

Machine Translation into Low-Resource Language Varieties

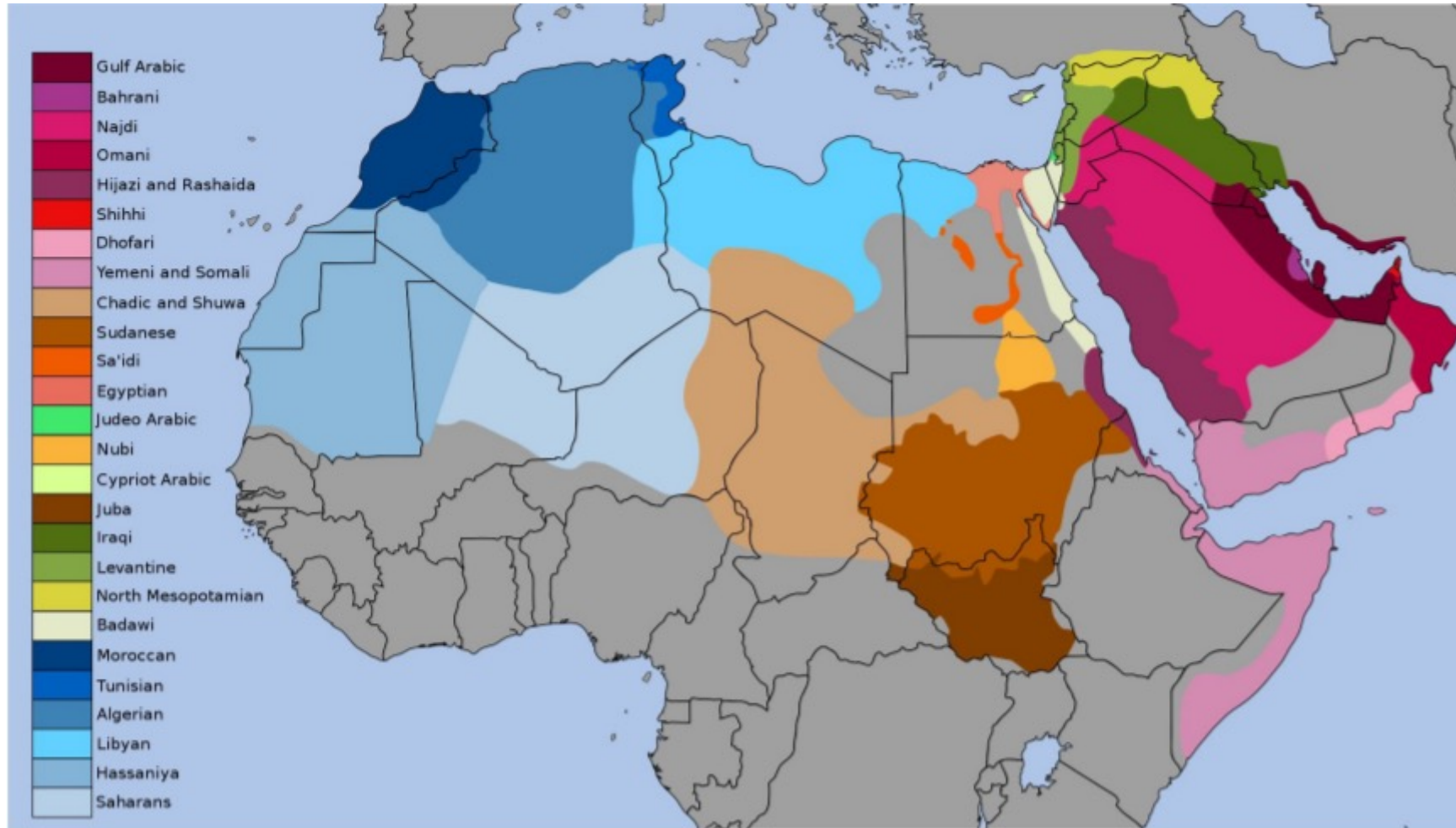
Kumar, Anastasopoulos, Wintner, Tsvetkov

Slides adapted by Farhan Samir

DL-NLP RG

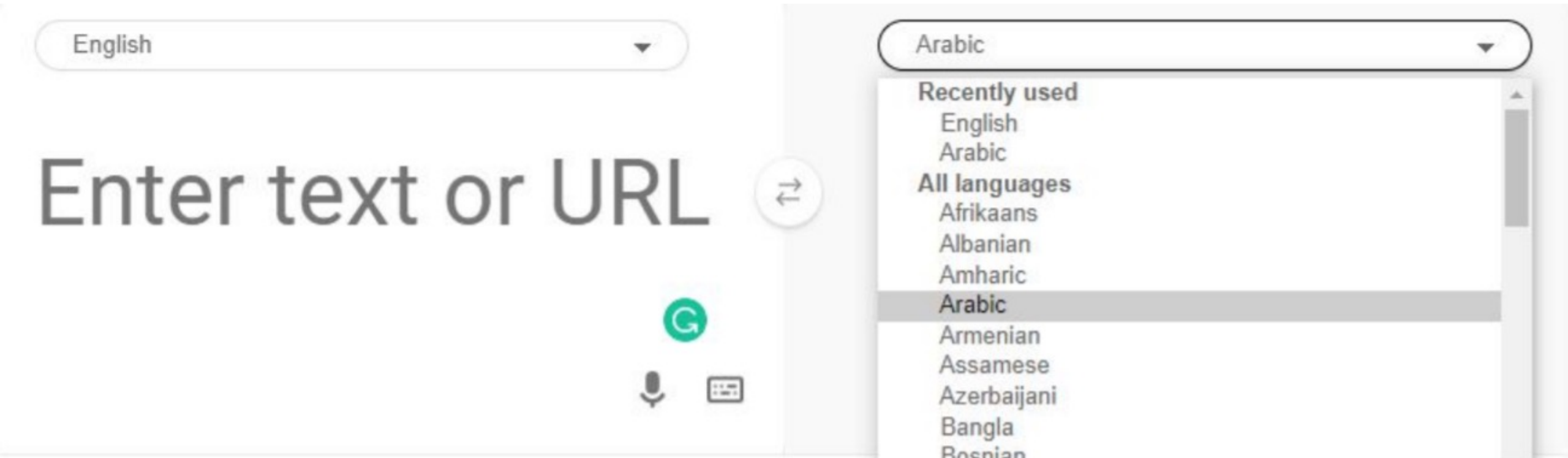
September 16, 2021

Language varieties



A map showing different varieties of Arabic

Language varieties



Snapshot from Bing Translate.

Goal

de quoi la plupart des gens
aujourd'hui ne sont-ils pas
conscients?

French

Source

what are most people today not
aware of?

English

Standard Variety

wetin many people today no
know?

West African Pidgin

whit ur maist fowk th'day nae
aware o'?

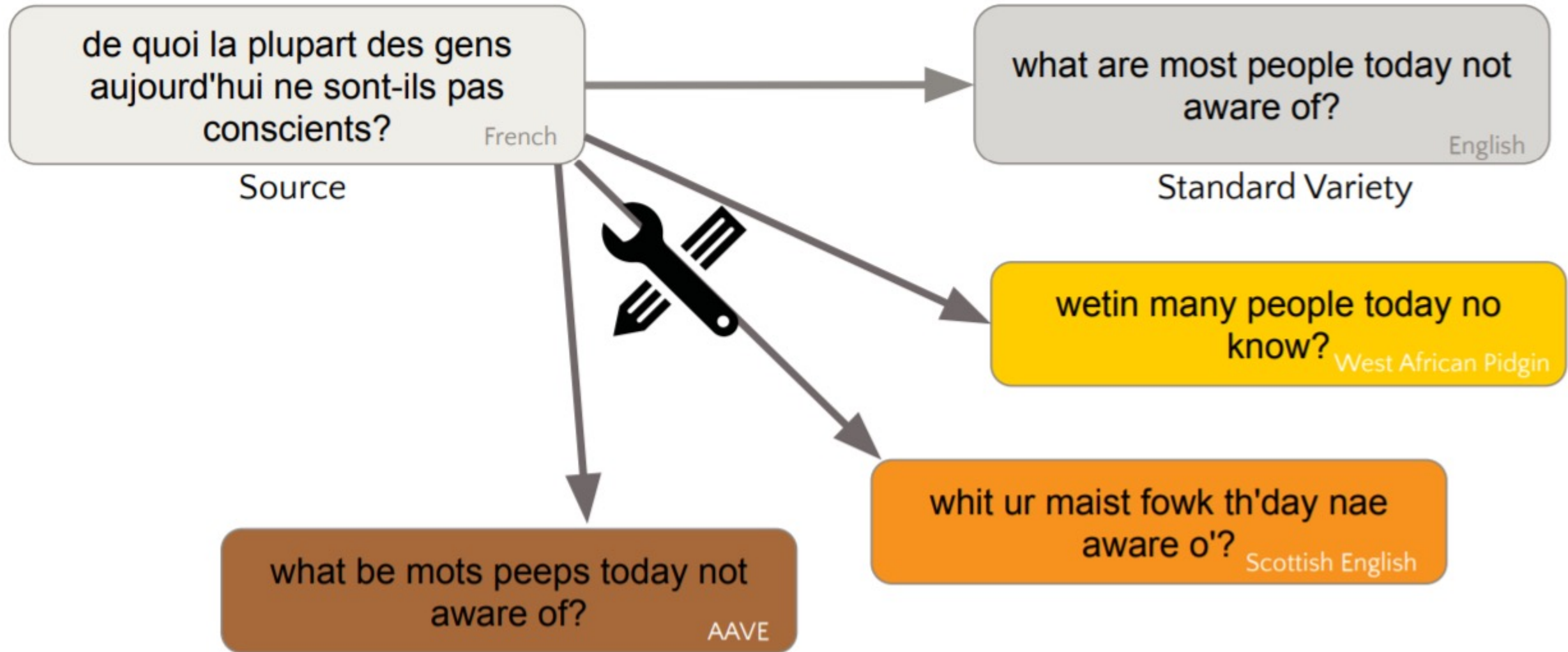
Scottish English

what be mots peeps today not
aware of?

