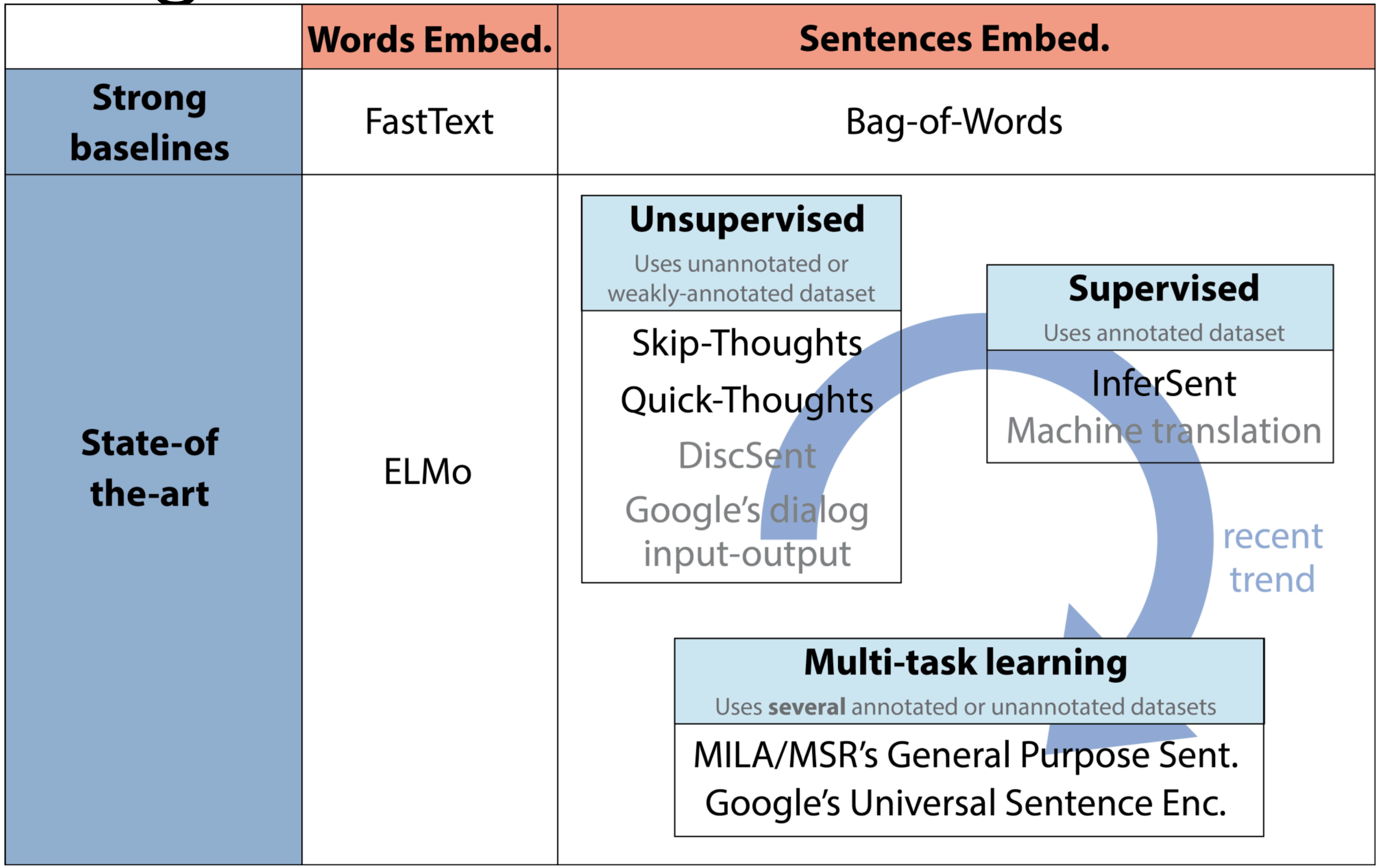
What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, Marco Baroni

Facebook AI Research Université Le Mans (LIUM)

