Rudolf Rosa, Jindřich Libovický, Tomáš Musil, David Mareček
rosa@ufal.mff.cuni.cz

Looking for linguistic structures in neural networks

Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

DeepLearn Open Session, Genova, 25 July 2018

Neural Machine Translation

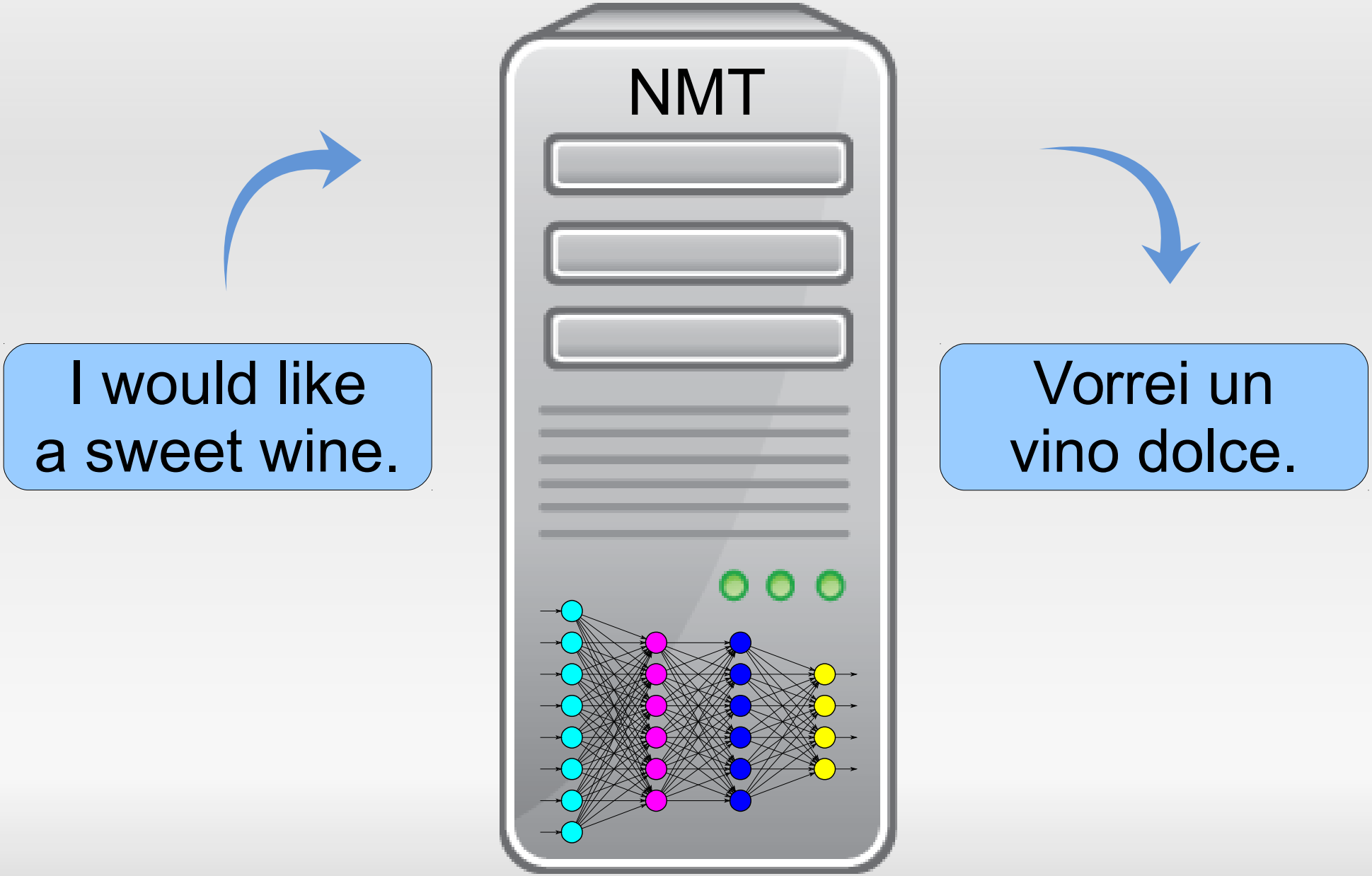
Neural Machine Translation

I would like
a sweet wine.

