FACULTY OF ARTS

KU LEUVEN

Classifying Northern and Southern Dutch

Tim Van de Cruys
KU Leuven

ELG workshop on Resources for Luxemburgish and Flemish – Thursday 8 July 2021

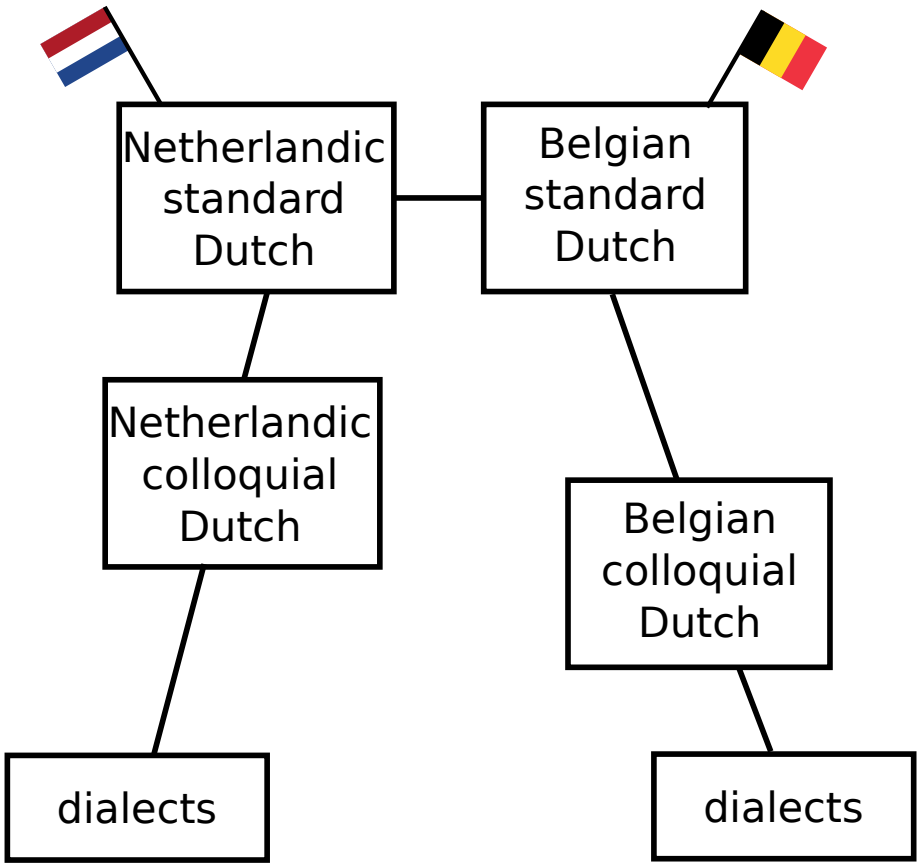
Introduction

Task

- Automatic classification of Dutch language varieties as they are used in the Netherlands (NL) and Belgium (BE)
- Similar to language identification, but more difficult due to similarity between variants
- How do novel transformer-based architectures with transfer learning fare?
- Can we deduce interesting linguistic features from the classification process?

Introduction

Linguistic situation



(Geeraerts, 2011)

Related Work

Discriminating between similar languages/varieties

- Zampiere & Gebre (2012): European and Brazilian Portuguese
- Lui & Cook (2013): American, Australian, and British English
- VarDial DSL shared tasks (2015–2021) for various language varieties
- Key takeaways:
  - Good performance with traditional feature-based machine learning (mostly words/character n-grams)
  - Often outperforms neural nets (Medvedeva et al. 2017)
  - Some researchers demonstrate increased performance with transformers; e.g. Bernier-Colborne et al. (2019) for cuneiform language identification

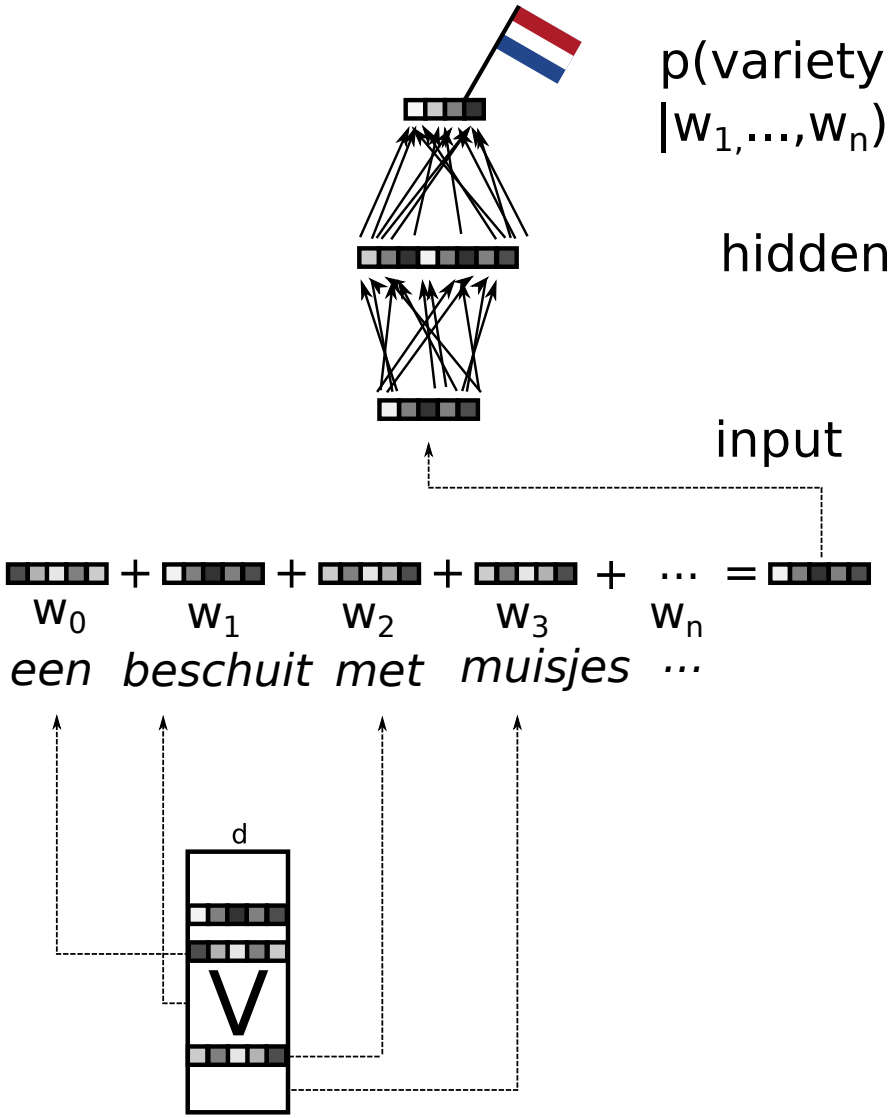
Related Work

DSL for Dutch

- van der Lee & van den Bosch (2017): SUBTIEL corpus (NL and BE subtitles)
- VarDial DSL 2018: shared task based on SUBTIEL corpus
  - Çöltekin et al. (2018)
  - van Halteren & Oostdijk (2018)
  - Kreutz & Daelemans (2018)
- Van Halteren (2020): controlled experiment in order to reduce domain bias

Models

fastText



(Joulin et al., 2017)

