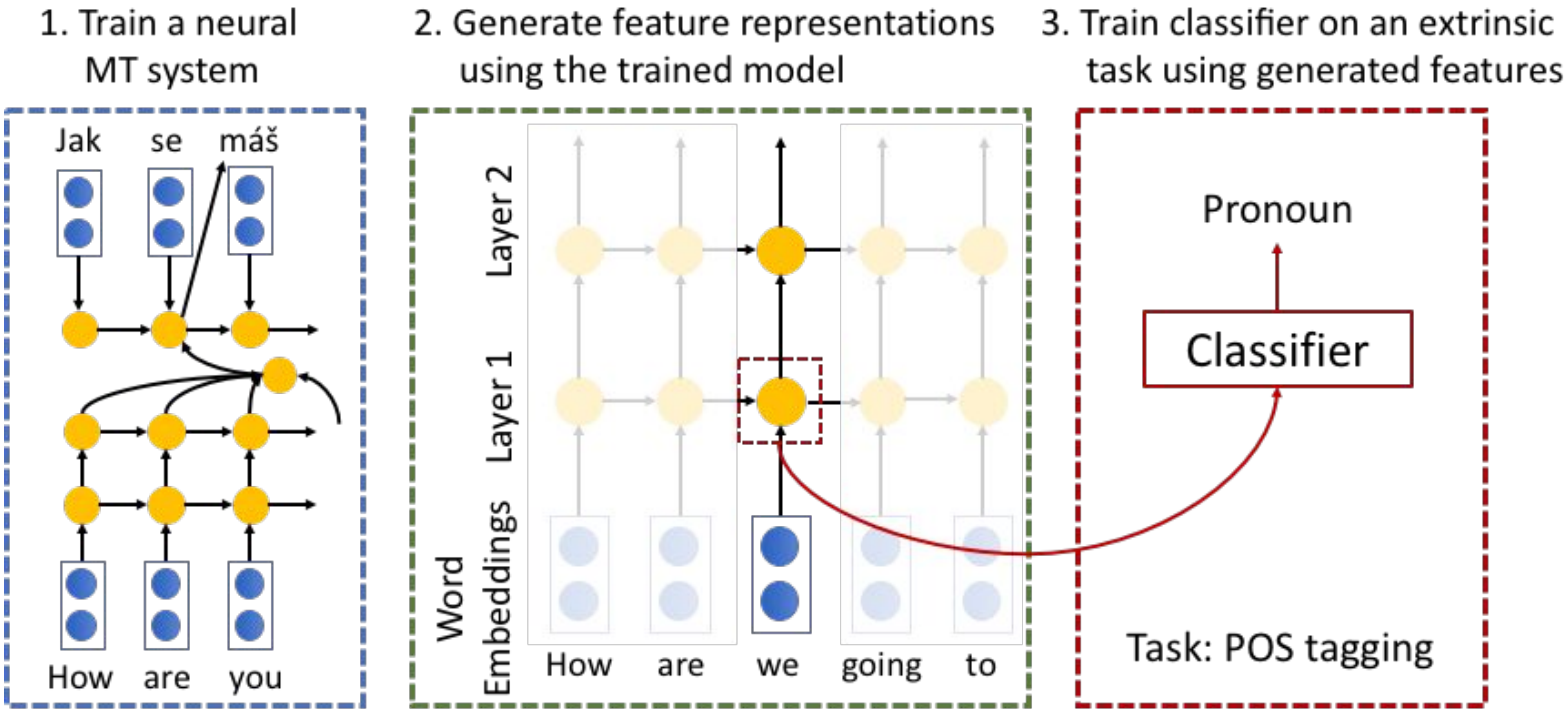
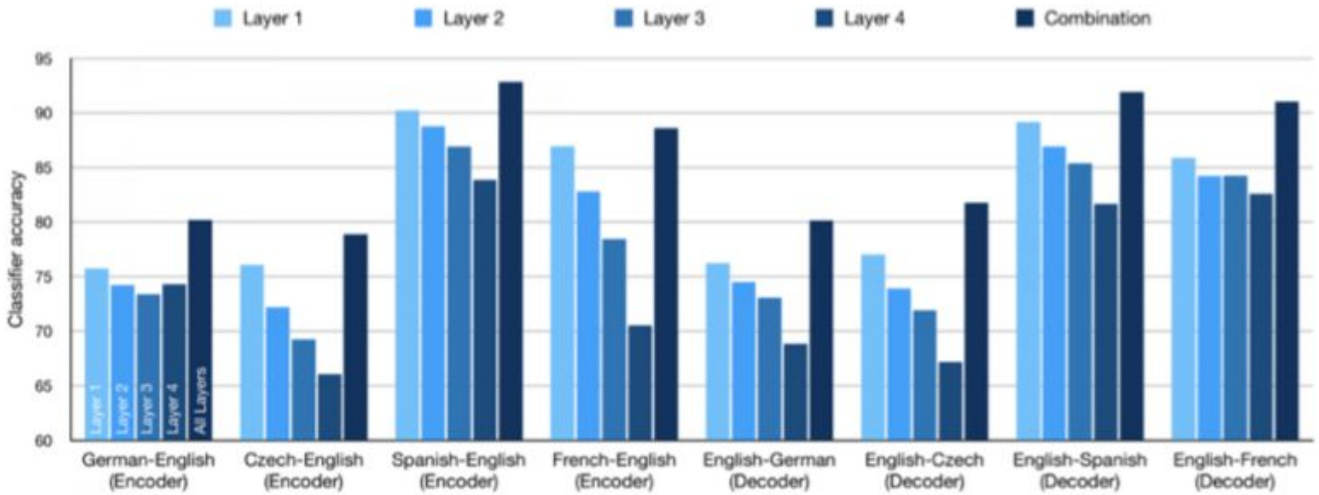
Example: Machine Translation

