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-

[1]tinyurl.com/linguinedemo2020, github.com/ritlinguine

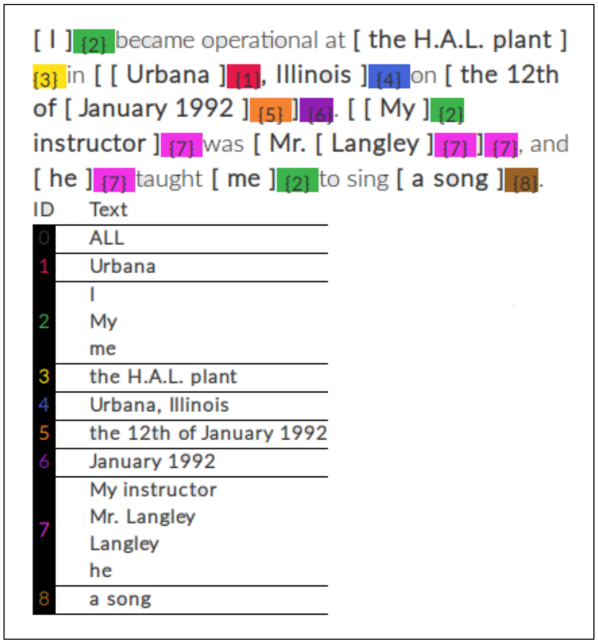

Figure 1: An example of our system's summarized and visual output of automated coreference analysis showing a passage from a dialogue turn from Kubrick's *2001: A Space Odyssey* with a list of predicted, isolated coreference chains. Items sharing color and index refer to the same referent. Coreference 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.

Suffixes	Stems (Sample)
NULL i	alik, atos, belem, bukil, desin, fomam, gümnad, kelos, lamerikänik, lifalejenöf, lukrayän, lulak, luruguyän, prepod, uyun

Suffixes	Stems (Sample)
NULL ün	balik, blefik, distik, donik, legretik, liegik, lunüpik, natöfik, siämik, vönädik, yunik

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.)

cessible. One example is the Speech Recognition Virtual Kitchen Toolkit, which contains virtual machine images with difficult-to-setup NLP software preconfigured (Bates and Kim 2016). However, this resource still requires setting up a virtual machine and having command-line user knowledge. More specifically for computing majors, the instructional barrier for NLP applications has been lowered by letting students work on development with proprietary software after completing prerequisite AI or data structures coursework (Wollowski 2016). For majors taking AI coursework, efforts have also been made to integrate viewable content and interactive demonstrations to foster learning (Singh and Riedel 2016).

Several studies discuss introductory natural language processing courses for interdisciplinary or other majors (Agarwal 2013; Cassell and Stone 2005; Hockey and Christian 2008; Liddy and McCracken 2005; Madnani and Dorr 2008). Yet, they tended to require prior programming knowledge, or alternatively spent class-time teaching higher-level programming languages such as Python or Prolog. Open-source software, such as NLTK (Bird, Klein, and Loper 2009a), was used but proprietary licenses or setup issues affected some courses by preventing straightforwardly using the programs. In contrast, this work avoids both issues by removing the hurdle of requiring programming knowledge and relying on open source libraries, web-based use, and visual output.

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.

Word 1	Word 2	Score
universities	colleges	0.92159140

Word 1	Word 2	Score
woman	doctor	0.72527343

Word 1	Word 2	Score
woman	nurse	0.71550202

Word 1	Word 2	Score
universities	fish	0.13112722

Word 1	Word 2	Score
man	doctor	0.71195793

Word 1	Word 2	Score
man	nurse	0.57187039

Transcription:

[SIL] that [SIL] that [SIL] first one goes to secretary Clinton because [SIL] you started out the last one to the audience [SIL] he wants to start he can start [SIL] no go ahead Donald [SIL] no I'm a [SIL] gentleman Hillary go ahead [SIL] secretary Clinton [SIL] well I think Donald was about to say he's going to solve it [SIL] by repealing it [SIL] and getting rid of uh [SIL] the affordable care act [SIL] and [SIL] I'm going to fix it [SIL] because I agree with you premiums have gotten too high copays deductibles prescription drug cost

Statistics:

Calculation	Value
Number of non-word fillers:	16
Number of words:	84
Total time:	27.5
Non-word filler time:	7.157
Word time:	19.693
Words per minute:	183.27
Syllables per minute:	270.55

Longest Words:

Word	Length (seconds)
deductibles	0.728
copays	0.651
prescription	0.5
cost	0.495
repealing	0.491
audience	0.473
premiums	0.422
uh	0.418
and	0.402
and	0.394

Figure 3: Top: Visualizing similarity for word pairs that are semantically similar (left-top) or dissimilar (left-bottom) or explore presence of gender bias (middle vs. right columns) based on contrasting similarity scores for pairs of word embeddings. Bottom: Summarizing features of transcribed and temporally word-aligned excerpt from a 2016 US presidential debate.

Word	Operation
woman	+
king	+
man	-

Word	Operation
strong	+
darker	+
dark	-

Word	Operation
woman	+
doctor	+
man	-

Word	Score	Word	Score	Word	Score
queen	0.85236037	stronger	0.83253789	nurse	0.84046429
throne	0.76643336	contrast	0.78543246	child	0.76632601
prince	0.75921446	robust	0.78284526	pregnant	0.75701302
daughter	0.74738830	reflected	0.76698452	mother	0.75174582
elizabeth	0.74602205	strongest	0.76484495	patient	0.75166631
princess	0.74245703	reflects	0.76359135	physician	0.75072813

Figure 4: Classical examples for nouns (left), positive and comparative adjectives (middle), and additionally for gender bias (right). Students formulate vector equations and see top-scoring similar words from pre-trained word embeddings.

System Overview and Extensions

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 know-how (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 temporally 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-