AAVE







Goal



Challenges

Lack of parallel source to target variety data


*de quoi la plupart
des gens aujourd'hui
ne sont-ils pas
conscients*  *what are most people
today not aware of?
(standard English)* 

*de quoi la plupart
des gens aujourd'hui
ne sont-ils pas
conscients*  *whit ur maist fowk
th'day nae aware o'?
(Scottish English)* 

Challenges

Lack of parallel source to target variety data

Little monolingual corpora in target variety



*Unsupervised machine translation data
(e.g., Artexe et al., 2017)*

Challenges

Lack of parallel source to target variety data

Little monolingual corpora in target variety

Vocabulary mismatch between standard variety and target variety

what are most people today not aware of?

whit ur maist fowk th'day nae aware o'?

Challenges

Lack of parallel source to target variety data

Backtranslate target to source, using a *standard-variety* to source MT model

Challenges

Lack of parallel source to target variety data

Backtranslate target to source, using a *standard-variety* to source MT model

Little monolingual corpora in target variety

Finetune a source to standard-variety MT model

Challenges

Lack of parallel source to target variety data

Backtranslate target to source, using a *standard-variety* to source MT model

Little monolingual corpora in target variety

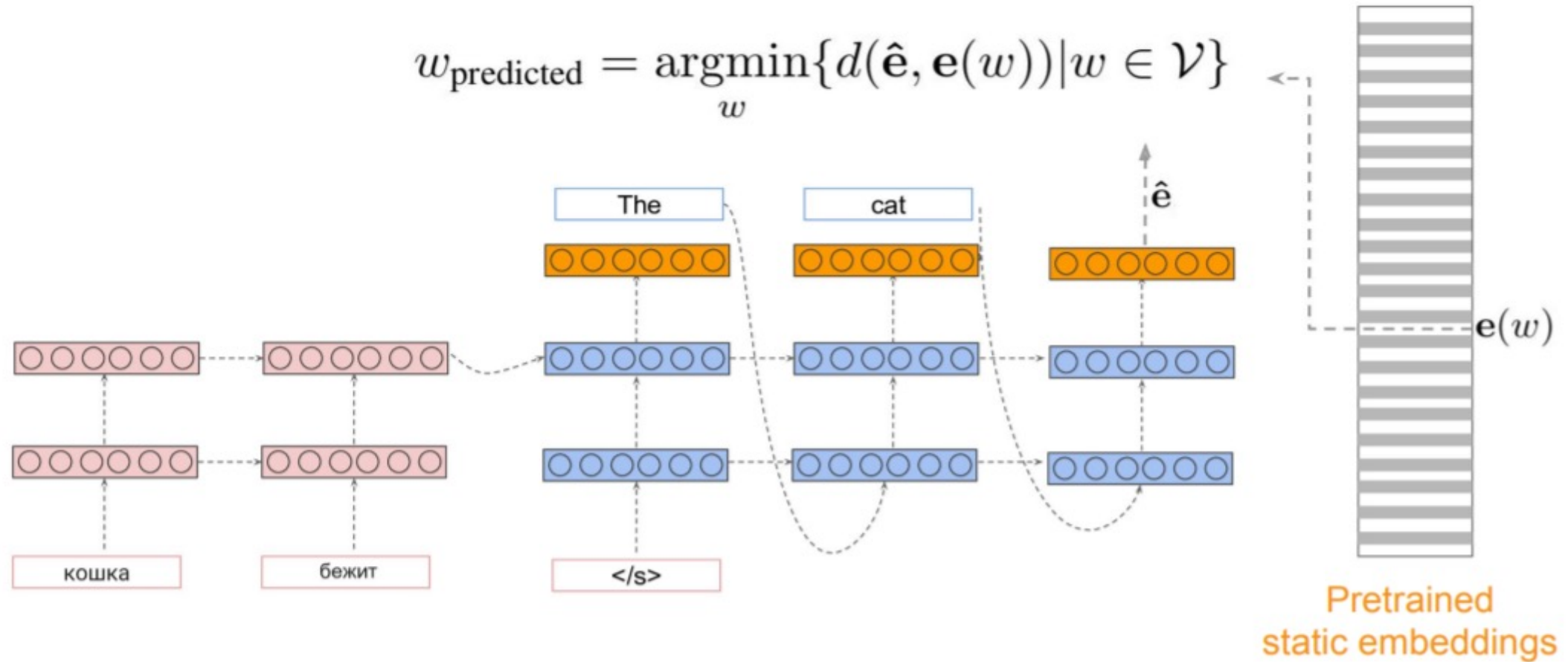
Finetune a source to standard-variety MT model

Vocabulary mismatch between standard variety and target variety

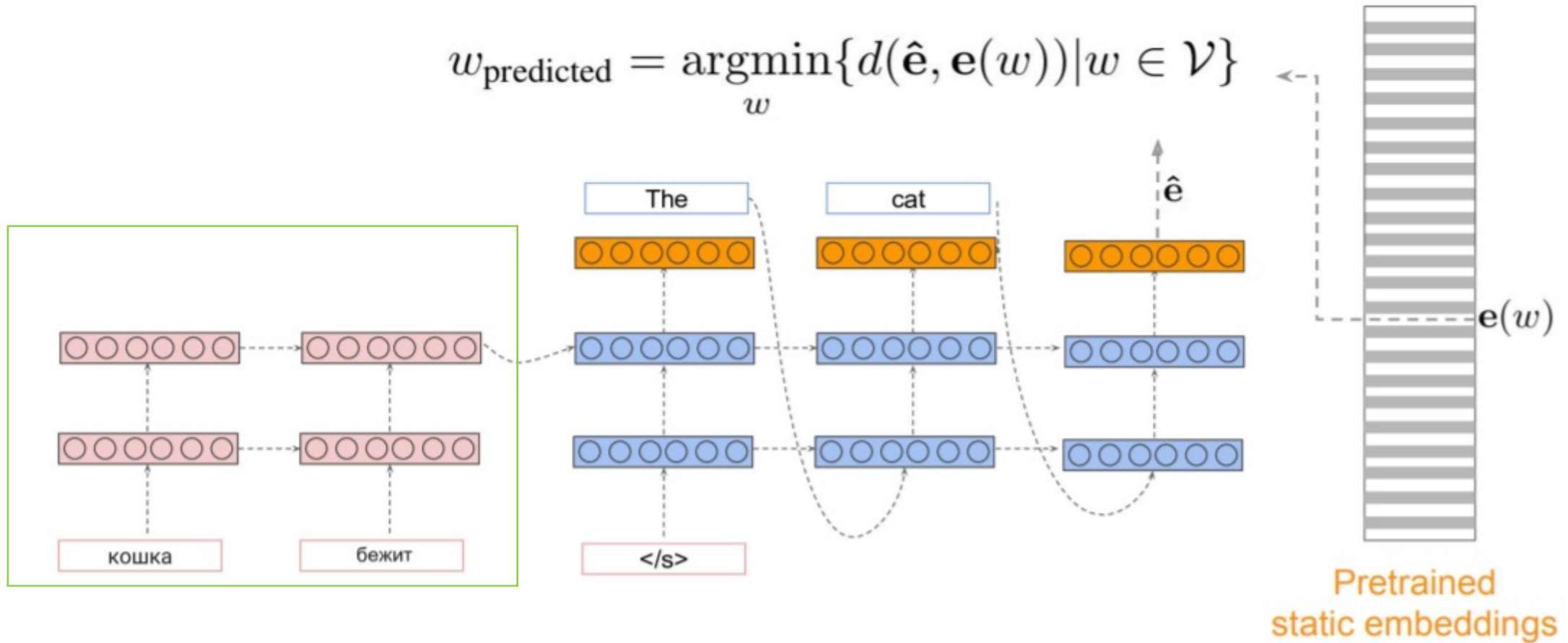
what are *most* people today not aware *of*?

whit ur maist fowk th'day nae aware *o'*?

