English – Afaan Oromoo Machine Translation: An Experiment Using a Statistical Approach

Sisay Adugna
Haramaya University
Ethiopia
sisayie@gmail.com

Andreas Eisele
DFKI GmbH
Germany
eisele@dfki.de

Outline

- Introduction
- Objectives
- Experiment
- Result and Discussion
- Conclusion
- Next Steps
- Acknowledgement

Introduction

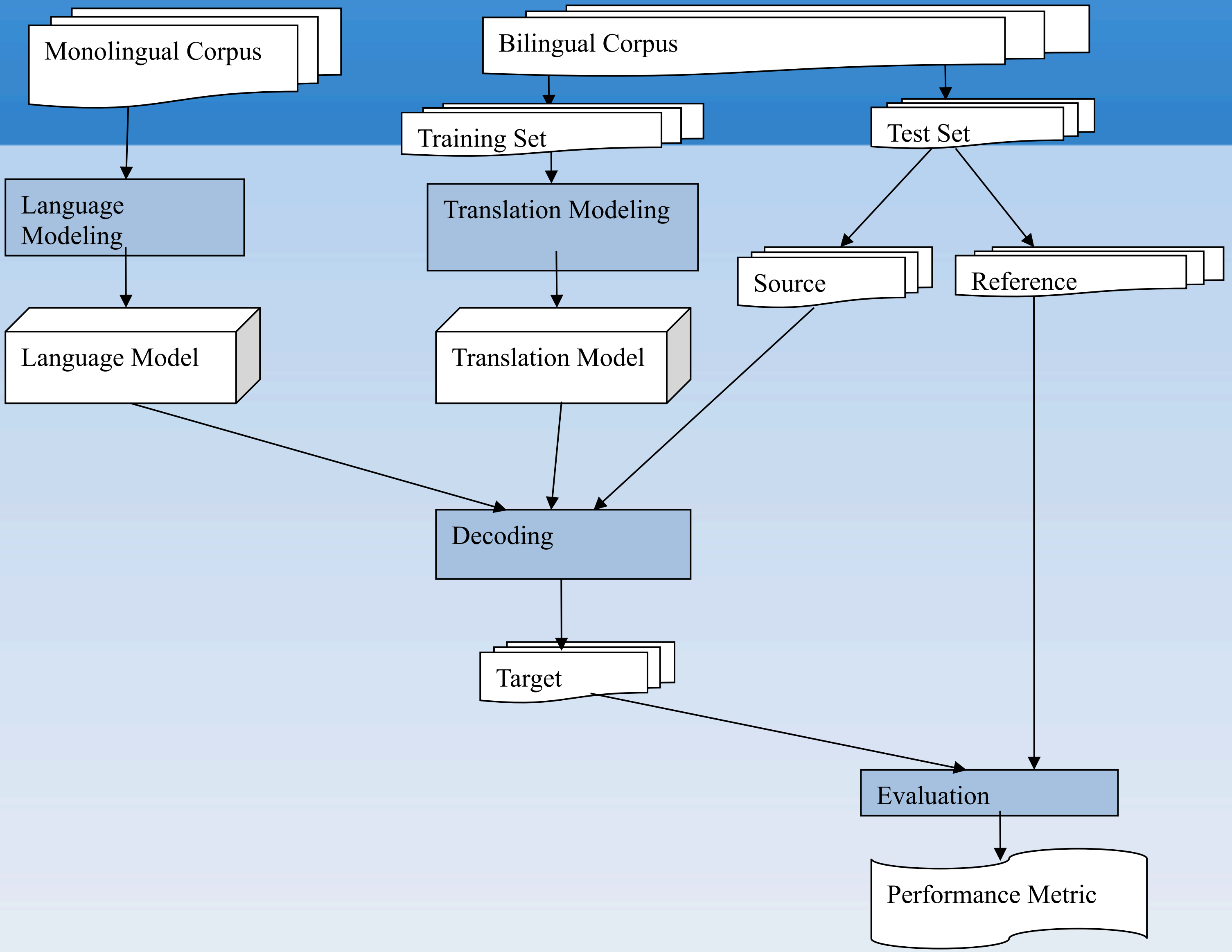
- *Afaan Oromoo* (ISO Language Code: om)
  - 17 million people's mother tongue – MS Encarta
  - 24,395,000 people's Official working language-CSA
  - Spoken also in Kenya and Somalia
- English (ISO Language Code: en)
  - Lingua franca of online information.
  - 71% of all web pages – www.oclc.org

Objectives

- The paper has two main goals:

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.

Experiment


Monolingual Corpus
Bilingual Corpus
Training Set
Test Set
Language Modeling
Translation Modeling
Source
Reference
Language Model
Translation Model
Decoding
Target
Evaluation
Performance Metric

Experiment ...

- 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

Experiment ...

- 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

Experiment ...