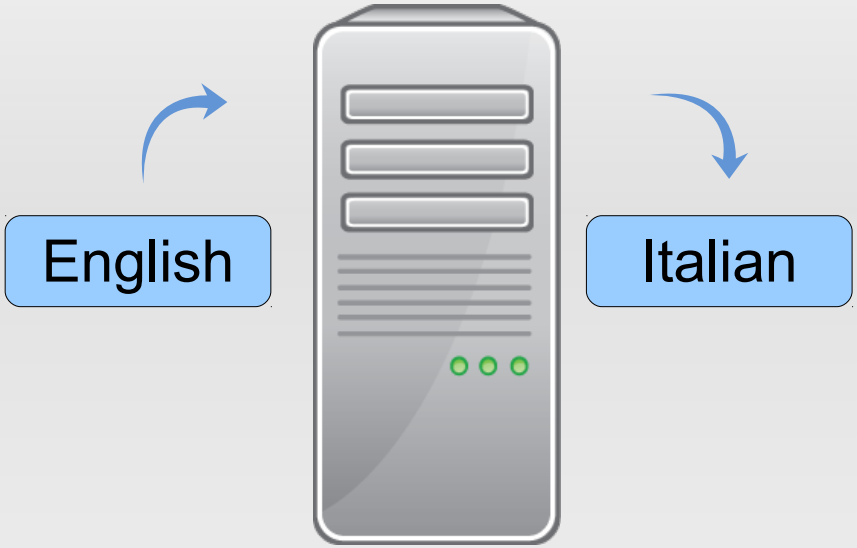
Neural Machine Translation

I would like
a sweet wine.

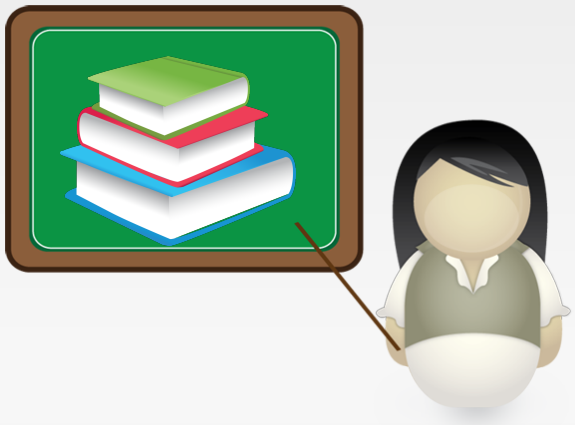
Neural Machine Translation



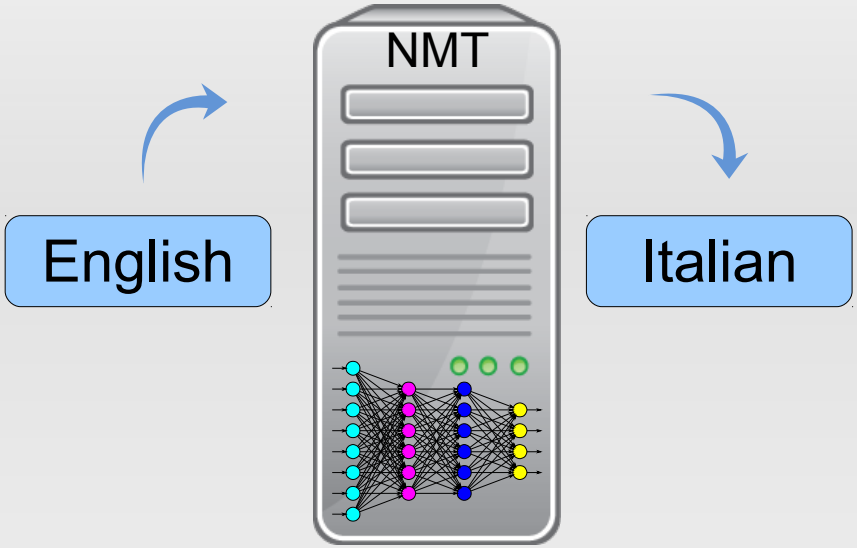
Explicit linguistic knowledge?

