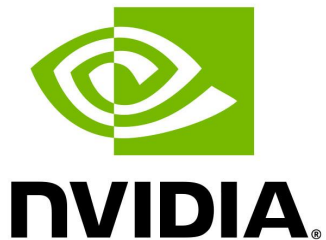


Language models inspired by human language understanding

Rob van der Goot

IT UNIVERSITY OF COPENHAGEN



CARLSBERG
FONDET



Bla bla bla ...

populære



Goeie

Hallo

Hi

你好

Hello

Me

- PhD: Normalization and parsing of social media data
- Postdoc – now: at ITU, investigating:
 - Robustness in NLP
 - NLP for low/mid-resource languages
 - Multilingual NLP
 - Multi-task learning

Me

- PhD: Normalization and parsing of social media data
- Postdoc – now: at ITU, investigating:
 - Robustness in NLP
 - NLP for low/mid-resource languages
 - Multilingual NLP
 - Multi-task learning
- Danish:
 - Multi-domain normalization,
 - Multi-domain NER
 - POS tagging of TikTok data
 - Dependency parsing of TikTok data
 - Infection classification
 - Morphemes -> LM's
 - SnakModel
 - Cultural evaluation of LM's

Today

- What is a language model
- SnakModel
- MorSed
- LMLM

What is a language model?

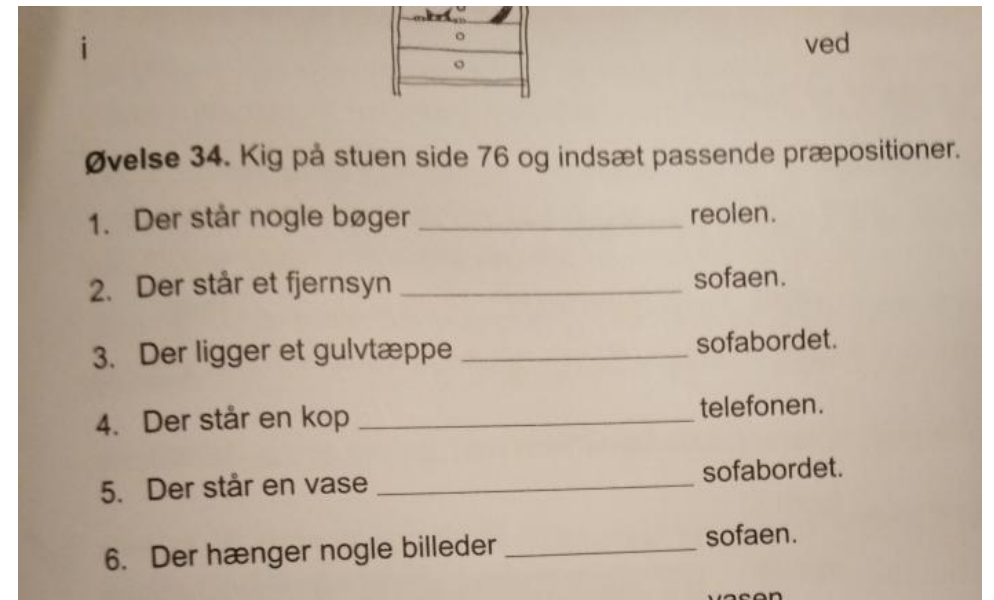
- A model that predicts the likelihood of a sequence of text
- Nowadays: usually transformer-based

What is a language model?

- A model that predicts the likelihood of a sequence of text
 - Nowadays: usually transformer-based and large
 - A text is a sequence of (sub)words, we transform the task into next word prediction
-
- https://books.google.com/ngrams/graph?content=Rob+is+a+*&year_start=1800&year_end=2022&corpus=en&smoothing=3

Predicting words

- Might happen in our brain all the ...
- Is used for language learning
- One way to do it well is to try and grasp the meaning of the surrounding context

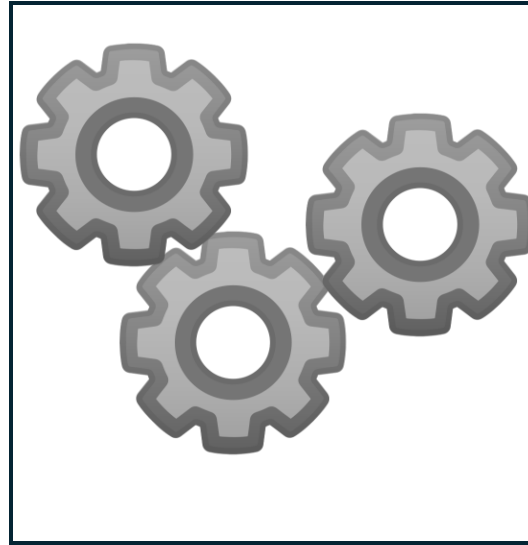


How do we train (L)LMs

Data



Language model



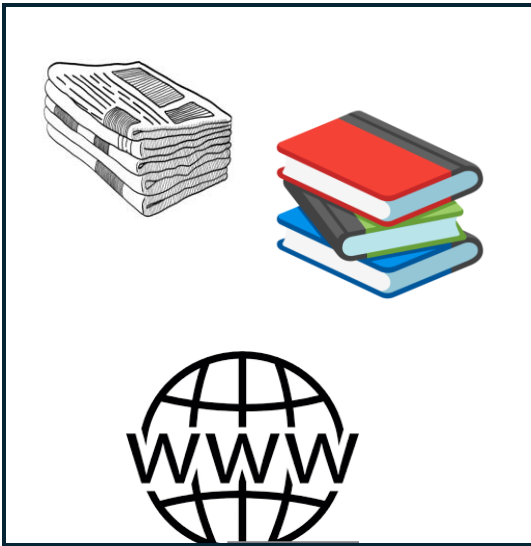
Hyggeligt
Hyggeligt vejr
Hyggeligt vejr i
Hyggeligt vejr i dag
Hyggeligt vejr i dag .

