REFLECTIONS OF SYNTACTIC STRUCTURES IN NON-AUTOREGRESSIVE LANGUAGE MODELS

Pletenev S.A.

Moscow, 2021

INTRODUCTION

Since the popularization of the Transformer as a general purpose feature encoder for NLP, many studies have attempted to decode linguistic structure from its novel multihead attention mechanism. However, much of such work focused almost exclusively on autoregressive style of decoding.

In this study, we present decoding experiments on Non-autoregessive models in order to test the generalizability of the claim that dependency syntax is reflected in attention patterns.

RELATED WORKS

Models

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. **Non-autoregressive machine translation with disentangled context transformer.**

Junliang Guo, Linli Xu, and Enhong Chen. **Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation**

Jungo Kasai, Nikolaos Pappas, Hao Peng et al**. Deep Encoder, Shallow Decoder: Reevaluating the SpeedQuality Tradeoff in Machine Translation**

Datasets

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. **Distilling the knowledge in a neural network**

Kim Yoon, Rush Alexander M. **SequenceLevel Knowledge Distillation**

Methodology

Voita Elena, Talbot David, Moiseev Fedor et al. **Analyzing Multi Head Self Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned.**

Michel Paul, Levy Omer, Neubig Graham. **Are Sixteen Heads Really Better than One?**

Olga Kovaleva, Alexey Romanov, Anna Rogers, Anna Rumshisky. **Revealing the Dark Secrets of BERT**

DATASETS

WMT16 EnDe test set
1000 examples of coreference
huggingface/neuralcoref

1000 examples of subject-verb-object triplets
AllenNLP OpenIE

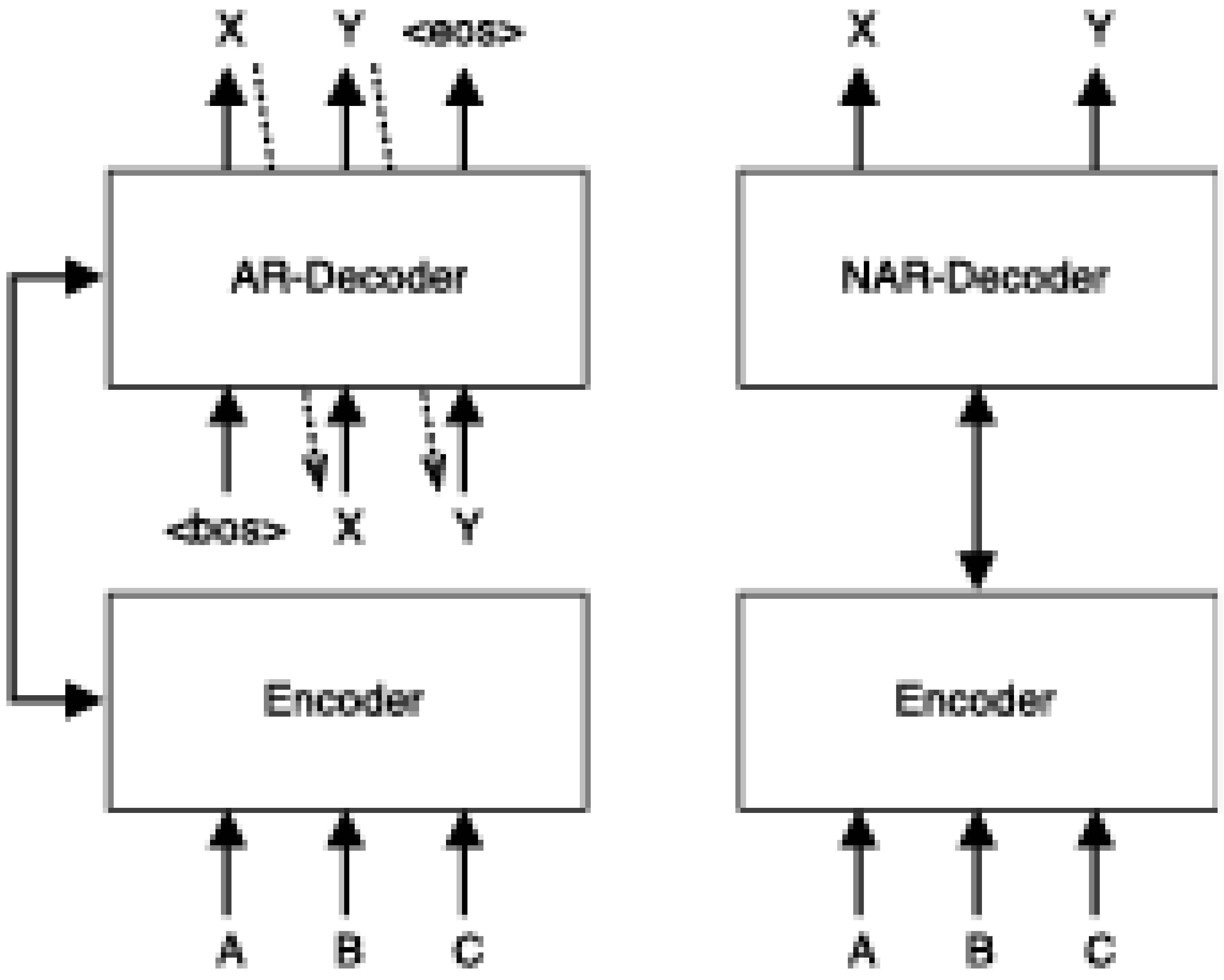
All models trained with
WMT16 EnDe distil dataset from Autoregressive transformer (BLEU ~28)

