Deep Learning for Social Media Processing

Malta, March 2018

Barbara Plank

Agenda

- Course logistics
- Introduction to Machine Learning (or recap) ;)
- Project topics

Course logistics

- Lecturer: Barbara Plank (bplank@gmail.com)
- Time: typically at 10:00 (check your schedule)
- Location: Block A, room 15

Introduction round

Course outcomes

- After completing this course, students are able to tackle a challenge in NLP, implement a solution, and critically assess their solution in the light of recent research papers in NLP
- Focus: Deep Learning for NLP

Prerequisites

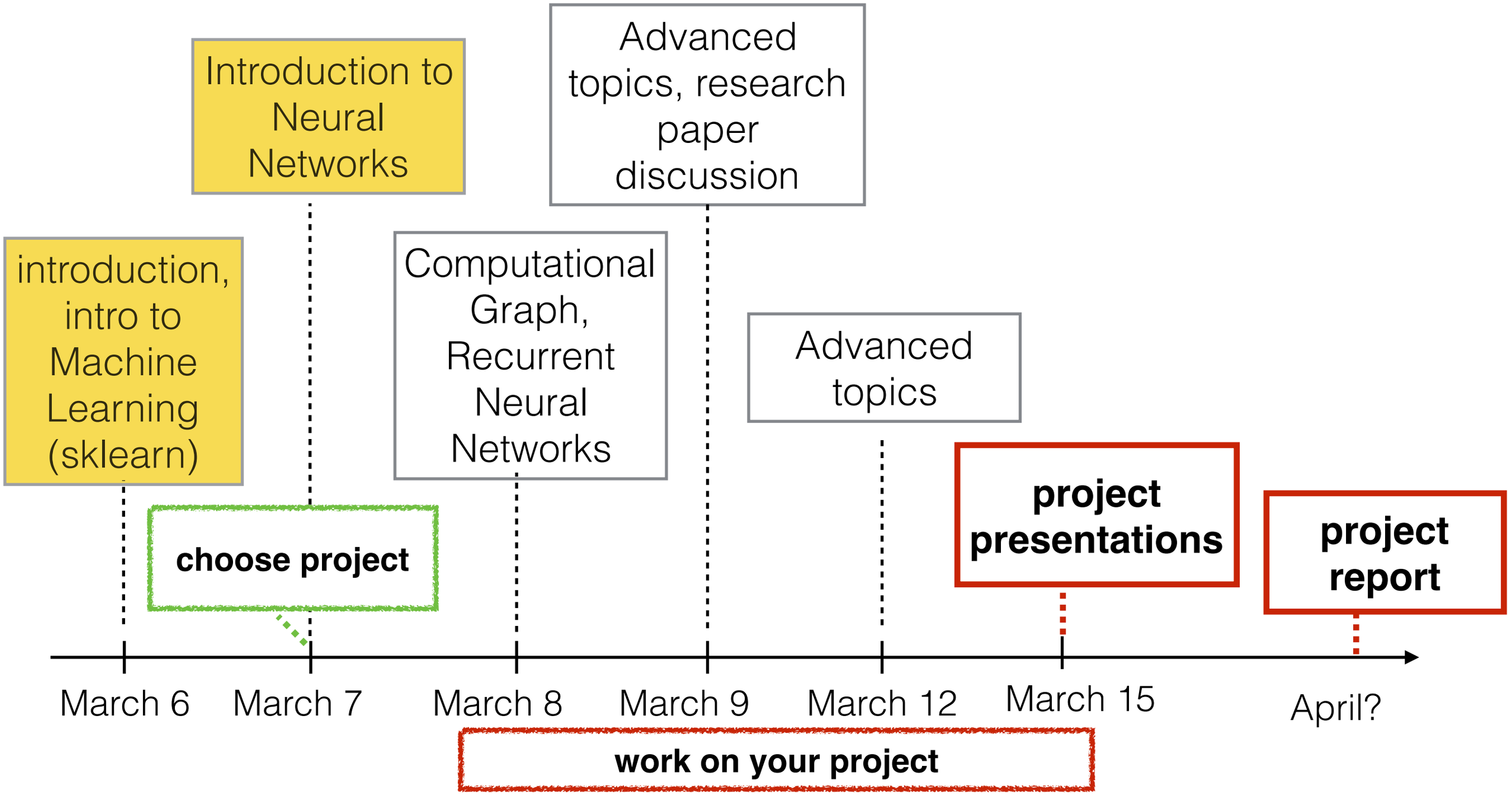
- Familiarity with Machine Learning and Natural Language Processing