Data I found

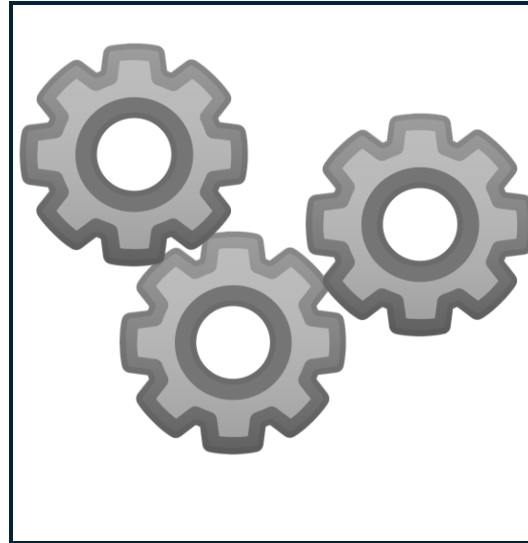
DATASET	ORIGINAL		+ FASTTEXT	
	Docs	Words	Docs	Words
Bookshop	8.65M	208M	6.80M	187M
CC-100	344M	7.82B	256M	7.16B
CulturaX	449M	14.8B	333M	13.7B
DaNewsroom	24.2M	391M	11.3M	369M
Dawiki	1.70M	62.4M	1.20M	57.3M
FTSpeech	2.03M	43.3M	1.69M	40.9M
Gigaword	62.0M	1.02B	39.3M	898M
OpenSubtitles	30.2M	207M	19.6M	156M
Reddit	4.50M	73.9M	2.37M	64.0M
Twitter	1.69M	21.9M	406K	6.61M
TOTAL	927M	24.6B	672M	22.6B
+ DEDUPLICATION			350M	13.6B

How do we train (L)LMs

Data



Language model



Hyggeligt
Hyggeligt vejr
Hyggeligt vejr i
Hyggeligt vejr i dag
Hyggeligt vejr i dag .

Input:

Hvor kommer julemanden fra?

Output:

Hvor kommer havfrue fra?

How do we train (L)LMs

Forklar forskellen mellem fotosyntese og respiration.

Fotosyntese og **respiration** er to modsatrettede biologiske processer, som tilsammen holder planter og mange andre organismer i live.

Fotosyntese foregår i planters grønkorn og bruger **lysenergi**, **vand** og **kuldioxid** til at danne **glukose** og **ilt**. Processen lagrer energi i sukker.

Respiration foregår i cellernes mitokondrier og bruger **glukose** og **ilt** til at frigive den energi, som organismen skal bruge til vækst, bevægelse og vedligeholdelse. Under respiration dannes **kuldioxid** og **vand** som affaldsprodukter.

Kort sagt: Fotosyntese **opbygger** energirige stoffer ved hjælp af sollys, mens respiration **nedbryder** dem for at frigive energi.

hvordan kan jeg skrive et resumé af en tekst

Et kort resumé gengiver **hovedpointerne** i en tekst uden detaljer, eksempler eller vurderinger.

Et typisk resumé beskriver: – hvad teksten handler om, – hvilke centrale ideer eller begivenheder der præsenteres, – og hvordan de hænger sammen.

Et eksempel på et kort resumé kunne være: *“Teksten forklarer, hvordan genbrug reducerer affald og sparer ressourcer. Den beskriver også, hvilke materialer der kan genanvendes, og hvorfor det er vigtigt for miljøet.”*

SnakModel

- Open Danish language model (7B)
- Retrained from an Englocentric model
- Our part of training data -> public
- Improvements in standard NLP tasks and culture

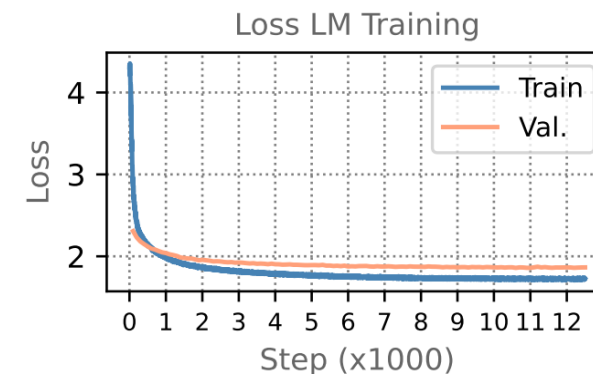
Why?

- 2 people with a 1 month gap
- 4 GPU's
- RQ: (how much) do language models become more culturally aware through retraining?

Training timeline (2024)

DATASET	ORIGINAL		+ FASTTEXT	
	Docs	Words	Docs	Words
Bookshop	8.65M	208M	6.80M	187M
CC-100	344M	7.82B	256M	7.16B
CulturaX	449M	14.8B	333M	13.7B
DaNewsroom	24.2M	391M	11.3M	369M
Dawiki	1.70M	62.4M	1.20M	57.3M
FTSpeech	2.03M	43.3M	1.69M	40.9M
Gigaword	62.0M	1.02B	39.3M	898M
OpenSubtitles	30.2M	207M	19.6M	156M
Reddit	4.50M	73.9M	2.37M	64.0M
Twitter	1.69M	21.9M	406K	6.61M
TOTAL	927M	24.6B	672M	22.6B
+ DEDUPLICATION			350M	13.6B

Parameter	Value
<i>Data Split</i>	
Training data	96.9%
Validation data	3.1%
<i>Training Configuration</i>	
Vocabulary size	32,000
Context length	4,096
Training steps	12,500
Warmup steps	1,250
Number of epochs	1
Global batch size	512
<i>Optimizer Parameters (AdamW)</i>	
$\beta_1; \beta_2$	0.9; 0.95
ϵ	10^{-5}
Peak learning rate	1.5×10^{-5}
Minimum learning rate	5×10^{-8}
Weight decay	0.1
Gradient clipping	1.0



Results

↓ MODEL	AVG.
LLAMA2-7B _{base}	46.48
LLAMA2-7B _{chat}	49.50
LLAMA2-7B _{base} + INST _{da}	49.35
LLAMA2-7B _{chat} + INST _{da}	51.39
VIKING-7B	37.82
SNAKMODEL-7B _{base}	46.45
SNAKMODEL-7B _{instruct}	56.63 ^{↑10.15}
MISTRAL-7B-v0.1	53.76
MUNIN-7B-ALPHA	57.48
MUNIN-7B-v0.1DEV0	61.28 ^{↑7.52}

Cultural evaluation

- Ask Danes to probe language models for cultural awareness

 **SnakModel**

Stil et spørgsmål eller giv en opgave om dansk kultur (fx. viden om samfund, normer og skikke) til de tre virtuelle assistenter nedenfor. Du får tre svar og skal give hver af dem 👍/👎.
([Samlede instruktioner](#))

Input

Hvad skal vi lave?

Svar 1

...

Svar 2

...

Svar 3

...

Det er et projekt af IT-Universitetets NLPnorth gruppe.

(c) Prompt Interface.

 **SnakModel**

Stil et spørgsmål eller giv en opgave om dansk kultur (fx. viden om samfund, normer og skikke) til de tre virtuelle assistenter nedenfor. Du får tre svar og skal give hver af dem 👍/👎.
([Samlede instruktioner](#))

Antal inputs: 1

Tak for dit første input! Prøv endelig med en ny forespørgsel. ✎

Input

Hvordan kommer man lettest rundt i København?