ACL 2018

The quest for universal sentence embeddings



*Courtesy: Thomas Wolf blogpost, Hugging Face

Now-famous Ray Mooney's quote

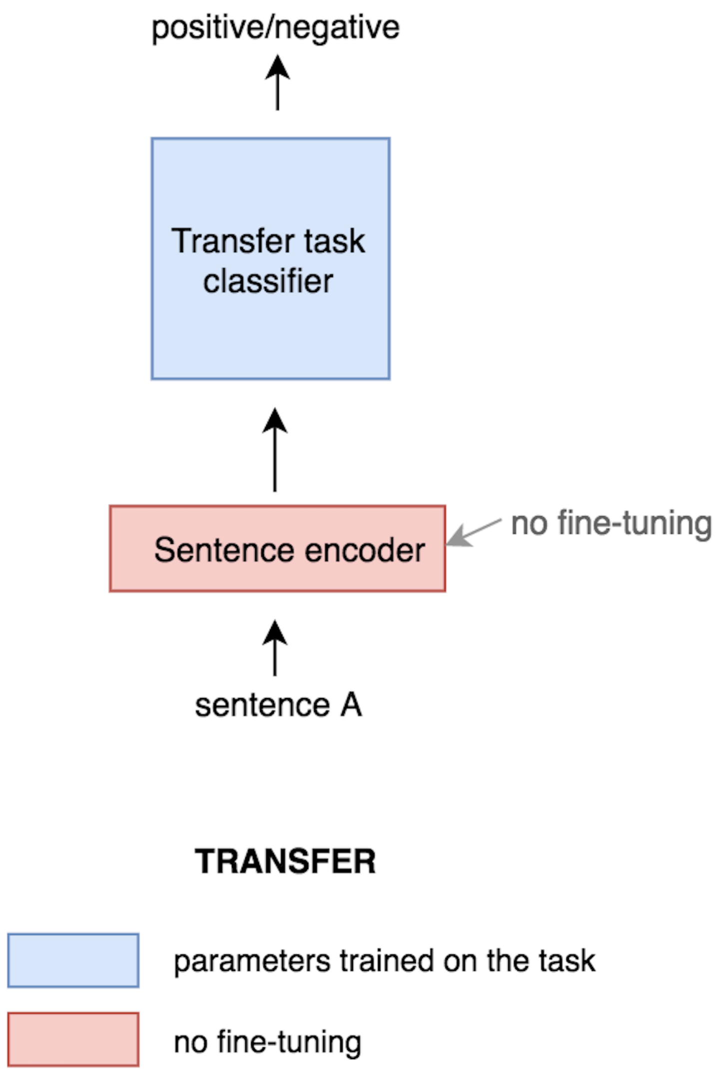
You can't cram the meaning of a single $&!#* sentence into a single $!#&* vector!

Professor Raymond J. Mooney

- While not capturing meaning, we might still be able to build useful transferable sentence features
- But what can we actually cram into these vectors?

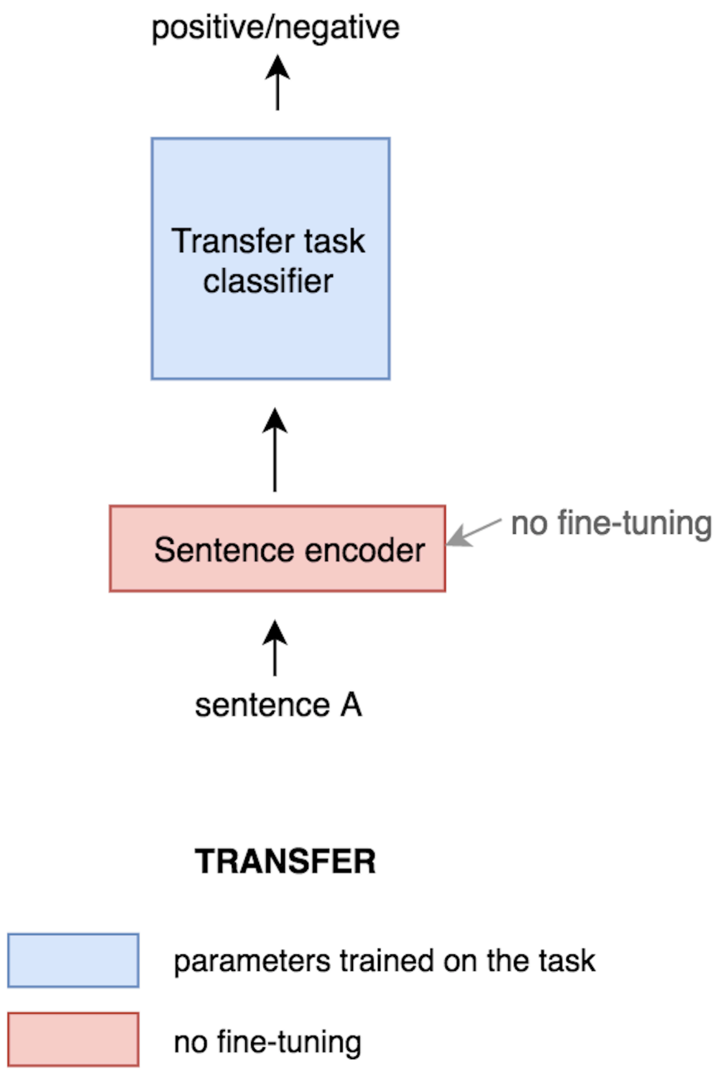
The evaluation of universal sentence embeddings

- Transfer learning on many other tasks
- Learn a classifier on top of pretrained sentence embeddings for transfer tasks
- SentEval downstream tasks:
  - Sentiment/topic classification
  - Natural Language Inference
  - Semantic Textual Similarity



The evaluation of universal sentence embeddings

- Downstream tasks are complex
- Hard to infer what information the embeddings really capture
- “**Probing tasks**” to the rescue!
  - designed for inference
  - evaluate simple isolated properties



Probing tasks and downstream tasks

Probing tasks are simpler and focused on a single property!

Subject Number probing task

Sentence: The hobbits waited patiently .

Label: Plural (NNS)

Natural Language Inference downstream task

Premise: A lot of people walking outside a row of shops with an older man with his hands in his pocket is closer to the camera .

Hypothesis: A lot of dogs barking outside a row of shops with a cat teasing them .