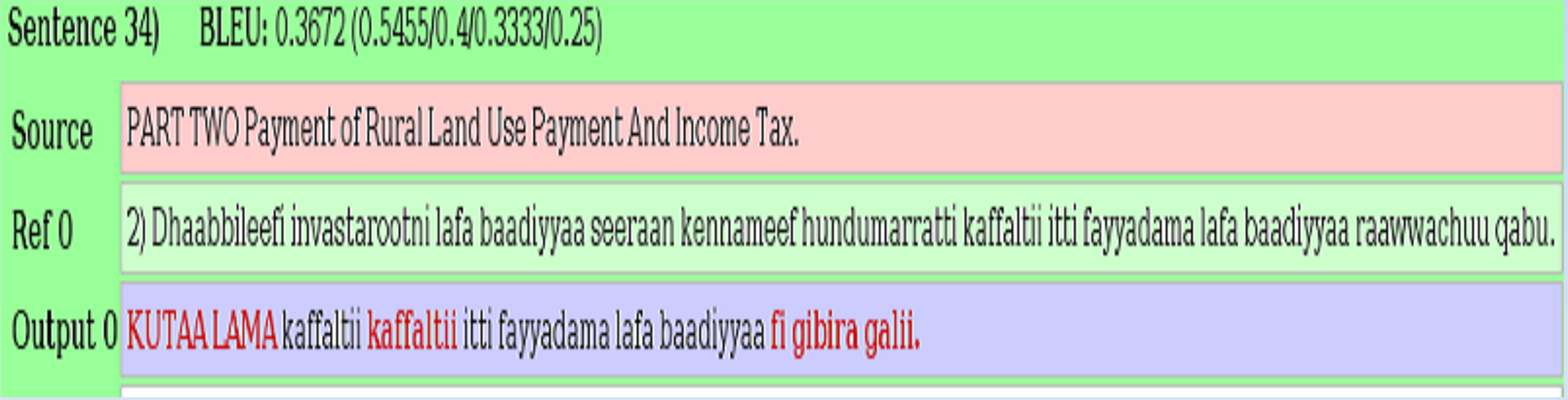
- Software tools used
  - Preprocessing : PERL and python scripts
  - Language Modeling: SRILM
  - Alignment: GIZA++
  - Phrase-based Translation Modeling: Moses
  - Decoding: Moses
  - Postprocessing: PERL scripts
  - Evaluation: PERL Script
  - Demonstration: Python Scripts

Result and Discussion

- Sentence aligner mistake in tokenization
  - Due to appostrophe called *hudhaa(`)| in Oromo*
  - Wrong tokenization bal'ina → bal ‘ ina
  - Results in wrong alignment

Result and Discussion ...

- Impurity in the data
  - mis-alligned sentences pairs were found to cause lower BLUE score of 5.06%
  - Example of wrongly aligned sentence pair



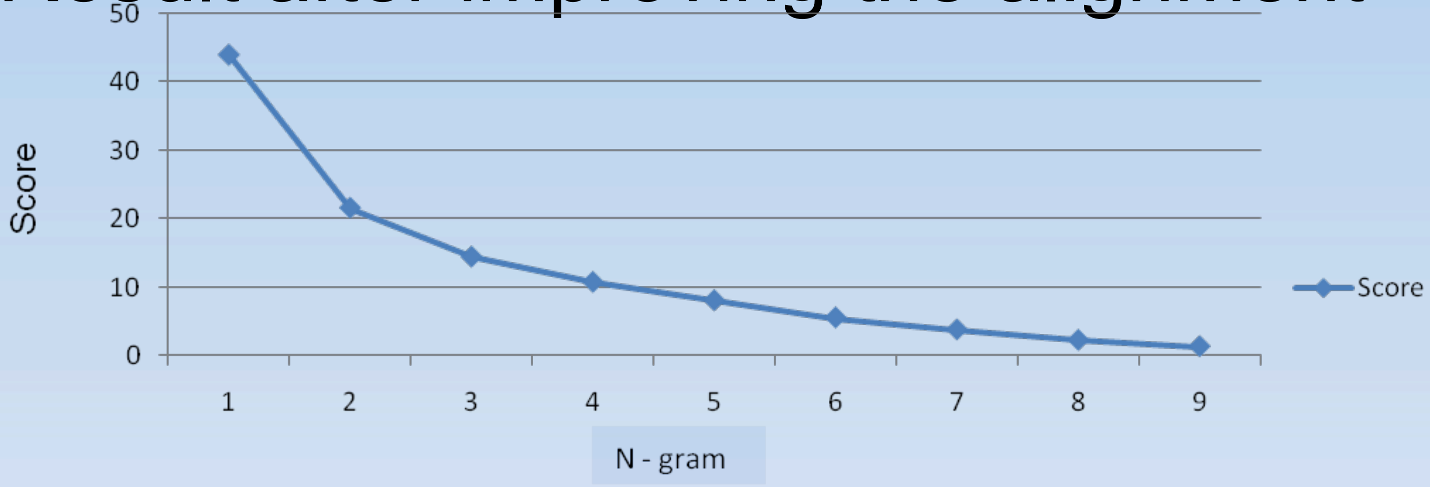
Sentence 34) BLEU: 0.3672 (0.5455/0.4/0.3333/0.25)

Source	PART TWO Payment of Rural Land Use Payment And Income Tax.
Ref 0	2) Dhaabbileefi invastarootni lafa baadiyyaa seeraan kennameef hundumarratti kaffaltii itti fayyadama lafa baadiyyaa raawwachuu qabu.
Output 0	KUTAA LAMA kaffaltii kaffaltii itti fayyadama lafa baadiyyaa fi gibira galii.