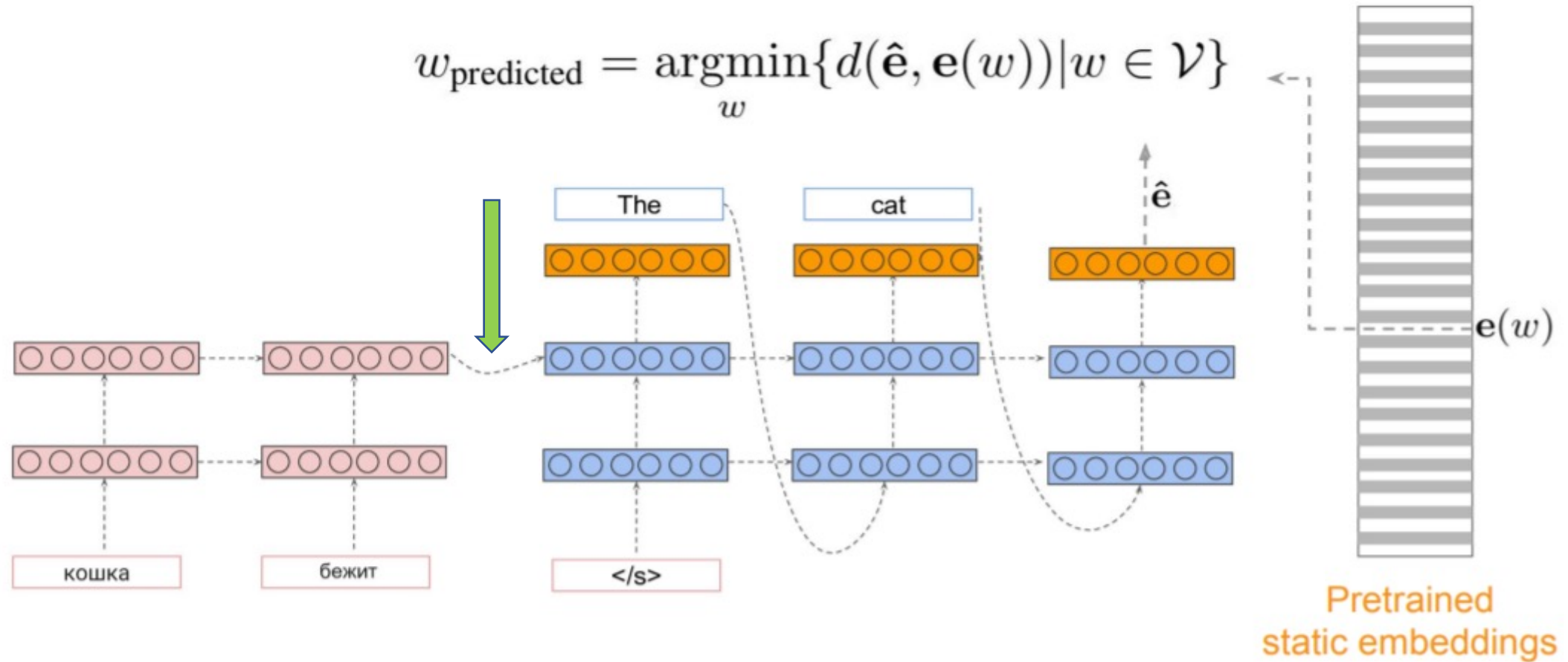
Background



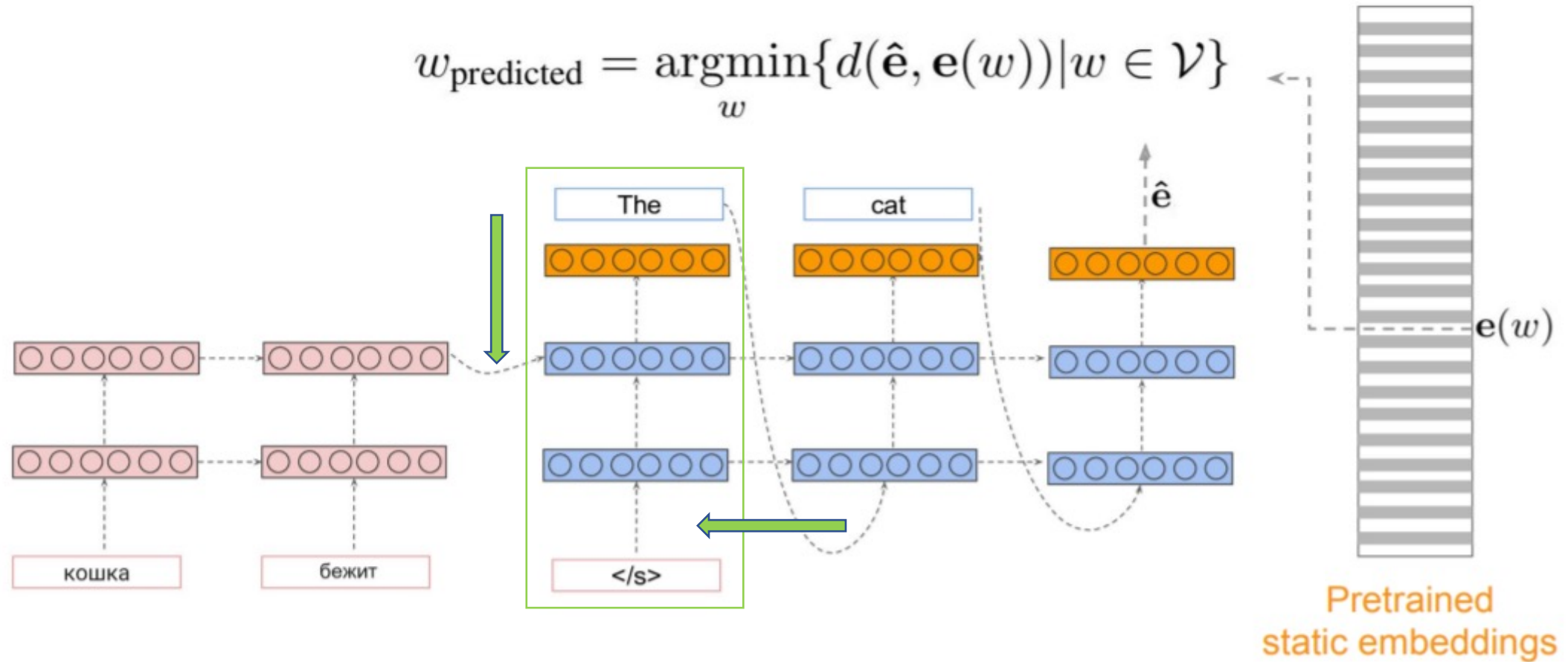
Background



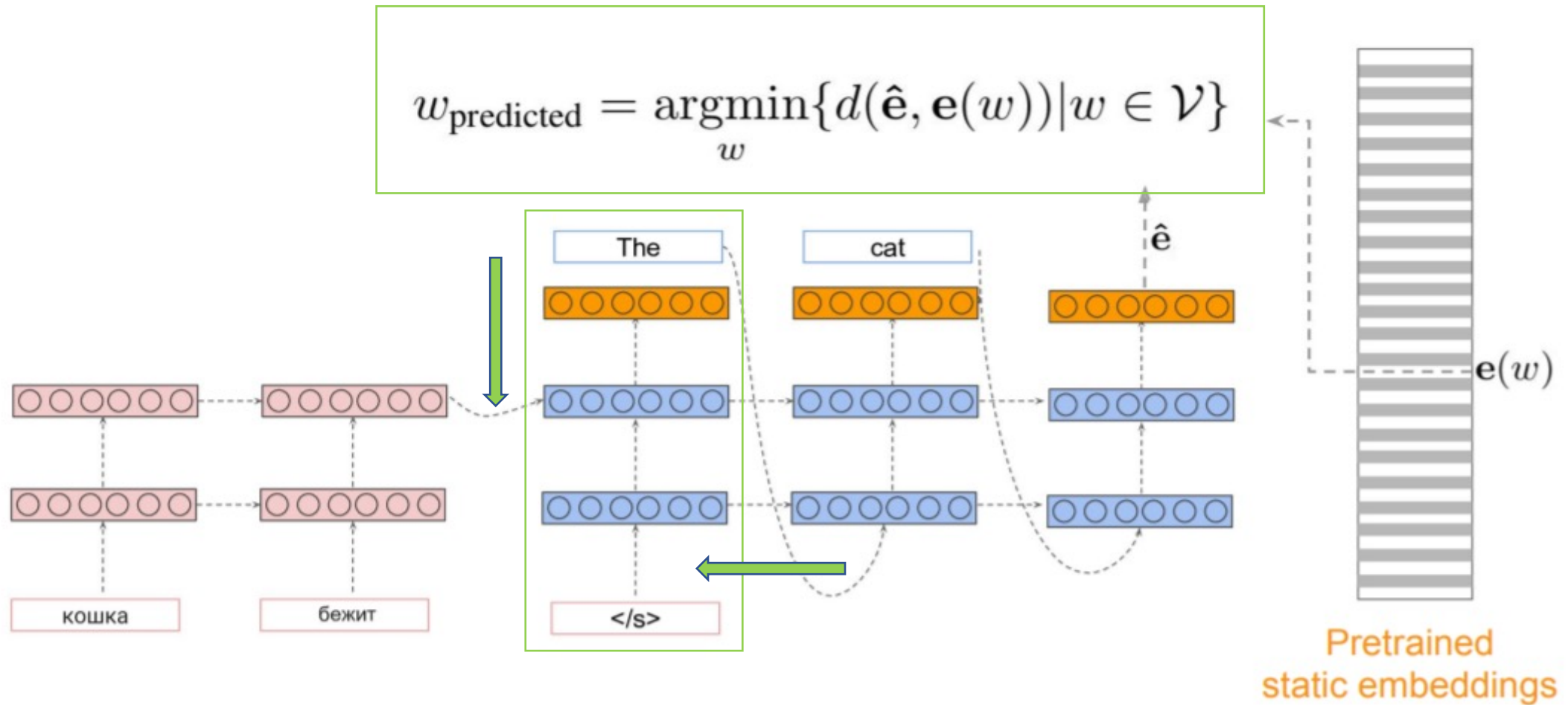
Background



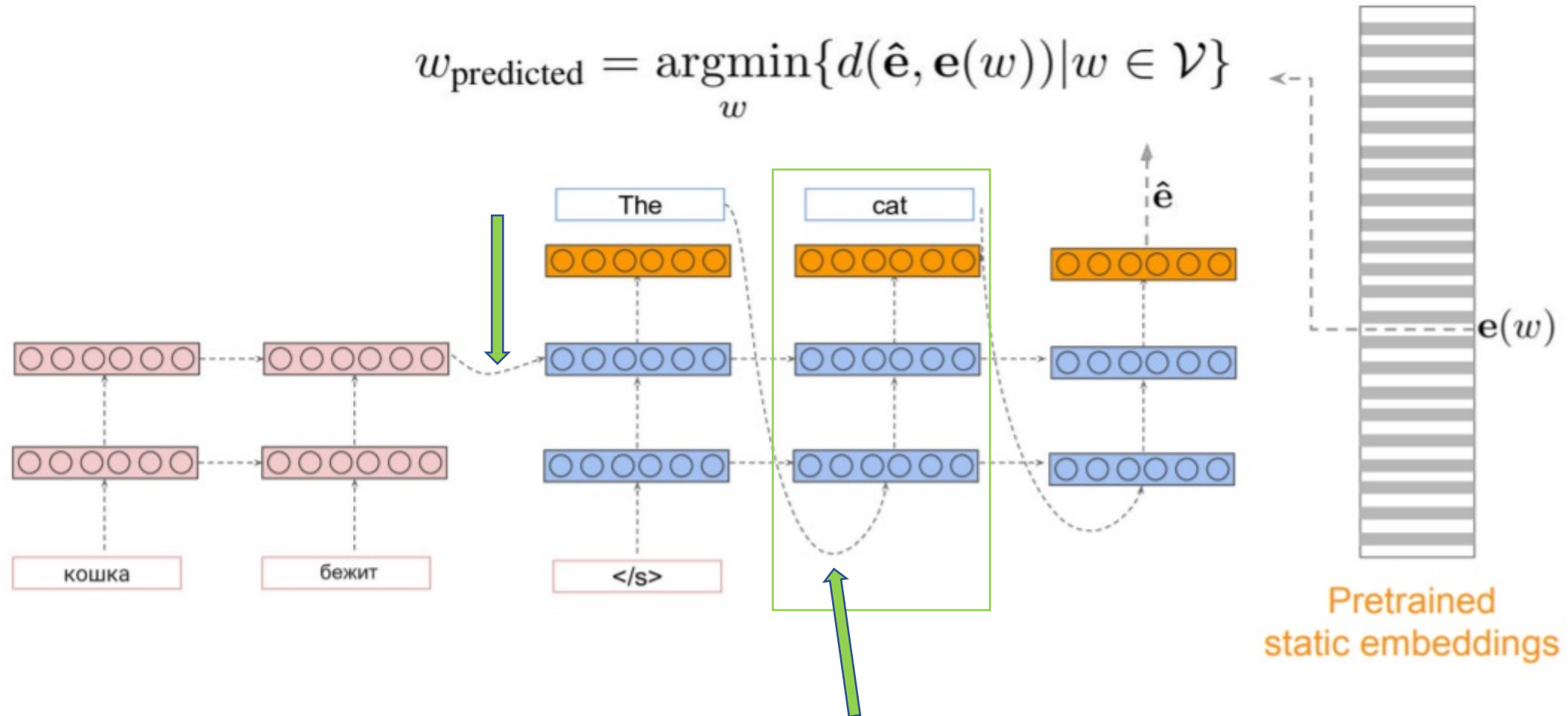