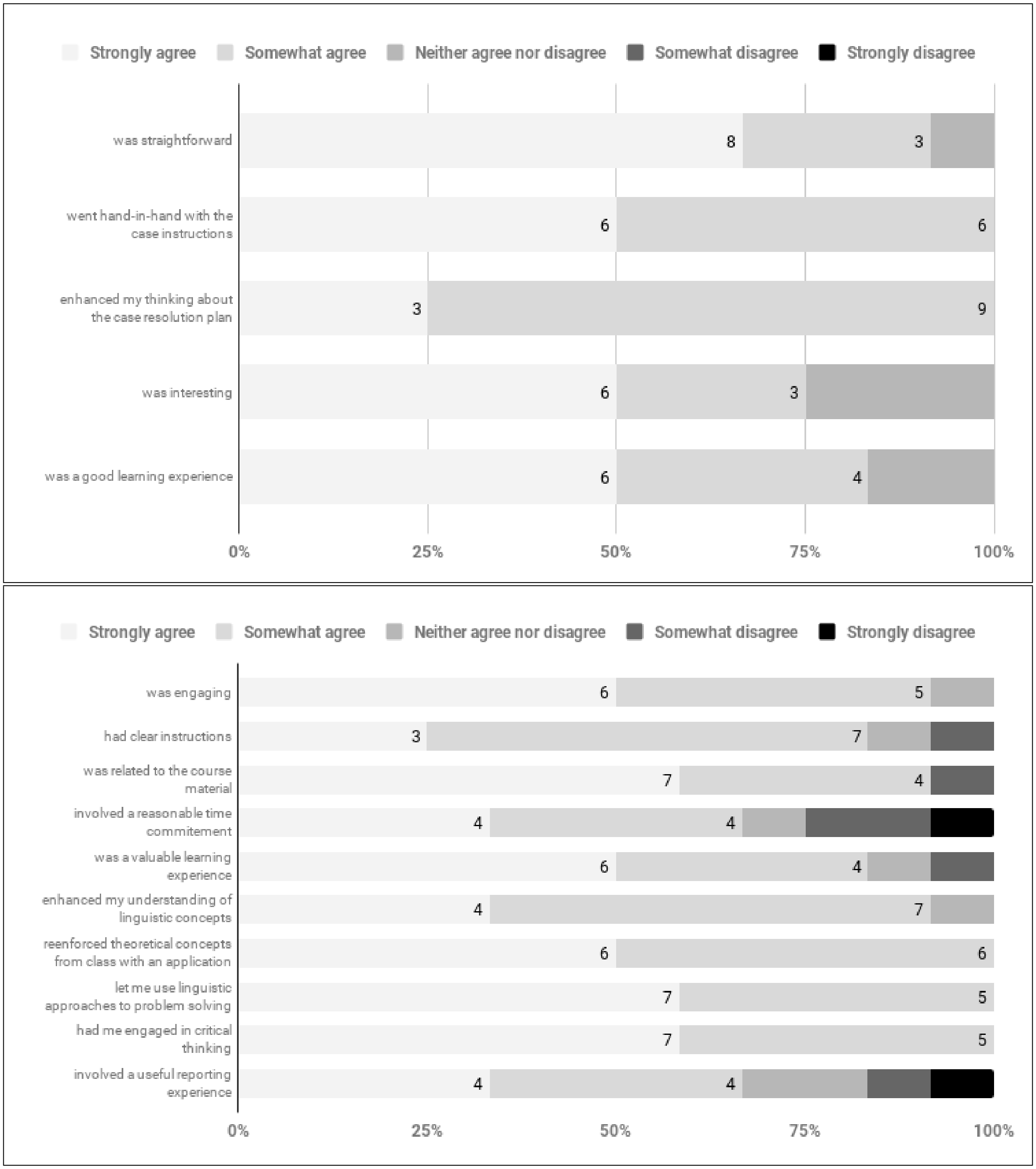

Figure 5: Response frequency to survey statements on a 5-point scale for *Writing System of an Unknown Language* that began with *Using the provided web tools and input...* (top) and *This case study activity...* (bottom). Both were dominated by *strongly agree* or *somewhat agree* responses indicating that these participating students reacted positively to their experience and interactions with the system.

ation and reflection both on natural language linguistics and the usefulness of the application, potentially also growing an interest in natural language processing and bringing students from multiple backgrounds into this AI discipline.

In each of two classes, students were presented with an activity shaped as a human-AI collaborative task, where the provided application analyzed data to support the students as they developed a recommended solution to the case problem. Since it was a case study, students were given a narrative scenario involving a practical application. They were also given data and instructions on how to use the application online to complete automated NLP analyses with guiding questions. Assessment instructions provided criteria for the written reporting and presentation that students were expected to produce, and task objectives for the case resolution. NLP analyses alone did not provide the solution; rather students used their output as data analysis combined with their own interpretation. They were encouraged to use their own creativity and reason over the results of the NLP system, deepening the system's analyses with human inference on top of automated results. They also decided how to present and discuss their solution. After the assignment, students completed a survey that was inspected only after the course had been fully completed to mitigate any potential influence on the perception of student performance. Students consented to participating in this IRB-approved study. The evaluation focused on the students' survey responses because confounding factors in a classroom setting makes assignment scores or course performance unsuitable measures. Next we introduce each case study and their corresponding results from student surveys.

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?	*What could be improved with this case study?*	*Any additional comments about the case study or technology?*
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.[TASK FORMAT]	I feel like a clearer goal would be helpful, like a chart at the end that needed to be filled out with you[r] assumptions or hypotheses.[DELIVERABLES]	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.
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.