Svar 1

De fleste af de offentlige transportsystemer i København giver en effektiv måde at komme rundt på, og der er en udbredt cykelkultur.

Kommentarer (valgfrit)...

Svar 2

At have en cykel

ikke muligt for alle

Svar 3

Hvis man ønsker at rejse rundt i København på den bedste måde, bør man sørge for at have en billet på metroen.

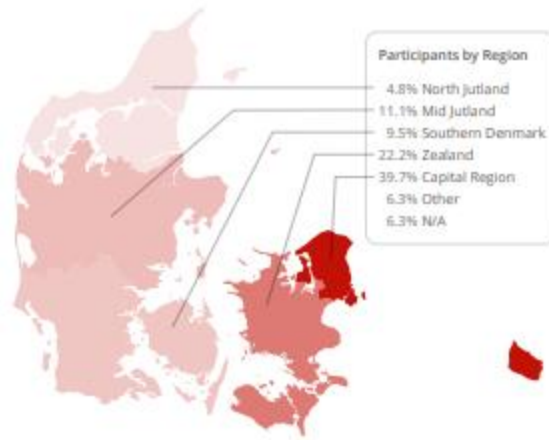
Kommentarer (valgfrit)...

Giv hvert svar 👍/👎 før du klikker fortsæt. **Fortsæt >**

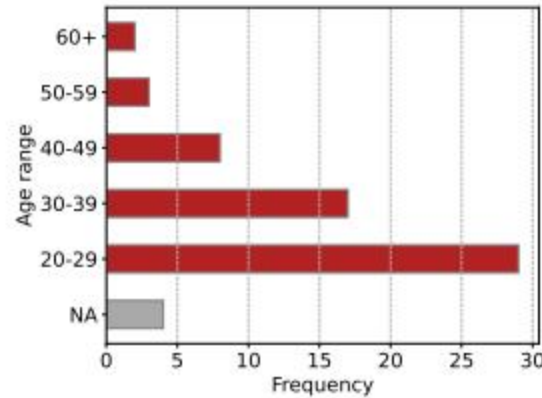
Det er et projekt af IT-Universitetets NLPnorth gruppe.

(d) Evaluation Interface.

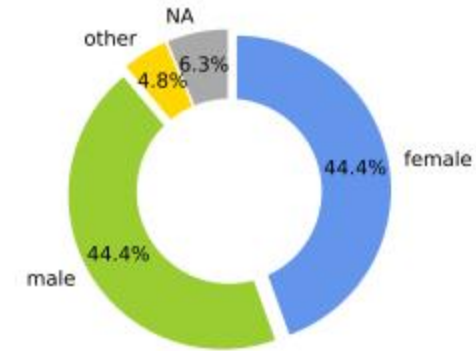
Cultural evaluation



(a) Participants by Region.



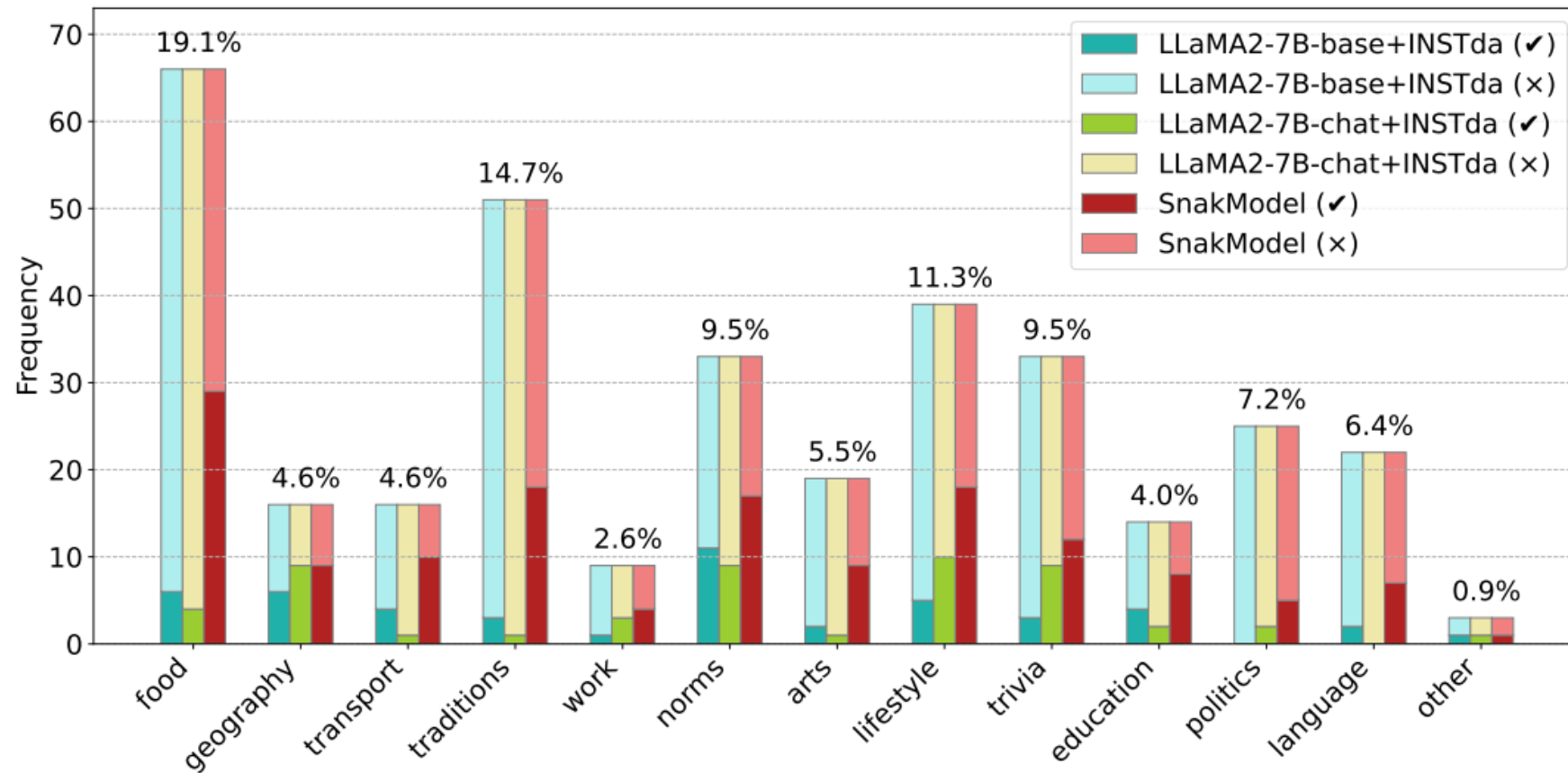
(b) Participants by Age Range.



