A better evaluation scheme for multilingual parsing

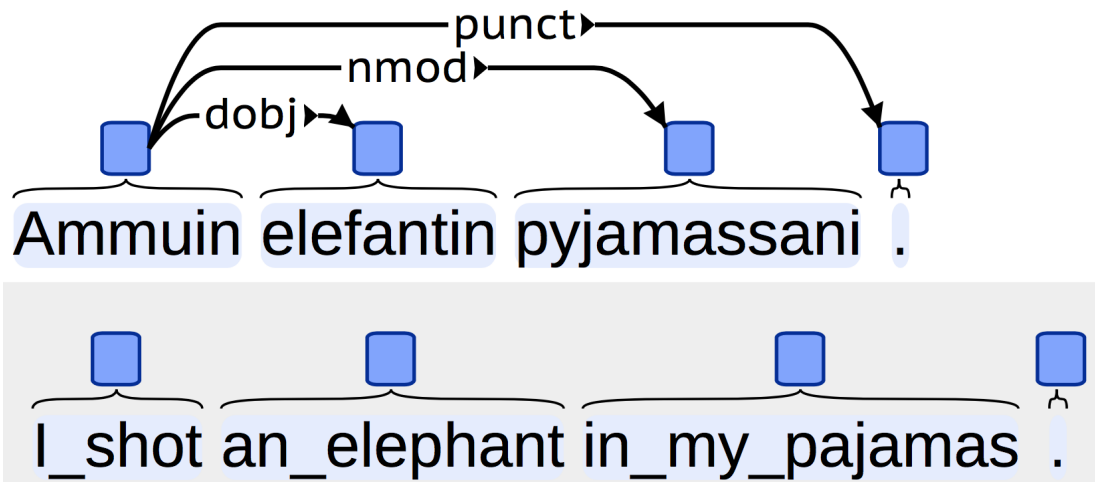
Jimmy Callin

October 21 2015

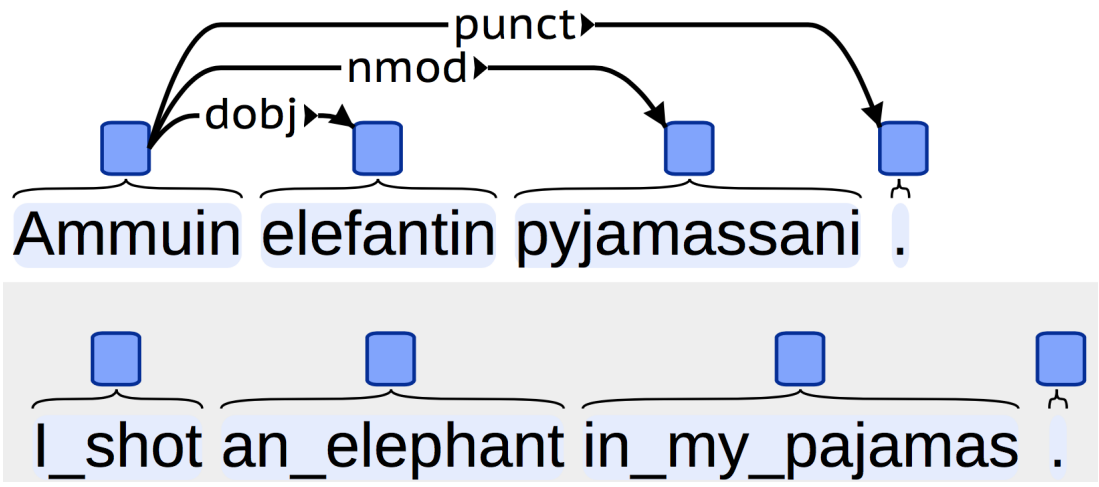
Current situation

- Evaluation
  - (Labeled/Unlabeled) Attachment Score
  - Assumes we don't really know anything about the framework.
- Are parsing results between languages equivalent?

What's wrong?


punct
nmod
dobj
Ammuin elefantin pyjamassani .
I_shot an_elephant in_my_pajamas .

What's wrong?



- Inject a faulty edge:
  - Finnish: 33 ppt error increase
  - English: 14 ppt error increase

Enter Universal Dependencies

- Labels mean something now!

Enter Universal Dependencies

- Labels mean something now!
- **Purpose:**
  - How much does languages with high lexicalization of grammar particles benefit from current evaluation schemes?
  - Is it better to focus on correct classification of content word dependencies?
- Can we use this to learn more about our models?

Next steps

1. Study the taxonomy of UD (2 weeks).
   - What dependencies are used for function words?
   - Create an initial manual separation of content/function dependencies.

Next steps

1. Study the taxonomy of UD (2 weeks).
   - What dependencies are used for function words?
   - Create an initial manual separation of content/function dependencies.

2. Study distribution of function dependencies (2 weeks).
   - Can we come up with a statistically motivated separation?
   - Would allow us to study varying degrees of granularity.

Next steps

1. Study the taxonomy of UD (2 weeks).
   - What dependencies are used for function words?
   - Create an initial manual separation of content/function dependencies.

2. Study distribution of function dependencies (2 weeks).
   - Can we come up with a statistically motivated separation?
   - Would allow us to study varying degrees of granularity.

3. Implement evaluation metric (2 weeks).
   - Apply on an appropriate parser with a suitable number of different languages.
   - Expectation: Performance gap on the analytic-synthetic spectrum to reduce.

4. Write (or rather compile) the thing (1 week).

Evaluation of evaluation

- How do we evaluate an evaluator?
- Compare with analytic-synthetic language spectrum.
  - Token/type ratio correlate with function dependency ratio.

HELP

- Looking for typological studies on function word frequency distribution
- Any papers on distributional properties of dependency relations?
- Has any parser model been trained and tested on all UD languages? Output, please?