Background



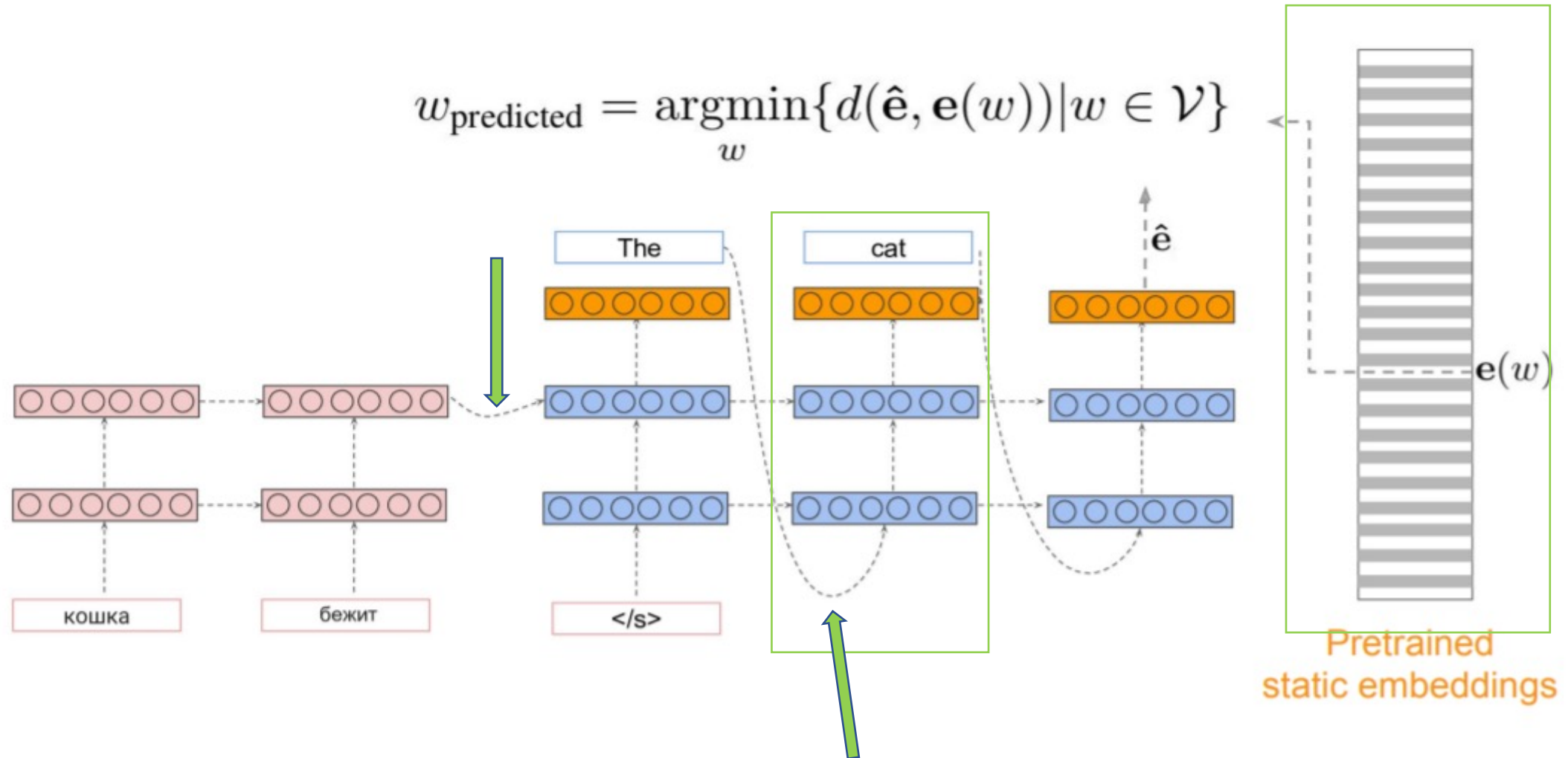
Background



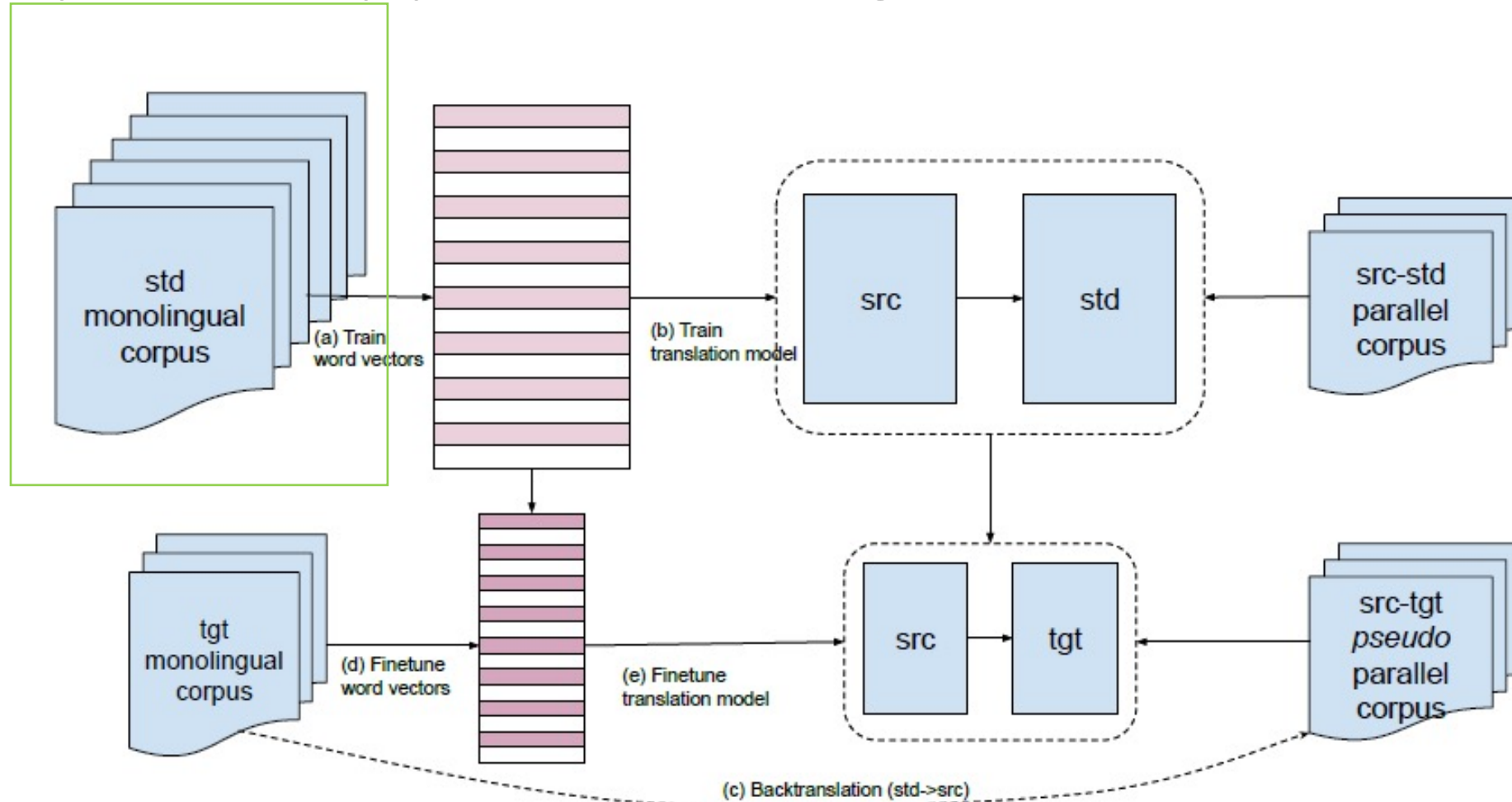
Background



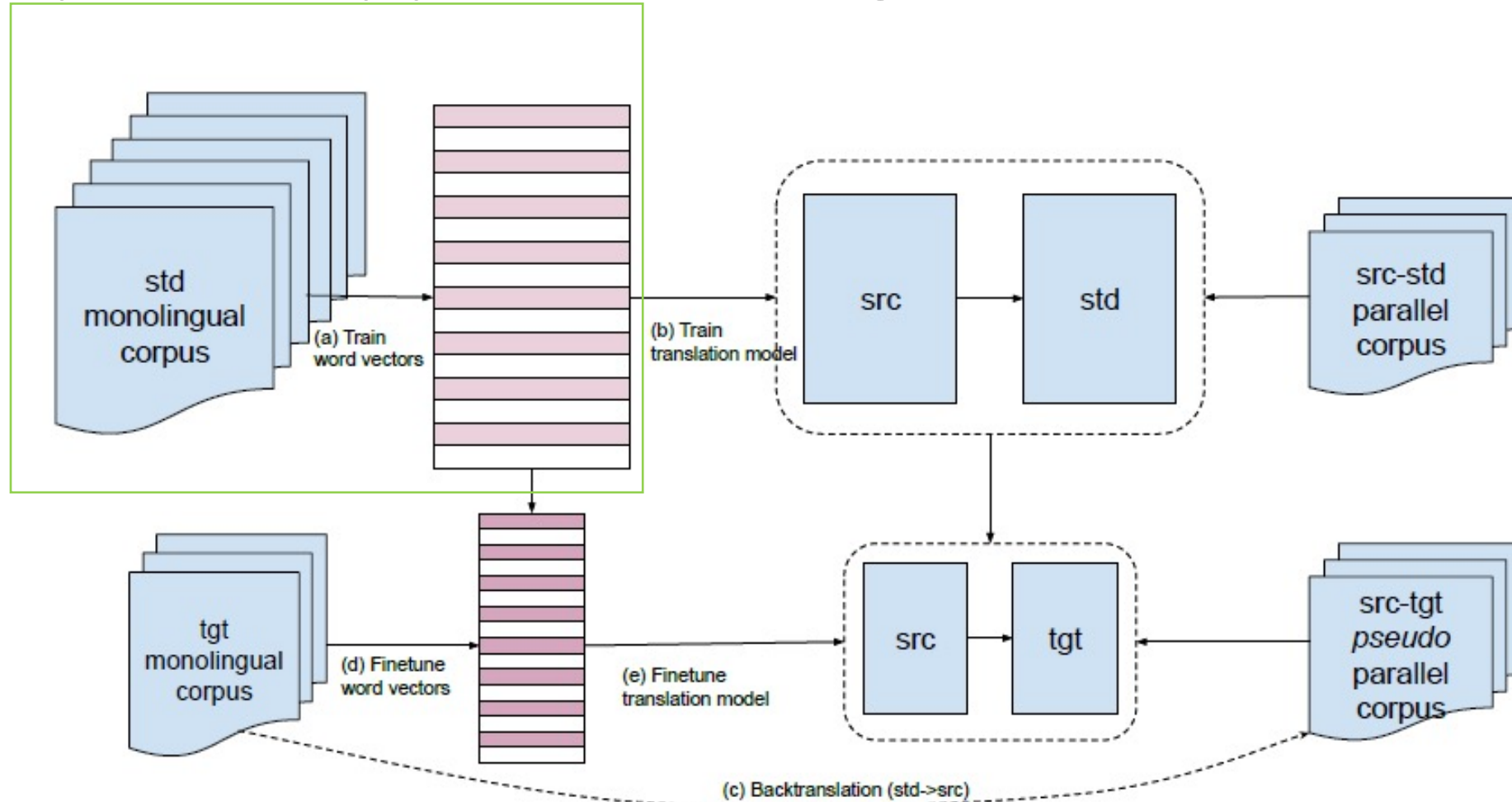
Background