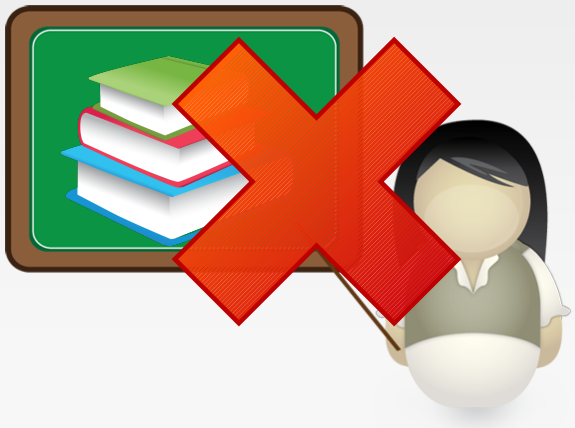


- Traditional translation systems: explicit linguistic knowledge
  - words, morphology, syntax, word order...



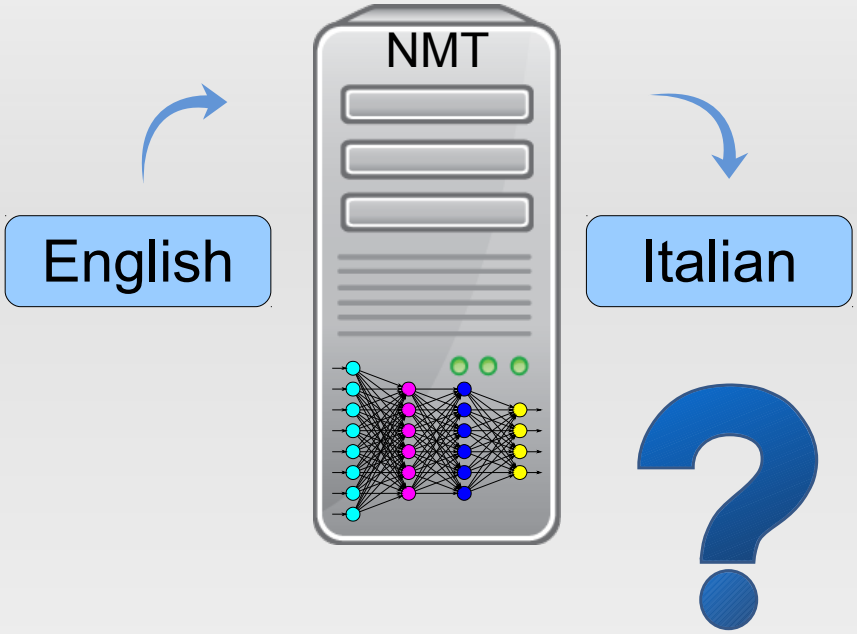
Explicit linguistic knowledge?





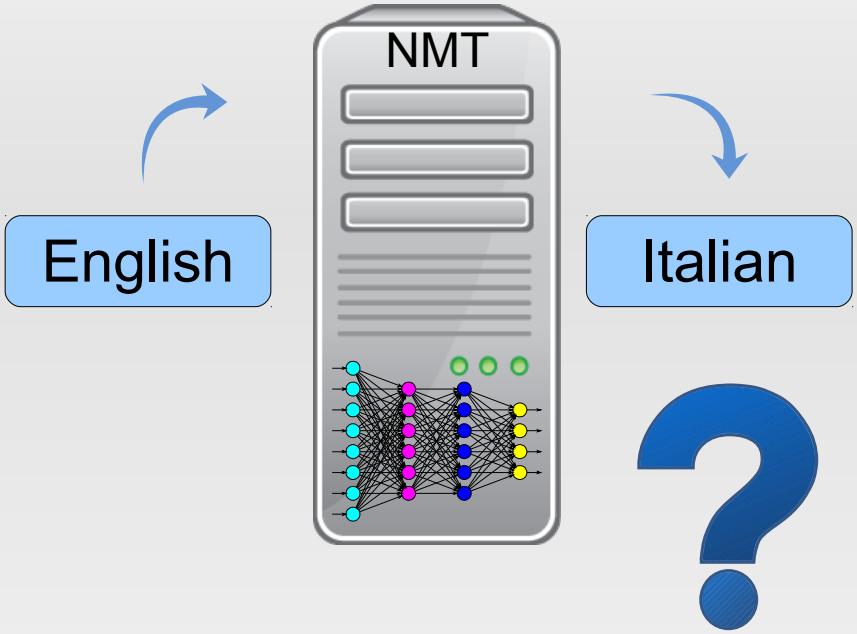
- Traditional translation systems: explicit linguistic knowledge
  - words, morphology, syntax, word order...
- NMT systems: **no** explicit knowledge
  - end-to-end systems
  - directly trained with plain texts on input

Language understanding in NMT?



