

The background is a dark blue-grey color. It is decorated with various geometric shapes in orange and white. There are circles of different sizes, some with dotted patterns inside. There are hexagons, some solid orange and some outlined in white. There are triangles, some solid orange and some outlined in white. There are also lines, some solid and some dotted. The overall style is modern and abstract.

Limburgish-Dutch Translation

Group 10

T. Debets, S. Jain, J. Kunnen,
K. Shcherbakov, K. Stessen

01.

Introduction

Introduction of the
project topic

02.

Data Collection and preprocessing

03.

Model

Explanation of Model

04.

Results & Discussion

Results of the experiments



Limburgish news article
publishing the
miscommunication
concern



Kennis van dialect van groot belang voor niet-Limburgse arts: 'Een onbegrepen woord kan leiden tot een verkeerde diagnose'

24-02-2021 om 07:45 door Benti Banach



Afbeelding: Peter Schols



Communicatie tussen arts en patiënt is van groot belang. Maar wat als een niet-Limburgse arts zich in deze regio vestigt? Verdiep je in het dialect, zegt psychiater in opleiding Mette Konings. Dan snap je wat iemand bedoelt als het 'ram sjleecht' gaat.

Introduction



Limburgish spoken language region

UNSUPERVISED NEURAL MACHINE TRANSLATION

2020 waor e
bezoonder jaor, ouch
veur Veldeke



*2020 was een
bijzonder jaar, ook
voor Veldeke*

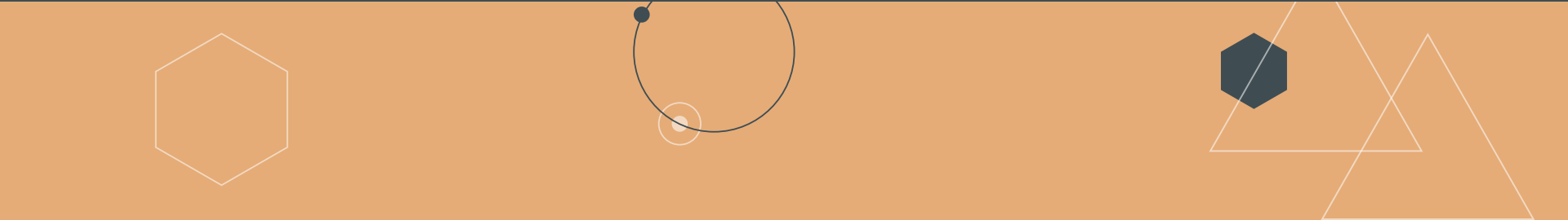




Research Questions

How can a database for Limburgish be created?

Which architectures and methods exist to build an Unsupervised Neural Machine Translator?





Data Collection

.....

.....

..... Accumulated Data

SENTENCES	Limburgish	Dutch
Total Training Data	274.315	18.899.914
Matched Training Data	274.315	5.535.535

Sources:

- Websites containing both dutch and limburgish texts
- li.wikipedia.org/
- nl.wikipedia.org/
- Independent authors and other publications
- Leipzig Database



Matching

vreugmoderne tied
(early modern age)

→

vrgmdrnn td

←

vroegmoderne tijd

trèkväögel
(migratory bird)

→

trkvgl

←

trekvogel

hieëring
(herring)

→

hrng

←

haring

Amount of Limburgish articles: 50.851

Amount of Dutch articles: 32.687

..... Accumulated Data

SENTENCES	Limburgish	Dutch
Total Training Data	274.315	18.899.914
Matched Training Data	274.315	5.535.535

SENTENCES	Short	Long
Validation Set	1141	2044
Test Set	1226	2000



Data Analysis



Limburgish wordcloud



Dutch wordcloud

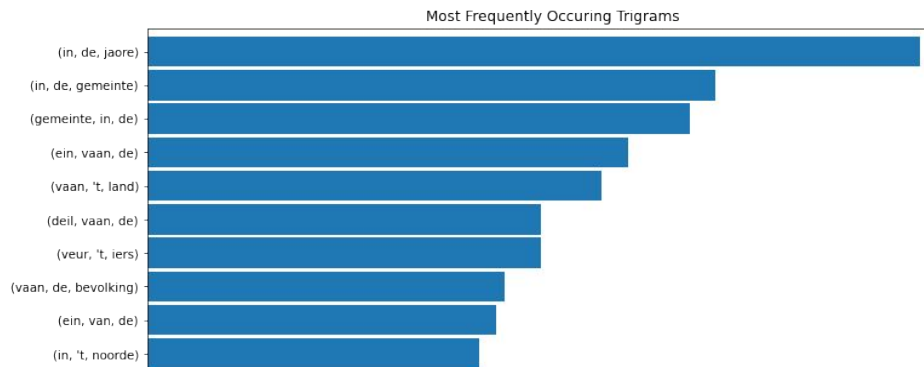
..... Multiple spellings

- 1) gemeente, gemeindje and gemeint (municipality, community)
- 2) hebbe, hobbe and hubbe (to have)
- 3) stad and sjtad (city)
- 4) ierste and eerste (first)