MODELS

AR vs NAR:

$$p_{\mathcal{AR}}(Y|X;\theta) = \prod_{t=1}^{T+1} p(y_t|y_{0:t-1}, x_{1:T'};\theta),$$

$$p_{\mathcal{NA}}(Y|X;\theta) = p_L(T|x_{1:T'};\theta) \cdot \prod_{t=1}^{T} p(y_t|x_{1:T'};\theta).$$



MODELS

CMLM:

src	we do not believe that we should cher_ r_ y-_ pick .									
1	[mask]	[mask]	nicht	[mask]	[mask]	[mask]	[mask]	[mask]	[mask]	sollten	.
2	[mask]	[mask]	nicht	,	[mask]	wir	[mask]	[mask]	[mask]	sollten	.
3	[mask]	glauben	nicht	,	dass	wir	[mask]	[mask]	[mask]	sollten	.
4	wir	glauben	nicht	,	dass	wir	uns	[mask]	[mask]	sollten	.
5	wir	glauben	nicht	,	dass	wir	uns	aus_	suchen	sollten	.

DiSCO:

src	cher_ r_ y-_ pic_ king is a practice of taking only the most beneficial items .															
0	Kir_	r_	y-_	pic_	pic_	ist	ist	Praxis	Praxis	Praxis	nur	nütz_	nütz_	nütz_	Gegenstände	Gegenstände	nehmen
1	Kir_	sch_	y-_	pic_	king	ist	eine	Praxis	,	,	,	die	lichsten	lichsten	lichsten	Gegenstände	aufzunehmen
2	Kir_	sch_	ern_	pic_	king	ist	eine	Praxis	,	nur	nur	nur	nütz_	nütz_	esten	Gegenstände	aufzunehmen
3	Kir_	sch_	ern_	pic_	king	ist	eine	Praxis	,	nur	die	die	die	haft_	esten	Gegenstände	aufzunehmen
4	Kir_	sch_	ern_	pic_	king	ist	eine	Praxis	,	nur	die	vorteil_	vorteil_	nütz_	esten	Gegenstände	aufzunehmen
5	Kir_	sch_	ern_	pic_	king	ist	eine	Praxis	,	nur	die	vorteil_	haft_	haft_	esten	Gegenstände	aufzunehmen
6	Kir_	sch_	ern_	pic_	king	ist	eine	Praxis	,	nur	die	vorteil_	haft_	wert_	esten	Gegenstände	aufzunehmen

MODELS

AR vs NAR:

Model	T	E-D	WMT17 EN→ZH		WMT17 ZH→EN			WMT14 EN→FR	
			BLEU	S_1	BLEU	S_1	S_{max}	BLEU	S_1
CMLM	4	6-6	33.58	**3.5×**	22.56	**3.8×**		40.21	**3.8×**
CMLM	10	6-6	34.24	1.5×	23.76	1.7×		40.55	1.7×
DisCo		6-6	34.63	2.5×	23.83	2.6×		40.60	3.6×
AR		6-6	**35.06**	1.0×	24.19	1.0×		41.98	1.0×
Dist. Teacher		6-6	35.01	–	24.65	–		42.03	–