Models

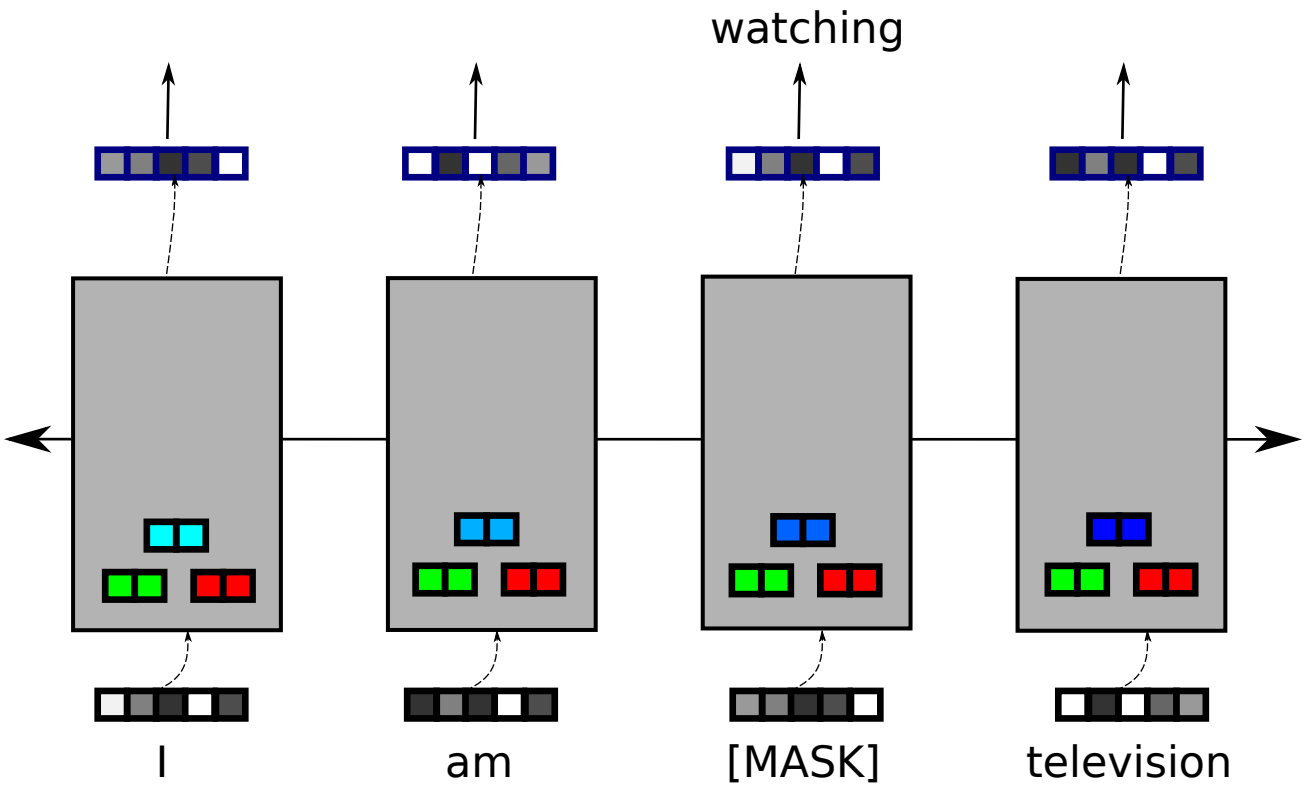
BERT

- Recent NLP model with state of the art results
- Representations based on bi-directional context
- Transformer architecture
- General training on language modeling task, finetuning on specific NLP task

(Devlin et al., 2019)

