NLP without Annotated Dataset

NLP and Language Learning

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

29 January 2021

NLP and Language Learning

What are some applications?

Broadly, four categories:

- learning assessment
- learning support
- support tool for research (e.g., on language acquisition)
- learning analytics

Learning Assessment

Language Assessment

- Automatic essay scoring systems (such as e-rater by ETS)
- Automatic speech scoring systems (SpeechRater by ETS)
- Automatic creation of test items: multiple choice questions, fill in the blanks etc
- Active area of research for non-English languages too, in the past few years.

Content Assessment

- Evaluating short answers for correctness in relation to the question asked (i.e., going beyond looking for fluency, grammar/spelling correctness etc)
- To an extent, tougher than assessing language form.
- It was an active area of research a few years ago, and SfS had a strong group.

@SfS

In the SFB 833-A4 project, we are developing automatic meaning assessment methods for short-answer reading comprehension. To collect a rich task-based corpus in a real-life teaching context, we created the WELCOME app (Ott et al., 2012) and obtained the CREG corpus (36k answers to 1.5k questions). Our research showcases the importance of interpreting data in context (Ziai & Meurers, 2014; De Kuthy et al., 2015, 2016a, b; Ziai et al., 2016).

source: http://www.sfs.uni-tuebingen.de/icall/

NLP and Language Learning

What are some applications?

Broadly, four categories:

- language and content assessment
- **learning support**
- support tool for research on language acquisition
- learning analytics

Learning support

- Reading support: TextEvaluator (ETS) like tools to choose texts appropriate for a reading level
- Writing support:Spelling/grammar check tools (e.g., grammarly); specialized writing support (e.g., WritingMentor from ETS for academic writing support)
- Language learning apps (e.g., duolingo)

@SfS

- "Portuguese Intelligent Tutoring System (ITS) TAGARELA (Amaral Meurers 2011) designed to complement university instruction"
- "in collaboration with a German school book publisher we created the FeedBook, an interactive workbook for English 7th grade in a DFG-funded transfer project."
- "In the new BMBF project AISLA, we develop an intelligent dialog system supporting the acquisition of English in authentic, spoken language contexts. "
- "developing Prosodiya, a mobile serious game for German dyslexic primary-school children currently"

source: http://www.sfs.uni-tuebingen.de/icall/

@SfS-2

We are developing linguistic complexity analyzers integrating a wide range of linguistic, psycholinguistic, and SLA complexity features for English (Vajjala & Meurers 12, 13, 14a, b, c, Chen & Meurers 2016a, b) and German (Hancke, Vajjala, Meurers 12; Hancke & Meurers 2013) — and tools such as CTAP making it easy to use these measures.

Applying these methods to education, we investigate the (in)appropriateness of textbooks for students of different grades and school types (Bryant et al. 2017, Berendes et al., in press).

To support teachers and learners in identifying texts that are both interesting and richly represent the language constructs to be acquired, we created the linguistically-aware search engine FLAIR (Chinkina & Meurers 16). On this basis, we collaborate in the BMBF-funded KANSAS project with the German Institute for Adult Education (DIE) and the Mercator Institute for Literacy and Language Education to build a tool supporting teachers of functional literacy courses.

Connecting foundational and applied issues, we are spelling out Krashen's i+1 input fostering learning in terms of linguistic complexity using SyB (Chen & Meurers 17), a syntactic benchmarking tool, and we investigate the impact of challenging learners with such input.

CTAP | SyB | FLAIR | KANSAS

source: http://www.sfs.uni-tuebingen.de/icall/

NLP and Language Learning

What are some applications?

Broadly, four categories:

- learning assessment
- learning support
- **support tool for research**
- learning analytics

Language Acquisition Research

- Using NLP tools to study specific linguistic phenomenon in large learner corpora, to understand language acquisition
- Dependency parsing of learner language
- A recent paper: "Subcategorization frame identification for learner English"

etc.

@SfS

Studying learner language:

- "With Katrin Wisniewski we explored linguistic correlates of the CEFR as part of the MERLIN project."
- "We characterize language development both for specific constructions, e.g., relative clauses (Alexopoulou, Geertzen, Korhonen & Meurers, 2015) and in terms of linguistic complexity, emphasizing the need to account for task effects (Alexopoulou, Michel, Murakami & Meurers, 2017)."
- "We also analyze L1 transfer effects using machine learning for Native Language Identification"

source: http://www.sfs.uni-tuebingen.de/icall/

NLP and Language Learning

What are some applications?