EXPERIMENTS

Methodology for AR models (Analyzing MultiHead SelfAttention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, Voita et al)

1) Mark up each of the transformer heads for positional and syntactic function

2) Sequentially remove the heads with the pruning mechanism

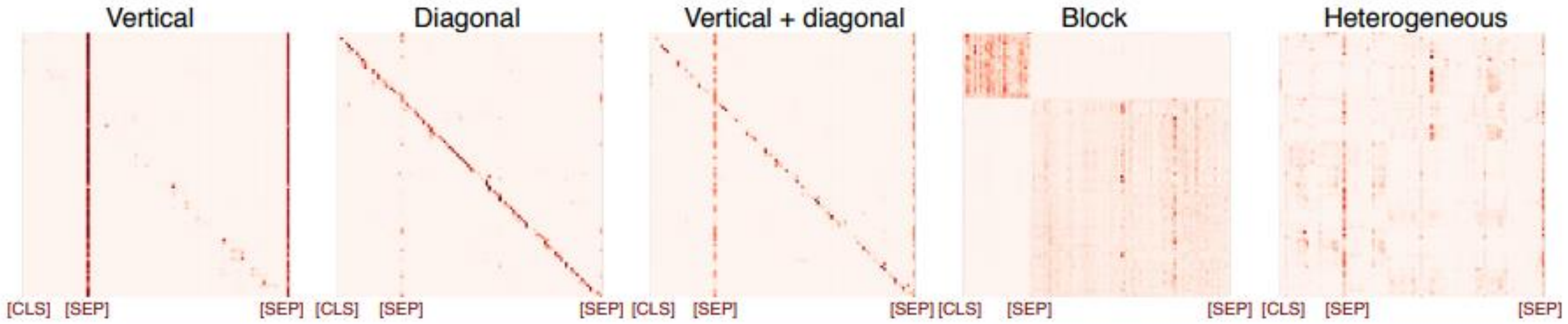
3) Compare the ranks of importance of the heads with the original markings

