of 17.74%
- As n increases, accuracy decreases sharply
In addition to limited size and impurity of the data, the BLUE score was affected by:

- Availability of a single reference translation
- Domain of the test data
  - the system performs better if it is tested on religious documents than documents from other domain
● How well has this system performed?
  ● Average score was 17.74%
● Compare?
  ● No MT for Oromoo
● Compared to other systems
  ● Fair score as shown in the tables on the following slide
• Size

<table>
<thead>
<tr>
<th>Language</th>
<th>Days</th>
<th>Chapters</th>
<th>Speaker Turns</th>
<th>Sentences</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danish (da)</td>
<td>492</td>
<td>4,120</td>
<td>90,017</td>
<td>1,032,764</td>
<td>27,153,424</td>
</tr>
<tr>
<td>German (de)</td>
<td>492</td>
<td>4,119</td>
<td>90,135</td>
<td>1,023,115</td>
<td>27,302,541</td>
</tr>
<tr>
<td>Greek (el)</td>
<td>398</td>
<td>3,712</td>
<td>66,928</td>
<td>746,834</td>
<td>27,772,533</td>
</tr>
<tr>
<td>English (en)</td>
<td>488</td>
<td>4,055</td>
<td>88,908</td>
<td>1,011,476</td>
<td>28,521,967</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>492</td>
<td>4,125</td>
<td>90,305</td>
<td>1,029,155</td>
<td>30,007,569</td>
</tr>
<tr>
<td>French (fr)</td>
<td>492</td>
<td>4,125</td>
<td>90,335</td>
<td>1,023,523</td>
<td>32,550,260</td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>442</td>
<td>3,627</td>
<td>81,370</td>
<td>941,890</td>
<td>18,841,346</td>
</tr>
<tr>
<td>Italian (it)</td>
<td>492</td>
<td>4,117</td>
<td>90,030</td>
<td>979,543</td>
<td>28,786,724</td>
</tr>
<tr>
<td>Dutch (nl)</td>
<td>492</td>
<td>4,122</td>
<td>90,112</td>
<td>1,042,482</td>
<td>28,763,729</td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td>492</td>
<td>4,125</td>
<td>90,329</td>
<td>1,014,128</td>
<td>29,213,348</td>
</tr>
<tr>
<td>Swedish (sv)</td>
<td>492</td>
<td>3,627</td>
<td>81,246</td>
<td>947,493</td>
<td>23,535,265</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source Language</th>
<th>da</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>fi</th>
<th>it</th>
<th>nl</th>
<th>pt</th>
<th>sv</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>-</td>
<td>18.4</td>
<td>21.1</td>
<td>28.5</td>
<td>26.4</td>
<td>28.7</td>
<td>14.2</td>
<td>22.2</td>
<td>21.4</td>
<td>24.3</td>
<td>28.3</td>
</tr>
<tr>
<td>de</td>
<td>22.3</td>
<td>-</td>
<td>20.7</td>
<td>25.3</td>
<td>25.4</td>
<td>27.7</td>
<td>11.8</td>
<td>21.3</td>
<td>23.4</td>
<td>23.2</td>
<td>20.5</td>
</tr>
<tr>
<td>el</td>
<td>22.7</td>
<td>17.4</td>
<td>-</td>
<td>27.2</td>
<td>31.2</td>
<td>32.1</td>
<td>11.4</td>
<td>26.8</td>
<td>20.0</td>
<td>27.6</td>
<td>21.2</td>
</tr>
<tr>
<td>en</td>
<td>25.2</td>
<td>17.6</td>
<td>23.2</td>
<td>-</td>
<td>30.1</td>
<td>31.1</td>
<td>13.0</td>
<td>25.3</td>
<td>21.0</td>
<td>27.1</td>
<td>24.8</td>
</tr>
<tr>
<td>es</td>
<td>24.1</td>
<td>18.2</td>
<td>28.3</td>
<td>30.5</td>
<td>-</td>
<td>40.2</td>
<td>12.5</td>
<td>32.3</td>
<td>21.4</td>
<td>35.9</td>
<td>23.9</td>
</tr>
<tr>
<td>fr</td>
<td>23.7</td>
<td>18.5</td>
<td>26.1</td>
<td>30.0</td>
<td>38.4</td>
<td>-</td>
<td>12.6</td>
<td>32.4</td>
<td>21.1</td>
<td>35.3</td>
<td>22.6</td>
</tr>
<tr>
<td>fi</td>
<td>20.0</td>
<td>14.5</td>
<td>18.2</td>
<td>21.8</td>
<td>21.1</td>
<td>22.4</td>
<td>-</td>
<td>18.3</td>
<td>17.0</td>
<td>19.1</td>
<td>18.8</td>
</tr>
<tr>
<td>it</td>
<td>21.4</td>
<td>16.9</td>
<td>24.8</td>
<td>27.8</td>
<td>34.0</td>
<td>36.0</td>
<td>11.0</td>
<td>-</td>
<td>20.0</td>
<td>31.2</td>
<td>20.2</td>
</tr>
<tr>
<td>nl</td>
<td>20.5</td>
<td>18.3</td>
<td>17.4</td>
<td>23.0</td>
<td>22.9</td>
<td>24.6</td>
<td>10.3</td>
<td>20.0</td>
<td>-</td>
<td>20.7</td>
<td>19.0</td>
</tr>
<tr>
<td>pt</td>
<td>23.2</td>
<td>18.2</td>
<td>26.4</td>
<td>30.1</td>
<td>37.9</td>
<td>39.0</td>
<td>11.9</td>
<td>32.0</td>
<td>20.2</td>
<td>-</td>
<td>21.9</td>
</tr>
<tr>
<td>sv</td>
<td>30.3</td>
<td>18.9</td>
<td>22.8</td>
<td>30.2</td>
<td>28.6</td>
<td>29.7</td>
<td>15.3</td>
<td>23.9</td>
<td>21.9</td>
<td>25.9</td>
<td>-</td>
</tr>
</tbody>
</table>

(From Koehn, 2005)
Grow of parallel corpora for this language pair using the output of the system

Consider collection and use of comparable corpora

Building linguistic models of Oromo morphology in a suitable finite-state formalism
**EuroMatrix Plus** plans to build

- easy-to-access MT engines for many EU language pairs
- a platform for translation and post-editing of Wikipedia articles

Languages like Oromoo could be easily incorporated

**ACCURAT** works on learning of MT models from comparable corpora, which would be highly applicable to Oromoo

We would need additional manpower to make this happen
EU projects EuroMatrix and EuroMatrix Plus
Saarland University
DFKI GmbH
Addis Ababa University
German Academic Exchange Service (DAAD)