- Knowledge of Python - there is a tutorial here

Grading

- 60% project report
- 40% project presentation

Course timeline



Tools

- python3 (suggestion: Anaconda)
- sklearn (recap today)
- Dynamic Deep Learning library: DyNet
  https://dynet.readthedocs.io/en/latest/

Course page

https://github.com/bplank/2018-ma-notebooks

Project topics

General Theme

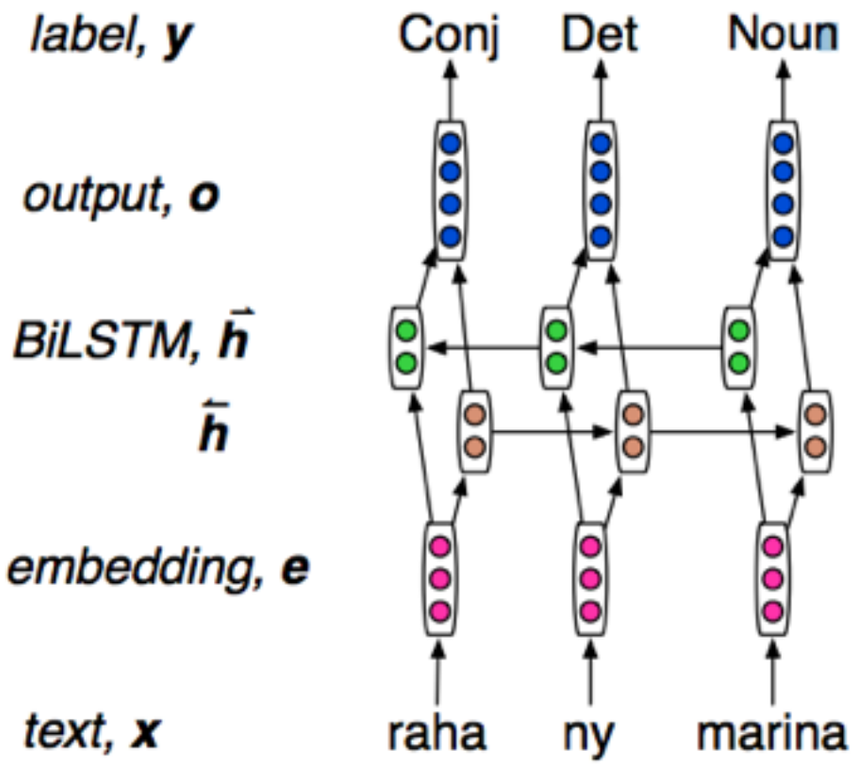
- **Comparison** of a traditional method (e.g. LogRegression, SVM, HMM/CRF…) to a neural network approach
    - Baseline: traditional ML approach
    - Your system: neural network
- You can work in groups (2 people)

Project Topics

- Low-resource POS tagging
- Neural NER for Twitter
- L2 Error detection
- Language-independent gender prediction
- Second-language acquisition modeling (most difficult - official shared task running now!)

Low-resource POS tagging

- How to build a POS tagger with minimal supervision?
  - information you can exploit: embeddings, auxiliary tasks, semi-supervised learning, Wiktionary? …



Li et al., 2012 http://www.aclweb.org/anthology/D12-1127
Fang & Cohn, 2017 http://aclweb.org/anthology/K16-1018.pdf