- Why does it work?
- Does the network “understand English”?

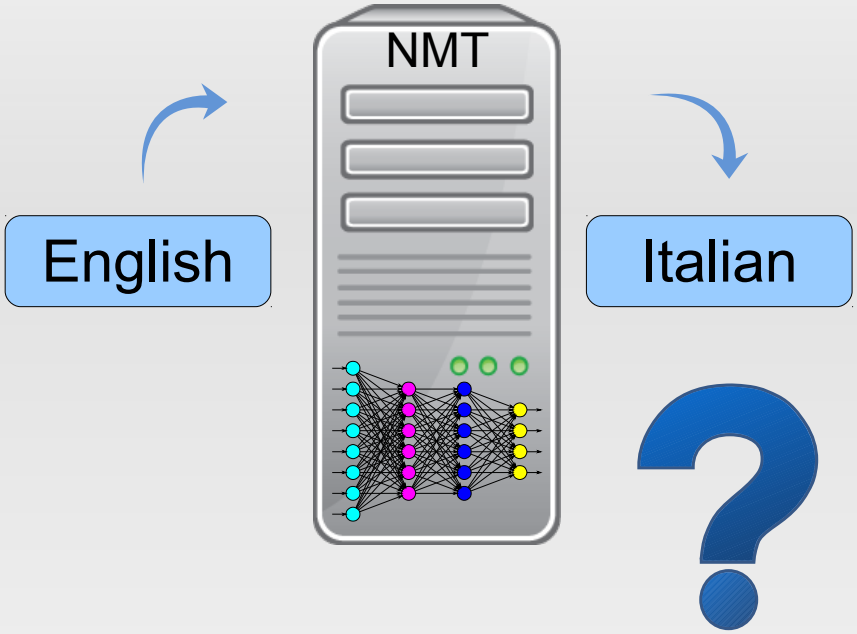
Language understanding in NMT?



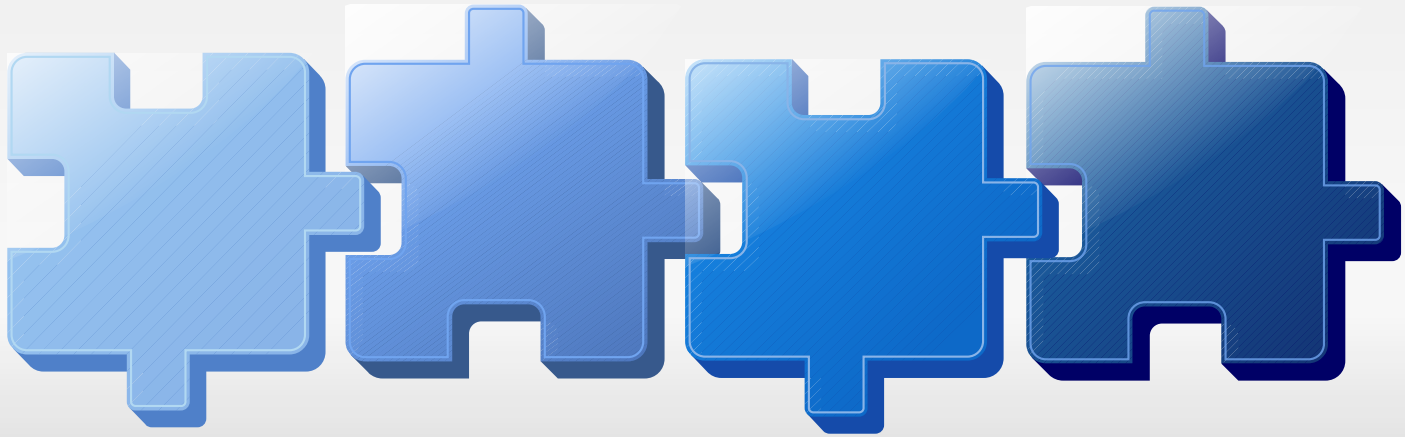
- Why does it work?
- Does the network “understand English”?

What does it mean to ”understand English”?