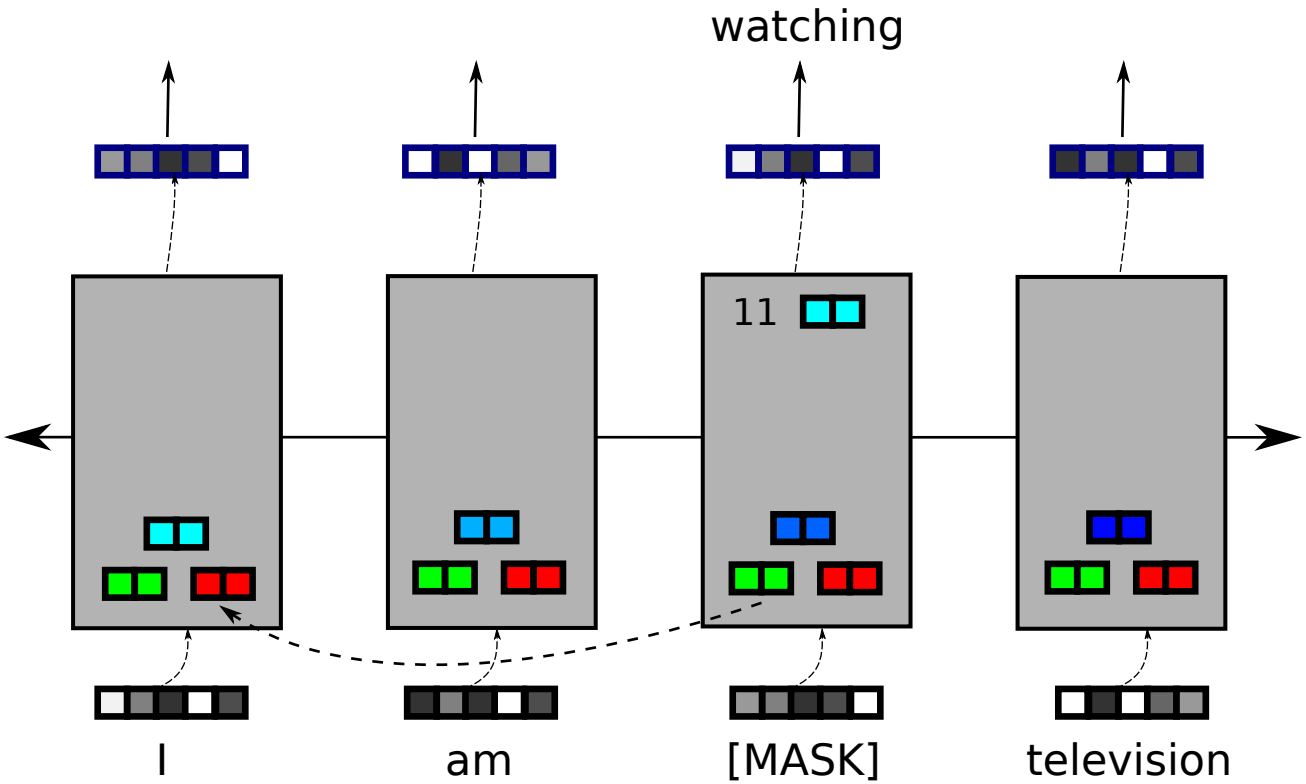
Models

BERT: self-attention



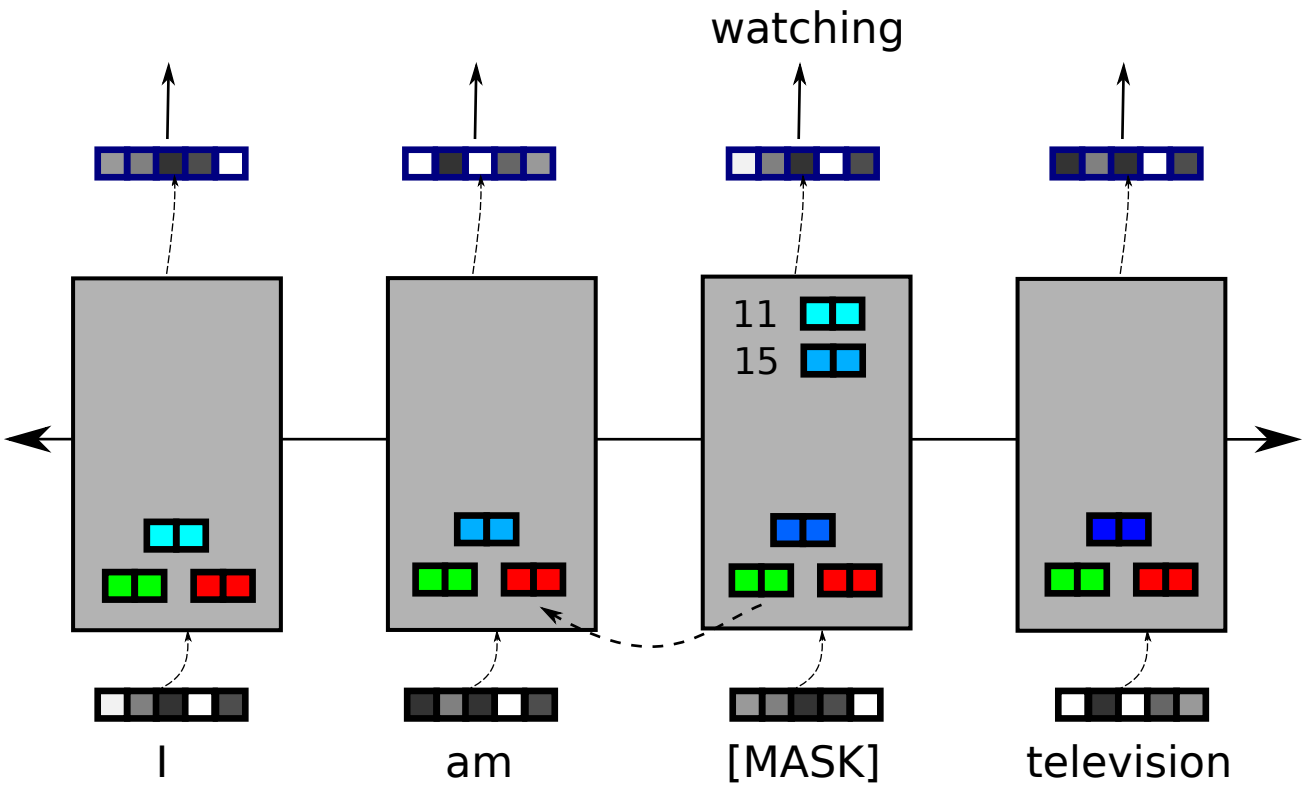
Models

BERT: self-attention



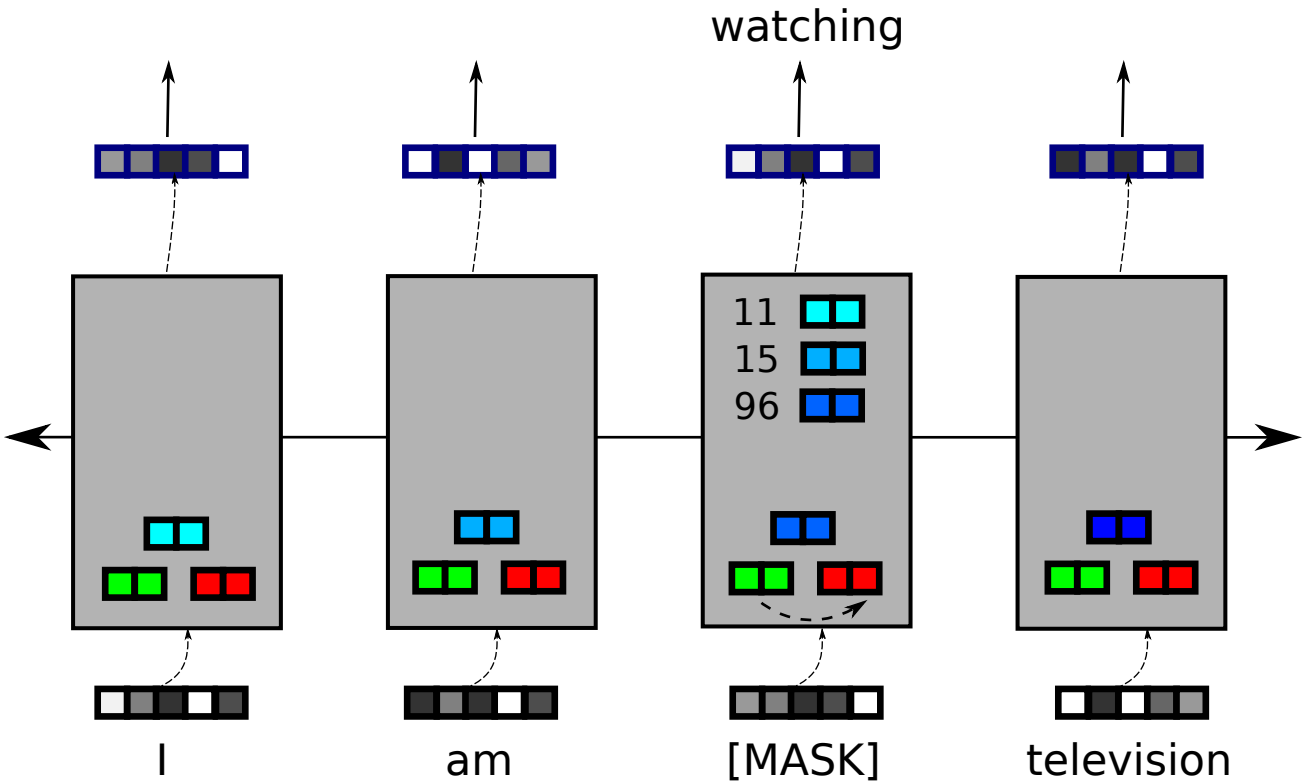
Models

BERT: self-attention



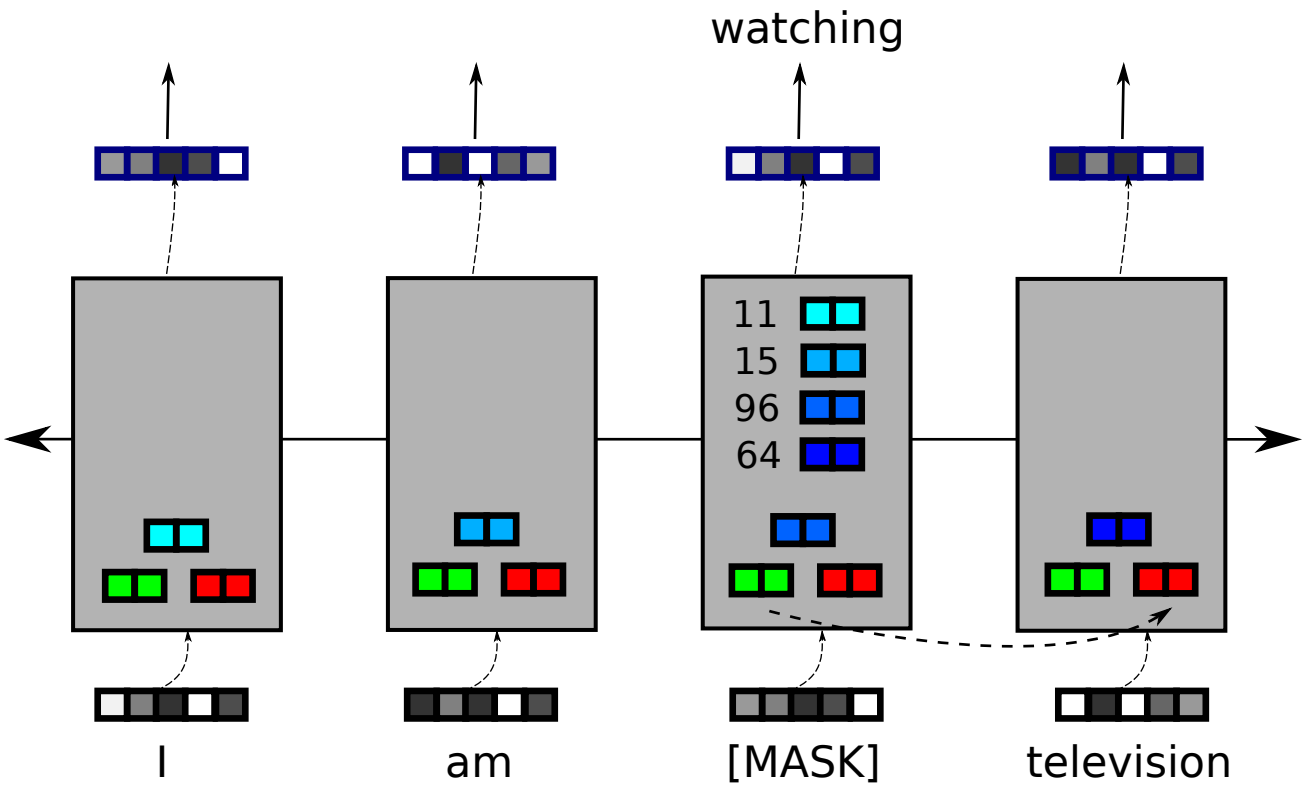
Models

BERT: self-attention