Label: contradiction

Our contributions

An extensive analysis of sentence embeddings using probing tasks

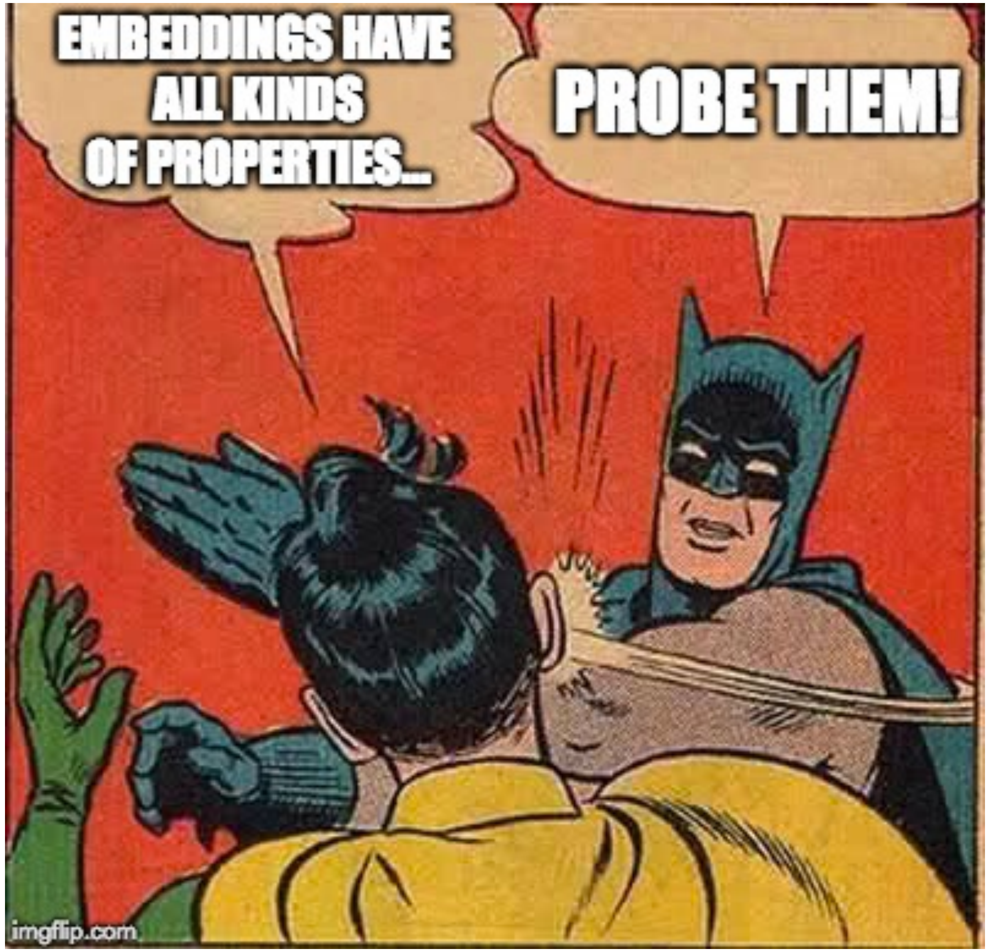
- We vary the architecture of the encoder (3) and the training task (7)

- We open-source 10 horse-free classification probing tasks.

- Each task being designed to probe a single linguistic property

Shi et al. (EMNLP 2016) – Does string-based neural MT learn source syntax?
Adi et al. (ICLR 2017) – Fine-grained analysis of sentence embeddings using auxiliary prediction tasks

Probing tasks: understanding sentence embeddings content



Probing tasks

What they have in common:

- Artificially-created datasets all framed as classification

- ... but based on natural sentences extracted from the TBC (5-to-28 words)

- 100k training set, 10k valid, 10k test, with balanced classes

- Carefully removed obvious biases (words highly predictive of a class, etc)

Probing tasks

Grouped in three categories:

- **Surface information**
- **Syntactic information**
- **Semantic information**

Probing tasks (1/10) – Sentence Length

input	MLP classifier	output
She had not come all this way to let one stupid wagon turn all of that hard work into a waste !	→	21-25

- Goal: Predict the length range of the input sentence (6 bins)
- Question: Do embeddings preserve information about sentence length?

Surface information

Probing tasks (2/10) – Word Content

Helen took a pen from her purse and **wrote** something on her cocktail napkin.

input

MLP classifier →

wrote

output

- <u>Goal</u>: 1000 output words. Which one (only one) belongs to the sentence?

- <u>Question</u>: Do embeddings preserve information about words?

Adi et al. (ICLR 2017) – Fine-grained analysis of sentence embeddings using auxiliary prediction tasks

Surface information

Probing tasks (3/10) – Top Constituents

MLP classifier

input		output
Slowly he lowered his head toward mine.	→	ADVP_NP_VP_.
The anger in his voice surprised even himself .	→	NP_VP_.