(c) Participants by Gender Identity.

Figure 1: **Demographic Statistics for Our 63 Study Participants**, who were asked to optionally provide the region, where one grew up (Fig. 1a), age range in decades (Fig. 1b), and gender identity (Fig. 1c). 94% of respondents opted to provide this information.

Cultural evaluation



SnakModel

Maximilian Müller-Eberstein, Mike Zhang, Elisa Bassignana, Rob van der Goot



<https://huggingface.co/datasets/NLPnorth/dakultur>

SnakModel

Hvad er egentlig
danskernes nationalret?



Syltet pandekage



Stegt flæsk med
persillesovs



Snakmodel evaluation

MODEL	LANGUAGE						CULTURE		
	LA	NER	SA	AS	CSR	QA	PE	CT	DK
LLAMA2-7B _{base}	33.4	23.7	61.5	65.5	29.8	63.5	38.6	57.7	—
+ INST _{da}	36.1	28.5	62.9	66.4	29.0	64.4	49.1	58.5	13.9
LLAMA2-7B _{chat}	47.4	24.6	66.2	66.3	32.2	61.3	46.7	55.2	—
+ INST _{da}	43.4	29.7	65.9	65.8	31.0	62.5	57.3	55.6	15.0
SNKMODEL	52.9	29.8	66.7	66.6	29.5	64.7	71.1	71.9	42.4

SnakModel

- Lets give it a try

MorSed

- Language models don't use characters or words as units, but subwords.
- What are subwords?

What are subwords?

- We train tokenizers on text, they attempt to identify frequent sequences of characters -> subwords
- Different algorithms: BPE/Unigram/Wordpiece
- Motivation: Characters do not have meaning, and lead to long sequences, words are too scarce, and have overlapping meaning.

What are subwords?

- Lets take a look at a 2023 SOTA model (Llama2):

thinking -> thinking

hovercraft -> hover_craft

tænke -> t_æ_n_ke

luftpude fartøj -> lu_ftp_ude_f_art_ø_j

What are subwords?

- Lets take a look at a 2023 SOTA model (Llama2):

thinking -> thinking

hovercraft -> hover_craft

tænke -> t_æ_n_ke

luftpude fartøj -> lu_ftp_u de_f_art_ø_j

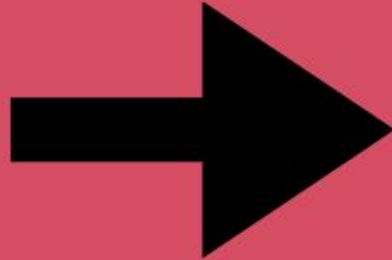
Llama 4 (2025):

t_æ_n_ke -> t_æn_ke

Why morphemes?



Smørrebrød

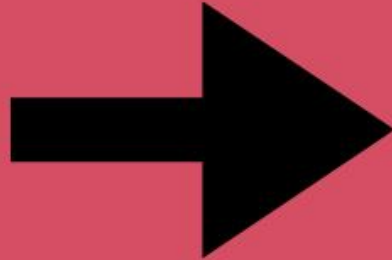


Smør-re-brød

Why morphemes?



Smørrebrød



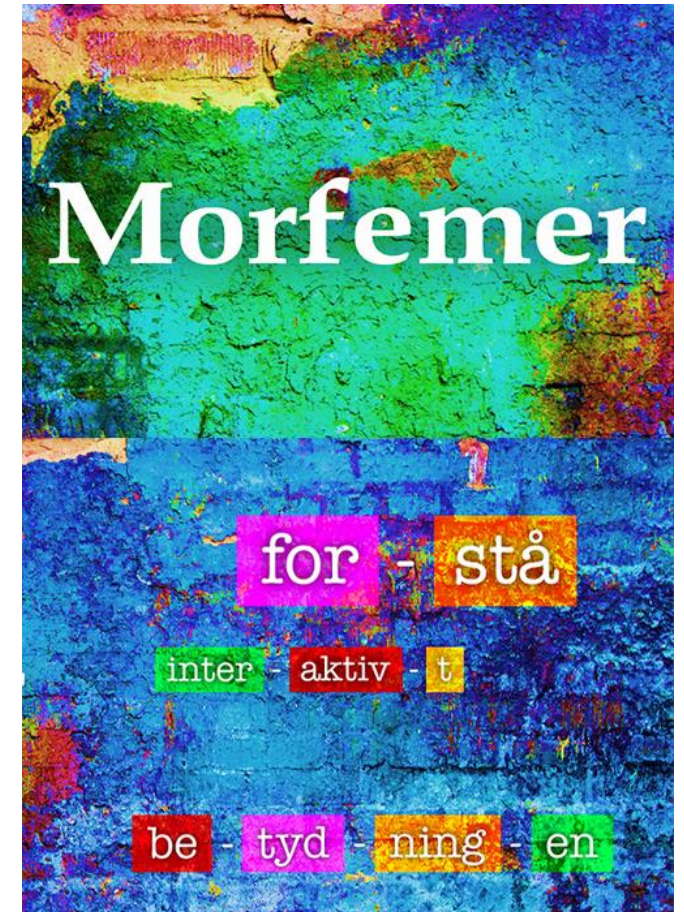
Smør-re-brød

Lets give it a try!

MorSed benchmark

- 800 words with morphological annotation

Word	Category	Split
Kranie	Root Morpheme	Kranie [Root]
Landstræner	Linking Morpheme	Land[Root] s[Link] træn[Root] er[Suff]
Lånte	Inflection	Lån[Root] te[Infl]
Bibringe	Prefix	Bi[Pref] bringe[Root]
Skoletaske	Compound	Skole [Root] taske [Root]
Venlig	Suffix	Ven[Root] lig[Suff]