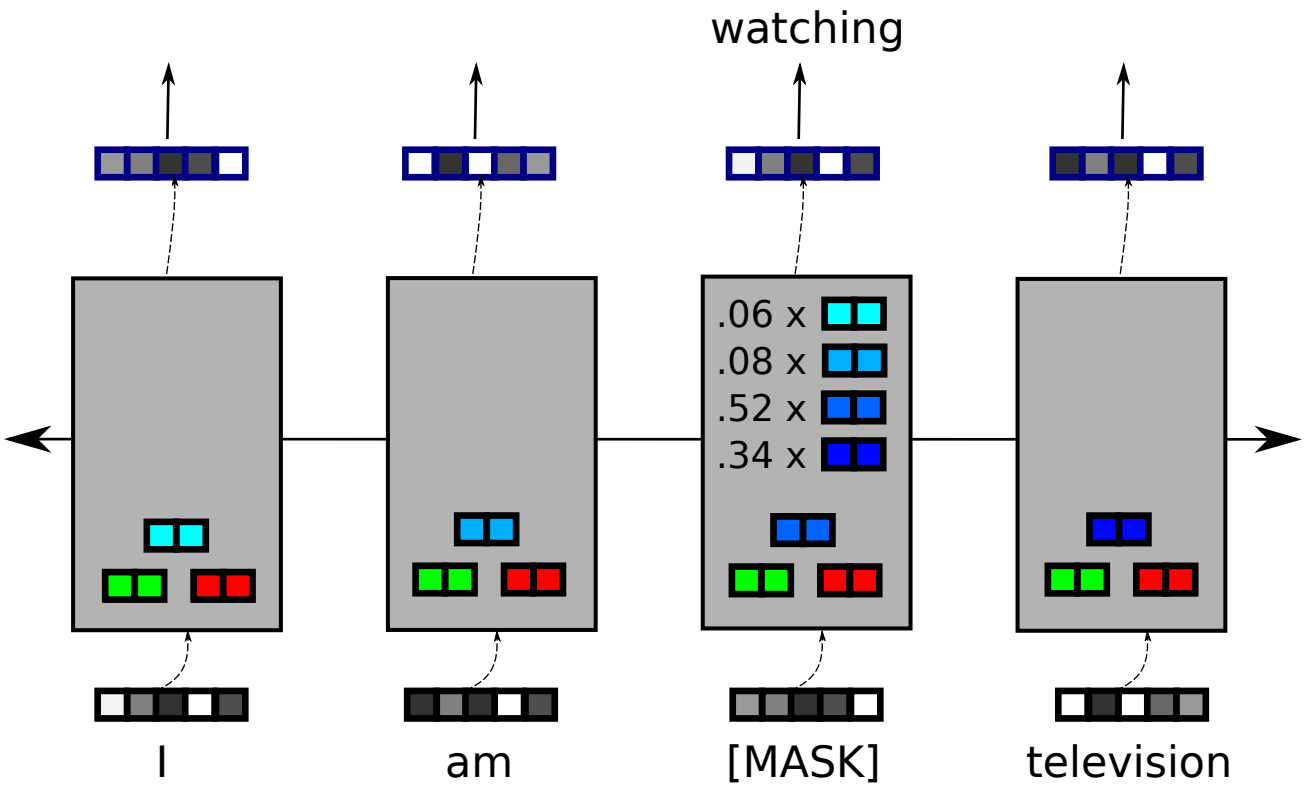
Models

BERT: self-attention



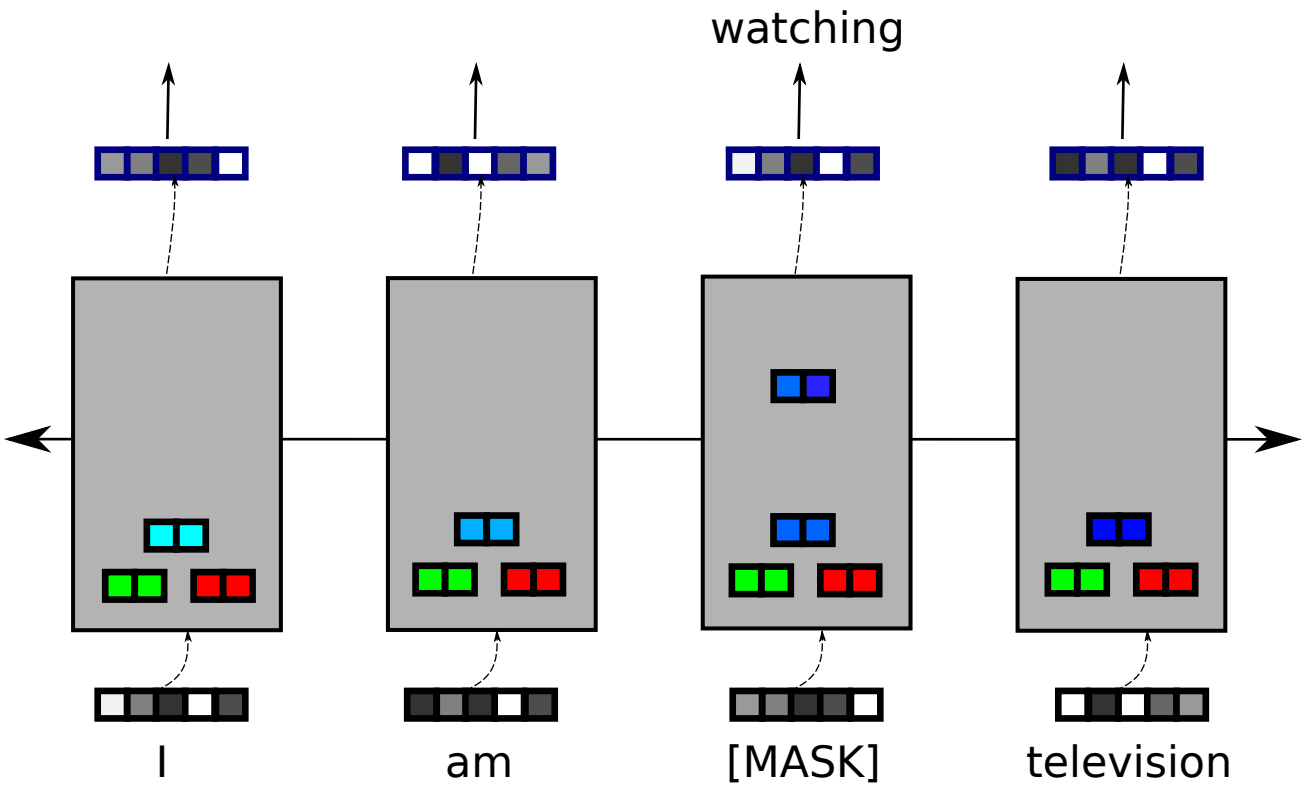
Models

BERT: self-attention



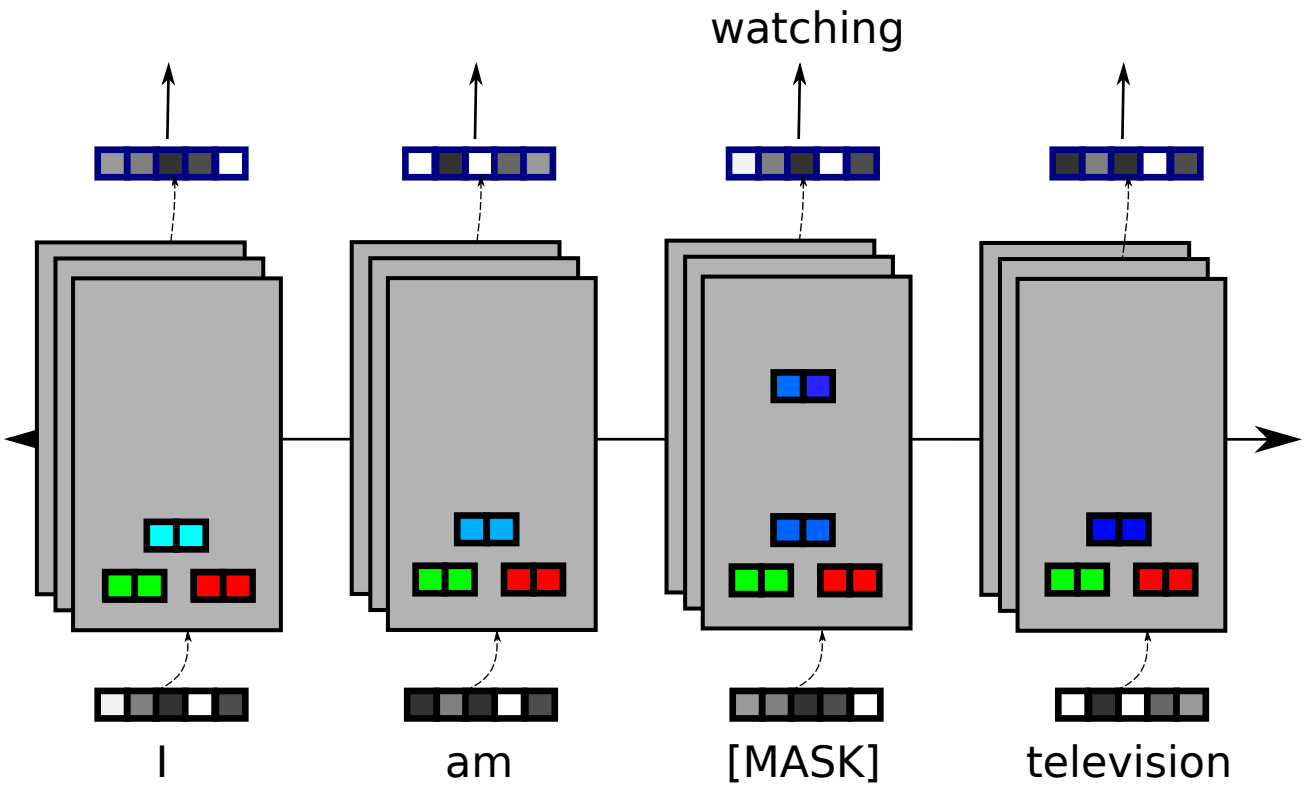
Models

BERT: self-attention

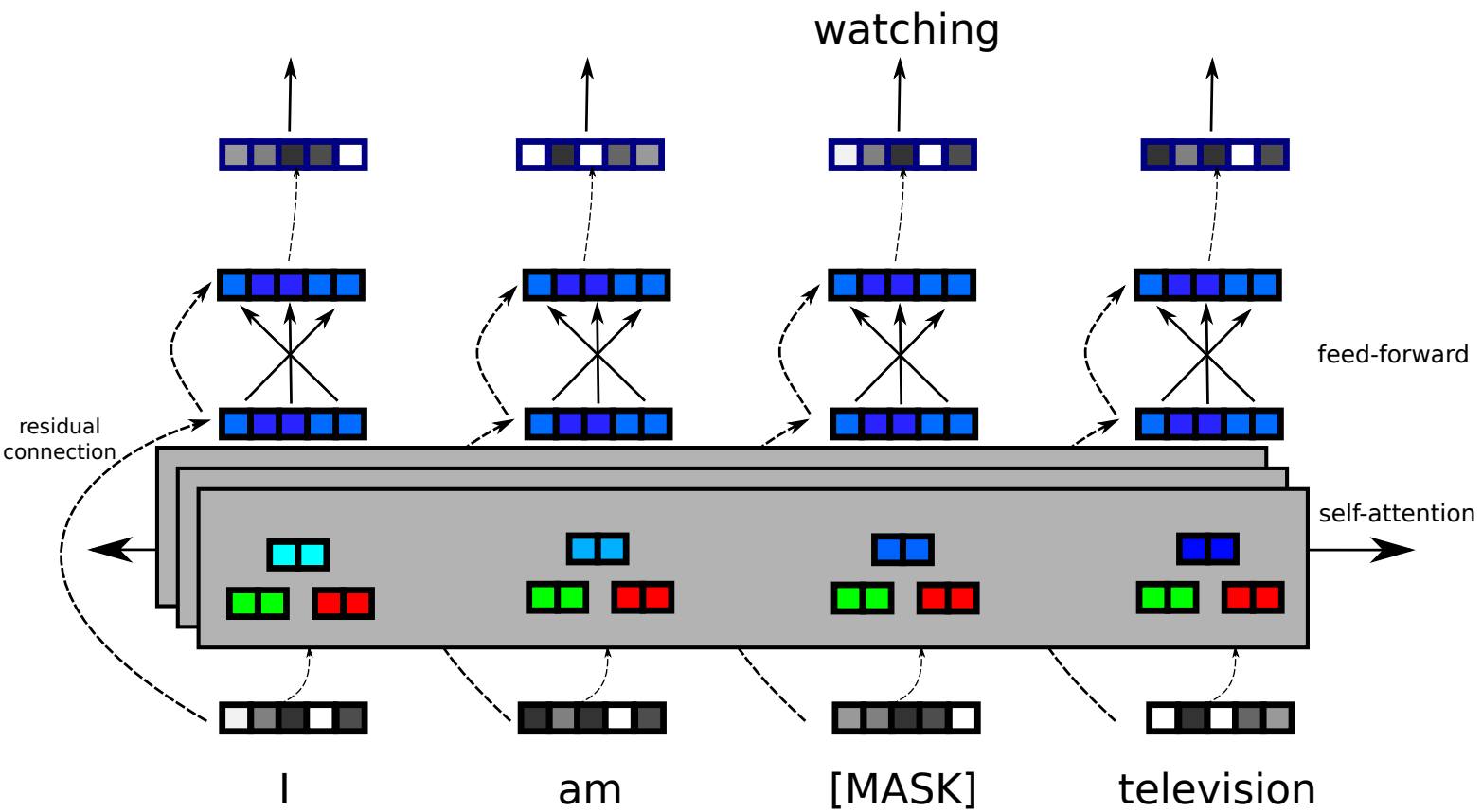


Models

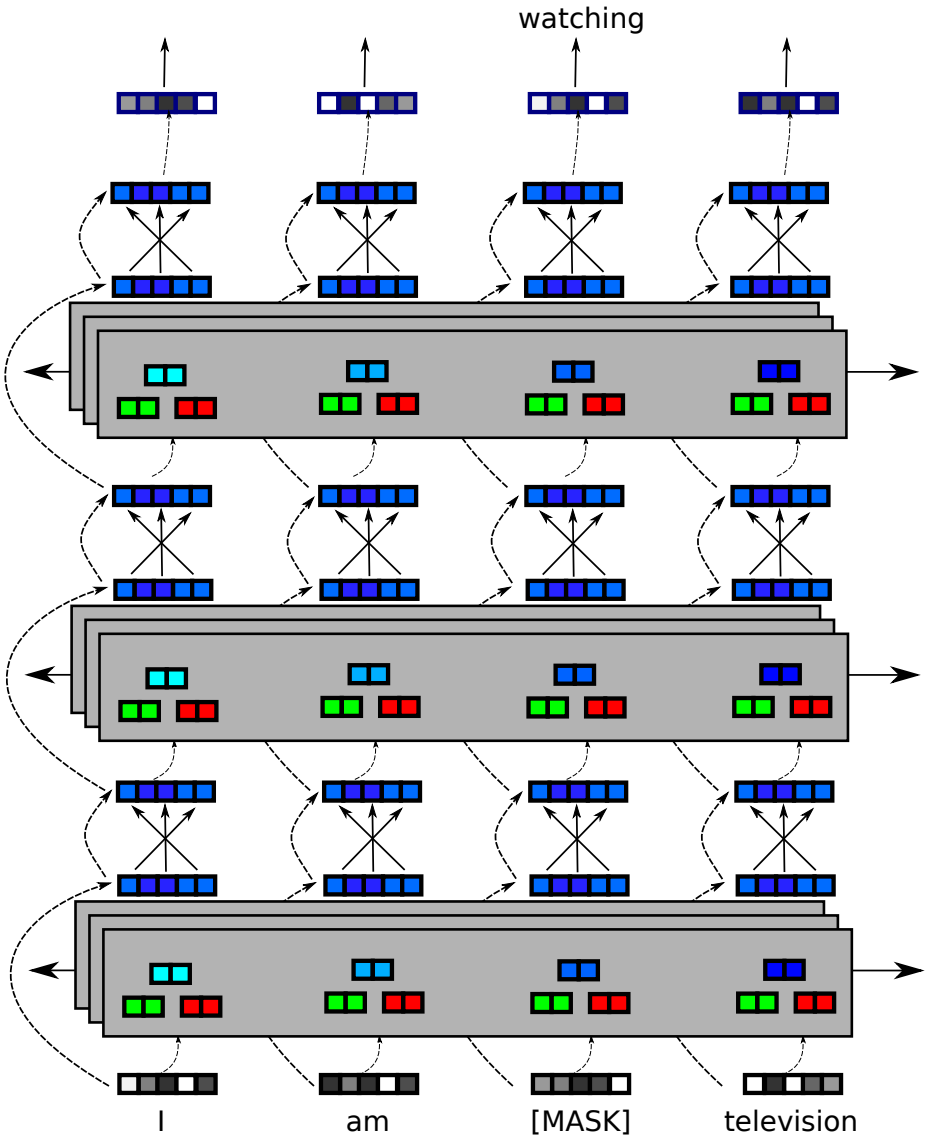