- Setup
  - *f* : an RNN encoder-decoder MT model
  - *x* and *y* are source and target sentences (lists of words)
  - *g*: a non-linear classifier (MLP with one hidden layer)
  - *z*: linguistic properties of words in *x* or *y*
- Morphology:
  - A challenge for machine translation, previously solved with feature-rich approaches.
  - Do neural networks acquire morphological knowledge?
- Experiment
  - Take $f^l(x)$, an RNN hidden state at layer *l*
  - Predict *z*, a morphological tag (*verb-past-singular-feminine*, *noun-plural*, etc.)
  - Compare accuracy at different layers *l*

Example: Machine Translation



Machine Translation: Morphology



- Lower is better
- But deeper models translate better → what's going on in top layers?

Example: Machine Translation

- Setup
  - *f* : an RNN encoder-decoder MT model
  - *x* and *y* are source and target sentences (lists of words)
  - *g*: a non-linear classifier (MLP with one hidden layer)
  - *z*: linguistic properties of words in *x* or *y*

Probing Classifiers Questionnaire

What is the goal of the study?

Scientific / Pedagogical / **Debugging** / Debiasing / **…**

Understanding model structure / model decisions / data / ...

How do you quantify an outcome? **Performance comparisons**

Who is your user or target group?

ML or NLP Expert / Domain Expert / Student / Lay User of the System ...

How much domain/ model knowledge do they have? **Enough to understand the model and problem domain**