Same for NAR

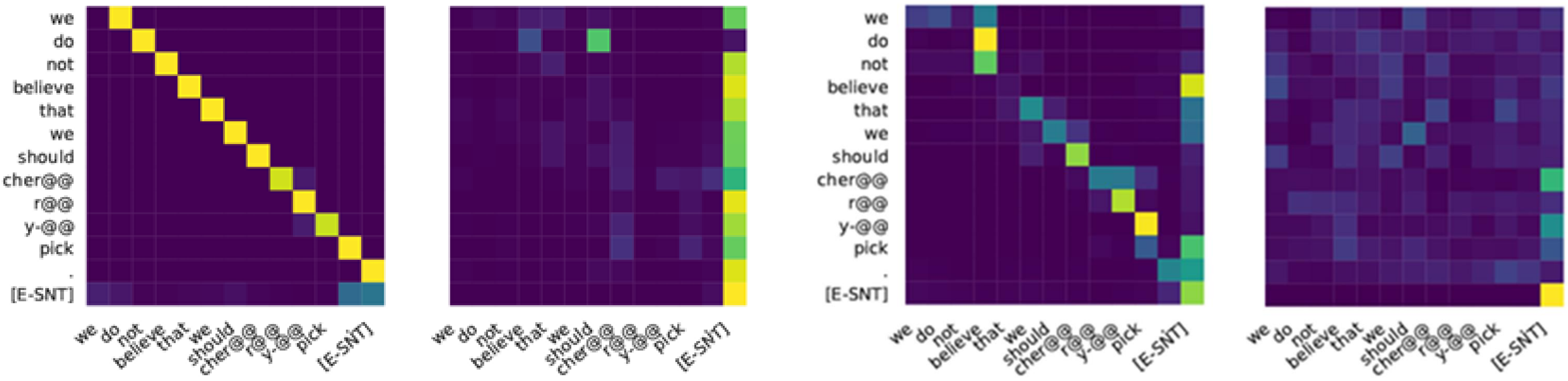
EXPERIMENTS

positional function

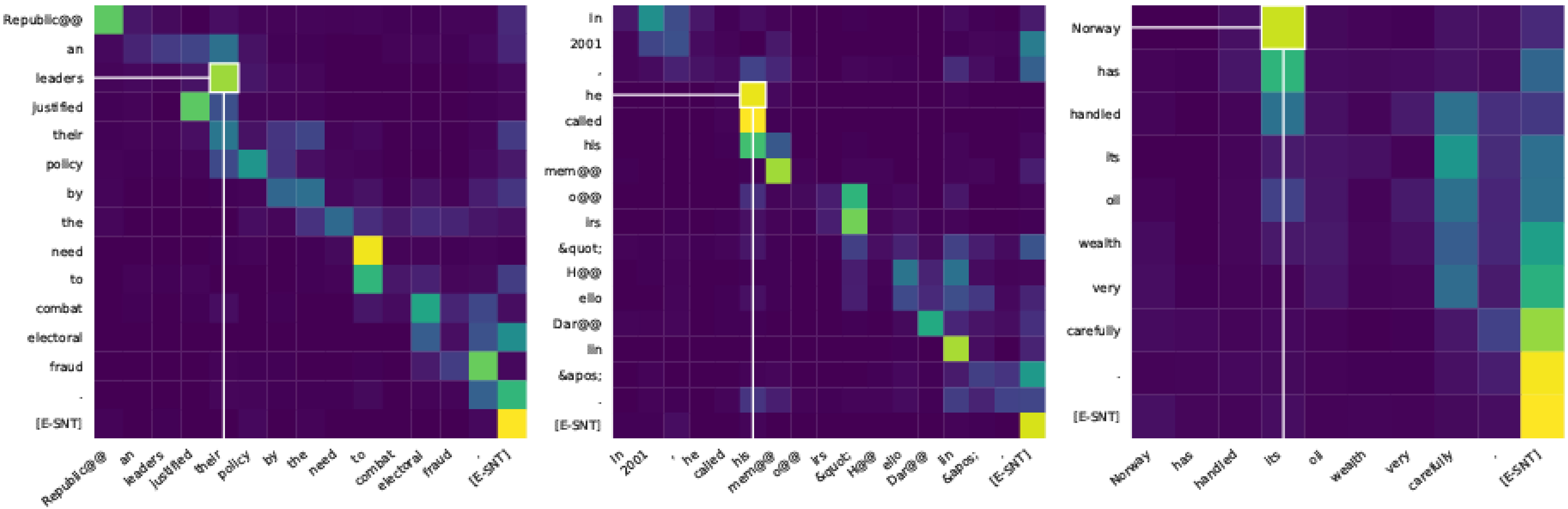
- Diagonal pattern - the highest weight is located on the main diagonal of the matrix, or on a +1/1 shift from the main diagonal
- Vertical pattern - the highest weight is located on one of the columns of the matrix
- Diagonal vertical pattern - combine of diagonal and vertical patterns
- Other pattern - all matrices that are not part of the first three patterns