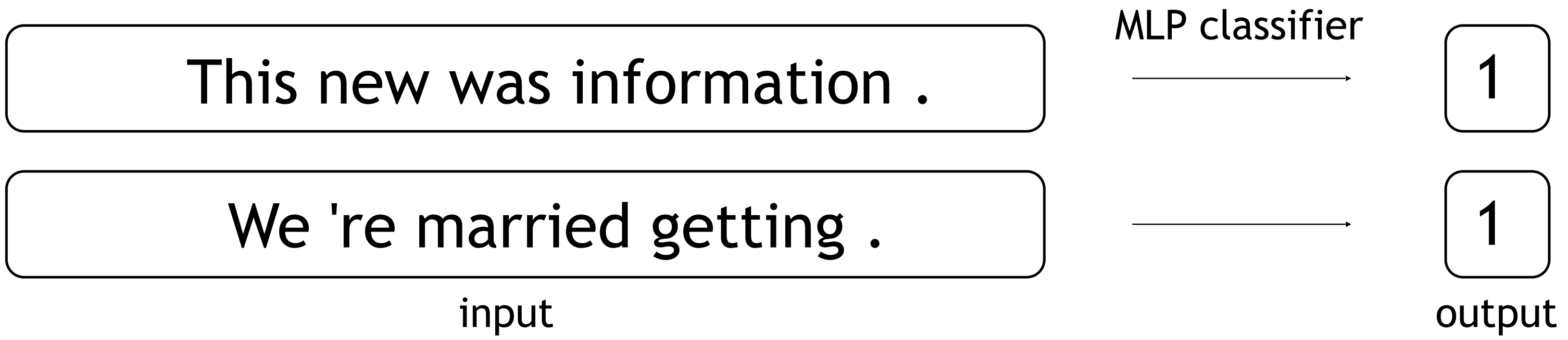
- Goal: Predict top-constituents of parse-tree (20 classes)

- Note: 19 most common top-constituent sequences + 1 category for others

- Question: Can we extract grammatical information from the embeddings?

Shi et al. (EMNLP 2016) – Does string-based neural MT learn source syntax?

Probing tasks (4/10) – Bigram Shift



- Goal: Predict whether a bigram has been shifted or not.

- Question: Are embeddings sensible to word order?

Syntactic information

Probing tasks – 5 more

- 5/10: **Tree Depth** (depth of the parse tree)
- 6/10: **Tense prediction** (main clause tense, past or present)
- 7-8/10: **Object/Subject Number** (singular or plural)
- 9/10: **Semantic Odd Man Out** (noun/verb replaced by one with same POS)

Probing tasks (10/10) – Coordination Inversion

They might be only memories, but I can still feel each one

MLP classifier →

O

I can still feel each one, but they might be only memories.

→

I

input

output

- <u>Goal</u>: Sentences made of two coordinate clauses: inverted (I) or not (O)?
- <u>Note</u>: human evaluation: 85%
- <u>Question</u>: Can extract sentence-model information?

Experiments and results

Experiments

We analyse almost 30 encoders trained in different ways:

- Our baselines:
  - Human evaluation, Length (1-dim vector)
  - NB-uni and NB-uni/bi with TF-IDF
  - CBOW (average of word embeddings)
- Our 3 architectures:
  - Three encoders: BiLSTM-last/max, and Gated ConvNet
- Our 7 training tasks:
  - Auto-encoding, Seq2Tree, SkipThought, NLI
  - Seq2seq NMT without attention En-Fr, En-De, En-Fi