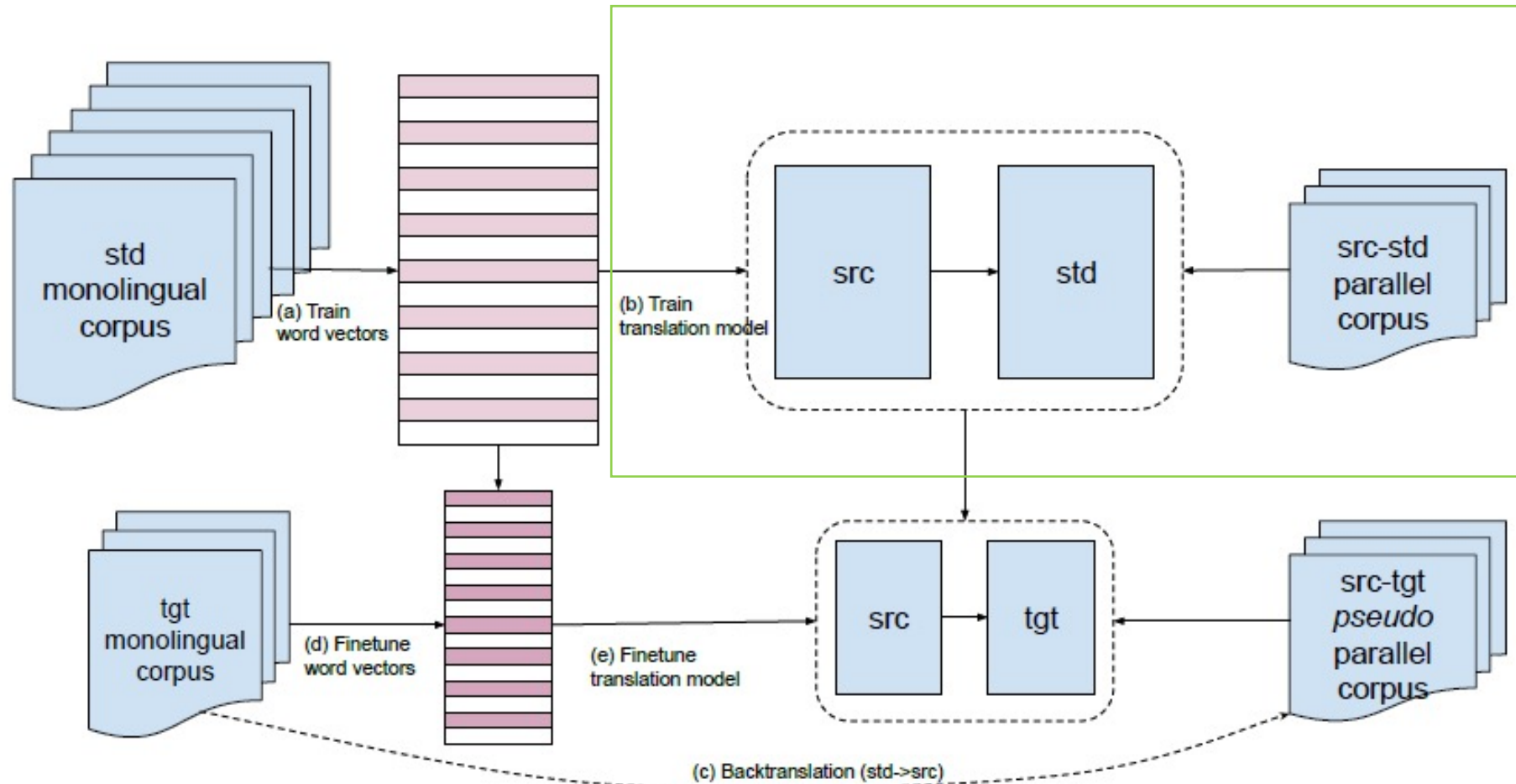
Proposed approach: LangVarMT



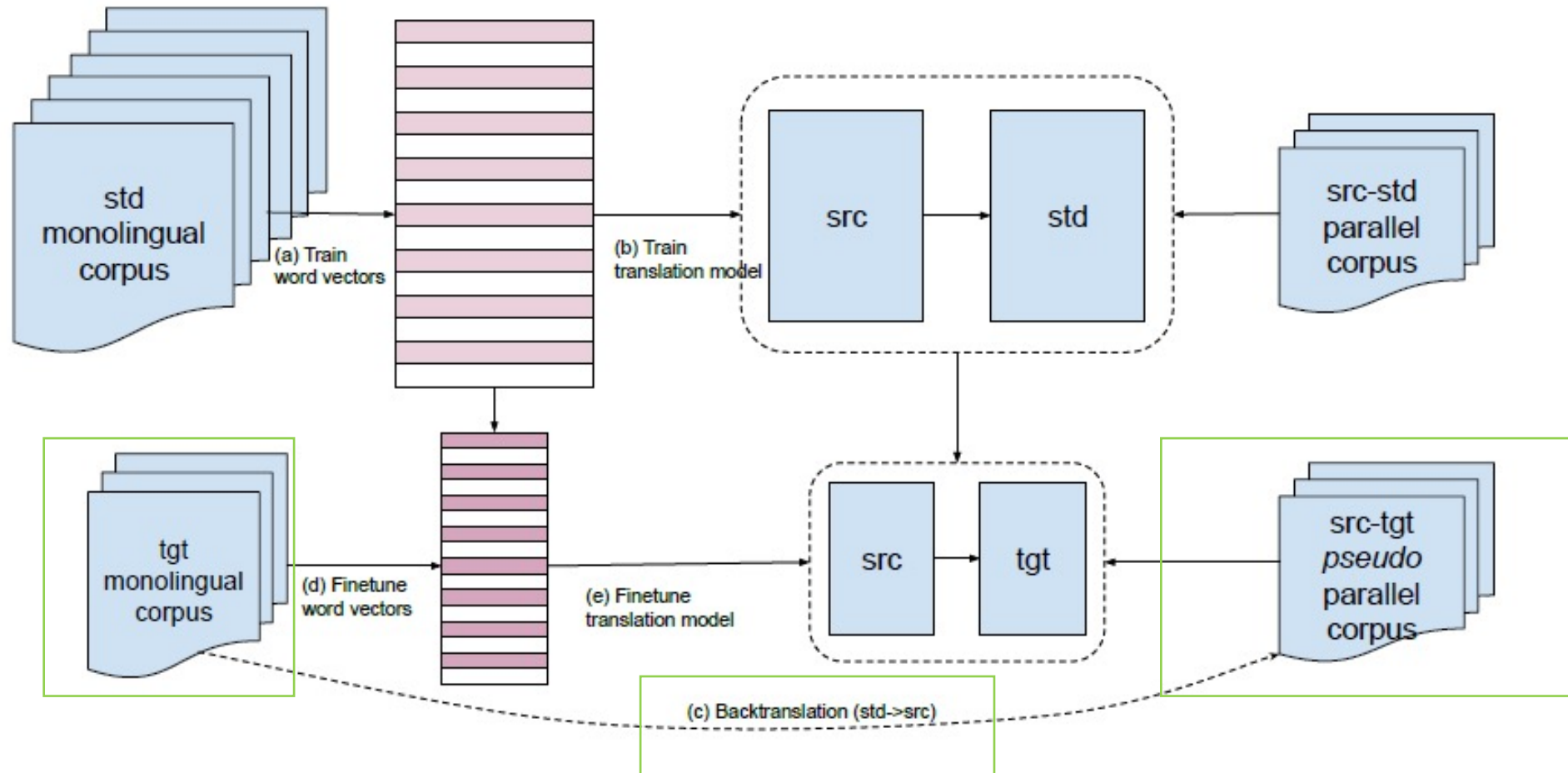
Proposed approach: LangVarMT



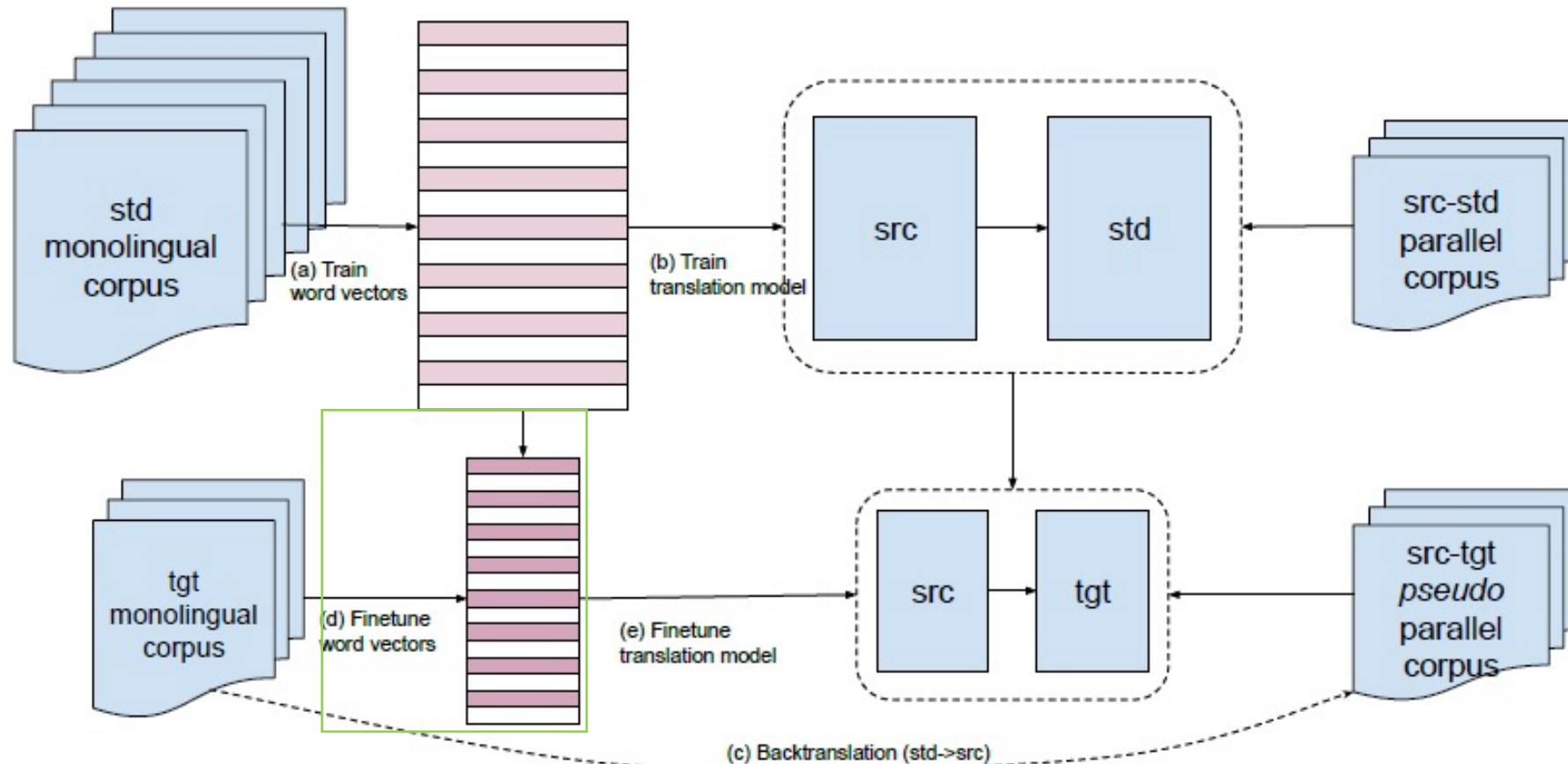