EXPERIMENTS

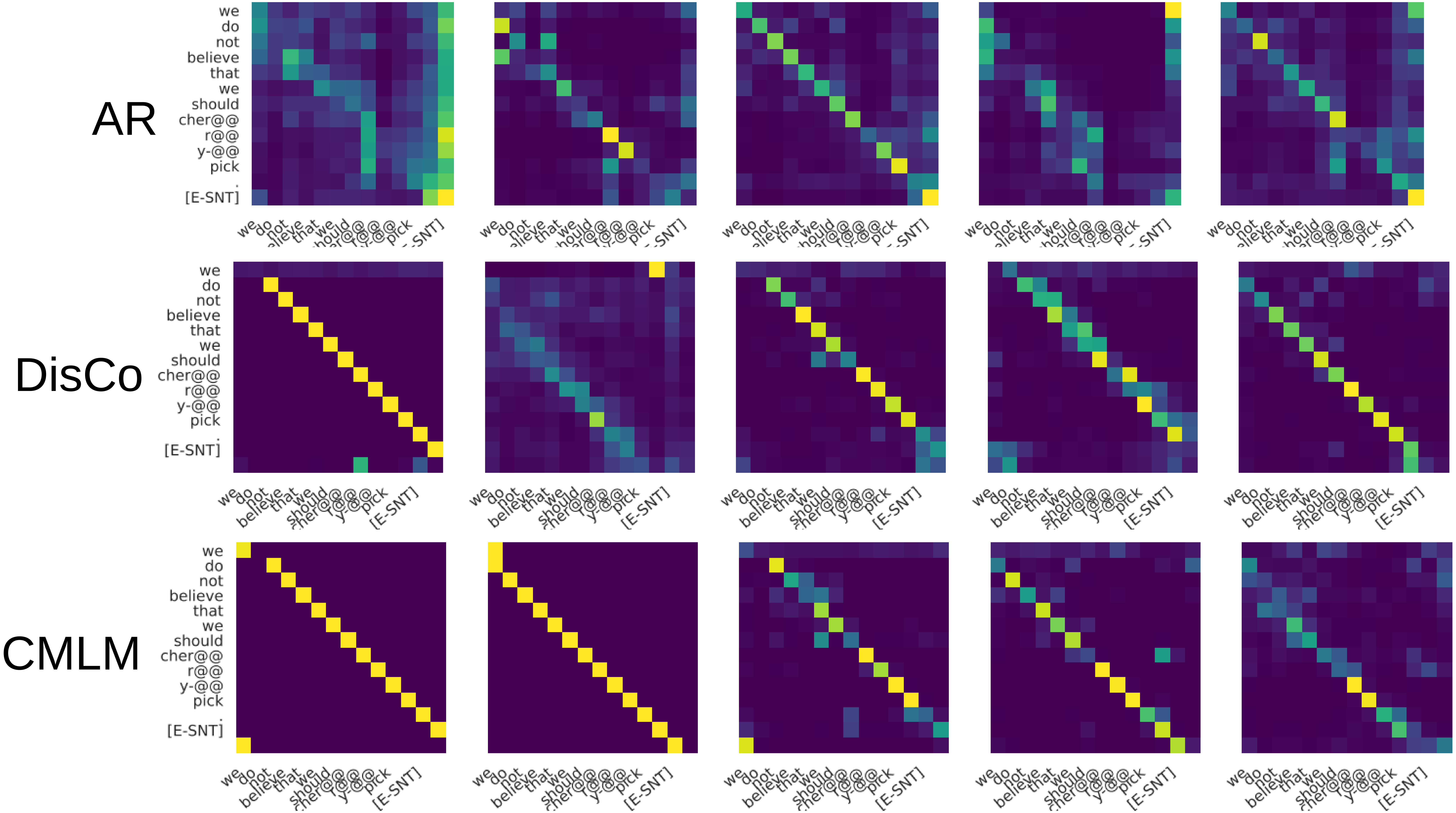


EXPERIMENTS



EXPERIMENTS

Pruning results



RESULTS

Модель	Точность	Количество голов		
		=0	<0.1	>0.1
Transformer E12-D1	0.06	84	0	12
CMLM E12-D1	0.17	58	26	11
DisCo E12-D1	0.19	53	32	11
Transformer E6-D6	0.04	39	0	9
CMLM E6-D6	0.16	35	9	4
DisCo E6-D6	0.16	31	13	4

Точность решения задачи кореференции

Модель	Точность	Количество голов		
		0	<0.1	>0.1
Transformer E12-D1	0.42	1	81	12
CMLM E12-D1	0.4	11	74	11
DisCo E12-D1	0.4	9	74	12
Transformer E6-D6	0.36	0	42	6
CMLM E6-D6	0.35	9	34	5
DisCo E6-D6	0.36	8	32	8

Точность нахождения SVO

RESULTS

	Позиционная			Синтаксическая	
Модель	diag	vertical	other	Кореференция	O-V-S
Transformer 12-1	6	2	2	0	0
CMLM 12-1	8	1	1	4	4
DiSco 12-1	7	0	3	3	4

CONCLUSIONS

We studied patterns of encoders of non-autoregressive models and proved that they are mostly similar to patterns from autoregressive models.

We have also found that non-autoregressive models are better at capturing some syntactic features of the language.

As a result, we refuted the original hypothesis that the quality of non-autoregressive models suffers due to model does not find similar language properties as autoregressive models

Pletenev Sergey

alex010rey@gmail.com

github.com/A1exRey/ReflectionOfNAR