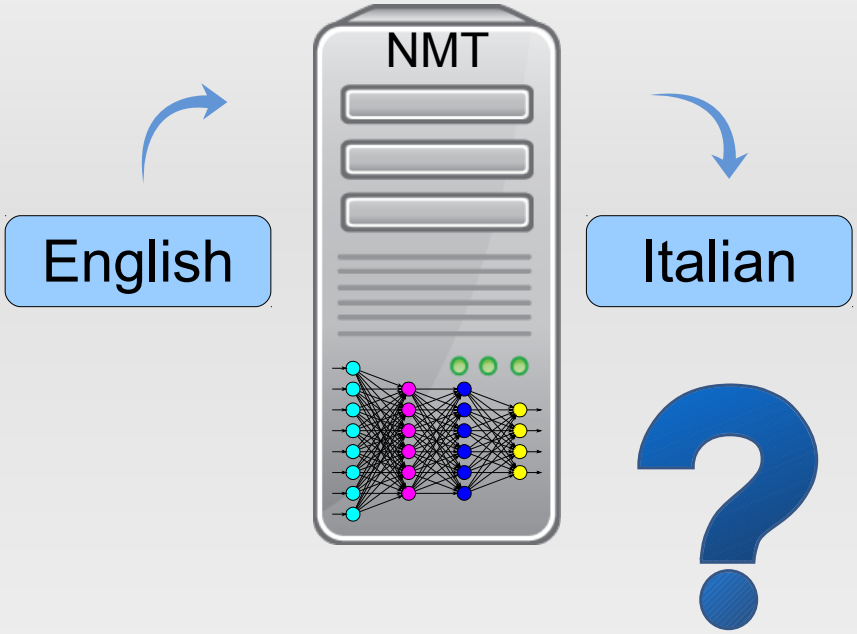
Language understanding in NMT?



- Why does it work?
- Does the network “understand English”?
- Does it know the syntax of English?



Language understanding in NMT?