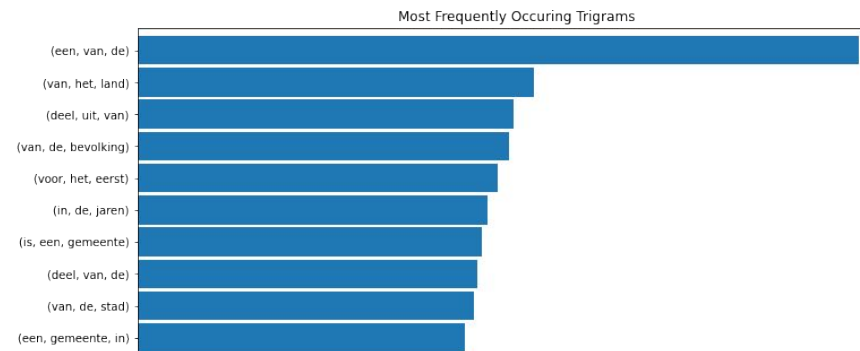


Domain Similarity

• • • • • • • • • • • •



Limburgish trigrams



Dutch trigrams

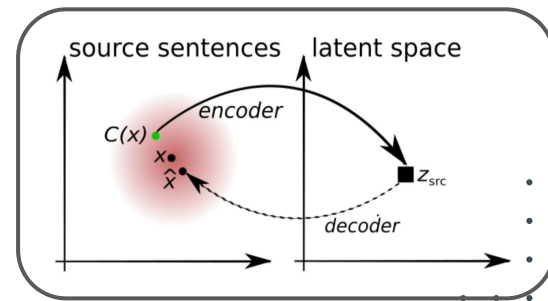
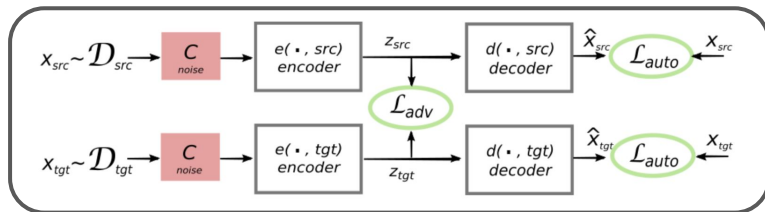




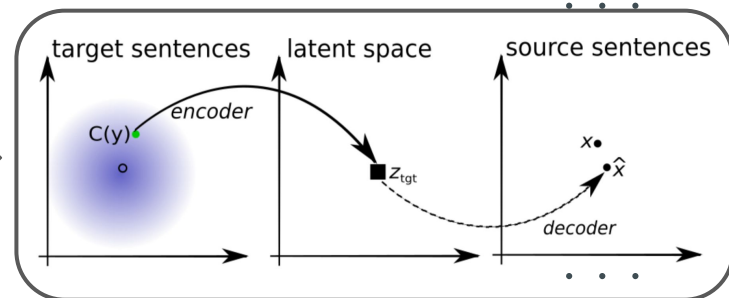
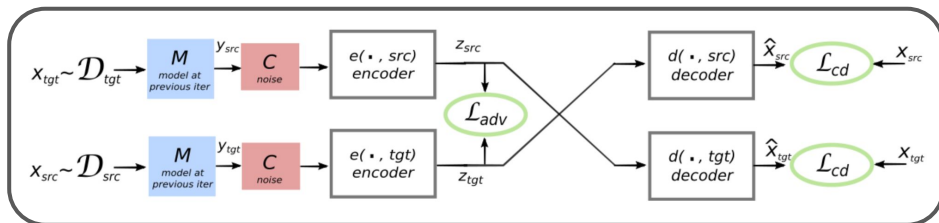
Model

PidginUNMT (Instadeep)