Proposed approach: LangVarMT



Proposed approach: LangVarMT



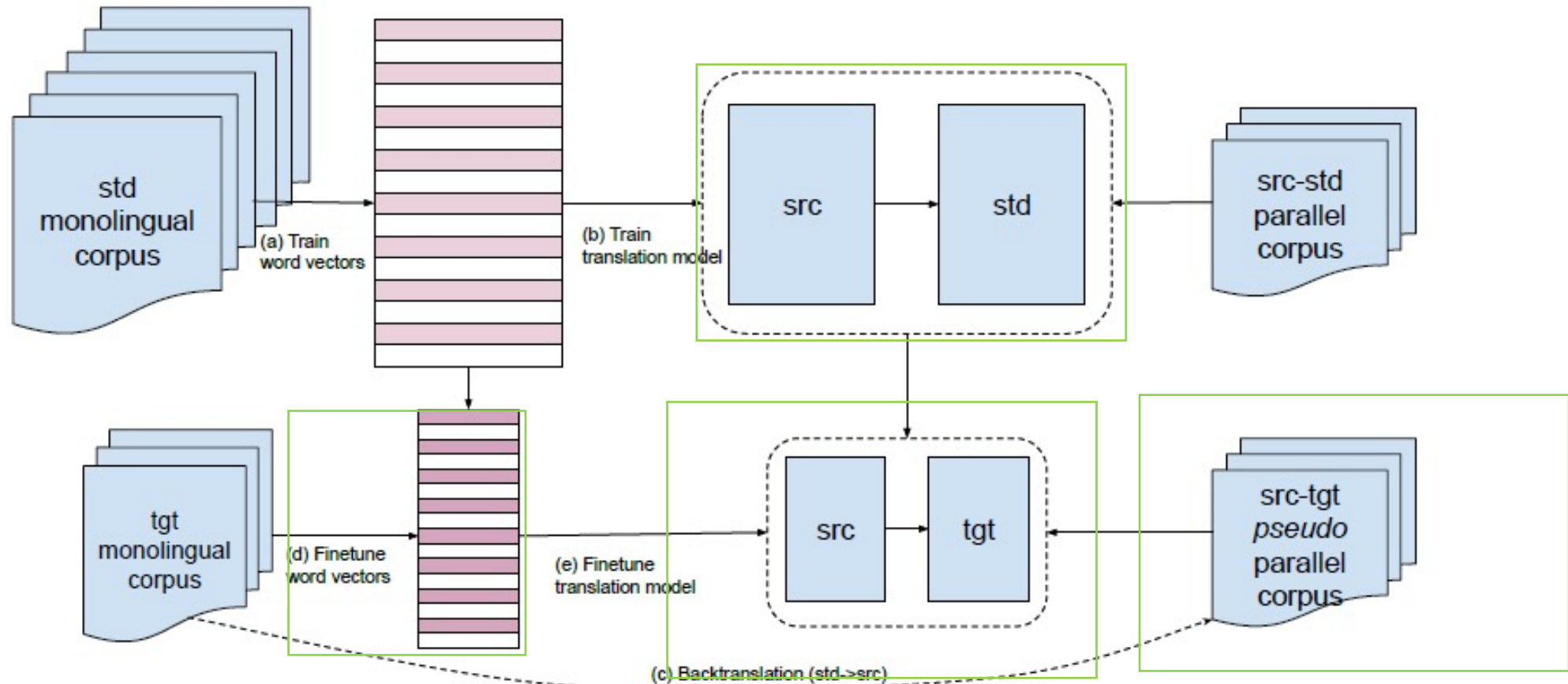
Proposed approach: LangVarMT



what are most people today not aware of?