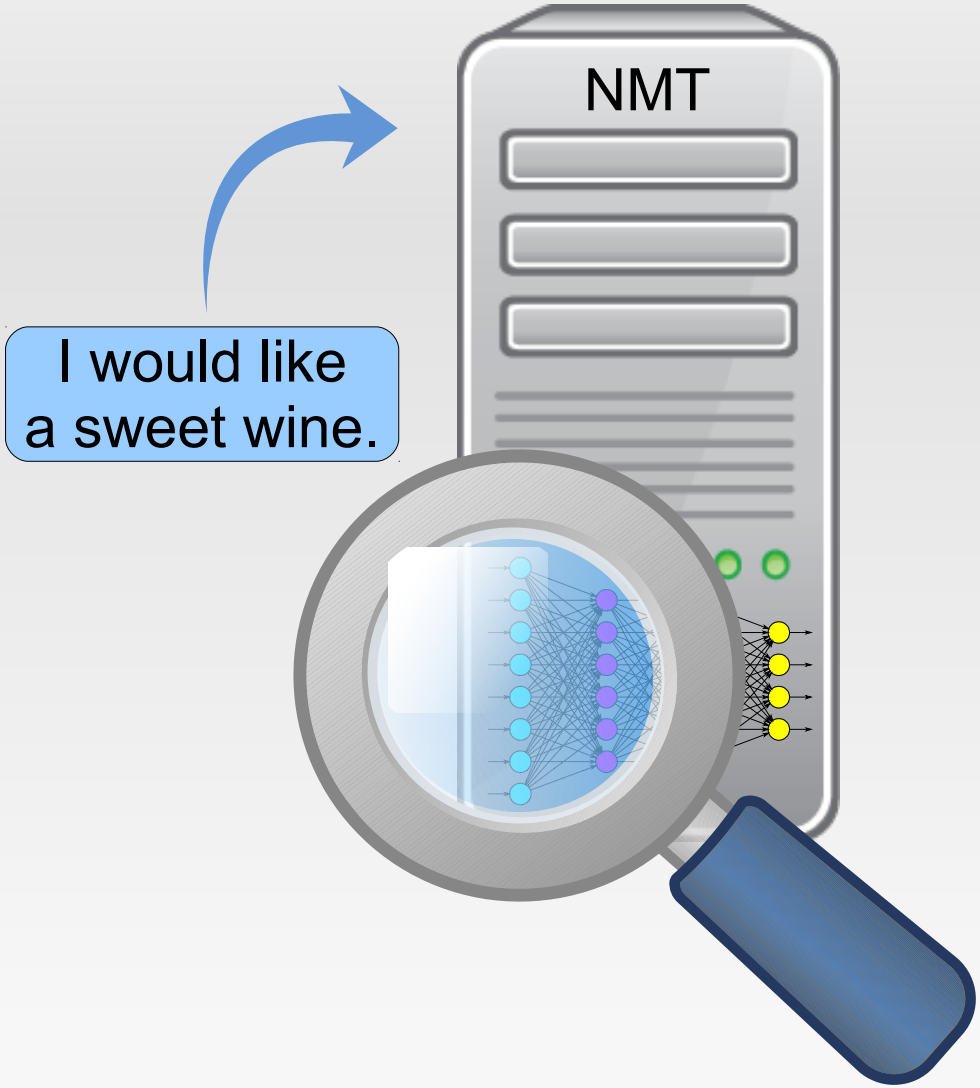


tree, horse, wine, car, watermelon, bed...

buy, eat, understand, sleep, read, relax...

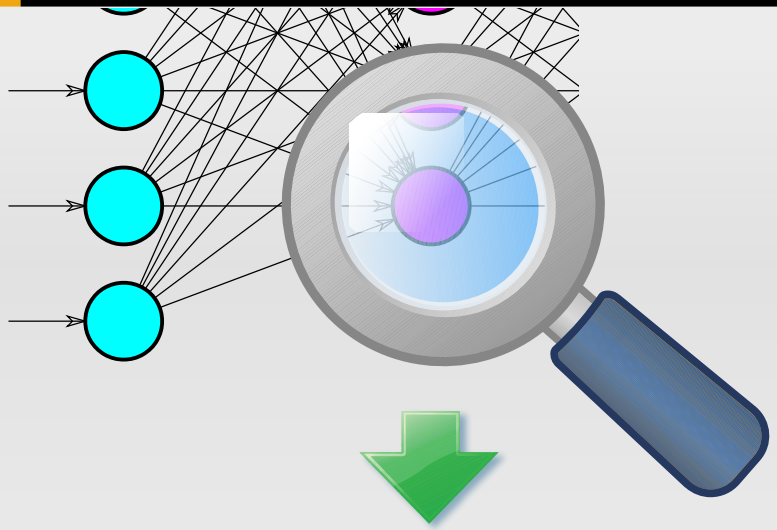
- Why does it work?
- Does the network "understand English"?
- Does it know the syntax of English?
- Does it know parts of speech, e.g. can it tell a noun from a verb?

Analysing the NMT encoder



- take internal representations from the encoder
  - word embeddings
  - hidden states

Analysing the NMT encoder

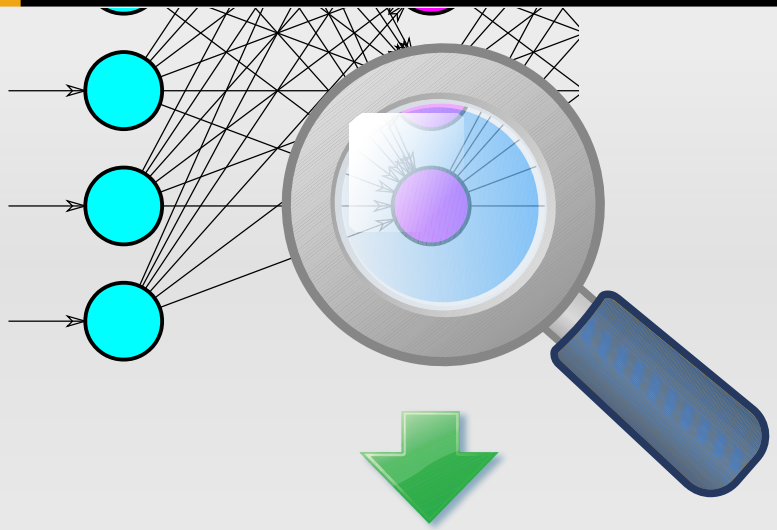


0.2 2.3 2.7 -5.4 … -0.7

wine

- take internal representations from the encoder
  - word embeddings
  - hidden states

