Xhosa-English Machine Translation for the Medical Domain

Proyag Pal

Research Assistant
University of Edinburgh

December 4, 2017
Outline

Data
- Existing parallel data
- Crawling for parallel texts
- Creating parallel data
- Monolingual data

Experiments
- Phrase-based systems
- Neural systems
- Comparison of sample outputs
Outline

Data

- Existing parallel data
- Crawling for parallel texts
- Creating parallel data
- Monolingual data

Experiments

- Phrase-based systems
- Neural systems
- Comparison of sample outputs
Existing Parallel Corpora

**Software Localisation**
- GNOME$^1$
- KDE4$^1$
- Ubuntu$^1$

**Other**
- Mobile Xhosa$^2$
- Tatoeba$^1$
- The Bible$^3$

---

$^1$From the OPUS collection: [http://opus.nlpl.eu](http://opus.nlpl.eu)

$^2$Thanks to Saadiq Moolla. [http://mobilexhosa.co.nf](http://mobilexhosa.co.nf)

$^3$The Bible Corpus: [https://github.com/christos-c/bible-corpus](https://github.com/christos-c/bible-corpus)
Mobile Xhosa

Examples

What are your symptoms?

I am going to examine your ears by looking inside.

Sit down and roll up your sleeve.

Number of parallel sentences: 471
Tatoeba

Examples

*Open your mouth!*

*How do you feel today?*

*Please close the window.*

Number of parallel sentences: 99
Bible Corpus

Examples

*It happened after the death of Saul, when David was returned from the slaughter of the Amalekites, and David had stayed two days in Ziklag;*

*The time that David was king in Hebron over the house of Judah was seven years and six months.*

Number of parallel sentences: 31065
Examples

*Insert Line Break*

*The message content was not accepted. %1*

*queen of diamonds*

Number of parallel sentences: **295286**
Crawling for Parallel Texts

UCT Clinical Skills Programme\(^4\): 181 Sentences

Measure the blood pressure

Correct placement of the cuff is important. Hold the patient’s arm out straight, palm upwards. (You can help by tucking his wrist under your elbow, as shown.) Start applying the cuff in the middle of the upper arm, with the tubes uppermost, and placed in line with the brachial artery.

AFRIKAANS    TRANSCRIPT    XHOSAN
I need to take your blood pressure. / Ek moet u bloeddruk meet. / Kufuneka ndithathe iBP yakho.

Please give me your arm. / Gee my u arm, asseblief. / Ndicela undinike inga yakho.

I?ll put this cuff around your arm. / Ek sal hierdie band om u arm sit. / Ndiza kufaka eli bhanti api apha engalweni.

It will be uncomfortable. / Dit sal ongemaklik wees. / Akuzokuva kamnandi,

I?ll try to be quick. / Dit sal nie lark neem nie. / Ndiza kuzama ukukhawuwezisa.


\(^4\)Thanks to UCT Department of Medicine. https://vula.uct.ac.za/access/content/group/9c29ba04-b1ee-49b9-8c85-9a468b556ce2/ClinicalSkills/index.html
Crawling for Parallel Texts

South African Constitution, Universal Declaration of Human Rights

Government Websites

- City of Cape Town\(^5\)
- Western Cape Government\(^6\)

Example

*Vacancies in a provincial legislature must be filled in terms of national legislation.*

\(^5\)http://www.capetown.gov.za/

\(^6\)https://www.westerncape.gov.za/
Additional sentence pairs collected: **67881**

Problems

- Still very out of domain
- Imperfect alignment
- PDFs are hard
Creating Parallel Corpora

10979 medical domain sentences being translated by professional translators.

Examples

People who are HIV positive are more susceptible to ordinary TB.

I am 45 years old and have diabetes.