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. Bender 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; Gonen and Goldberg 2019). Figure 3 (right, top image) suggests lower similarity for the word embeddings for *man* and *nurse* than for *woman* 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

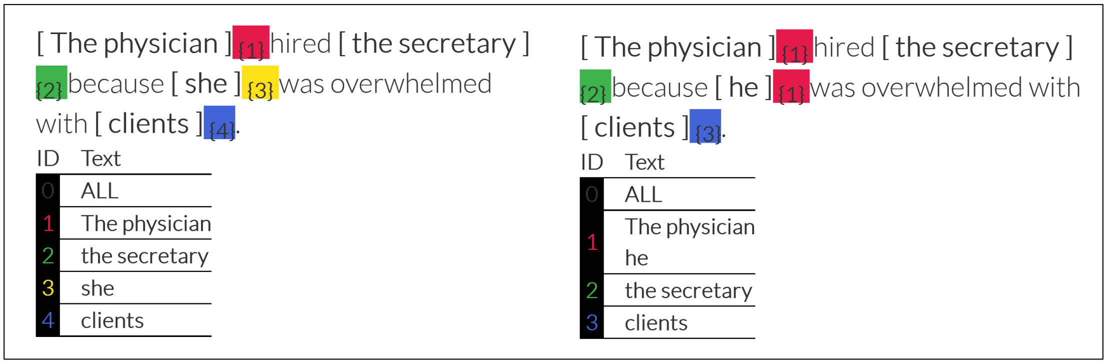

Figure 6: The coreference resolution model identifies *he* as a *physician* but not *she*, suggesting a gender bias in the underlying coreference resolution model that generated these predictions. Sentences adopted from a presentation by Goldwater (2019).

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

Agarwal, A. 2013. Teaching the Basics of NLP and ML in an Introductory Course to Information Science. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, 77–84. Sofia, Bulgaria: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3412.

Alm, C. O.; Meyers, B.; and Prud'hommeaux, E. 2017. An Analysis and Visualization Tool for Case Study Learning of Linguistic Concepts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 13–18. Association for Computational Linguistics. doi:10.18653/v1/D17-2003. URL http://aclweb.org/anthology/D17-2003.

Bates, R.; and Kim, E. 2016. The Speech Recognition Virtual Kitchen Toolkit. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, 162–162. New York, NY, USA: ACM. ISBN 978-1-4503-3685-7. doi:10.1145/2839509.2850506. URL http://doi.acm.org/10.1145/2839509.2850506.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of* *the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://www.aclweb.org/anthology/2020.acl-main.463.

Bhaskaran, J.; and Bhallamudi, I. 2019. Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. *CoRR* abs/1906.10256. URL http://arxiv.org/abs/1906.10256.

Bird, S.; Klein, E.; and Loper, E. 2009a. *Natural Language Processing with Python*. O'Reilly Media, Inc. ISBN 0596516495, 9780596516499.

Bird, S.; Klein, E.; and Loper, E. 2009b. *Natural Language Processing with Python [NLTK data: Switchboard corpus excerpt]*. O'Reilly Media, Inc. ISBN 0596516495, 9780596516499.

Bozhanov, B.; and Derzhanski, I. 2013. Rosetta Stone Linguistic Problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, 1–8. Sofia, Bulgaria: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3401.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science* 356(6334): 183–186. ISSN 0036-8075. doi:10.1126/science.aal4230. URL https://science.sciencemag.org/content/356/6334/183.

Cassell, J.; and Stone, M. 2005. Teaching Dialogue to Interdisciplinary Teams through Toolkits. In *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL*, 9–14. Ann Arbor, Michigan: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-0102.

Conati, C.; Carenini, G.; Toker, D.; and Lallé, S. 2015. Towards User-Adaptive Information Visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 4100–4106. AAAI Press. ISBN 0-262-51129-0. URL http://dl.acm.org/citation.cfm?id=2888116.2888291.