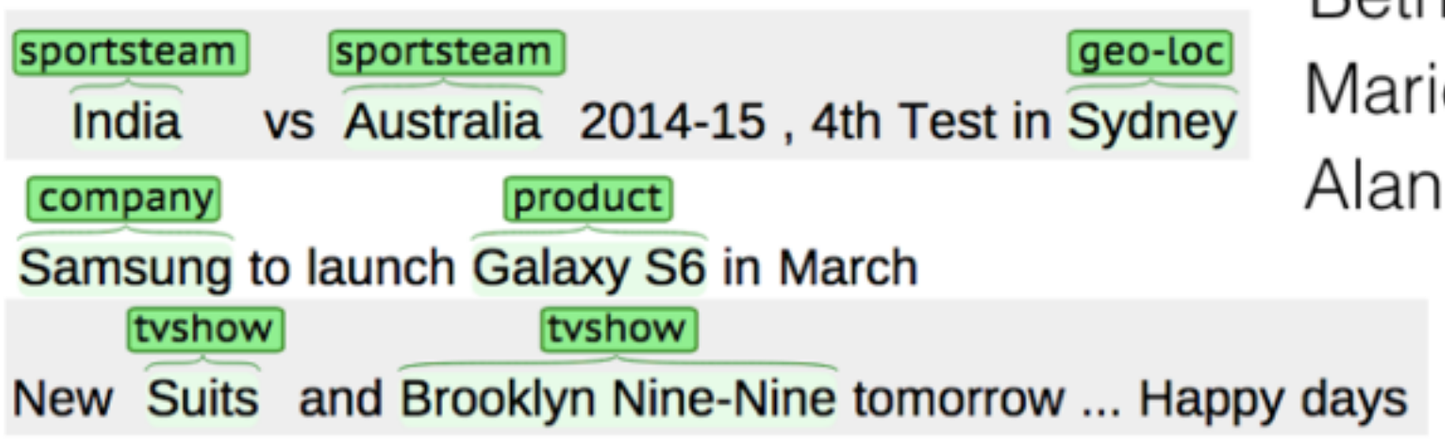
Neural NER on Twitter

W-NUT 2016 Shared Task

Twitter Named Entity Recognition

- Training and Development data available now
- Evaluation Period: **Sept 9 – 16, 2016**

Organizers: Benjamin Strauss
Bethany E. Toma
Marie de Marneffe
Alan Ritter



The Ohio State University

Language-independent Gender Prediction on Twitter

A user has posted the following tweets:

- USER sterkte en snel veel beterschap !
- Dio Oberon overleden: 'de beste nar van Nederland' - Show - Algemeen - bndestem URL
- USER dat is jullie pap op tv! Leuk!
- ja hoor dit gaan we winnen #songfestival ! ;-(((
- USER haha precies wat ik ook dacht ! Watje :-)
- USER hoe moet dat nou een feestje in Rooi zonder jou? beterschap USER xxx ook van katjazondertwitter :-)
- op de bank (alleen) :-(
- USER hè toch! Beterschap!
- Ik heb op deze video gestemd bij #Dichtbijfonds, jij ook? URL
- USER mooi ! ik wil ook nog deze week maar weet niet of ik de tijd kan vinden ;-(vrijdag of zat. hoop ik !!!
- MooiRooi.nl - Sint-Oedenrode - U kunt zich nog aanmelden als wensvervuller URL
- Shit gestoken door een #wesp au!
- USER vermaak je je een beetje?
- Kliknieuws.nl | Sint-Oedenrode - Foto's Hans Vervloed in bibliotheek URL
- USER USER hoe is het met jullie? nog gegaan en drank gewonnen? Ik kon dus echt niet komen #pijn ! morgen over hoop ik :-)
- USER USER was het een leuk feestje ? nog een fijne avond x
- Beleef een professionele fotoshoot bij Lachebekkies. Inclusief foto's op verschillend forma... URL Kei leuk !!!!!!
- USER USER USER USER weer naar huis :-(bedankt allemaal voor de gezelligheid xxx URL
- baaah dan blijf je echt liever alleen !!
- Film tip ! Pina #Pinabausch in 3D #modernballet. heel indrukwekkend !