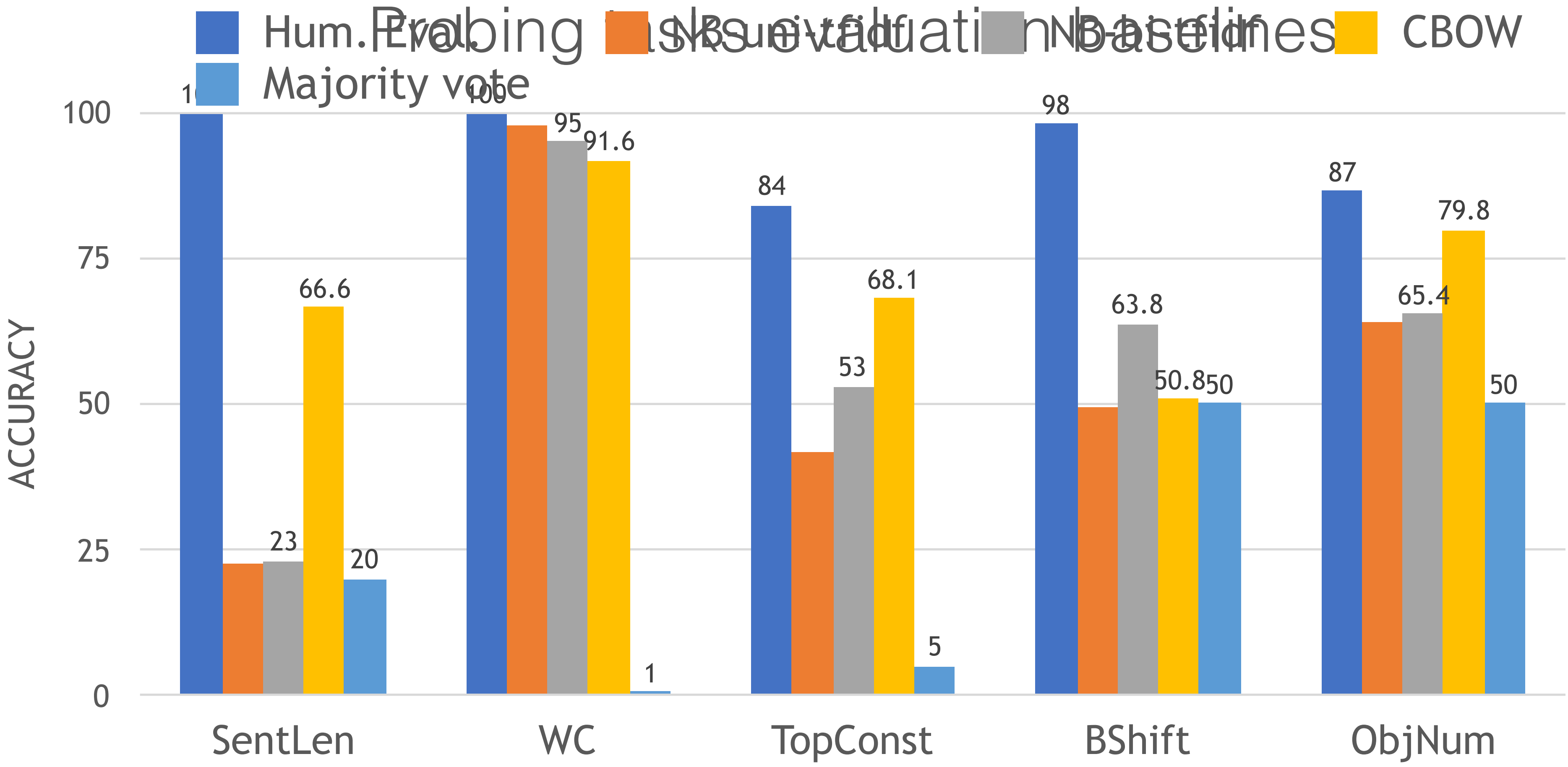
Experiments – training tasks

task	source	target
AutoEncoder	I myself was out on an island in the Swedish archipelago , at Sandhamn .	I myself was out on an island in the Swedish archipelago , at Sand@ ham@ n .
NMT En-Fr	I myself was out on an island in the Swedish archipelago , at Sandhamn .	Je me trouvais ce jour là sur une île de l' archipel suédois , à Sand@ ham@ n .
NMT En-De	We really need to up our particular contribution in that regard .	Wir müssen wirklich unsere spezielle Hilfs@ leistung in dieser Hinsicht aufstocken .