After skeletal surgery, exercises are usually prescribed.
Monolingual Data: Available Corpora

- NCHLT Corpus
- Leipzig CURL
- Wikipedia
- Xhosa Genre Classification Corpora

Total: Around 500k Sentences

---

7South African Centre for Digital Language Resources (SADiLaR): https://rma.nwu.ac.za/index.php/

8Crawling Underresourced Languages, Leipzig University: http://curl.corpora.uni-leipzig.de/languages/xho
Monolingual Data: Common Crawl

Text from the Common Crawl, classified using CLD2. Thought this was Xhosa:

\[
\begin{align*}
14 \\
0 \\
CG(10) \\
.797
\end{align*}
\]

Reclassified using NCHLT South African Language Identifier\(^9\), with strong confidence threshold.

Around 250k more sentences (instead of 5.8 million)

Outline

Data

- Existing parallel data
- Crawling for parallel texts
- Creating parallel data
- Monolingual data

Experiments

- Phrase-based systems
- Neural systems
- Comparison of sample outputs
Experiments

- Experiments in the xh→en direction.
- Phrase-based baseline.
- Our focus on creating neural models.
- Evaluated with detokenised BLEU scores.
Baseline

Highest scoring phrase-based model chosen as baseline for our experiments.
Baseline

Highest scoring phrase-based model chosen as baseline for our experiments.

BLEU score: 7.00
Baseline

Highest scoring phrase-based model chosen as baseline for our experiments.

BLEU score: 7.00

Byte Pair Encoding (BPE)

Apply BPE to reduce size of vocabulary.
Baseline

Highest scoring phrase-based model chosen as baseline for our experiments.

**BLEU score:** 7.00

**Byte Pair Encoding (BPE)**

Apply BPE to reduce size of vocabulary.

**BLEU scores**

- 30k vocab: 4.52
- 10k vocab: 6.03
Neural MT Systems

Score to beat: 7.00

Neural Baseline

- Single layer RNN encoder-decoder model with attention
- Dropout
- Byte Pair Encoding (30k wordpiece vocabulary)
Neural MT Systems

Score to beat: 7.00

Neural Baseline
- Single layer RNN encoder-decoder model with attention
- Dropout
- Byte Pair Encoding (30k wordpiece vocabulary)

BLEU score: 2.31
Neural MT Systems

Monolingual Data

- Monolingual data in the target language.
- Create training pairs where source and target sentences are identical.
- Mix these with the parallel corpus, in varying proportions.
Neural MT Systems

Monolingual Data

- Monolingual data in the target language.
- Create training pairs where source and target sentences are identical.
- Mix these with the parallel corpus, in varying proportions.

BLEU scores

1:1 mix: 4.37
1:3 mix: 3.91
Neural MT Systems

Deep RNNs

- BiDeep architecture
- 4-layer stacks of 2-layer deep transition GRUs for both encoder and decoder.
Neural MT Systems

Deep RNNs

- BiDeep architecture
- 4-layer stacks of 2-layer deep transition GRUs for both encoder and decoder.

BLEU scores
1:1 mix: 6.63
1:3 mix: 6.80
Scores So Far

Phrase-based 7.00
Neural 6.80
Scores So Far

Phrase-based  7.00
Neural        6.80

BLEU Scores aren’t perfect.
Especially untrustworthy when they’re low - not a very useful indicator in these cases.
## Summary of Models

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-based</td>
<td>7.00</td>
</tr>
<tr>
<td>Neural Baseline</td>
<td>2.31</td>
</tr>
<tr>
<td>Neural + Copied Data</td>
<td>4.37</td>
</tr>
<tr>
<td>Deep Neural + Copied Data</td>
<td>6.80</td>
</tr>
</tbody>
</table>
## Summary of Models

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-based</td>
<td>7.00</td>
</tr>
<tr>
<td>Neural Baseline</td>
<td>2.31</td>
</tr>
<tr>
<td>Neural + Copied Data</td>
<td>4.37</td>
</tr>
<tr>
<td>Deep Neural + Copied Data</td>
<td>6.80</td>
</tr>
</tbody>
</table>
Sample Translations

**Source**
Thatha inqolowa kwaye usele amanzi imihla ngemihla.

**Reference**
Take fiber and drink plenty of water on a daily basis.

**PBMT**
Take a inqolowa and drink water daily.

**NMT**
Take grain and drink water daily.
Sample Translations

Source: Ingaba kukhona okunye endinokwenza?
Reference: Can I do something more?