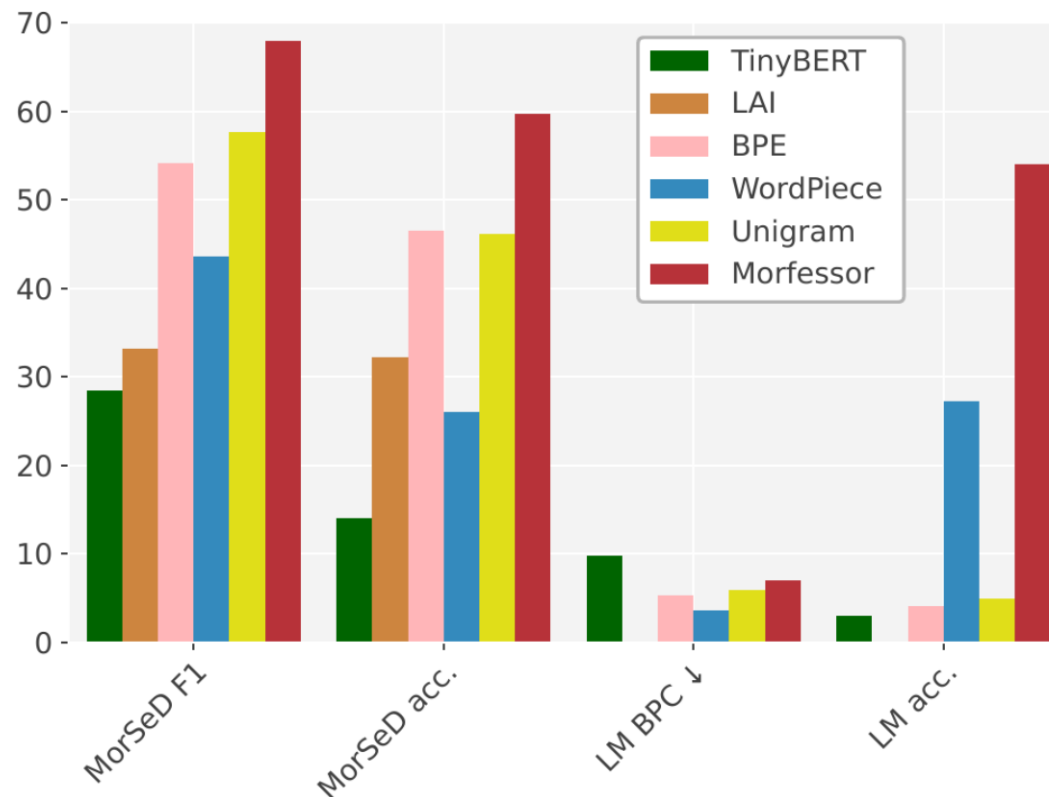


MorSed benchmark

- Research question: Is segmenting LM inputs on morpheme boundaries beneficial for performance?

MorSed benchmark

- Research question: Is segmenting LM inputs on morpheme boundaries beneficial for performance?

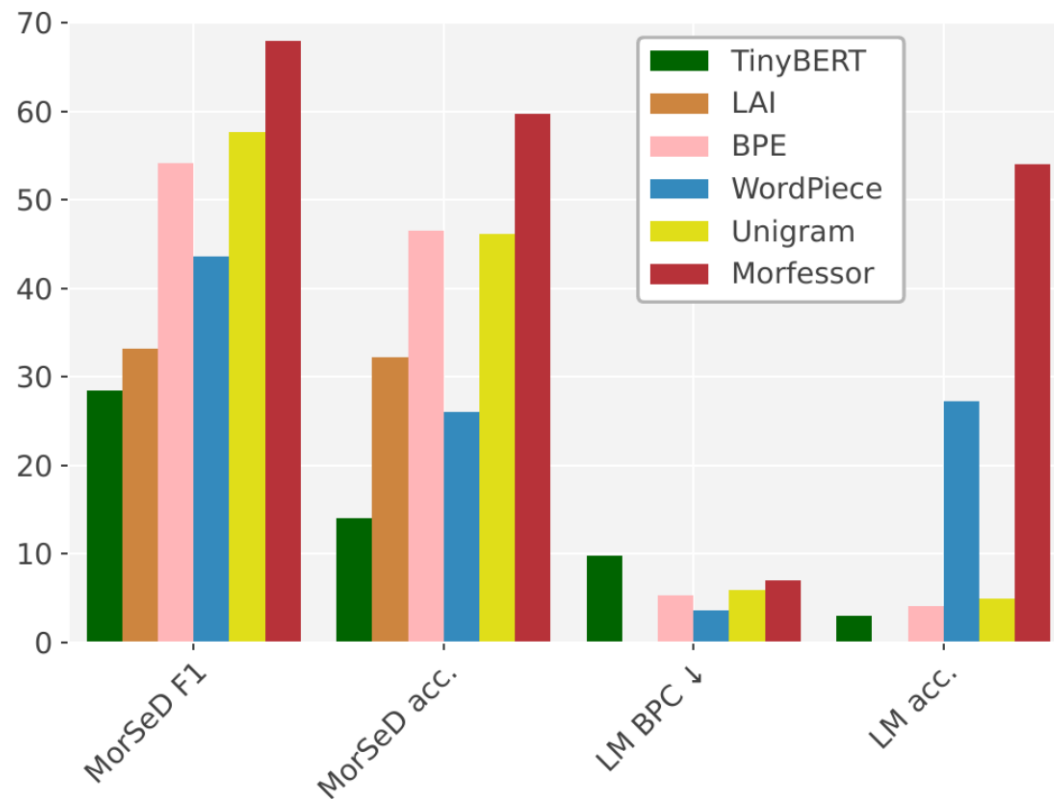


Model	Conditional Sentence			Dialog with Emotion			Descriptive Language			Historical Narratives			Average Score
	F	N	G	F	N	G	F	N	G	F	N	G	
Std. BPE. Cerebras	2,4	1,6	3,6	4,0	4,2	4,2	2,6	3,4	3,4	4,0	3,2	4,0	3,4
Da. BPE. Cerebras	2,2	2,0	2,2	2	3,0	3,8	3,0	3,2	3,4	2,2	1,6	2,2	2,6
Da. Morph. Cerebras	3,2	3,0	3,8	2,4	2,2	2,8	3,6	3,0	4,0	2,8	2,2	3,4	3,0
Da. Mixed. Cerebras	4,6	2,0	4,6	5,0	2,0	4,6	4,0	4,0	4,6	4,0	2,0	4,2	3,8

Table 9: Results on Fluency (F), Naturalness (N) and Grammar (G) in the human evaluation experiment.

MorSed benchmark

- Research question: Is segmenting LM inputs on morpheme boundaries beneficial for performance?
- Answer: Probably



Model	Conditional Sentence			Dialog with Emotion			Descriptive Language			Historical Narratives			Average Score
	F	N	G	F	N	G	F	N	G	F	N	G	
Std. BPE. Cerebras	2,4	1,6	3,6	4,0	4,2	4,2	2,6	3,4	3,4	4,0	3,2	4,0	3,4
Da. BPE. Cerebras	2,2	2,0	2,2	2	3,0	3,8	3,0	3,2	3,4	2,2	1,6	2,2	2,6
Da. Morph. Cerebras	3,2	3,0	3,8	2,4	2,2	2,8	3,6	3,0	4,0	2,8	2,2	3,4	3,0
Da. Mixed. Cerebras	4,6	2,0	4,6	5,0	2,0	4,6	4,0	4,0	4,6	4,0	2,0	4,2	3,8

Table 9: Results on Fluency (F), Naturalness (N) and Grammar (G) in the human evaluation experiment.