Analysing the NMT encoder

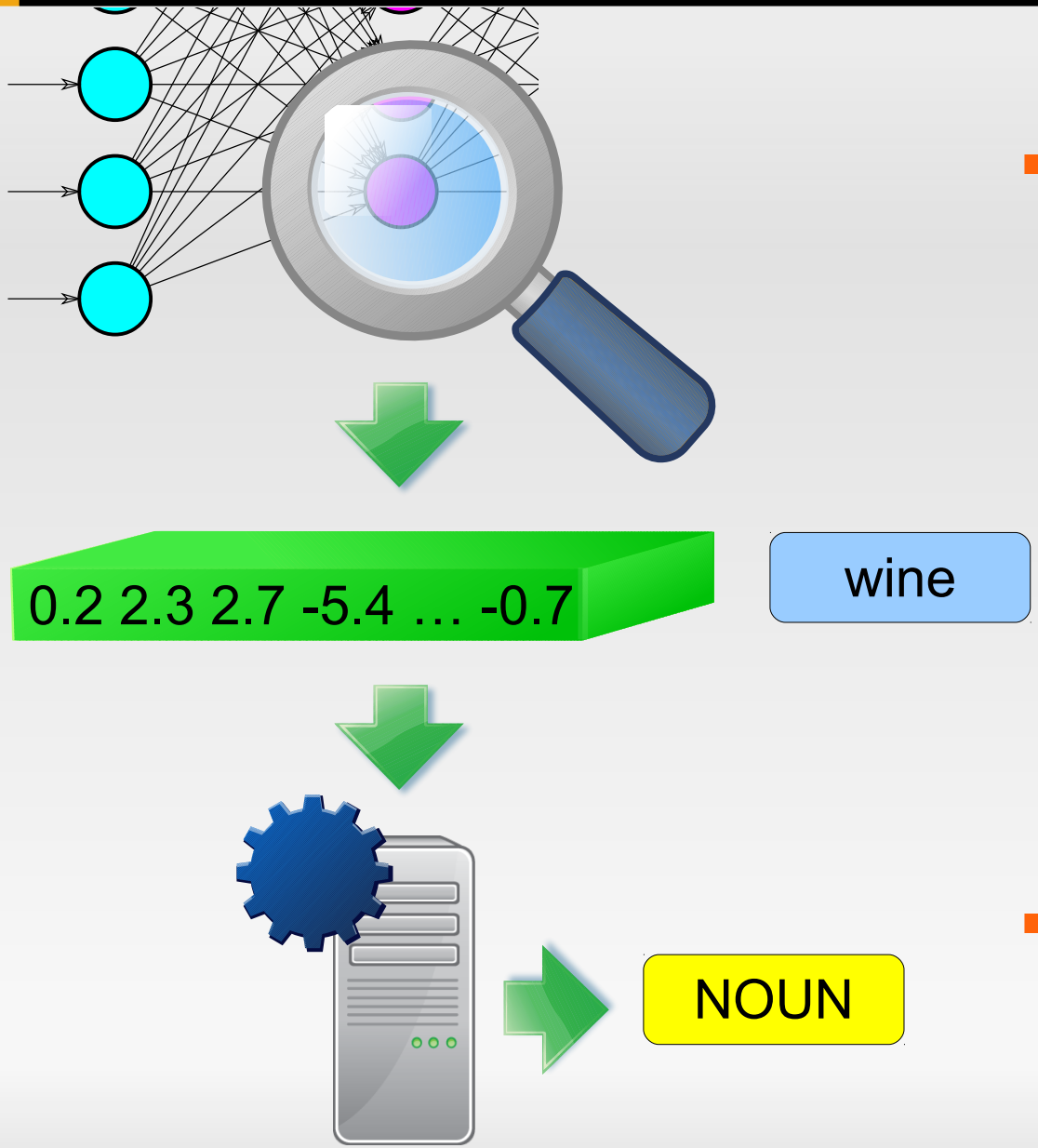


wine

NOUN

- take internal representations from the encoder
  - word embeddings
  - hidden states
- train a ML model to predict PoS from it
  - idea: if the vectors capture PoS, we can learn to predict PoS from the vectors

No conclusive results yet



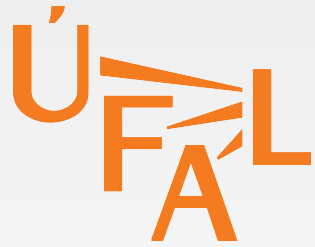
- may be influenced by
  - network architecture
  - network layer
  - words/subwords
  - language(s) used
  - …
  - our mistakes
- current&future work
  - get conclusive results
  - also look at syntax

Thank you for your attention

Rudolf Rosa, Jindřich Libovický, Tomáš Musil, David Mareček
rosa@ufal.mff.cuni.cz

Looking for linguistic structures in neural networks

Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



http://ufal.mff.cuni.cz/rudolf-rosa/