whit ur maist fowk th'day nae aware o'?

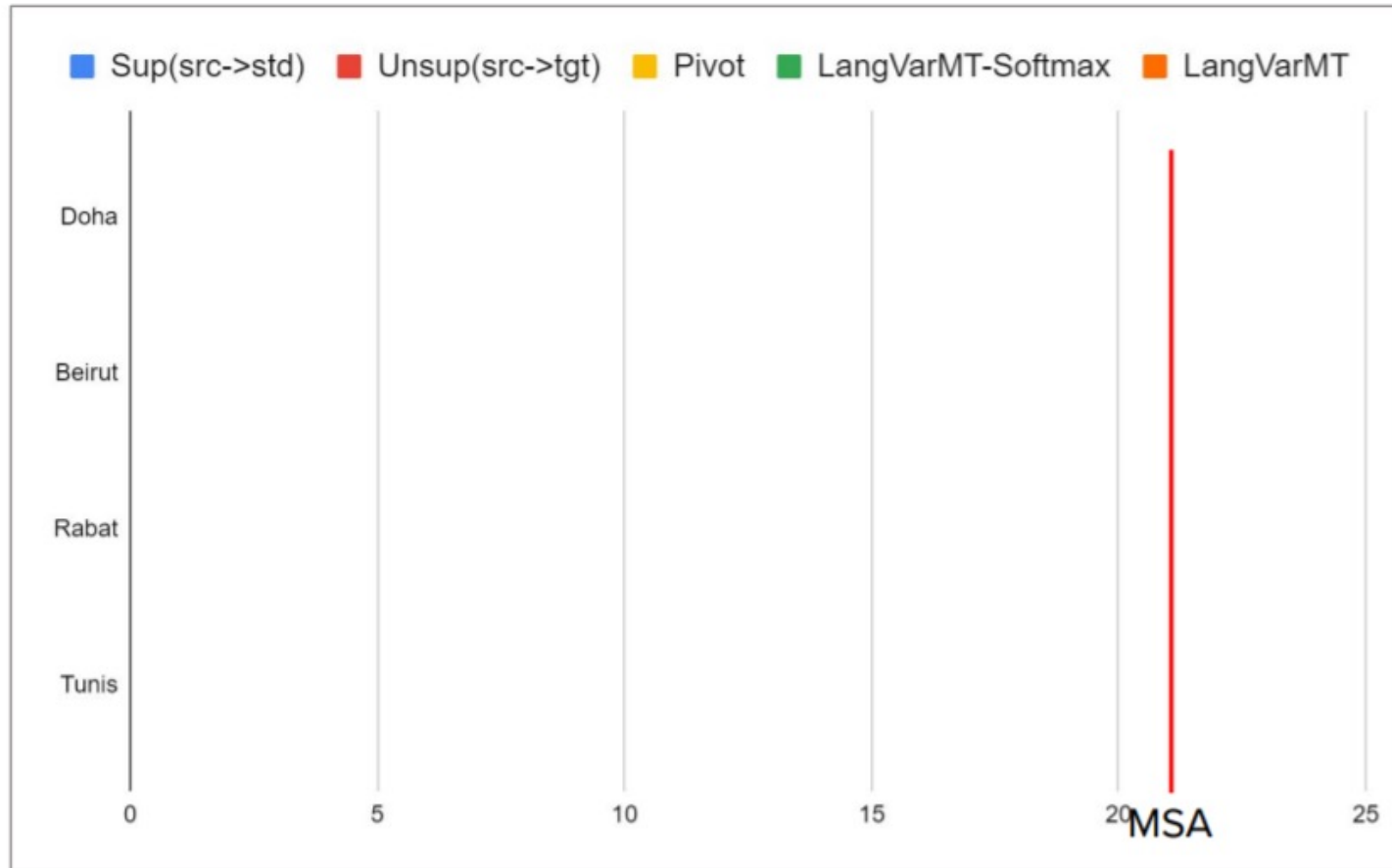
Proposed approach: LangVarMT



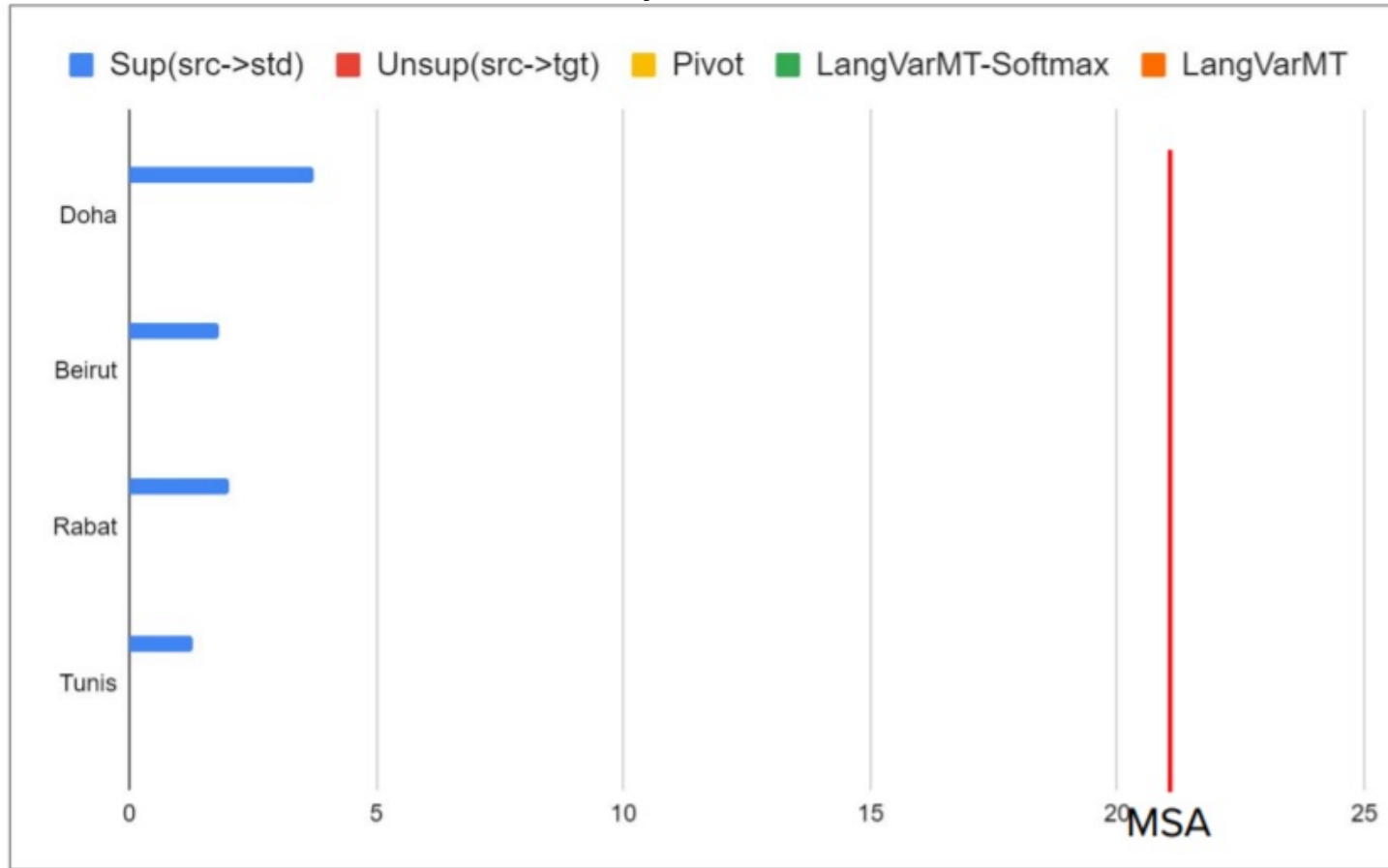
what are most people today not aware of?