Broadly, four categories:

- language and content assessment
- learning support for reading, writing, speaking, and listening
- support tool for research on language acquisition, learner corpora etc.
- **student data analytics**

Student data analytics

- modeling student engagement through their activity (incl. postings etc)
- summarizing course feedback given by students
- visualization of learning program etc.
- recent work from SfS: "Enhancing a Web-based Language Tutoring System with Learning Analytics"

Summary

- NLP is used in a wide range of topics related to human language learning.
- From research to industry, there are many interesting problems to study and solve.
- At SfS, there is a strong group focusing on this kind of research - so talk to them!

Where to look to know more

- BEA workshop series
- NLP4CALL workshop series
- Talk to Prof. Meurers and team.
- Two summary articles (by prof and ex-advisee):
  - Detmar Meurers (2013, 2020). Natural Language Processing and Language Learning. The Encyclopedia of Applied Linguistics, edited by Carol A. Chapelle. Wiley.
  - Sowmya Vajjala (2018). Machine Learning in Applied Linguistics. The Encyclopedia of Applied Linguistics (ed: Carol Chapelle). Wiley. (Okay, I am not sharing out of vanity. It really gives an overview)

NLP without Annotated Dataset

NLP Careers

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

29 January 2021

Why now?

"Perhaps one topic that could be covered is company expectations when hiring and where things stand industry-wise."

I will try to ...

give a quick overview of...

- the kind of jobs out there
- typical expectations, interview processes
- how to keep track of what is going on

... based on my very limited knowledge of north american practices, but to German students!

Typical career paths SfS

(that I know of)

- Getting into a PhD program
- Joining IBM, Mercedes etc. (tech. companies, with potentially NLP work)
- Joining translation services companies (e.g., Transline in Reutlingen)
- Joining in general software engineer roles
- Linguist roles in tech companies or related ones

(SfS alumni are into all of these, in and outside Germany. Try to reach out for help/referrals!)

SfS Alumni

... are at:

- IBM, Amazon, Elastic (Elastic Search), Microsoft
- Mercedes, Daimler, Volkswagen
- Software AG
- Transline (at least used to be)
- Almato, GFai, Docyet, Carmeq etc
- Explosion.AI (Spacy)

Expectations: PhD applications

- Some background working in research projects
- General interest in pursuing research, reading up papers etc
- Some ideas to pursue research on
- Good written and spoken communication
- Reasonbably good programming skills.

PhD life

- Some course work (optional)
- work as a research assistant in some funded project
- think about what related problem can you solve for phd thesis
- write papers
- read papers
- give talks, attend conferences etc.

Expectations: Linguist Roles

- I have no direct idea.
- Companies like Appen have roles for linguists, focusing on data annotation.
- Companies like Grammarly have roles involving rule engineering (which needs some linguistics knowledge)
- Duolingo has some roles related to language learning exercises
- Google, Microsoft etc all hire linguists for data annotation, evaluation, error analysis kind of profiles

note: I know these through past students who applied to such positions. No direct experience.

Expectations: Tech companies

(for NLP projects)

- Very good programming skills (which language is often not too important)
- Problem solving skills: how do you approach a problem?
- Some understanding of Linguistics, NLP, Machine Learning, Deep Learning.
- Some prior experience working on some NLP research project

Expectations: Software Engineer roles

(more generic than the previous slide)

- Very good programming skills.
- Ability to convert requirements to code
- Ability to understand and follow good coding practices (version control, code review, code profiling, documentation etc)
- Learn and adapt quickly

Life in a company

- Initial years: you may write a lot of code every day, attend some meetings, probably read a lot
- mid-years: reduced coding, more design, more meetings, probably some presentations, conferences etc.
- later: further reduced hands on work, and more management work.

What interviewers may look for in companies

What I look for, when I interview:

- Can this candidate explain their course projects clearly?
- If I pose a similar (but not same) problem, can they propose a solution, relating this to those problems?
- If I ask for an alternative solution, can they think and give some ideas?
- Do they have some idea about different "methods", as well as different "application" scenarios, based on course work?

