Visualizing NLP in Undergraduate Students’ Learning about Natural Language

Cecilia O. Alm, Alex Hedges
Rochester Institute of Technology
One Lomb Memorial Drive
Rochester, NY 14623
coagla@rit.edu, aph3032@rit.edu

Abstract
We report on the use of open-source natural language processing capabilities in a web-based interface to allow undergraduate students to apply what they have learned about formal natural language structures. The learning activities encourage students to interpret data in new ways, think originally about natural language, and critique the back-end NLP models and algorithms visualized on the user front end. This work is of relevance to AI resources developed for education by focusing on inclusivity of students from many disciplinary backgrounds. Specifically, we comprehensively extended a web-based system with new resources. To test the students’ reactions to NLP analyses that offer insights into both the strengths and limitations of AI systems, we incorporated a range of automated analyses focused on language-independent processing or meaning representations which still represent challenges for NLP. We conducted a survey-based evaluation with students in open-ended case-based assignments in undergraduate coursework. Responses indicated that the students reinforced their knowledge, applied critical thinking about language and NLP applications, and used the application not to solve the assignment for them, but as a tool in their own effort to address the task. We further discuss how using interpretable visualizations of system decisions is an opportunity to learn about ethical issues in NLP, and how making AI systems interpretable may broaden multidisciplinary interest in AI in early educational experiences.

Introduction
It is of growing importance in AI fields to ensure broad access to deployed models, libraries, and other resources for users who can leverage their capabilities or explore their limitations. By enabling users to apply an automated system to analyze language data, without requiring technical knowledge in natural language processing, machine learning, or even computer programming experience, these resources can be scrutinized and used creatively in new areas for problem solving. One domain of use that promises impact is education in the cognitive sciences, where language is a primary data source, yet where students must face a shift in thinking about language and learn about the complexities of the underlying structures of natural language that are not overtly apparent to us as language speakers and users.

Litman (2016) noted the benefits for evaluating education-oriented NLP systems extrinsically in class settings, as opposed to intrinsic examination of the performance of a system component, or in a hypothetical user lab experiment. Our application use case includes language science undergraduate coursework, with the goal to understand whether the extended Linguine system\(^1\) is perceived as an effective learning mechanism when used with new and originally designed case study assignments. This study expanded on prior work, which provided a detailed system description and introduced the use of case studies (Alm, Meyers, and Prud’hommeaux 2017).

The system’s analyses apply open technologies, allowing students to access automated natural language processing resources. Students put them to use creatively with corpus data provided along with assignments. They used the resulting summaries about language features in their data or visual outputs with interpretable annotations to reflect on assigned problems, alongside the AI models, including their benefits, possible applications, and limits. Automated and machine learning-based analyses of corpus input data are at the heart of the system, whose outputs are displayed to the user as archivable visuals. The summaries and visuals intend to facilitate students’ own creative reasoning over the outputs and are created to be intuitively interpretable with brief clarifying text, e.g. when hovering over elements. Users may also download output for continued self-exploration.

This work makes two main contributions: (1) We describe our system, substantially extended for processing language-independent input and meaning-oriented natural language tasks. (2) We evaluate the learning experience by studying the robustness of the system innovations from the student perspective using novel learning materials in an IRB-approved study. In addition to reporting on the results from evaluation with pilot assignments, we further discuss and illustrate how students can explore critical issues, such as models’ prejudice and unfairness, when using this framework in their learning.

Relevant Prior Work
Few examples of prior work have tried to make computational tools that process language more widely ac-

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

\(^1\)tinyurl.com/linguinedemo2020, github.com/ritlinguine
Figure 1: An example of our system’s summarized and visual output of automated coreference analysis showing a passage from a dialogue turn from Kubrick’s *2001: A Space Odyssey* with a list of predicted, isolated coreference chains. Items sharing color and index refer to the same referent. Coreference analysis allows students to explore the often complex structure of referents in natural language discourse and to consider how this applies to variation and ambiguity in language use. Moreover, while the shown output is arguably correct, the coreference task involves many challenges for NLP systems, revealing to students the limitations of current NLP, and also inviting discussion of such systems’ consideration of or insensitivity to non-binary gender.