NMT En-Fi	It is too early to see one system as a universal panacea and dismiss another .	Nyt on liian aikaista nostaa yksi järjestelmä jal@ usta@ lle ja antaa jollekin toiselle huono arvo@ sana .
SkipThought	the old sami was gone , and he was a different person now .	the new sami didn 't mind standing barefoot in dirty white , sans ra@ y-@ bans and without beautiful women following his every move .
Seq2Tree	Dikoya is a village in Sri Lanka .	(ROOT (S (NP NNP $)_{NP}$ (VP VBZ (NP (NP DT NN $)_{NP}$ (PP IN (NP NNP NNP $)_{NP}$ $)_{PP}$ $)_{NP}$ $)_{VP}$. $)_{S}$ $)_{ROOT}$

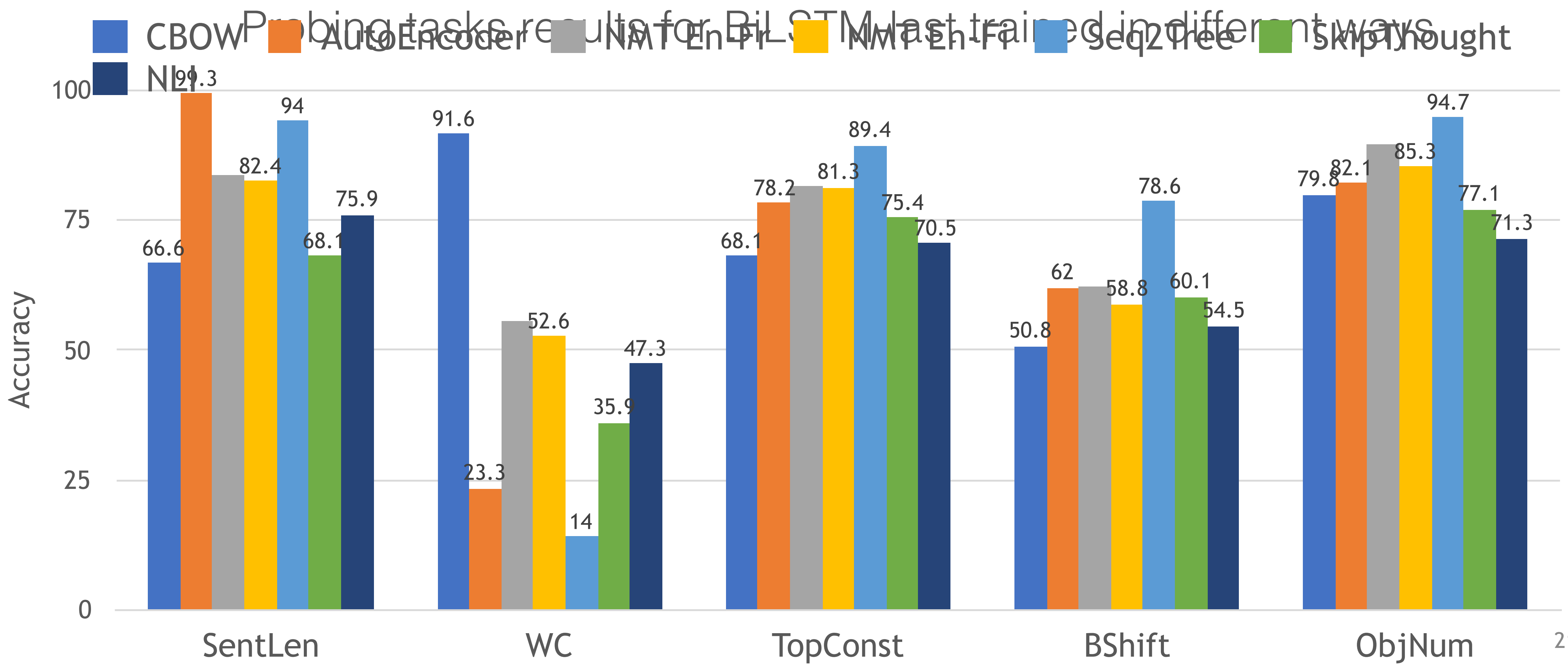
Source and target examples for seq2seq training tasks