whit ur maist fowk th'day nae aware o'?

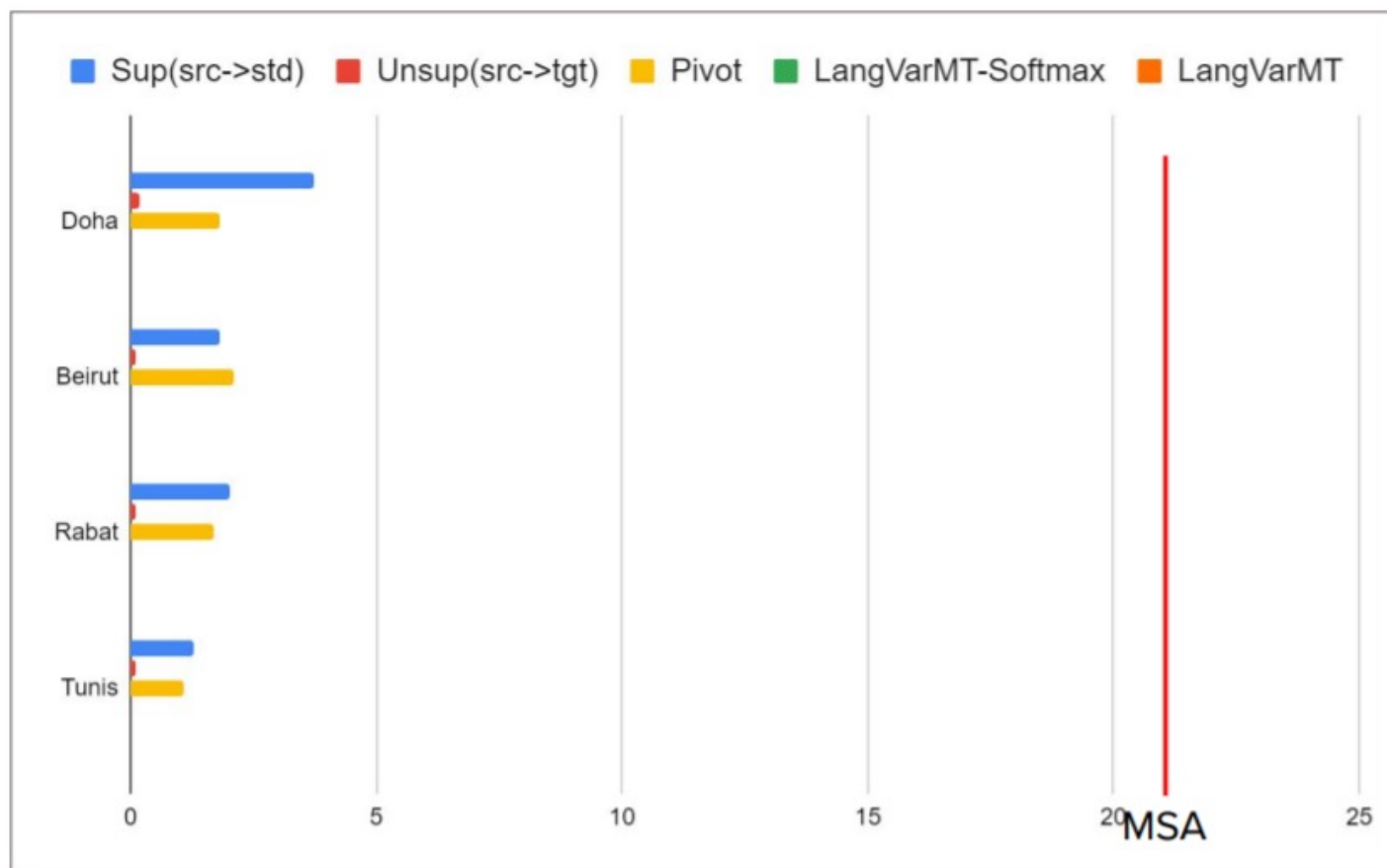
English to Arabic Varieties (with only 10K dialectal sentences)



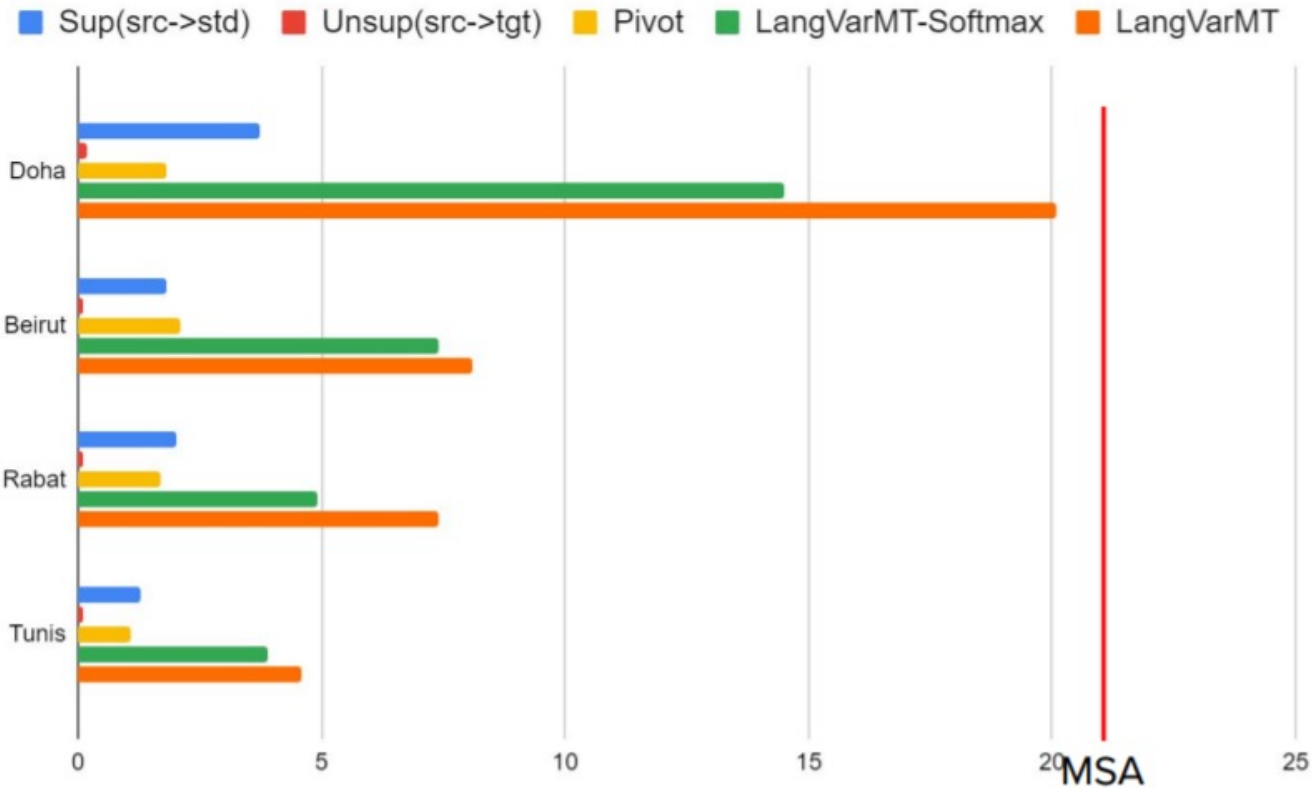
English to Arabic Varieties (with only 10K dialectal sentences)



English to Arabic Varieties (with only 10K dialectal sentences)



English to Arabic Varieties (with only 10K dialectal sentences)



Takeaways/conclusion

Rapidly adapt machine translation models to generate different language varieties

Lexical transfer through static subword embeddings

Model transfer through finetuning (on synthetic corpora)

Limitation: Back-translation of target using std->src model can be noisy if the varieties are not that close.

Failed experiments: English to Thai/Lao, English to Amharic/Tigrinya