BERT: self-attention



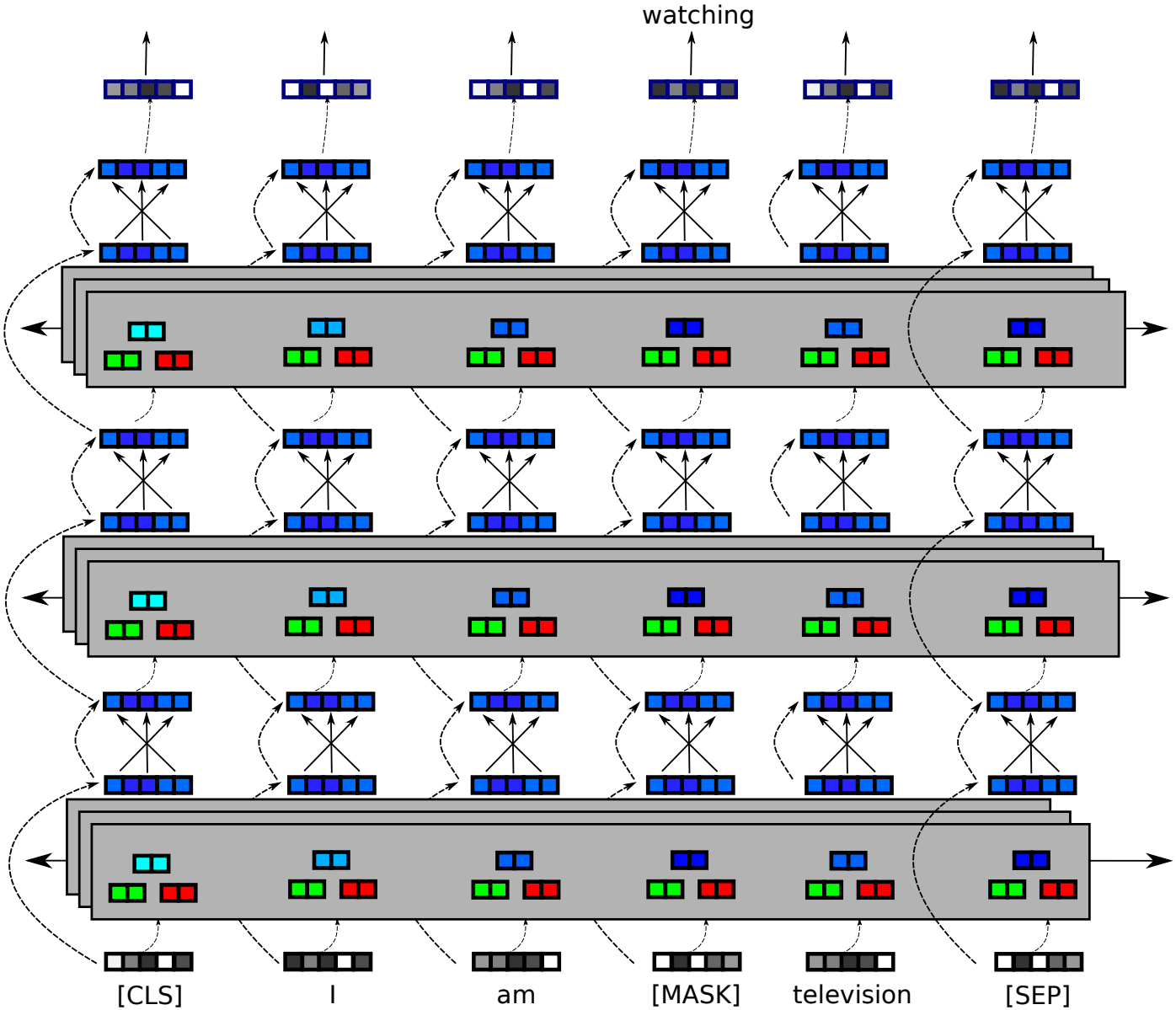
BERT


watching
feed-forward
residual
connection
self-attention
I
am
[MASK]
television

BERT



BERT



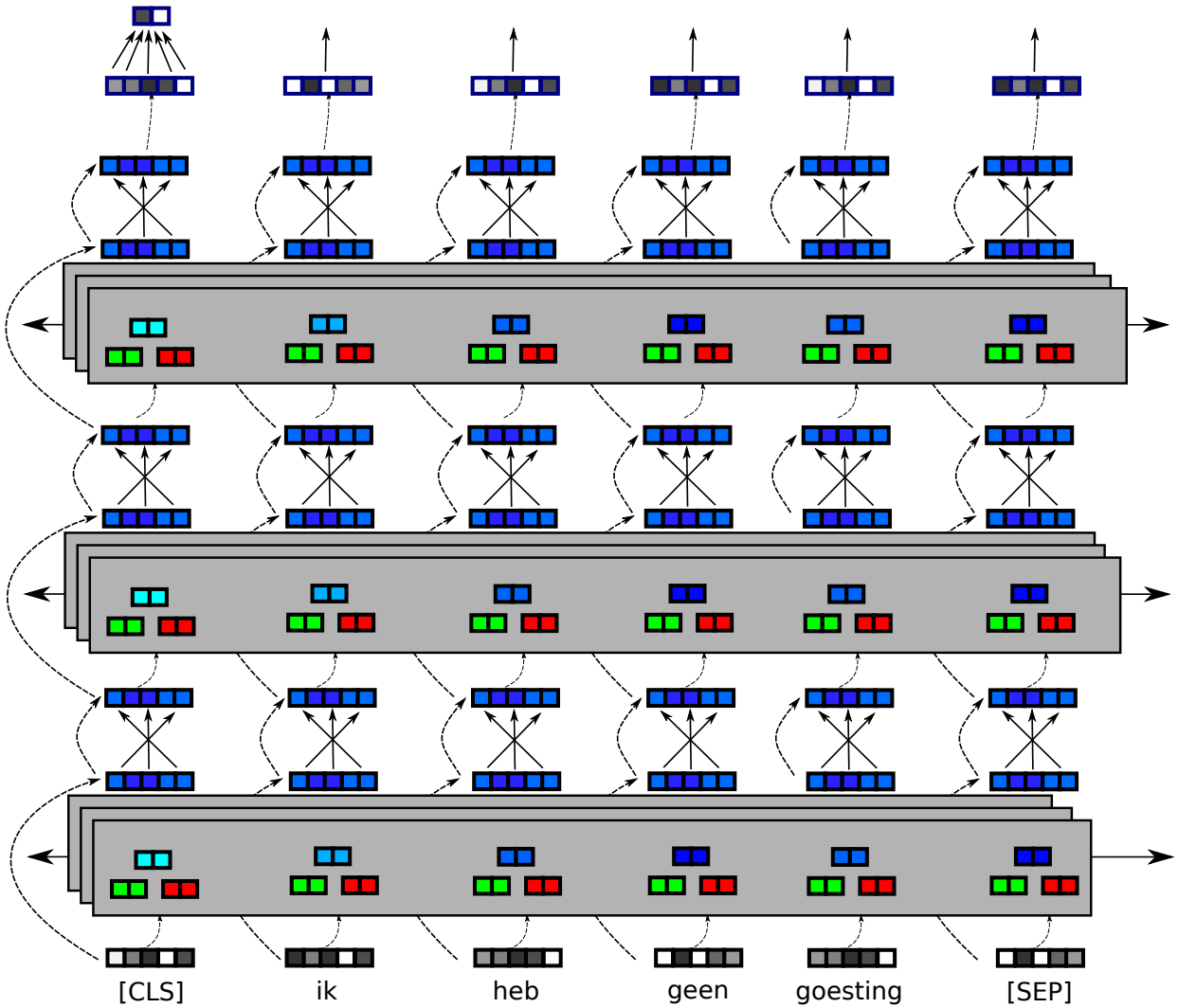
BERT

Pre-training and finetuning

- BERT is pre-trained on the masked language modeling task using a large corpus
- Once the model is pretrained for language modeling, it can be finetuned for a special natural language processing task
- Special, reserved token <CLS> is added to the start of each sentence
- The resulting embedding is considered as a representation of the entire sentence
- The representation can be used to perform a final classification task (by adding a softmax classification layer on top)
- The representation can be used as is, but the most common practice is to **finetune all the parameters of the model** to the task at hand