Sutskever et al. (NIPS 2014) – Sequence to sequence learning with neural networks
Kiros et al. (NIPS 2015) – SkipThought vectors
Vinyals et al. (NIPS 2015) – Grammar as a Foreign Language

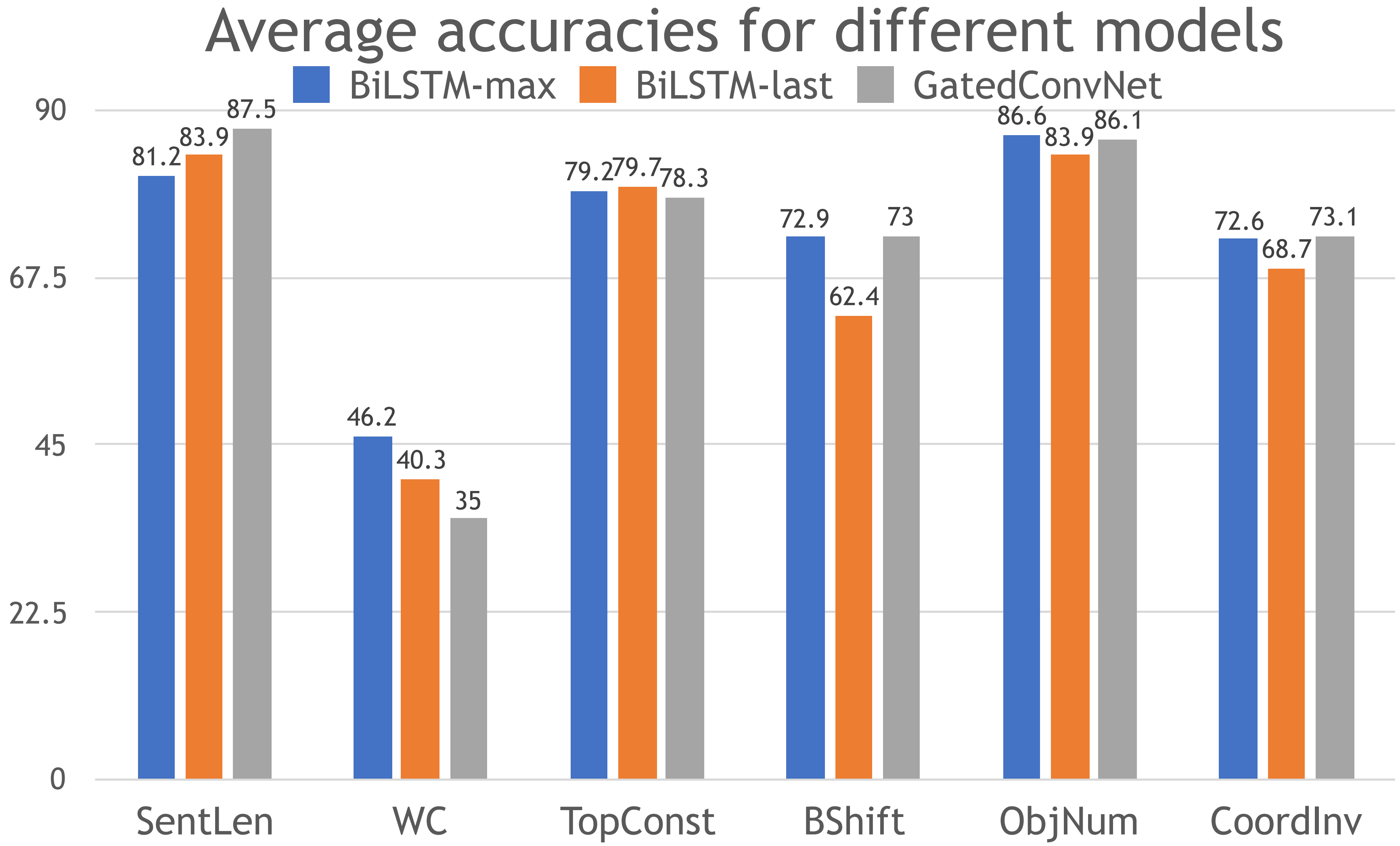
Baselines and sanity checks



Impact of training tasks



Impact of model architecture

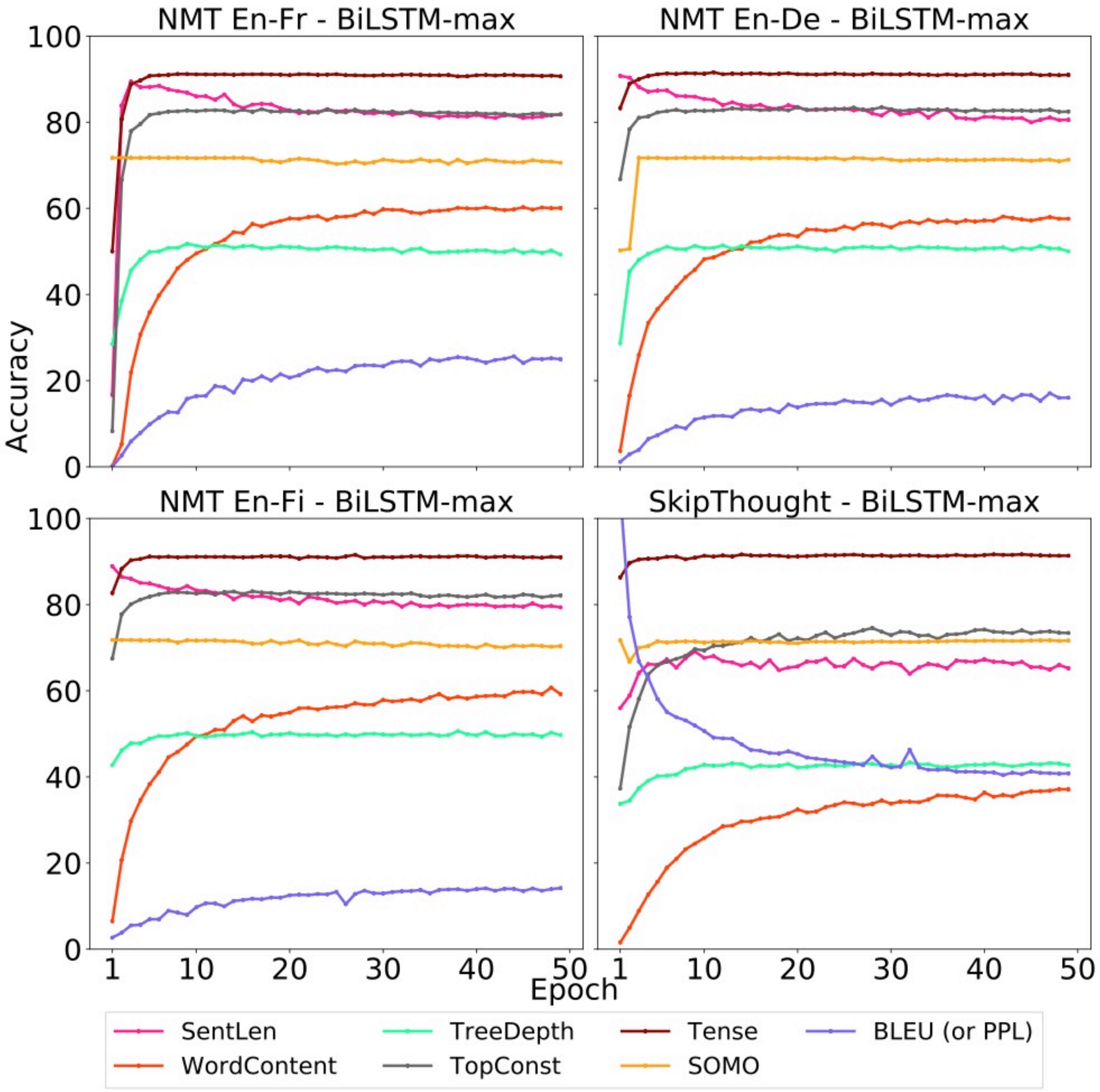


Evolution during training

- Evaluation on probing tasks at each epoch of training

- What do embeddings encode along training?