Figure 2: As an additional example, students can apply unsupervised morphology induction on a sufficiently large corpus and then visually inspect inferred suffixes with example word lists, two of which are shown here. Students can further reason with the system results and the broader data that -i and -ün represent meaningful grammatical affixes in the language under analysis (accusative case and superlative form, respectively) and begin to construct a grammar of an unknown language. (For some figures in the paper, results are transposed horizontally. The colors have been adjusted.)

Some open-source software, such as AllenNLP and Stanford CoreNLP (Gardner et al. 2018; Manning et al. 2014), provide online demos for their tools that display transient visualizations for short passages entered in a text box. This work differs by providing a web-based application that provides students the ability to store entire documents or concatenated text collections for analyses. The system incorporates multiple software packages for applying analyses, and stores a student’s resulting analyses and their visualized outputs for their later usage.

Litman (2016) surveyed three roles that NLP has played in the educational domain. Our evaluation study did not neatly fit into any category and rather straddles two of them (“Teaching and learning language-related subject matter” and “Using language to teach any subject”; p. 4170). We also focus on enhanced interpretability with visual output, an important interaction feature for human-attuned AI systems and especially important for enabling broader, non-expert student groups access as AI users in educational environments.
The Linguine application has a web front end that provides students access to tools and models, intuitive input of language data, and selection of analysis and pre-processing options, resulting in understandable output (Alm, Meyers, and Prud'hommeaux 2017). It is accessed through a browser and only requires point-and-click interaction, thus enabling access to open resources or system-customized programs that otherwise would demand that users have technical knowledge (Bird, Klein, and Loper 2009a; Lee and Goldsmith 2016; Manning et al. 2014; Meyers 2017; Řehůřek and Sojka 2010). Users with some technical background may transfer the results from completed automated analyses system-externally as JSON files for further self-programmed statistical or computational analysis and processing. Thus, in the employed educational use case, the application makes available data and data analysis functionalities, while case study assignments provide students with the task they are instructed to resolve.

The application incorporates new analyses with original visual summaries or visuals of linguistic sequences or examples to allow users to explore several computational natural language semantics tasks: coreference resolution (Manning et al. 2014) (Figure 1), topic modeling (Řehůřek and Sojka 2010), and word embedding operations (Pennington, Socher, and Manning 2014; Řehůřek and Sojka 2010) such as similarity scores (Figure 3, top) and vector equations (Figure 4).

Examples of language-independent analyses include unsupervised morpheme induction (Lee and Goldsmith 2016) (Figure 2), topics inferred from long texts, summary statistics on linguistic features, and glyph-based bigram arrays for studying written or transcribed spoken language. Students can also explore a range of features in temporarily word-aligned transcribed spoken language (Figure 3, bottom), including multi-party dialogues. That analysis works using a time-stamped input representation of transcribed language. As a pre-processing step for English recordings, we used CMU Sphinx-4 (Sphinx-4 Team 2016) to automatically transcribe recordings, called with the Sphinx4-HTTP-server wrapper (Jitsi 2017).

**Case Study Evaluation**

As a pedagogical tool, case studies use situation-based narratives to elicit problem-solving performance (Alm, Meyers, and Prud’hommeaux 2017). For evaluation, we used case study tasks allowing students in mixed-major coursework to practice theoretical concepts, methods, and tasks from language science while providing opportunities for idea gener-
Figure 5: Response frequency to survey statements on a 5-point scale for *Writing System of an Unknown Language* that began with *Using the provided web tools and input...* (top) and *This case study activity...* (bottom). Both were dominated by *strongly agree* or *somewhat agree* responses indicating that these participating students reacted positively to their experience and interactions with the system.
C1: Writing System of an Unknown Language