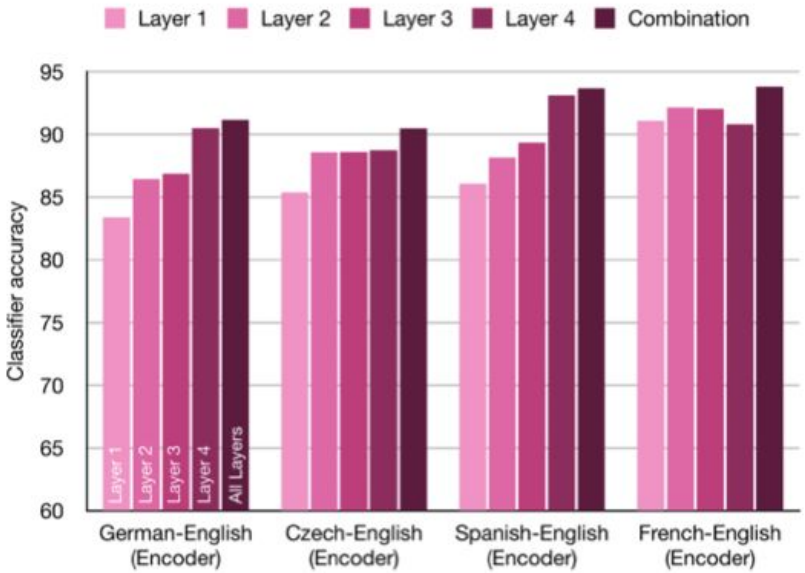
Example: Machine Translation

- Setup
  - *f* : an RNN encoder-decoder MT model
  - *x* and *y* are source and target sentences (lists of words)
  - *g*: a non-linear classifier (MLP with one hidden layer)
  - *z*: linguistic properties of words in *x* or *y*
- Syntax:
  - A challenge for machine translation, previously solved with hierarchical approaches.
  - Do neural networks acquire syntactic knowledge?

Example: Machine Translation

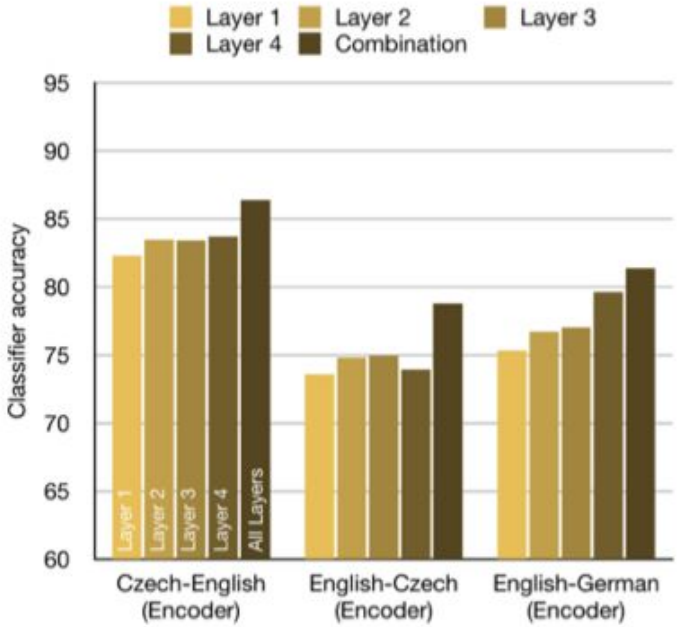
- Setup
  - *f* : an RNN encoder-decoder MT model
  - *x* and *y* are source and target sentences (lists of words)
  - *g*: a non-linear classifier (MLP with one hidden layer)
  - *z*: linguistic properties of words in *x* or *y*
- Syntax:
  - A challenge for machine translation, previously solved with hierarchical approaches.
  - Do neural networks acquire syntactic knowledge?
- Experiment
  - Take $[f^l(x_i) ; f^l(x_j)]$, RNN hidden states of words x_i and x_j, at layer *l*
  - Predict *z*, a dependency label (*subject, object*, etc.) between words x_i and x_j
  - Compare accuracy at different layers *l*

Machine Translation: Syntactic Relations



- Higher is better

Machine Translation: Semantic Relations



- Higher is better

Hierarchies

Language Hierarchy

Semantics

Discourse

Propositions

Roles

Syntax

Trees

Phrases

Relations

Morpho-Syntax

Parts-of-speech

Morphology

Lexicon

Hierarchies

Speech Hierarchy

Words

Syllables

Phonemes

Complex

Simple

Articulatory features

Place

Manner

⋮

Language Hierarchy

Semantics

Discourse

Propositions

Roles

Syntax

Trees

Phrases

Relations

Morpho-Syntax

Parts-of-speech

Morphology

Lexicon

Vision Hierarchy

Scenes

Objects

Object parts

Motifs

Edges