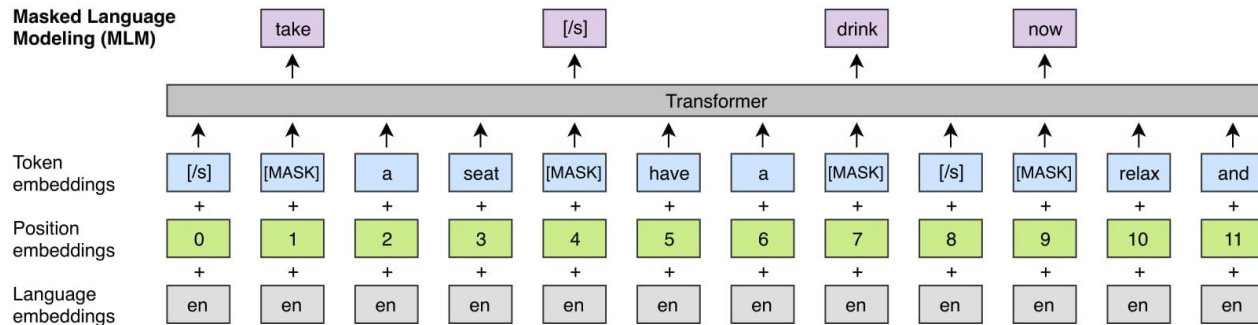
Auto-encoding



Translation



XLM(Facebook)



- 1) Building our own BPE codes and vocabulary.
- 2) The number of BPE tokens have been reduced to only 10.000 to make the model less complicated.
- 3) Pretraining both the encoder and decoder with MLM objective
- 4) Use the pretrained model for Machine translation and train the model with a denoising auto-encoding loss along with an online back-translation loss.
- 5) The parameters of our Transformer model are identical to the ones used for pretraining

• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •
• • •

Experiments

Experiment	Data set	Validation set	Test set
Experiment 1	5 million sentences WMT'07 + WMT'08 German data set + 5 million sentences WMT'07 + WMT'08 English data set	WMT'13 parallel data set	WMT'16 parallel data set
Experiment 2	complete Limburgish + complete Dutch data set	Smaller validation set	Smaller test set
Experiment 3	complete Limburgish + matched Dutch data set	Smaller validation set	Smaller test set
Experiment 4	complete Limburgish + matched Dutch data set	Larger validation set	Larger test set

1) Experiment 1 is not conducted.

2) the MLM model is trained for 10 hours or if the validation perplexity does not improve for 25 epochs.

3) the XLM model is trained for 10 hours or if the validation BLEU score does not improve for 10 epochs.

. . .
. . .
. . .
. . .
. . .
. . .
. . .
. . .
. . .
. . .

Results

Experiment 3 best performing

Domain similarity important

Overtraining

Experiment 1: Approx. 8 BLEU Score
Low-resource Language Modelling

Experiment 4: Not any big differences

Dutch → Limburgish: more difficult

BPE codes influenced perplexity

Influence of number of GPUs → larger minibatches → larger training efficiency

XML: DE-EN: 34.3 BLEU -- EN-DE 26.4 BLEU

	Validation set					
	Limburgish-Dutch			Dutch-Limburgish		
	ppl	acc	BLEU	ppl	acc	BLEU
Experiment 2	53.29	59.97	31.31	48.70	59.38	22.25
Experiment 3	20.33	67.46	37.25	28.26	61.00	25.11
Experiment 4	25.56	65.25	35.53	31.52	59.25	24.19
	Test set					
	Limburgish-Dutch			Dutch-Limburgish		
	ppl	acc	BLEU	ppl	acc	BLEU
Experiment 2	77.45	55.62	25.39	73.39	53.69	18.21
Experiment 3	35.55	62.19	31.73	43.92	55.43	20.21
Experiment 4	43.31	60.48	30.85	47.46	54.34	20.43

Results – Textual Dutch Translation

Dutch Original	Dutch Complete	Dutch Matched
2020 was een bijzonder jaar, ook voor Veldeke	2020 was een bijzonder jaar , ook voor Veluwe	2020 was een bijzonder jaar, ook voor Veldeke
Meertaligheid is een verschijnsel van alle tijden	Mietas is een verschijnsel van alle tijden	Menselijke ontwikkeling is een verschijnsel van alle tijden
Na 2013 is gekozen voor een ietwat andere opzet	Na 2010phoridae is besloten voor ' een andere manier	Na 2013 is gekozen voor ' een andere manier '
Zo zijn er tal van kerstliedjes in het Limburgs bekend	Aanvankelijk zijn d 'r vanno Spanjaarl Limburgse kersleedjes	Zo zijn d ' r v��l Limburgse kersliedjes
Hij is 24 jaar en is docent in Roermond	Hij is 24 jaar en wJohrk als docent in Remund	Hij is 24 jaar en wel als docent in Remund