The students completing this case with the NLP-driven application took part in an entry-level non-technical course about language technologies. The assignment required them to apply concepts seen in lectures or textbooks (Dickinson, Brew, and Meurers 2013; Sproat 2010). The case narrative with the task statement was as follows, and after it came checklist-like instructions about which NLP analyses to complete:

Historians are doing archival research in a library when they come across a corpus of texts in an unknown language. The books are written in Latin script with text that resembles the writing system of a language, but there is no record of the language these texts are written in. You are contacted to consult on fundamental linguistic characteristics about the writing system and the language, such as letter-sound units, syllable structure and phonotactics, and initial work on the underlying language’s morphology and language typology before decipherment is attempted.

This case resembles the “Rosetta Stone” exercise format where features of an unknown linguistic system are derived with analysis or comparisons to translated examples (Bozhanov and Derzhanski 2013). The case study asks students to explore features related to linguistic units’ length, different types of n-grams, and morphology induction.

Students were given short and long texts in a constructed language they did not know but which also had a substantial corpus in the open domain (Wikipedia Contributors 2017), to enable them to apply automated and unsupervised language-independent NLP techniques. Students applied a suite of language-independent analyses in the system, while they considered questions that guided them through data exploration. They were instructed not to search the web because that could reveal the identity of the language. To kick-start problem-solving, news items and scholarly readings were offered (Hardesty 2010; Hermjakob et al. 2018; Hohn 2013; Snyder, Barzilay, and Knight 2010). Thus, with this assignment students revisited concepts on writing systems, which had introduced notions associated with language technologies early in the term, yet re-enforced the understanding they had built later in the term about modern language technology.

Survey responses (n = 12) revealed that the students perceived the system was useful (Figure 5). For example, all strongly agreed or somewhat agreed that using the system for the case re-enforced theoretical concepts from class with an application, let them use linguistic approaches to problem solving, and had them engaged in critical thinking. In addition, 83% strongly agreed or somewhat agreed that it was a valuable learning experience. However, some disagreed that the case task involved a reasonable time commitment and involved a useful reporting experience. Open-ended comments suggested that the task and system enhanced their learning, enabled students to pursue independent thinking, and apply previously learned concepts, as articulated in student quotes in Table 1. Furthermore, students requested having more time for discussion with their peers.

C2: Dialogue Adaptation

Applying mostly language understanding analyses, students in a small English language history class comparatively examined Present-Day English (PDE) natural dialogues with Early Modern English (EME) Shakespearean literary dialogues (Bird, Klein, and Loper 2009b; Shakespeare 1997). They were also given recommended readings (Delabastita 2017; Shapiro 2015), and their assignment was introduced with a case study narrative, followed by guided application of NLP-based semantic and unsupervised language-independent analyses with this exploration leading into creative dialogue adaptation from EME to PDE. The case study narrative follows:

The studio Peter Quince Pictures is creating a new adaptation of Romeo and Juliet. The film team wants it to be set in the modern day, but unlike the similar 1996 adaptation Romeo + Juliet, which retained the Early Modern English (EME), they want the dialogue to use Present-Day English (PDE). You are hired as a linguistic consultant to help with this adaptation process. Using information from your analysis of Shakespeare’s text and real PDE dialogue, you are tasked to make recommendations to the film team about how to accomplish this adaptation to current language.

The case study asks students to use analyses for length statistics of linguistic units, word frequency, topic modeling, coreference resolution, and word vector operations.
What did you like about this case study?
I appreciated how the steps escalate from visual gleaning, to n-gram analysis, to full-blown sentence structure analysis, and how a minimal amount of hand-holding is done along the way, allowing us to explore the writing system to our content. [KNOWLEDGE GAIN]

What could be improved with this case study?
I feel like a clearer goal would be helpful, like a chart at the end that needed to be filled out with your [r] assumptions or hypotheses. [DELIVERABLES]

Any additional comments about the case study or technology?
It was a bit unclear what the “Solution” we were attempting to find was; [I] assumed it was just an overview of certain aspects of the language, but it might be good to revise some of the language around the case study, as it was unclear if there was a single deliverable solution to be found.