  - 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

Result and Discussion ...

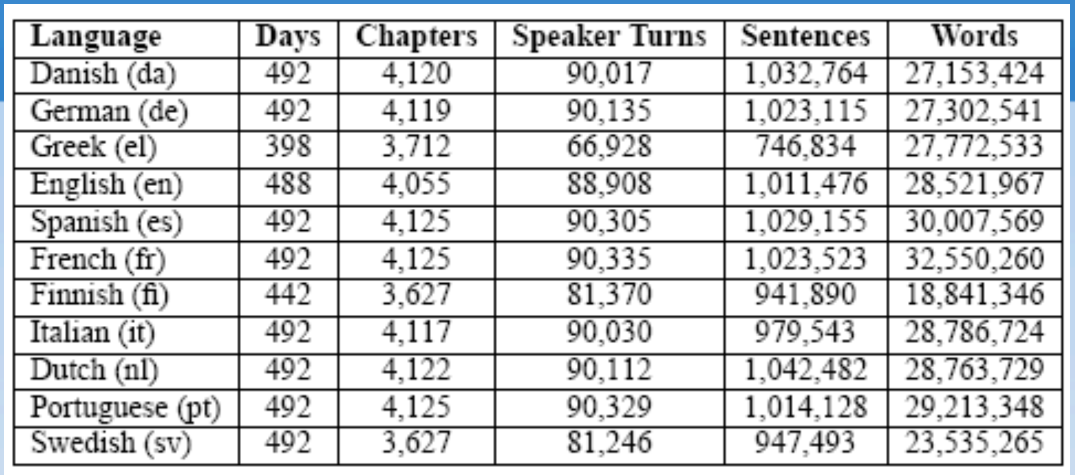
- In addition to limited size and impurity of the data, the BLUE score was affected by:
  - Availability of a single reference translation
  - Domain of the test data
    - the system performs better if it is tested on religious documents than documents from other domain

Conclusion

- 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

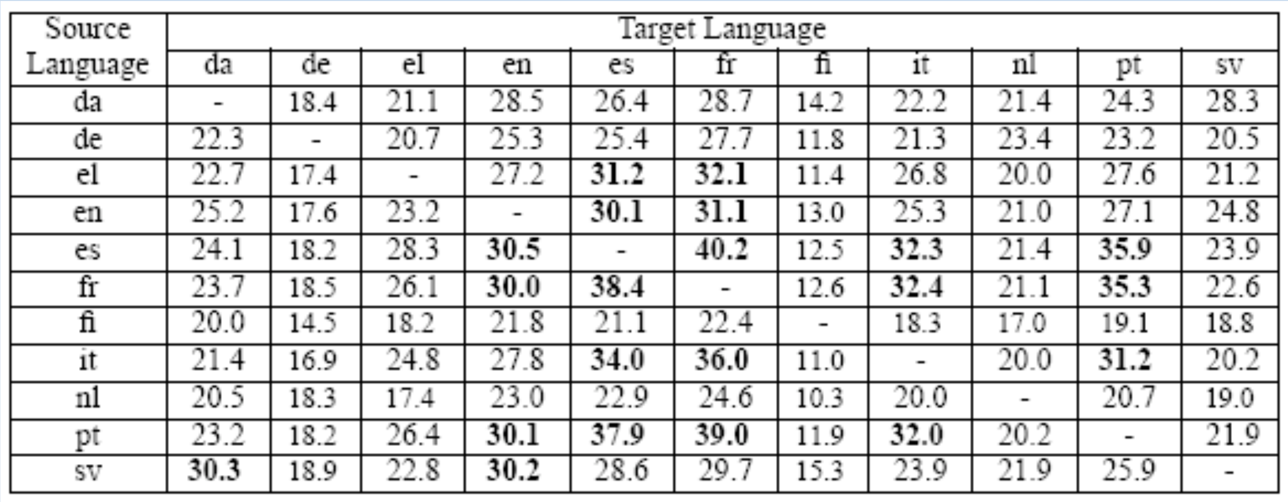
Conclusion (Cont.)

- Size



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

- Score



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

(From Koehn, 2005)

Next Steps

- 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

Relation to ongoing projects

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

Acknowledgement

- EU projects EuroMatrix and EuroMatrix Plus
- Saarland University
- DFKI GmbH
- Addis Ababa University
- German Academic Exchange Service (DAAD)