Do you think that the poster of these tweets is male or female? (required)

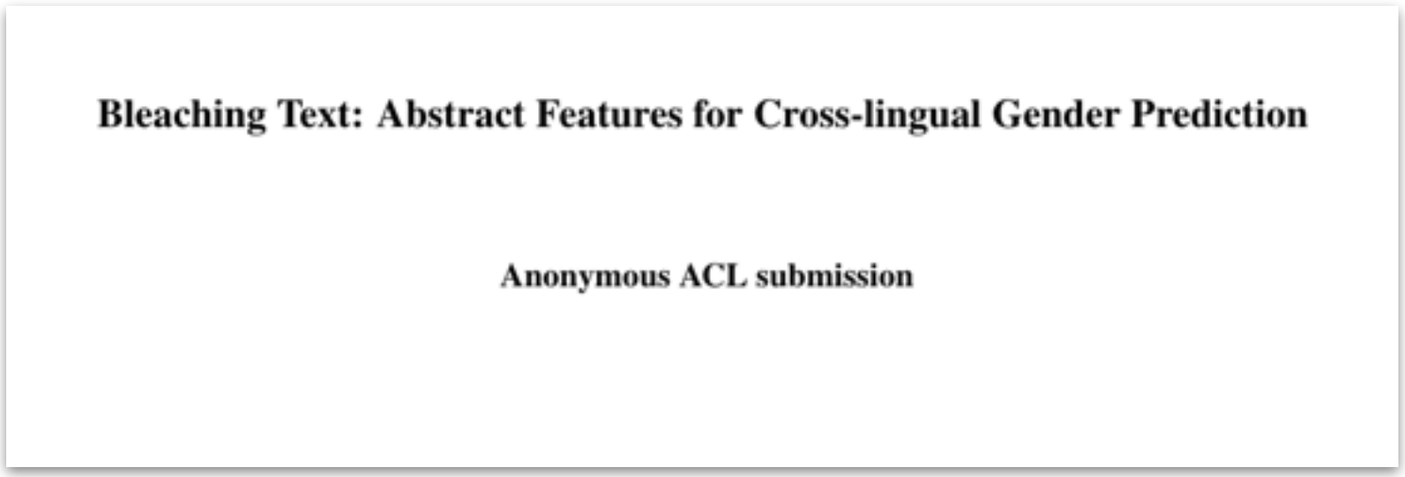
- ○ Female
- ○ Male

Language-independent Gender Prediction on Twitter

- Data: Subset of TwiSty corpus (prepared for you) covering EN, NL, FR, ES, PT

- Ljubesic et al., 2017

- van der Goot et al., (under submission): using abstract features (send me a mail if it interests you)



Bleaching Text: Abstract Features for Cross-lingual Gender Prediction

Anonymous ACL submission

Error detection

- Error detection as a sequence labeling task:

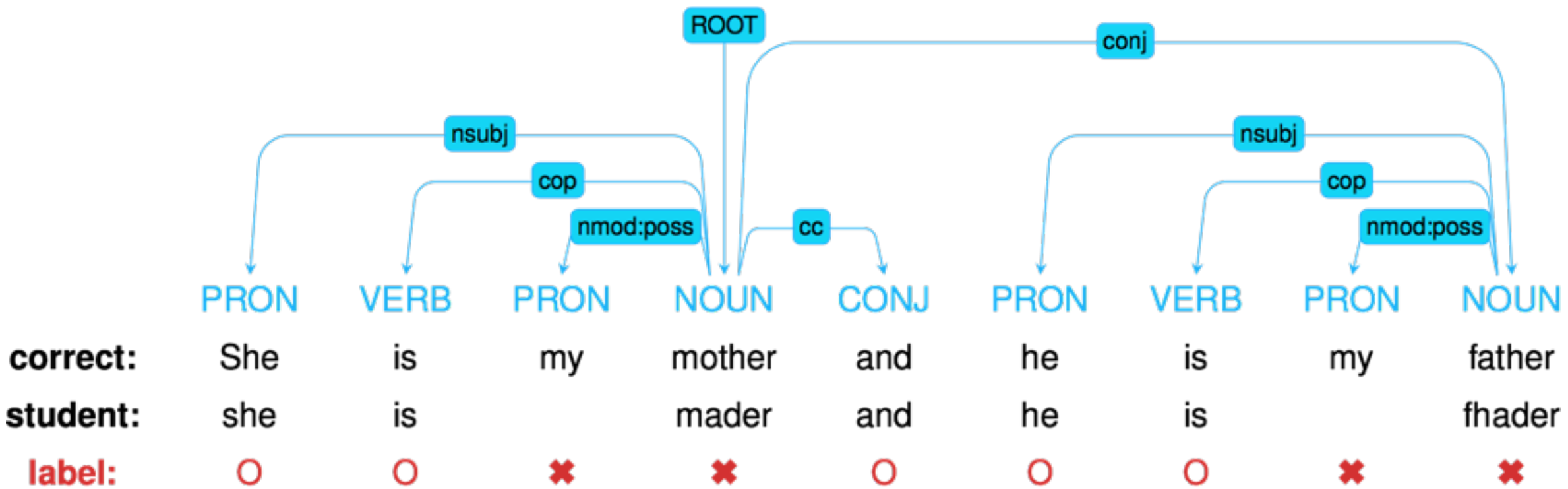
```
+ +    +  x       +   +      +   +    +    x      +
I like to playing the guitar and sing very louder .
```