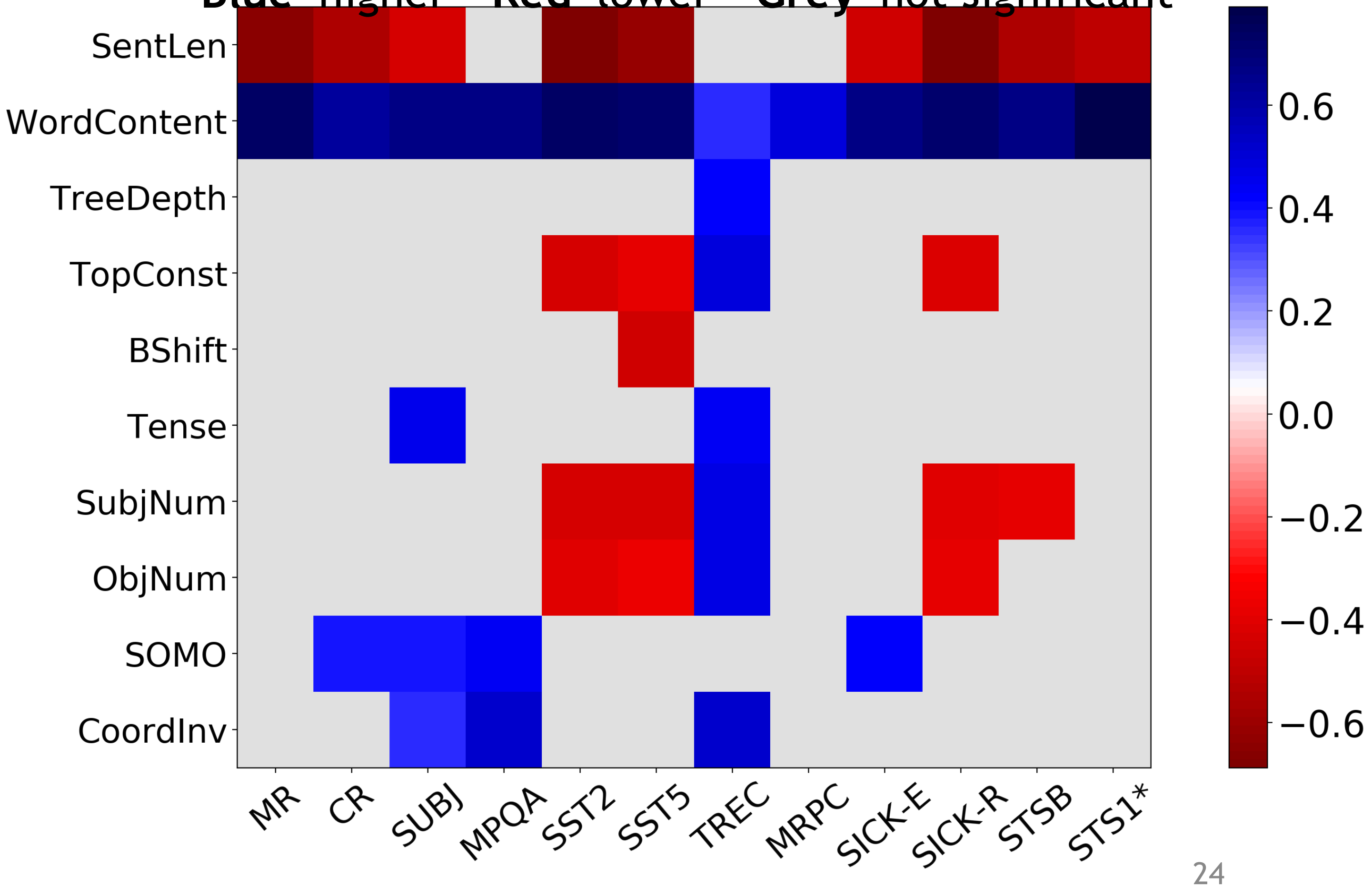
- NMT: Most increase and converge rapidly (only SentLen decreases). WC correlated with BLEU.



Correlation with downstream tasks

- Strong correlation between WC and downstream tasks
- Word-level information important for downstream tasks (classification, NLI, STS)
- If WC good predictor -> maybe current downstream tasks are not the right ones?

Correlation between probing and downstream tasks
Blue=higher - **Red**=lower - **Grey**=not significant



Take-home messages and future work

- Sentence embeddings need not be good on probing tasks
- Probing tasks are simply meant to understand what linguistic features are encoded and to designed to compare encoders.
- Future work
  - Understanding the impact of multi-task learning
  - Studying the impact of language model pretraining (ELMO)
  - Study other encoders (Transformer, RNNG)

Thank you!

Thank you!

- Publicly available in SentEval
- Automatically generated datasets (generalize to other languages)
- Natural sentences from Toronto Book Corpus
- Used Stanford parser for grammatical tasks

Task	Type	#train	#test
SentLen	Length prediction	100k	10k
WC	Word Content analysis	100k	10k
TreeDepth	Tree depth prediction	100k	10k
TopConst	Top Constituents prediction	100k	10k
BShift	Word order analysis	100k	10k
Tense	Verb tense prediction	100k	10k
SubjNum	Subject number prediction	100k	10k
ObjNum	Object number prediction	100k	10k
SOMO	Semantic odd man out	100k	10k
CoordInv	Coordination Inversion	100k	10k

https://github.com/facebookresearch/SentEval/tree/master/data/probing

Probing tasks – Semantic Odd Man Out

No one could see this Hayes and I wanted to know if it was real or a **spoonful** (orig: "ploy")

MLP classifier →

M

- <u>Goal</u>: Predict whether a sentence has been modified or not: one verb/noun randomly by another verb/noun with same POS
- <u>Note</u>: preserved bigrams frequency, human eval.: 81.2%
- <u>Question</u>: Can we identify well-formed sentences (sentence model)?