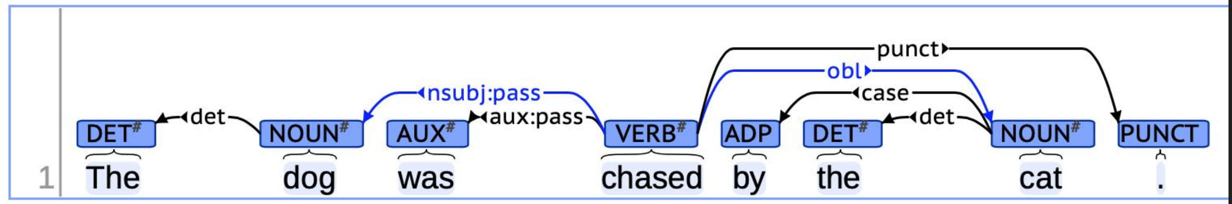
Every time I hire a *properly trained* linguist...

Nikola Ljubešić
Jožef Stefan Institute, Slovenia

The current state of natural language processing

Current natural language processing is based on learning from examples where specific problems are solved by humans.



Such examples are called training data. We obtain the training data by organizing annotation campaigns and hiring linguists to perform annotations.

The training data is used to train a system to solve such problems automatically on previously unseen data.

The current state of natural language processing

Three types of problems we are solving:

1. linguistic problems - part-of-speech tagging, syntactic analysis
2. end-to-end problems - hate speech detection, machine translation, speech-to-text, question answering, text summarization
3. model benchmarking - measure how well a model solves types of problems

We need linguists for:

1. manual annotation of training data, error analysis
2. error analysis
3. benchmark construction, error analysis

The current state of natural language processing

Three types of problems we are solving:

1. linguistic problems - part-of-speech tagging, syntactic analysis
2. end-to-end problems - hate speech detection, machine translation, speech-to-text, question answering, text summarization
3. model benchmarking - measure how well a model solves types of problems

We need linguists for:

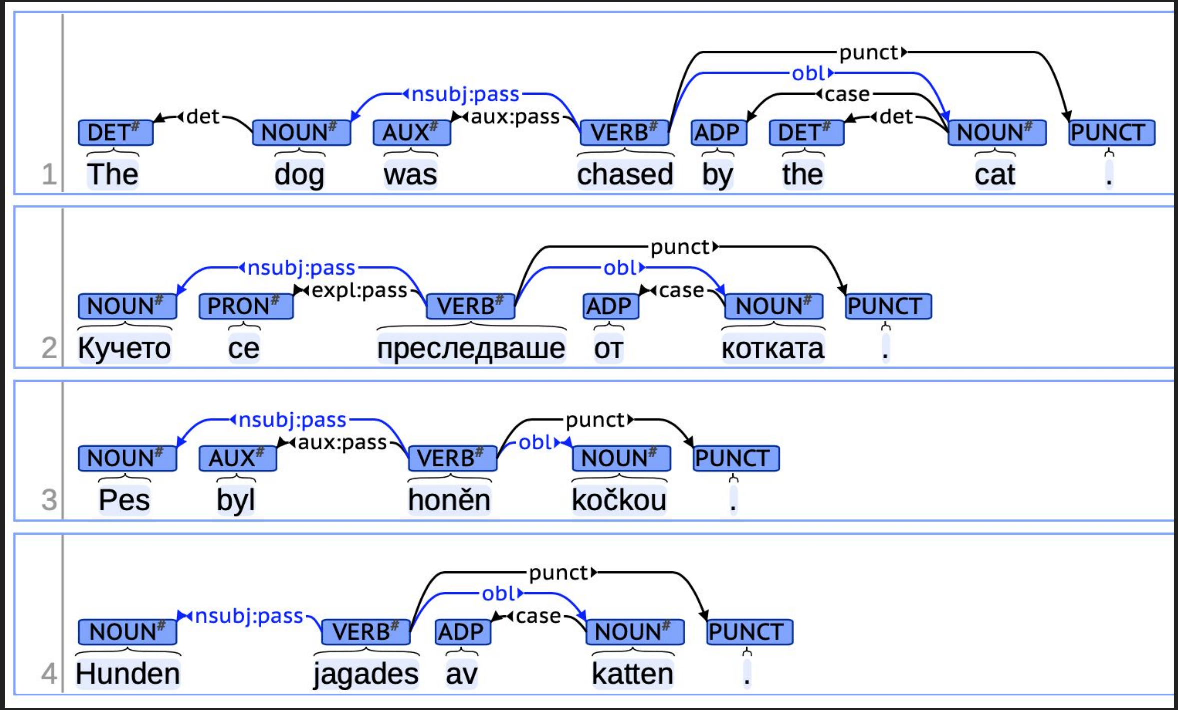
1. **manual annotation of training data**, error analysis
2. error analysis
3. **benchmark construction**, error analysis

Linguistic problem - dependency parsing

Requirements for human annotators:

Strong theoretical background in morphology and syntax (multiple theories - flexibility!)

Strict following of annotation guidelines, especially for edge cases



Benchmarking - Choice of plausible alternatives (COPA)

Dataset for measuring how well models cover commonsense causal reasoning.

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Model performance on English: 2013: 51%, 2019: 70%, 2021: 98%

The dataset translated into multiple languages - the translator is supposed to

1. translate by following the translation guidelines (with feedback)
2. take the test themself (to prime the translator and test for cultural bias)

Benchmarking - Choice of plausible alternatives (COPA)

Dataset for measuring how well models cover commonsense causal reasoning.

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Model performance on English: 2013: 51%, 2019: 70%, 2021: 98%

The dataset translated into multiple languages - the translator is supposed to

1. translate by following the translation guidelines (with feedback)
2. take the test themself (to prime the translator and test for cultural bias)

My vision on the skills required

- Linguistic annotation
  - Strong theoretical background
  - BUT flexibility to apply a specific theory / formalism (with all its imperfections)
  - Strict following of annotation guidelines (but also feedback)
  - Training in setting up annotation guidelines and running annotation campaigns
- Benchmarking
  - Strong theoretical background - what language models should be able to do, how to measure how good they are at a specific phenomenon
  - Good understanding of the types of models currently out there (and general understanding of their inner workings)
  - Strict following of annotation / translation guidelines (but also feedback)

Every time I hire a *properly trained* linguist...

The performance of my system goes up!

I understand better what the system decisions are based on.

I know what my system is GOOD and what it is BAD at.

I identify promising directions to further improve my system.

Every time I hire a *properly trained* linguist...

Nikola Ljubešić
Jožef Stefan Institute, Slovenia