Probably is not enough!

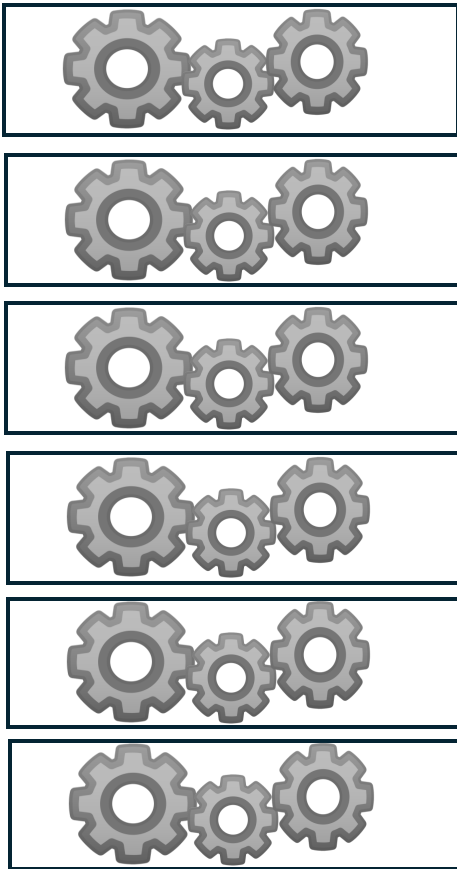
LMLM: Linguistically Motivated Language Models

- Language is mostly compositional
- Language models might learn this



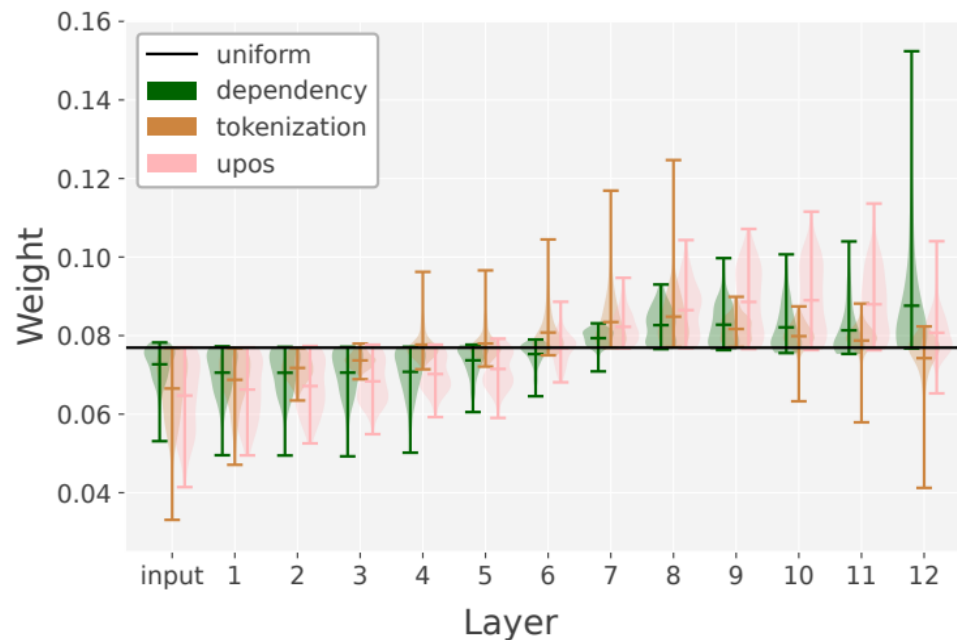
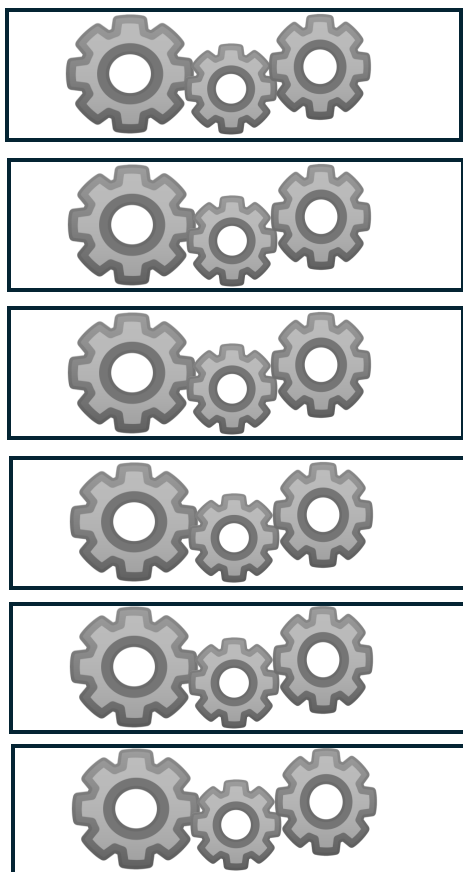
The plan

Language model



The plan

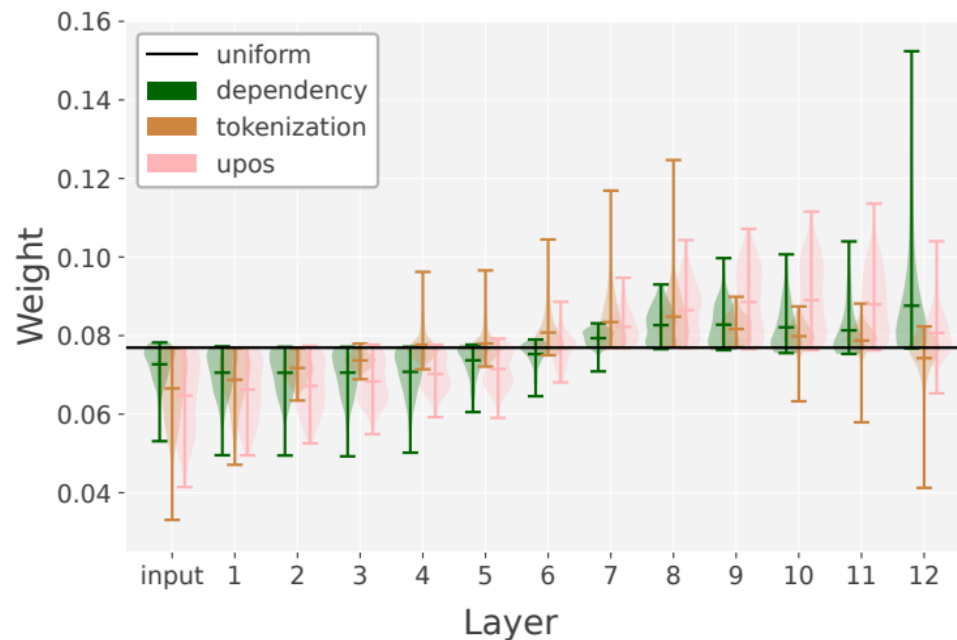
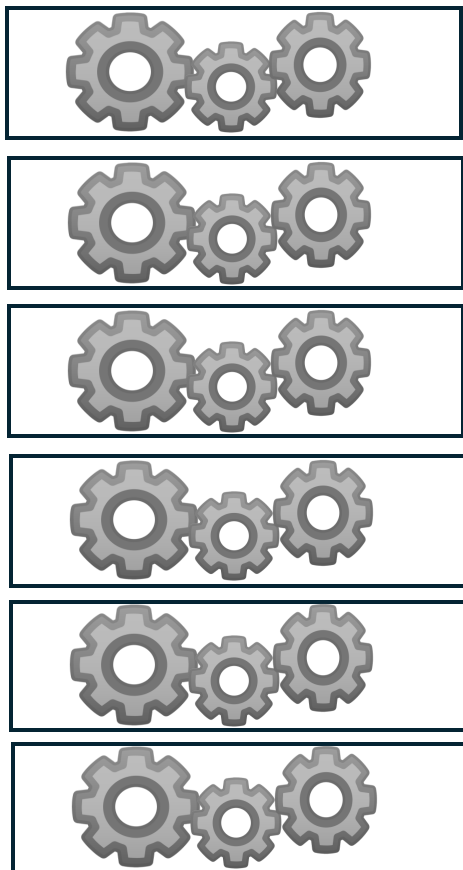
Language model