- CLC FCE Dataset: https://ilexir.co.uk/datasets/index.html

- Rei et al., 2016: https://aclweb.org/anthology/C/C16/C16-1030.pdf

Second Language Acquisition Modeling

2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)

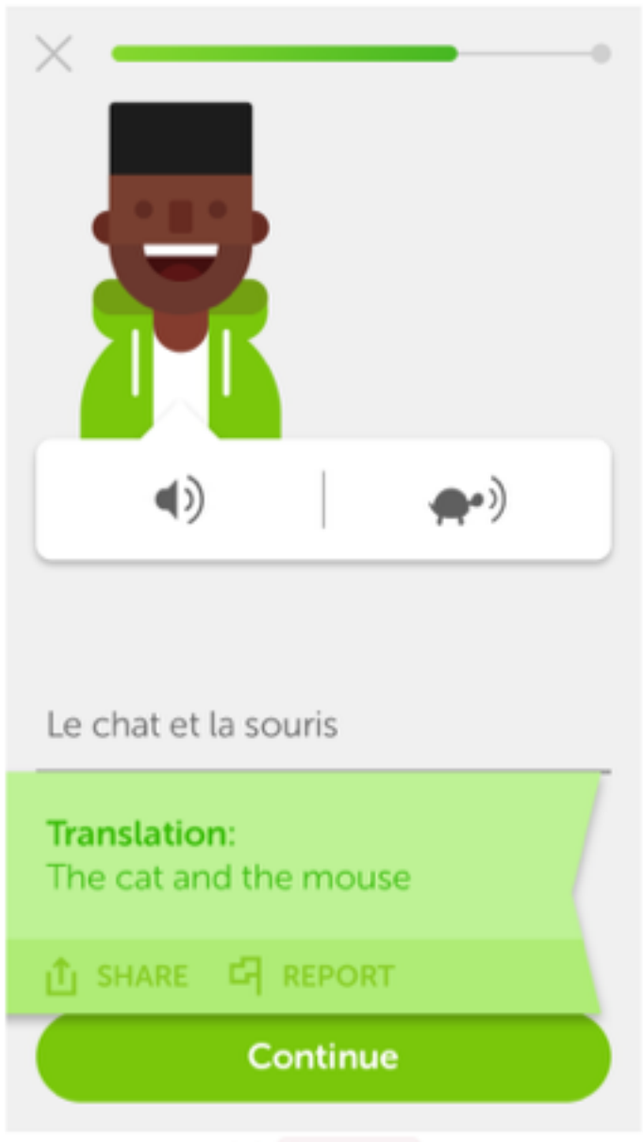


http://sharedtask.duolingo.com/

Second Language Acquisition Modeling

(a) `reverse_translate`

(b) `reverse_tap`



(c) `listen`

EN_ES
ES_EN
FR_EN

http://sharedtask.duolingo.com/