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`

[2] Thanks to Saadiq Moolla. `http://mobilexhosa.co.nf`

[3] The 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**

GNOME/KDE4/Ubuntu

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

Contents:

- Getting your patient into the right position
- General assessment
- Assess the pulse
- Measure the blood pressure
- Measure the jugular venous pressure
- Palpation of the praecordium
- Palpation of the apex
- Auscultation of the heart
- Other signs of heart failure

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	XHOSA
	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 ingalo yakho.td>	
	I?ll put this cuff around your arm./ Ek sal hierdie band om u arm sit. / Ndiza kufaka eli bhanti apha engalweni.	
	It will be uncomfortable. / Dit sal ongemaklik wees. / Akuzokuva kamnandi,	
	I?ll try to be quick. / Dit sal nie lank neem nie. / Ndiza kuzama ukukhawulezisa.	
	Thank you, that?s fine. / Reg so. Dankie. / ... Enkosi. Kulungile.	

[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[7]
- Leipzig CURL[8]
- Wikipedia
- Xhosa Genre Classification Corpora[7]

Total: Around 500k Sentences

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

[8] Crawling 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:

14
0
CG(10)
.797

Reclassified using NCHLT South African Language Identifier[9], with strong confidence threshold.

Around 250k more sentences (instead of 5.8 million)

[9] From SADiLaR. `https://rma.nwu.ac.za/index.php/resource-catalogue/nchlt-south-african-language-identifier.html`

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

Model	BLEU Score
Phrase-based	7.00
Neural Baseline	2.31
Neural + Copied Data	4.37
Deep Neural + Copied Data	6.80

Summary of Models

Model	BLEU Score
Phrase-based	7.00
Neural Baseline	2.31
Neural + Copied Data	4.37
Deep Neural + Copied Data	6.80

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

Source	Ndinabantwana abathathu ababezalwe ngokuqhelekileyo/ngemvelo.
Reference	I have three children that were born naturally.
PBMT	The three Ndinabantwana ababezalwe normally / by nature.
NMT	I have three children who have been born with a novel.

Sample Translations

Source	Xa sisaqala isifo sethayirodi egqithisileyo kungenzeka singabinazimpawu.
Reference	In its early stages, hyperparathyroidism may not have any symptoms.
PBMT	When the sisaqala sickness sethayirodi egqithisileyo might be singabinazimpawu.
NMT	When a folder was a regular expression, it may have a character set.

Effect of Training Data: Software Localisation

Source	Le nto ayiqhelekanga 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	Ugqirha 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 en$\rightarrow$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

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." (2014) https://arxiv.org/pdf/1409.0473.pdf
- Gal, Yarin, and Zoubin Ghahramani. "A theoretically grounded application of dropout in recurrent neural networks." Advances in neural information processing systems (2016) https://arxiv.org/pdf/1512.05287.pdf
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." (2015) https://arxiv.org/pdf/1508.07909.pdf
- Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield. "Copied Monolingual Data Improves Low-Resource Neural Machine Translation." Proceedings of the Second Conference on Machine Translation (2017) https://kheafield.com/papers/edinburgh/copy_paper.pdf

References II

- Barone, Antonio Valerio Miceli, et al. "Deep architectures for neural machine translation." (2017) https://arxiv.org/pdf/1707.07631.pdf
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving neural machine translation models with monolingual data." (2015) https://arxiv.org/pdf/1511.06709.pdf
- Zoph, Barret, et al. "Transfer learning for low-resource neural machine translation." (2016) https://arxiv.org/pdf/1604.02201.pdf
- Nguyen, Toan Q., and David Chiang. "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation." (2017) https://arxiv.org/pdf/1708.09803.pdf
- Johnson, Melvin, et al. "Google's multilingual neural machine translation system: enabling zero-shot translation." (2016) https://arxiv.org/pdf/1611.04558.pdf