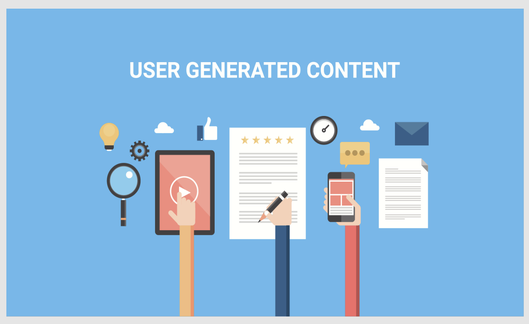
FACULTY OF ARTS AND PHILOSOPHY

LT³, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

Claudia Matos Veliz, Orphée De Clercq & Véronique Hoste

COMPARING MT APPROACHES FOR TEXT NORMALIZATION

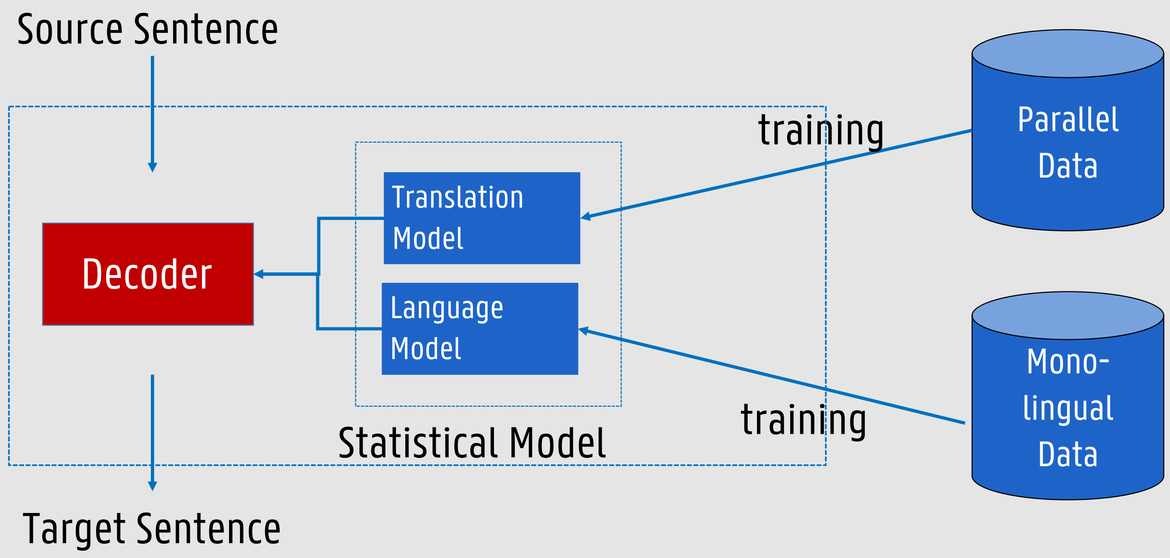


Source sentence
iz da muzieksgool vnavnd ? kwt da niemr .
@renskedemaessc dm me je gsmnummer eens ;-)

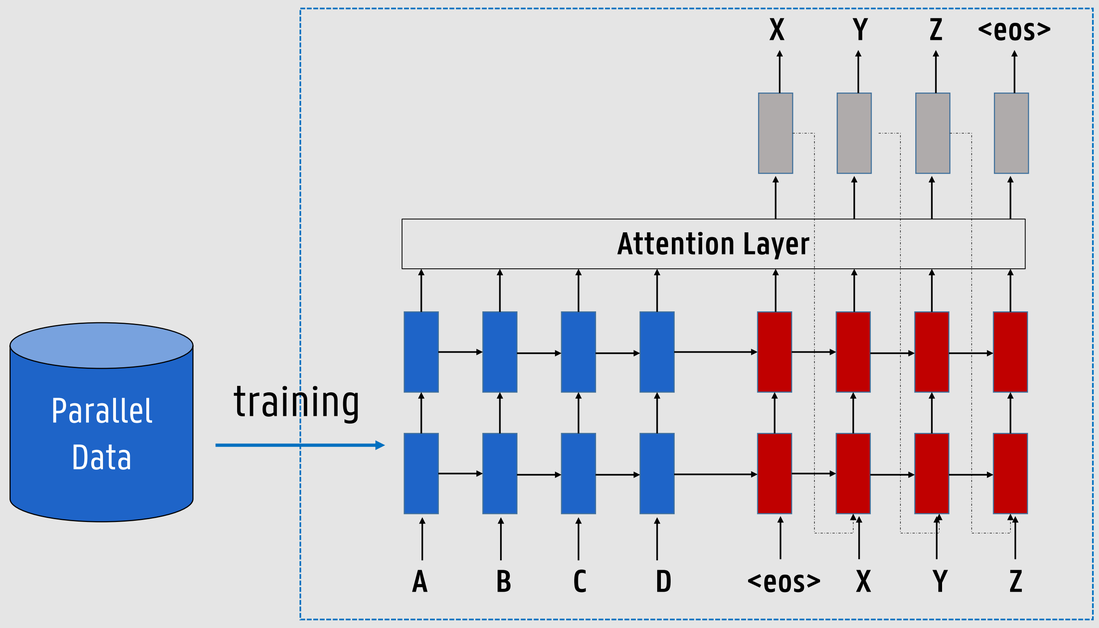
Normalization

Target sentence	Translation
is dat muziekschool vanavond ? ik weet dat niet meer .	is that music school tonight? I don't know that anymore.
<user> doormail me je gsm-nummer eens <emoji>	<user> mail me your cellphone number once <emoji>