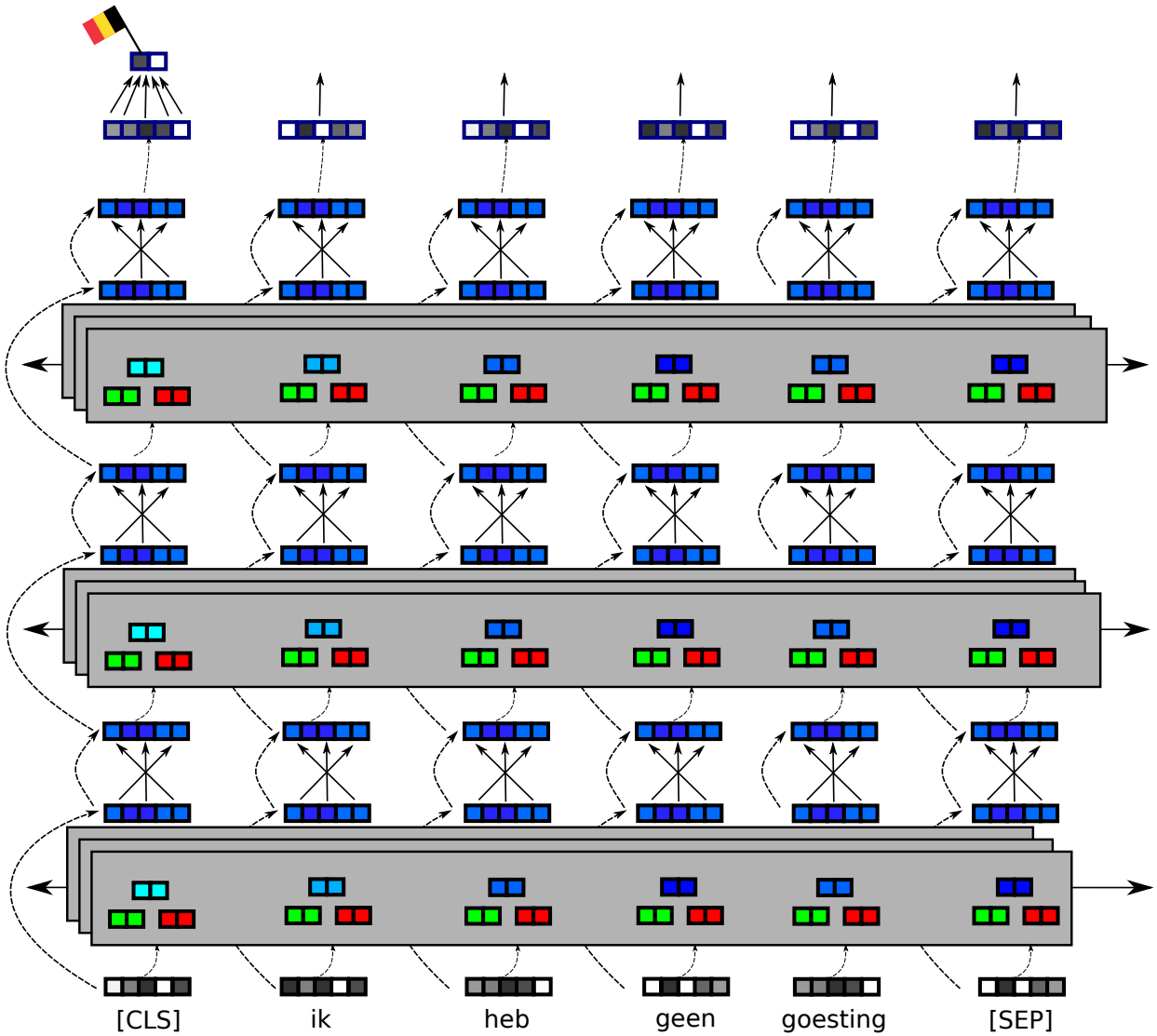
BERT

For NLP tasks



BERT

For NLP tasks



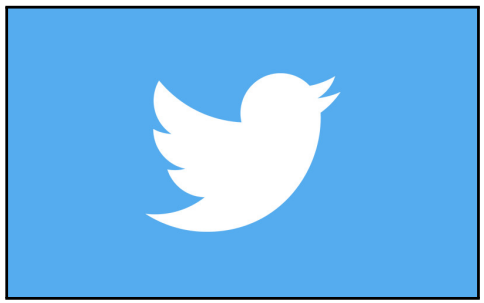
BERT for Dutch

Two models

- **BERTje** (de Vries et al., 2019): trained on 2.4 billion tokens (diverse data: TwNC, SoNaR, ...)
- **RobBERT** (Delobelle et al., 2020): trained on 6.6 billion tokens (Dutch part of Oscar corpus, i.e. CommonCrawl data)

Data

- Social media
  - 1 million tweets written in Dutch (2015–2021)
  - semi-automatically labeled based on user-defined location (BE or NL)
- Newspapers
  - 1 million sentences from *NRC* (NL) and *Standaard* (BE) (2016–2018)



Results

Twitter

	Twitter			
	acc	prec	rec	F1
baseline	.72	.00	.00	.00
fastText	.87	.81	.70	.75
BERTje	.87	.83	.68	.75
RobBERT	.89	.85	.74	.79

- precision/recall/F1 computed for minority class

Results

News

	News			
	acc	prec	rec	F1
baseline	.50	.00	.00	.00
fastText	.80	.80	.80	.80
BERTje	.83	.83	.83	.83
RobBERT	.83	.83	.83	.83

Results

Analysis: Twitter with fastText

$p(nl)$	$p(be)$	tweet
1.00	0.00	Neerslag Gisteren #neerslag #Drenthe #Emmen #Nederland
1.00	0.00	#zoektwerk #vacature Bijbaan postbezorger in Zierikzee ...
0.75	0.25	Je merkt echt al dat de dagen lengen! We gaan weer richting zomer!
0.50	0.50	Ik hoor gewoon haar lach als ik deze foto zie
0.04	0.96	Joepie, het is weekend
0.00	1.00	Beringen | HLN “Sorry voor de veroorzaakte overlast” ...
0.00	1.00	Is dat een smartwatch? @crevits #deafspraak

Results

Analysis: Twitter with RobBERT

$p(nl)$	$p(be)$	tweet
1.00	0.00	#zzp markt-update: Gemeente biedt ZZP'ers ontbijt aan - ...
1.00	0.00	Tilburgers Bas en Noortje bedachten een supergaaf idee voor ...
0.79	0.21	Oh en ik mis ook mijn courgetteplant. Verder niet echt iets, ...
0.52	0.48	Eet vezelrijke groenten, die helpen bij het voldoen aan je ...
0.28	0.71	Die ballonnetjes zijn te leuk haha
0.00	1.00	Zeg mij aub da ik ni de enige ben die gwn boven op zijn/haar ...
0.00	1.00	Als het school belt woensdagvoormiddag betekent da ik ...

Results

Analysis: newspapers with fastText

$p(nl)$	$p(be)$	sentence
1.00	0.00	„ Dat zit in alle poriën van de voorstelling .
1.00	0.00	Waar staat Rutte III voor ?
0.80	0.20	Lekker , maar ook een beetje eentonig .
0.66	0.34	Spanningsdips zijn niet uitzonderlijk , zegt Slootweg .
0.51	0.49	Het was te laat , het doek was al doorweekt .
0.01	0.99	Aan Franstalige zijde heeft , rara , vooral het CDH ...
0.00	1.00	’ Zeker .

Results

Analysis: newspapers with RobBERT

$p(nl)$	$p(be)$	sentence
1.00	0.00	„ En het is heel mooi dat de supporters dit zelf hebben geregeld . ”
1.00	0.00	De toename van afbreekbare bioplastics heeft volgens Bergsma ...
0.87	0.13	Perfecte combinatie om iets zinnigs over het artikel te zeggen , lijkt me
0.51	0.49	Soms moet je domweg geluk hebben .
0.50	0.50	De glans is er onherstelbaar af .
0.00	1.00	Het parket is van plan te kijken of er sprake was van huisjesmelkerij .
0.00	1.00	En ik denk dat het werk de jongere generatie kan aanspreken . ’

Results

Analysis: difference fastTtext/BERT

correlation	Kendall's τ
Twitter	.60
News	.71

fT $p(be)$	RB $p(be)$	instance
.27	.99	Fobby kijkt tv met Noortje. Want ik lig asociaal in mijn zetel.
.39	.99	Lachwekkend noem ik u
.46	.95	We willen hier ook regelmatig een optreden organiseren
.46	.98	De sociale-inspectiediensten kunnen zien welke zaken met een witte kassa werken , maar gaan ook daar nog controleren

Conclusion

- Identification of varieties of Dutch is a highly lexicalized task
- Strong performance of lexical classification method, viz. fastText
- Moderate but consistent improvement using fine-tuned transformer architecture
- Qualitative analysis: fastText strongly influenced by lexical features, transformer more interesting for linguistics

Future work

- Combination of classifiers: best of both worlds
  - Decision based on probabilities of both classifiers
  - Adversarial setting: train transformer on examples that are difficult for fastText
- Preprocessing of corpora
  - Construction of clean and balanced corpora
  - Masking of named entities
- Comparison with traditional feature-based on SUBTIEL corpus

References I

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger, *Improving cuneiform language identification with bert*, Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, 2019, pp. 17–25.

Çagrı Çöltekin and Taraka Rama, *Tübingen-oslo at semeval-2018 task 2: Svms perform better than rnns in emoji prediction*, Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 34–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim, *Bertje: A dutch bert model*, arXiv preprint arXiv:1912.09582 (2019).

References II

Pieter Delobelle, Thomas Winters, and Bettina Berendt, *Robbert: a dutch roberta-based language model*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 3255–3265.

Dirk Geeraerts, *Een zondagspak? het nederlands in vlaanderen: gedrag, beleid, attitudes*, Ons erfdeel **44** (2001), no. 3, 337–343.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, *Bag of tricks for efficient text classification*, arXiv preprint arXiv:1607.01759 (2016).

Tim Kreutz and Walter Daelemans, *Exploring classifier combinations for language variety identification*, Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), 2018, pp. 191–198.

Marco Lui and Paul Cook, *Classifying english documents by national dialect*, Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013), 2013, pp. 5–15.

References III

Maria Medvedeva, Martin Kroon, and Barbara Plank, *When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages*, Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), 2017, pp. 156–163.

Chris van der Lee and Antal van den Bosch, *Exploring lexical and syntactic features for language variety identification*, Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial), 2017, pp. 190–199.

Hans van Halteren, *Domain bias in distinguishing flemish and dutch subtitles*, Natural Language Engineering **26** (2020), no. 5, 493–510.

BJM van Halteren and NHJ Oostdijk, *Identification of differences between dutch language varieties with the vardial 2018 dutch-flemish subtitle data*.

Marcos Zampieri and Binyam Gebrekidan Gebre, *Automatic identification of language varieties: The case of portuguese*, KONVENS2012-The 11th Conference on Natural Language Processing, Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), 2012, pp. 233–237.