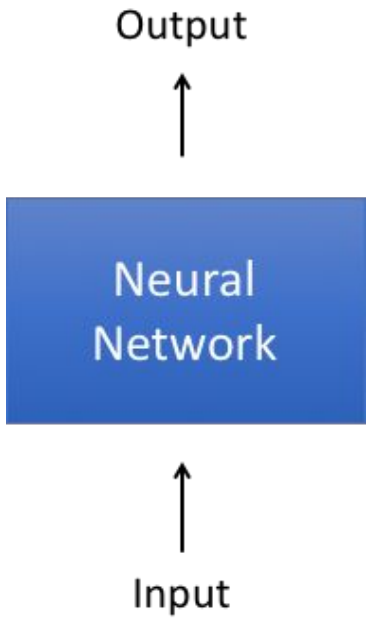
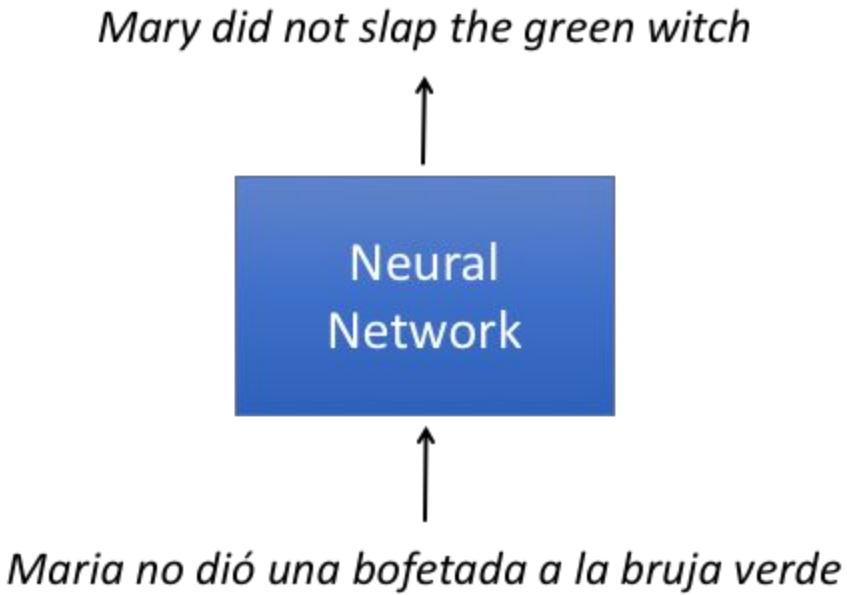
End-to-End Learning

- The predominant approach in NLP these days is end-to-end learning
- Learn a model $f : x \rightarrow y$, which maps input x to output y



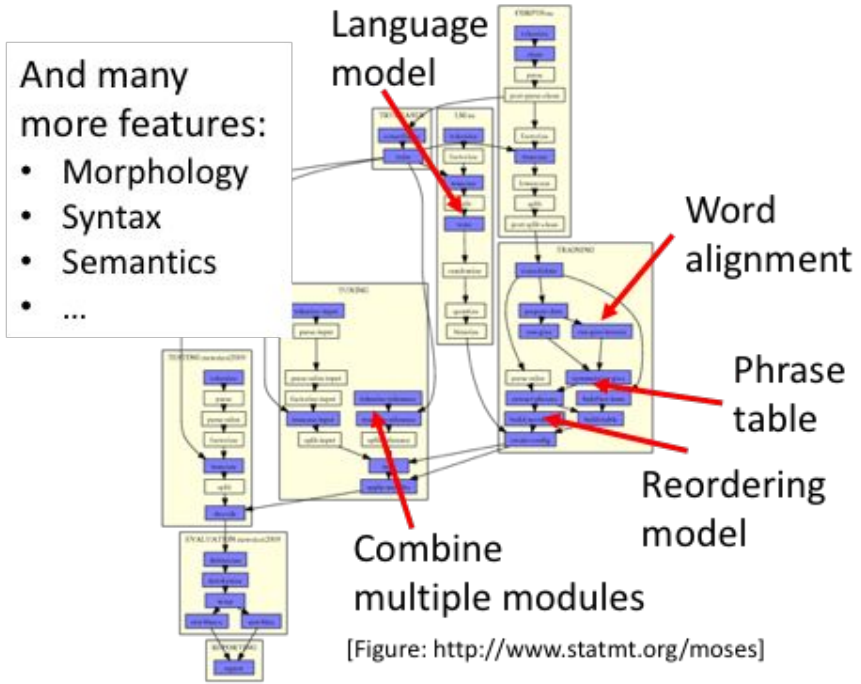
End-to-End Learning

- For example, in machine translation we map a source sentence to a target sentence, via a deep neural network:



A Historical Perspective

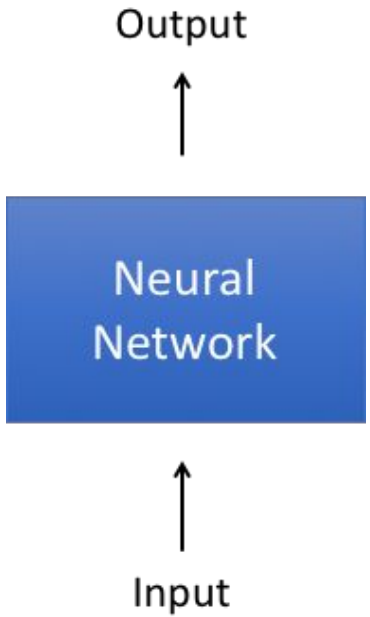
- Compare this with a traditional statistical approach to MT, based on multiple modules and features:



[Figure: http://www.statmt.org/moses]

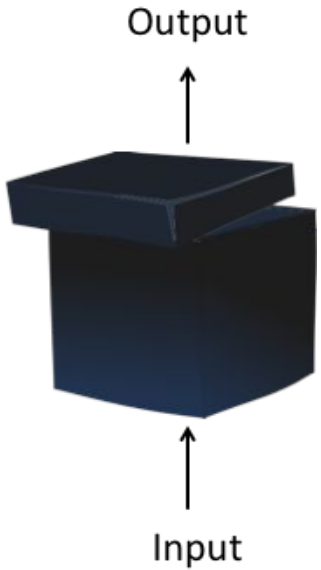
End-to-End Learning

- The predominant approach in NLP these days is end-to-end learning, where all parts of the model are trained on the same task:



How can we open the black box?

- Given $f : x \rightarrow y$, we want to ask some questions about f
  - What is its internal structure?
  - How does it behave on different data?
  - Why does it make certain decisions?
  - When does it succeed/fail?
  - ...



Why should we care?

- Much deep learning research:
  - Trial-and-error, shot in the dark
  - Better understanding → better systems

Why should we care?

- Much deep learning research:
  - Trial-and-error, shot in the dark
  - Better understanding → better systems

- Accountability, trust, and bias in machine learning
  - “Right to explanation”, EU regulation
  - Life threatening situations: healthcare, autonomous cars
  - Better understanding → more accountable systems

Why should we care?

- Much deep learning research:
  - Trial-and-error, shot in the dark
  - Better understanding → better systems

- Accountability, trust, and bias in machine learning
  - “Right to explanation”, EU regulation
  - Life threatening situations: healthcare, autonomous cars
  - Better understanding → more accountable systems
- Neural networks aid the scientific study of language (Linzen 2019)
  - Models of human language acquisition
  - Models of human language processing
  - Better understanding → more interpretable models

Goal for today

1. Understand the toolbox of interpretability methods in NLP

2. Have an idea which tool to apply to a problem