- Language models process language in a somewhat logical order.

The plan

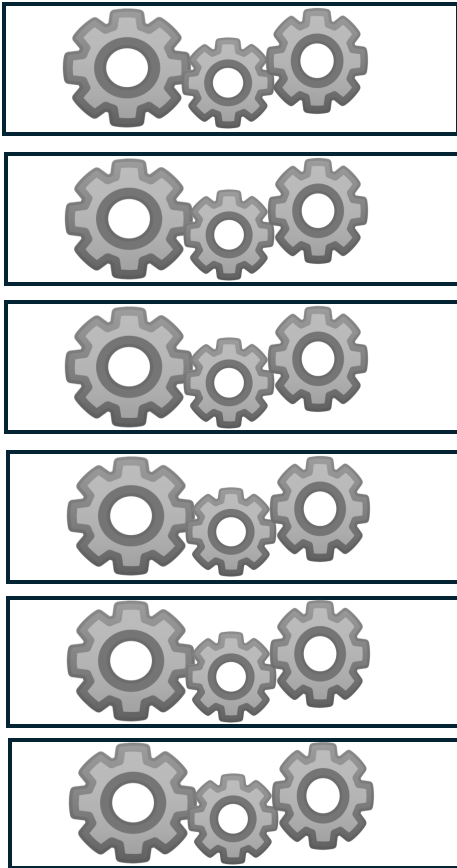
Language model



- Language models process language in a somewhat logical order.
- What happens if we enforce this?