Specific Skills to prepare for

- questions related to resume/past work
- programming and software engineering questions
- machine learning/Deep learning/NLP: conceptual questions
- problem solving questions (mostly related to company's domain)
- statistics questions
- behavioral questions
- work culture related questions
- questions we can/should ask the interviewer

Source - my blog post about my own job search

General Notes

- Good programming skills are a must when starting out. Practice on websites like hackerrank.
- Be willing to keep learning and adapting to change more, at least in initial years out of Uni.
- Don't stick to NLP - see where else can you apply what you learnt in school and apply there too.
- It is useful to have a github profile with some project codes etc. if possible

Useful links

- https://nlppeople.com/ - I used to look here for jobs.
- Twitter - following NLP companies, researchers provide good source of information about current trends, job openings, careers etc.
- I follow newsletter.ruder.io (among others) for getting news about recent NLP research updates
- TowardsDataScience blog for more practical articles (how to use a library to do X etc)

NLP without Annotated Dataset

Course Review

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

29 January 2021

Topics we covered

1. NLP Overview
2. NLP system development pipeline
3. Corpus collection, extraction, exploration
4. Automatically labeling data
5. Snorkel: Spam Classification without annotated data, Data augmentation with annotated data
6. Text Embeddings and Transfer Learning: An overview
7. Other shorter topics: NLP for Endangered Languages, NLP for Language Learning
8. Group Discussions on various topics.

NLP Overview

- Different faces of NLP: Research, Industry, Other disciplines
- Various day to day applications
- Challenges with NLP
- Some common tasks and
- Degrees of language processing

NLP Pipeline

- How does one build NLP based software?
- How do we represent text for NLP?
- A real world case study
- A code walk through of the pipeline

Corpus collection, extraction, exploration

- How can we get data for NLP?
- How do we extract text from various file formats?
- What is inside our corpus?
- What are some issues to consider while developing/using a corpus?

Automatic Data Creation

- Different ways of labeling data
- Weak supervision
- Data augmentation
- Other topics:
  - Text anonymization
  - Text classification

Snorkel

- Generating training data for spam classification
- Using text augmentation for improving spam classification

Text Embeddings and Transfer Learning

- Bag of words to BERT
- What happens during fine-tuning?

Other Topics

- NLP for Endangered Languages
- NLP for Language Learning
- NLP careers
- Lot of interesting discussion papers

Graded Part

- Assignments
- Group discussion (DONE!)
- Term paper
- Classroom participation (DONE!?)

Some Ideas for Termpaper

I am taking text classification as an example a I covered it in class

- Compare different data augmentation methods for a simple text classification task
- Explore different *BERT models for a simple task
- Generating a labeled dataset for some new classification/IE task with Snorkel
- Surveying existing resources and software for a new language, and listing a few directions on how to extend them
- Exploring already existing datasets for its coverage, identifying potential issues with using it etc (e.g., following datasheets paper)
- Evaluating an existing tool (e.g., Spacy NER) in terms of how it does for various categories of text, using standard evaluation sets.

....

Initial ideas: Course Objectives

- Provide an overview of NLP system development pipeline
- Discuss some common approaches for collecting, cleaning and exploring text data
- Introduce some methods to develop labeled data for NLP

Initial ideas: Learning Outcomes

Students should be able to:

- Understand the end to end NLP system development pipeline
- Compile and explore labeled/annotated corpora for NLP
- Build some basic text classification and information extraction systems

... upon successful completion of the course..

Initial ideas: What the course can't do

- Don't expect to become an NLP expert with one compact course.
- Contents may not always meet your own expectations, but there is a term paper and a group discussion, which gives you opportunities to explore your specific interests related to this topic.
- The course won't teach you programming.

Upcoming Deadlines

- March 1st: Assignments, Term paper
- Today: Make a decision on whether or not you want to submit a term paper, and let me know! (thanks for those who already informed!)
- Soon: Start working on assignments and term paper!

How we learn and grow

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

One fourth from the teacher, one fourth from own intelligence,
One fourth from classmates, and one fourth only with time.

AchAryAt pAdamAdatte, pAdam shiShyaH swamedhayA |
sa-brahmachAribhyaH pAdam, pAdam kAlakrameNa cha ||

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

Source

Thank you!

contact: sowmya.vajjala @ nrc-cnrc.gc.ca

Feel free to connect on social networks (Linkedin, twitter etc) if you want to (finding me is up to you)

Wish you all good luck!!