PBMT: Are there are other endinokwenza?
NMT: Is there any other I can do?

PBMT just copies unknown words.
### Sample Translations

#### Source
Rhoqo, ndifumana iintlungu phezu kweenyawo zam, ingakumbi ekuqaleni kwentsasa.

#### Reference
I often have pain on the tops of my feet, particularly first thing in the morning.

#### PBMT
Often, as well as the The pain on my feet, especially from the beginning of the morning.

#### NMT
I have found the pain on my feet, especially at the beginning of the fury.

NMT is fluent, but *fury* is completely unrelated to the source.
## Sample Translations

<table>
<thead>
<tr>
<th>Source</th>
<th>Ndinabantwana abathathu ababezalwe ngokuqhelekileyo/ngemvelo.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>I have three children that were born naturally.</td>
</tr>
<tr>
<td>PBMT</td>
<td>The three Ndinabantwana ababezalwe normally / by nature.</td>
</tr>
<tr>
<td>NMT</td>
<td>I have three children who have been born with a novel.</td>
</tr>
</tbody>
</table>
## Sample Translations

<table>
<thead>
<tr>
<th>Source</th>
<th>Xa sisaqala isifo sethayirodi egqithisileyo kungenzeka singabinazimpawu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>In its early stages, hyperparathyroidism may not have any symptoms.</td>
</tr>
<tr>
<td>PBMT</td>
<td>When the sisaqala sickness sethayirodi egqithisileyo might be singabinazimpawu.</td>
</tr>
<tr>
<td>NMT</td>
<td>When a folder was a regular expression, it may have a character set.</td>
</tr>
</tbody>
</table>
Effect of Training Data: Software Localisation

Source: Le nto ayiqhelelekanga kodwa ngokwenene iyenzeka.

Reference: This is not very common, but certainly happens.

NMT: The file is not valid but cannot be deleted.
Effect of Training Data: Government Texts

Source
Uirqirha wabantwana wesifo samalungu kufuneka axilonge ukuba simandundu yaye siphila kangakanani na isifo sokuqaqamba kwamalungu ebantwaneni (iJIA) aze athabathe isigqibo ngawona mayeza anokusinceda.

Reference
Your pediatric rheumatologist will need to assess how severe and how active the systemic JIA is at a given point, and then decide which drug is best.

NMT
The doctor of the majority (85%) of the majority of the members of the members of the
Effect of Training Data: The Bible

Source  Ndifuna ukwazi ngendawo egcina igazi lenkaba.

Reference  Also, I would like to know about the cord blood bank.

NMT  I want knowledge about the blood of the wicked.
Next Steps

How Might the Neural Models be Improved?

A deep network with copied monolingual data has given us the best results among the neural models so far.

Main problem: Lack of high-quality in-domain data.
Future Experiments I

- **Obvious** - Use all of the data being translated. Only a few hundred lines used so far (what we had when experiments were started). Much of what has been translated since then has not been used to ensure fair comparisons.

- **Back-translation** - Use \( \text{en} \rightarrow \text{xh} \) phrase-based/neural model to create artificial training pairs.

- **Domain Adaptation** - Fine-tuning on medical domain parallel data.
Future Experiments II

- **Morphological Splitting for isiXhosa**
- **Transfer Learning** - Train model for high-resource pair with same target language (e.g. de→en), initialise xh→en model with those parameters.
- **Multilingual Translation** - Single model to translate from (or to) multiple languages. Source languages use common BPE vocabulary. Prepended language token to enable multiple target languages.
- **Pre-training Decoder with LM** - Initialise the decoder RNN parameters with those of an RNN-LM trained on the target language.
Conclusion

- Very low-resource problem.
- No/little in-domain training data.
- Neural systems show promise in adapting to medical domain.
- Work very much in progress.
Thank You
References I


References II