The plan

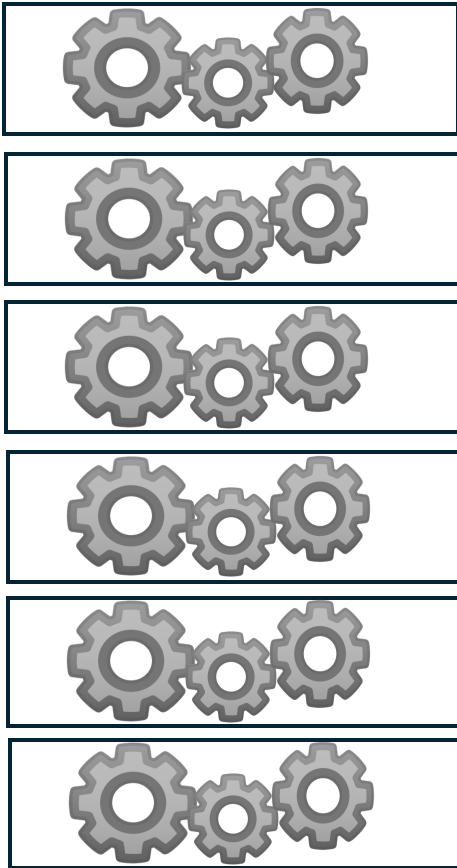
Language model



- Morphemes
- Words
- Chunks
- Syntax
- Knowledge,
reasoning etc

The plan

Language model



- Morphemes

- Words

- Chunks

- Syntax

- Knowledge,
reasoning etc

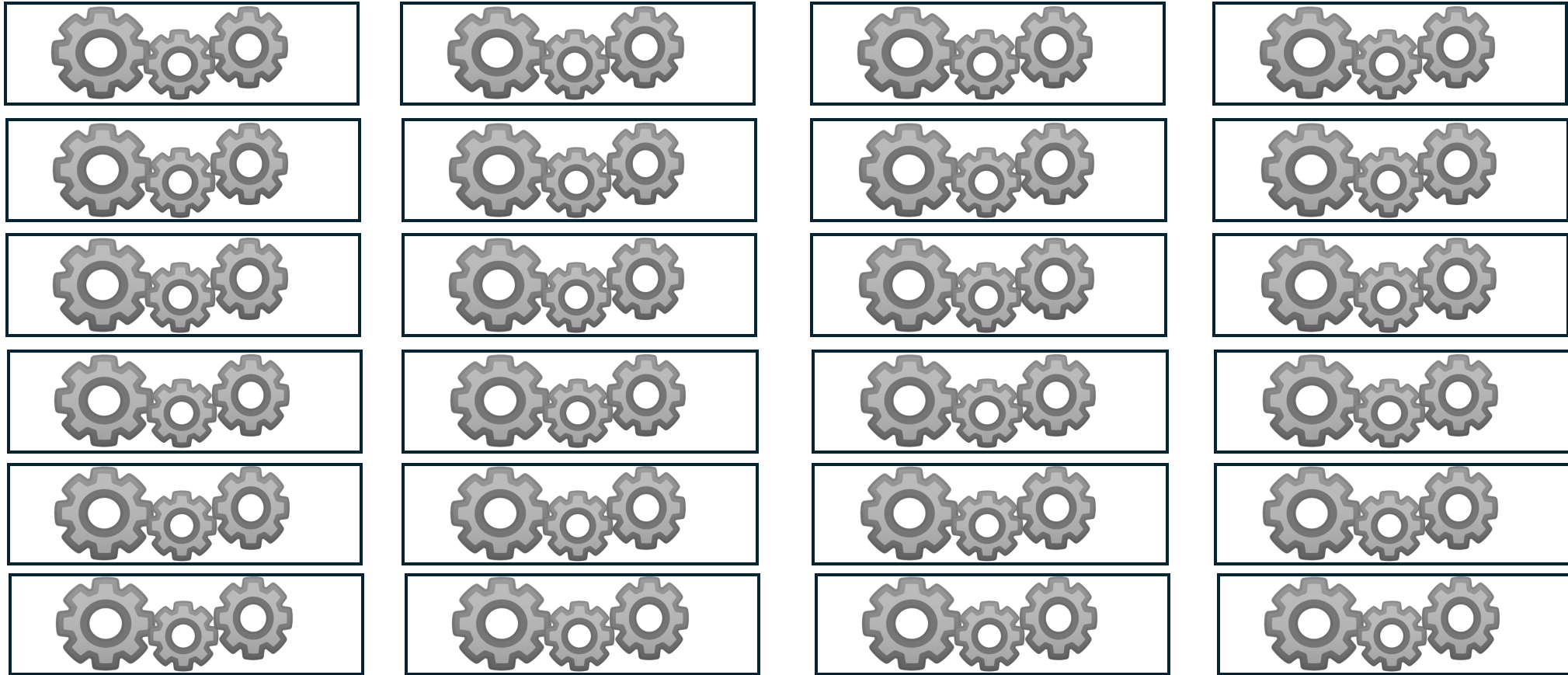
x

- Segmentation
- Labeling
- Level specific tasks

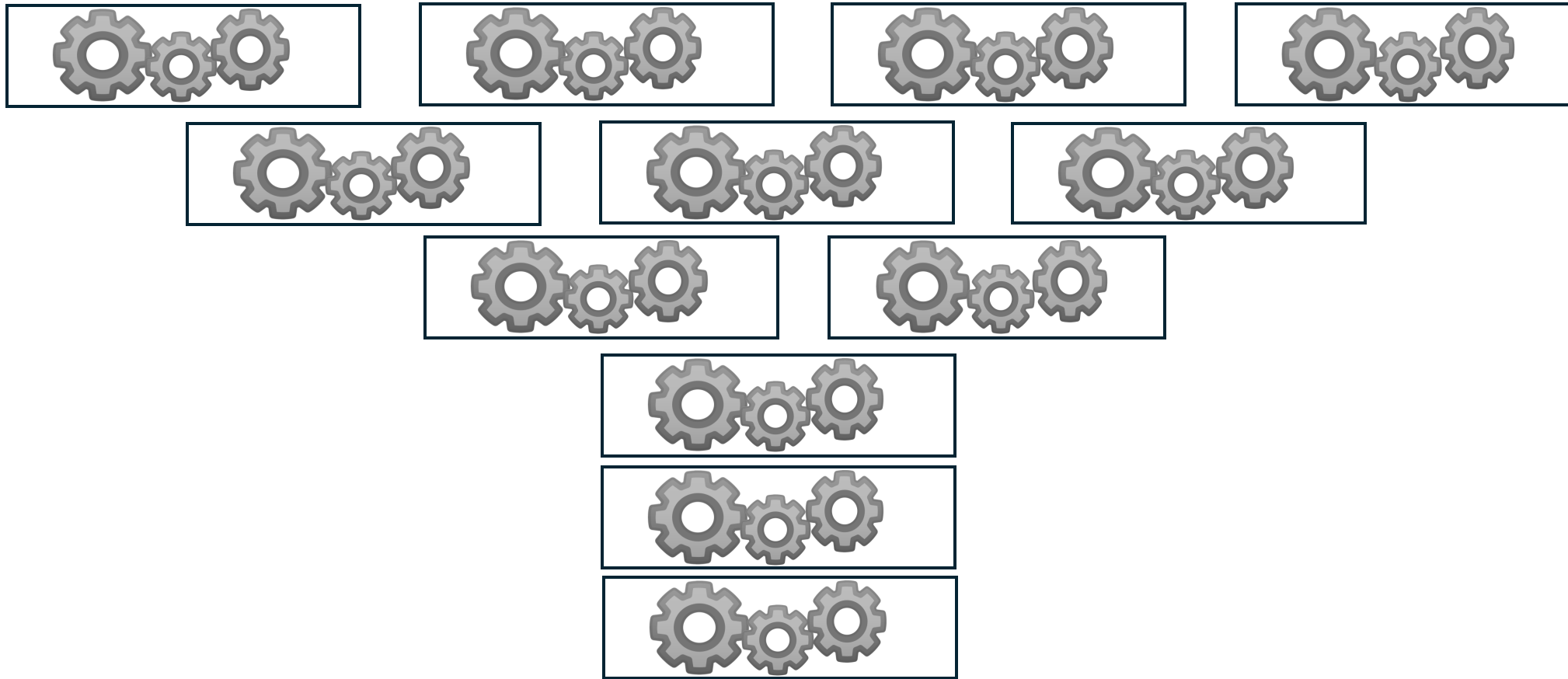
Besides training objective

- Architectures do not have to be identical
- Inputs can be hierarchical:
- Characters – morphs – words - chunks

Hierarchical inputs



Hierarchical inputs



How do we define success?

- LM loss: How good can it predict the next (sub)word
- LM performance: How good can it answer to queries
- Efficiency: Training/Using
- Language understanding: shortcuts, compositionality, grammar

Will we be successful?

How do we define success

- LM loss: previously 100% of weights dedicated, now ~60%?
- LM performance: same
- Efficiency
- Language understanding

Compositionality benchmarking

Novel Combination of Familiar Primitives and Grammatical Roles

Subject → Object (common noun)	A hedgehog ate the cake.	The baby liked the hedgehog .
--------------------------------	---------------------------------	--------------------------------------

Novel Combination Modified Phrases and Grammatical Roles

Object modification → Subject modification	Noah ate the cake on the plate .	The cake on the table burned.
--	---	--------------------------------------

Deeper Recursion

Depth generalization: PP modifiers	Ava saw the ball in the bottle on the table .	Ava saw the ball in the bottle on the table on the floor .
------------------------------------	--	---

Verb Argument Structure Alternation

Passive → Active	The book was squeezed .	The girl squeezed the strawberry.
------------------	--------------------------------	--

Thousand thanks!

- Anette Jensen
- Emil Allerslev Schledermann, Mikkel Wildner Kildeberg, Nicolaj Larsen
- Mike Zhang, Elisa Bassignana, Maximilian Müller-Eberstein
- Barbara Plank

Thousand thanks!

IT UNIVERSITY OF COPENHAGEN



CARLSBERG
FONDET