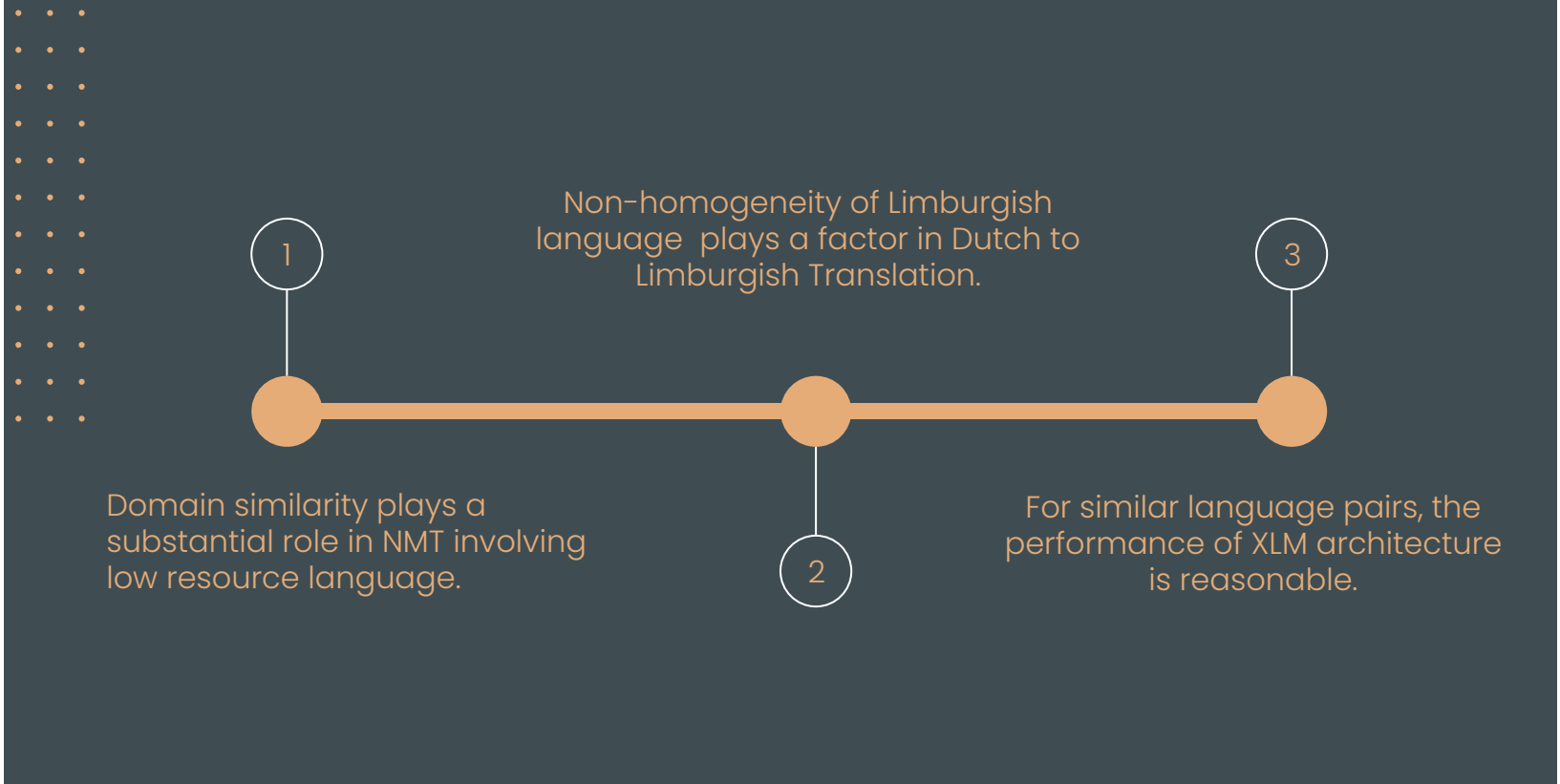
Results – Textual Limburgish Translation

Limburgish Original	Limburgish Complete	Limburgish Matched
2020 waor e bezunder jaor, ouch veur Veldeke	2020 waor e bezunder jaor , ouch veur Veldeke	2020 waor e bezoonder jaor, ouch veur Veldeke
Mietaolegheid is e versjijsel vaan alle tieje	Meertaligheid ies n versjiensel van alle tieje	Meertaligheid is n versjiensel van alle tieje
Nao 2013 is gekaoze veur 'ne andere meneer	Nao 201v-2012 ies gekoze veur n ietwat ander opzet	Nao 2014 is gekoze veur n ietwat aander opzat
Zoa zeen d'r vääöl Limburgse kersleedjes	Zo zeen dr tal van kerkelijke tekste in t Limburgs bekend	Zoa zint dr tal van kerstleedsjes in t Limburgs bekend
Heer is 24 jaor en werk es docent in Remund	Hae ies 24 jaor en ies docent in Roermond	Hae is 24 jaor en is docent in Roermond



Conclusion

Highlights





Thank you for your attention.

Questions?

References

1. Benti Banach. "Kennis van dialect van groot belang voor niet-Limburgse arts: 'Een onbegrepen woord kan leiden tot een verkeerde diagnose'". In: Limburger.nl (2021).
2. Guillaume Lampe and etc. "Unsupervised Machine Translation Using Monolingual Corpora Only"