I was able to see real progress in figuring out key characteristics about the unknown writing system. [KNOWLEDGE GAIN]

It felt a bit rushed, so maybe two full weeks would have been a bit better. [TASK TIME]

It was good! I think we should do multiple, it was great practice of what we’ve learned. [PEER LEARNING]

It was cool to see how each group had a slightly different interpretation, despite having looked at the exact same data. [CRITICAL THOUGHT]

More discussion after presentations, we all had unique approaches and it would have been cool to explore the benefits of each more. [PEER LEARNING]

Overall a great experience, I hope [the university] continues to develop [the system] in the future to make it more robust and feature-full.

Table 1: Select open-ended replies for C1 Writing System of an Unknown Language suggest general student satisfaction and point to improvements. Other example improvement suggestions involved readings or the optional JSON output export feature.

<table>
<thead>
<tr>
<th>What did you like about this case study?</th>
<th>What could be improved with this case study?</th>
<th>Any additional comments about the case study or technology?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I appreciated how the steps escalate from visual gleaning, to n-gram analysis, to full-blown sentence structure analysis, and how a minimal amount of hand-holding is done along the way, allowing us to explore the writing system to our content.</td>
<td>I feel like a clearer goal would be helpful, like a chart at the end that needed to be filled out with your [r] assumptions or hypotheses.</td>
<td>It was a bit unclear what the “Solution” we were attempting to find was; [I] assumed it was just an overview of certain aspects of the language, but it might be good to revise some of the language around the case study, as it was unclear if there was a single deliverable solution to be found.</td>
</tr>
<tr>
<td>I was able to see real progress in figuring out key characteristics about the unknown writing system.</td>
<td>It felt a bit rushed, so maybe two full weeks would have been a bit better.</td>
<td>It was good! I think we should do multiple, it was great practice of what we’ve learned.</td>
</tr>
<tr>
<td>It was cool to see how each group had a slightly different interpretation, despite having looked at the exact same data.</td>
<td>More discussion after presentations, we all had unique approaches and it would have been cool to explore the benefits of each more.</td>
<td>Overall a great experience, I hope [the university] continues to develop [the system] in the future to make it more robust and feature-full.</td>
</tr>
</tbody>
</table>

Among these survey respondents (n = 5), 80–100% strongly agreed or somewhat agreed on 13 of the 15 statements (same as in Figure 5), 60% concurred it was a valuable learning experience, but only 20% agreed with the statement had clear instructions. The latter may reflect the task’s open-ended nature or its focus on linguistic meaning as opposed to linguistic structure. Three students commented on what integrating the use of the NLP-based system afforded: I learned different methods of [automatically] analyzing languages, including topic modeling, word frequency, sentence length, and word vectors. I also learned what a coreference is, including anaphora, cataphora, and split antecedent, which were all new concepts to me. And I liked using the [...] tool and writing custom commands for the word vector analysis. As well as I liked being able to think about and apply the tool to a literary work like Romeo and Juliet. I also liked the helpful representations of analysis provided by [the system] [...]. Students’ suggested improvements for the case included extending the 3-page report limit and providing clearer instructions for the assignment.

Collectively, students using the NLP-driven system felt that it was interesting and advanced understanding of concepts and data methods for linguistics. It motivated them to engage in critical, creative, and collaborative problem-solving, and it stimulated an interest in discussing with peers the different problem-solving and decision-making paths that could be followed. The assignment assessment also stimulated reflections about language technology applications, including outside linguistics. Engagement with the visual output and data representations derived secondarily by students from system output helped students compare the distinct approaches they took in interpreting data and seeking solutions using evidence-based, data-driven, and visual observations. However, the more open-ended assignment (C2) was perceived a bit less successful, and some students wanted more information on what solving the task meant for the case studies.