Normalization Using SMT Approach



Normalization Using NMT Approach[1]



Data

Parallel Corpora for SMT-Based Normalization (In-house[2])

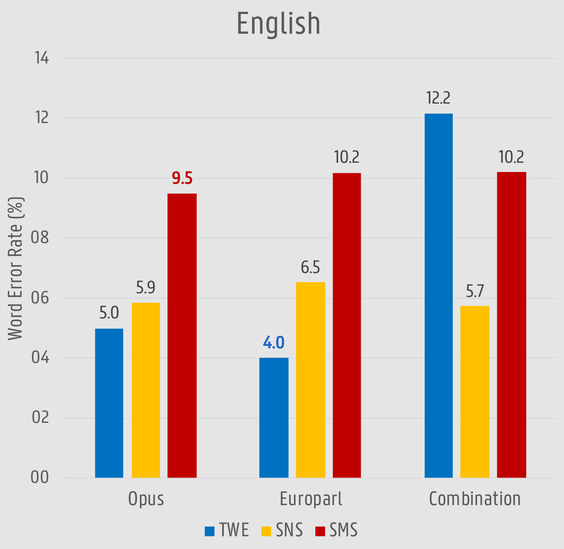
	English (EN)				Dutch (DU)			
Genre	#	Before	After	%	#	Before	After	%
TWE	810	13477	13545	**0.5**	842	13013	13024	**0.1**
SNS	2592	26881	27713	3.0	770	11670	11913	2.1
SMS	1435	220663	22946	**3.9**	801	13063	13610	**4.1**

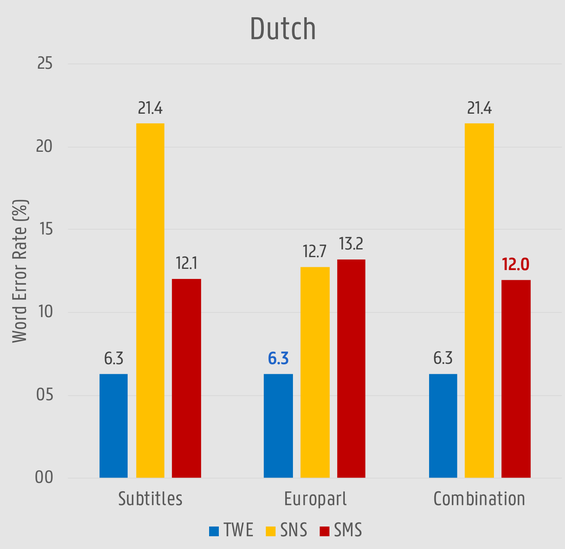
Background Corpora for the Language Model (3 corpora)

	English (EN)		Dutch (DU)	
Corpus	Sentences	Tokens	Sentences	Tokens
OPUS / Subtitles	22512649	18608349	7360869	41681454
Europarl	2005395	54745273	2000113	55132329
Combination	OPUS+Europarl		OPUS+Europarl	

Results and Discussion

SMT (Token Level)





NMT

Neural approaches require big amounts of parallel data = bottleneck

Data augmentation = solution? ➔ Preliminary tests on Dutch SMS

① Naïve data augmentation ② Annotate more data

NMT using data augmentation	
src.	dne dvd vn is ni goe ze . ge kunt nx zien . mt betale . x .
norm.	het dvd hem is niet niet het maar wat wat in ik . niet betale . x x x x x
tgt.	die dvd van is niet goed ze . ge kunt niks zien . moet betalen . x
NMT using new annotations	
src.	zal dan eentje v mzelf sturen . zorgen we morgenavond dan voor verrassing v kareltje ?
norm.	zal dan eentje van mag sturen . zorgen we morgenavond dan voor cocktail van droomt ?
tgt.	zal dan eentje van mezelf sturen . Zorgen we morgenavond dan voor verrassing voor kareltje ?

Conclusion

- Important to make **variations** in the **background data** for building the LM, depending on the amount of noise and vocabulary present in the genre.
- For a low-resource language like Dutch **adding additional training data** works **better** than artificially augmenting the data.

Future Work

- Using a **modular approach** instead of only using SMT could lead to a much better performance.
- Using **more sophisticated data augmentation** techniques could lead to better results.

With the support of Ghent University BOF research fund.

References

[1] Image taken from Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation.arXiv preprint.

[2] Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, and Lieve Macken. 2016. Multimodular Text Normalization of Dutch User-Generated Content. ACM Transactions on Intelligent Systems and Technology, 7(4):1–22.

Contact

Claudia.MatosVeliz@Ugent.be

https://www.lt3.ugent.be/people/claudia-matos-veliz/

GHENT UNIVERSITY