A limitation of the present evaluation study was that the two groups of students participating in the evaluation were modest (12 and 5 students, respectively), and thus results must be interpreted with caution. In addition, the evaluation did not include a control group that did not receive the educational NLP-driven application intervention to compare with quantitatively. Nevertheless, the groups represented two different undergraduate linguistics course settings and qualitative responses suggested that the system afforded an original and generally positive learning experience.

Educational Exploration of NLP Model Bias

Developing critical thinking about the affordances and limitations of NLP-based models is important, e.g. to avoid overly optimistic characterizations of NLP systems; cf. Bendor and Koller (2020). This is essential both for future users of AI applications or AI-based decision-support systems and for computing majors before they enter into the community as professionals and developers, or continue on to graduate school and research education. Human bias, prejudice, and stereotypes in automated models are receiving attention across NLP and AI communities (Bhaskaran and Bhallamudi 2019; Sun et al. 2019).

The visual component of our system enables users to qualitatively explore problematic bias in machine learning-based models. For example, as fundamental units of many predictive NLP systems, word embeddings have come under scrutiny for gender bias and other forms of prejudice (Caliskan, Bryson, and Narayanan 2017; Oonen and Goldberg 2019). Figure 3 (right, top image) suggests lower similarity for gender-biased result for (man) and (woman) than for (doctor) and (nurse); though this gender-bias discrepancy is not noted when (nurse) is replaced by (doctor) (middle, top image). Nonetheless, as another example, Figure 4 (right) shows a gender-biased result for (doctor) when used in other vector operations. Returning to automated coreference, Figure 6 also shows how the model used in the analysis fails to link
The coreference resolution model identifies he as a physician but not she, suggesting a gender bias in the underlying coreference resolution model that generated these predictions. Sentences adopted from a presentation by Goldwater (2019).

physician to the female pronoun she (left), but not to the male pronoun he (right).

Exploring questions of demographic bias in NLP output can help students develop critical thinking about the ethical issues impacting data-driven NLP and machine learning applications, and comparatively examine how NLP systems learn from different data sources. We can use the case study-based process applied in the present work to prepare assignments that allow students to explore model prejudice, and we can add quantitative analysis into the NLP-driven system of bias severity and use visual indicators to aid interpretation. The example of ethical exploration also highlights a potential for our application’s use in undergraduate project-based learning.

Conclusion

While non-technical users may focus on consumption of AI technologies rather than their production—cf. Langley (2019) for a discussion in the context of AI coursework—the expanded system we described has potential to stimulate curiosity in and nurture thoughtful reflection and realistic expectations about NLP.

Summatively, students using the system with case assignments appreciated the system and felt it was interesting and advanced understanding of concepts and the use of data-driven methods for studying natural language forms and functions in language science classrooms. It also enabled students to engage in critical exploration and problem-solving. The system’s use stimulated reflections about intelligent systems, and engagement with the visual data representations in the system and those derived secondarily by students from system output fed into students’ considering how to communicate observations and the importance of visualizing data.

A potential limitation of this study is that survey instruments used quite broad as opposed to more specific questions; in the future, questions about particular system functionalities or case study assignment components could be added. Additional inquiries left for future work include whether visual output formats support students’ learning, as well as to which degree system components (such as different NLP analysis alternatives), or which types of NLP-based visuals, support or hinder interpretation of data or students’ learning process. In addition, other work points to the effectiveness of adjusting visual information based on the user (Conati et al. 2015), to aid students to take better advantage of the visual content. Future system evaluation can also measure how users attend to visuals by leveraging combined analysis of peer-to-peer discussions and eye-tracking.

Another step to explore is what it takes to expand to new learning contexts, possibly involving other age groups, where intuitive broad access to exploring AI systems could constitute a vehicle toward educational uses that convey or contest some of the “big ideas” of AI (Touretzky et al. 2019).

Acknowledgments

Many thanks to all collaborators on prior iterations of Linguine. An RIT Provost’s Learning Innovations Grant award supported the current work.

References


Madnani, N.; and Dorr, B. J. 2008. Combining Open-Source with Research to Re-Engineer a Hands-on Introductory NLP