Delabastita, D. 2017. 'He shall signify from time to time'. *Romeo and Juliet* in Modern English. *Perspectives* 25(2): 189–213. doi:10.1080/0907676x.2016.1234491.

Dickinson, M.; Brew, C.; and Meurers, D. 2013. *Language and Computers*. West Sussex, UK: Wiley-Blackwell.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/W18-2501. URL https://www.aclweb.org/anthology/W18-2501.

Goldwater, S. 2019. Gender Bias in NLP. URL https://www.inf.ed.ac.uk/teaching/courses/anlp/slides/anlp29-2x2.pdf. [Presentation accessed online in January 2020.].

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 609–614. URL https://aclweb.org/anthology/papers/N/N19/N19-1061/.

Hardesty, L. 2010. Computer Automatically Deciphers Ancient Language. In *MIT News*. Massachusetts Institute of Technology. URL https://news.mit.edu/2010/ugaritic-barzilay-0630.

Hermjakob, U.; May, J.; Pust, M.; and Knight, K. 2018. Translating a Language You Don't Know in the Chinese Room. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, 62–67. Association for Computational Linguistics. URL http://aclweb.org/anthology/P18-4011.

Hockey, B. A.; and Christian, G. 2008. Zero to Spoken Dialogue System in One Quarter: Teaching Computational Linguistics to Linguists Using Regulus. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, 80–86. Columbus, Ohio: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W08-0210.

Hohn, D. 2013. Known Unknowns. In *The New York Times*. The New York Times Company. URL https://www.nytimes.com/2013/06/16/books/review/margalit-foxs-riddle-of-the-labyrinth.html.

Jitsi. 2017. Sphinx4-HTTP-server. URL https://github.com/jitsi/Sphinx4-HTTP-server. [Online; last build May 2017].

Langley, P. 2019. An Integrative Framework for Artificial Intelligence Education. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1): 9670–9677. doi:10.1609/aaai.v33i01.33019670. URL https://aaai.org/ojs/index.php/AAAI/article/view/5032.

Lee, J. L.; and Goldsmith, J. A. 2016. Linguistica 5: Unsupervised Learning of Linguistic Structure. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, 22–26. San Diego, California: Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-3005.

Liddy, E. D.; and McCracken, N. J. 2005. Hands-on NLP for an Interdisciplinary Audience. In *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, TeachNLP '05, 62–68. Stroudsburg, PA, USA: Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1627291.1627305.

Litman, D. 2016. Natural Language Processing for Enhancing Teaching and Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 4170–4176. AAAI Press. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12310.

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

Course. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, 71–79. Columbus, Ohio: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W08-0209.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. doi:10.3115/v1/P14-5010. URL http://aclweb.org/anthology/P14-5010.

Meyers, B. 2017. SPLAT: Speech Processing & Linguistic Analysis Tool. URL https://github.com/meyersbs/SPLAT. [Online; last build November 2017].

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. URL http://www.aclweb.org/anthology/D14-1162.

Řehůřek, R.; and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: Elra. URL http://is.muni.cz/publication/884893/en.

Shakespeare, W. 1997. *The Tragedy of Romeo and Juliet*. Project Gutenberg. URL https://www.gutenberg.org/ebooks/1112.

Shapiro, J. 2015. Shakespeare in Modern English? In *The New York Times*. The New York Times Company. URL https://www.nytimes.com/2015/10/07/opinion/shakespeare-in-modern-english.html.

Singh, S.; and Riedel, S. 2016. Creating Interactive and Visual Educational Resources for AI. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 4101–4106. AAAI Press. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11794.

Snyder, B.; Barzilay, R.; and Knight, K. 2010. A Statistical Model for Lost Language Decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 1048–1057. Stroudsburg, PA, USA: Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858681.1858788.

Sphinx-4 Team. 2016. Sphinx-4 Speech Recognition System. URL https://github.com/cmusphinx/sphinx4. [Online; last build May 2016].

Sproat, R. 2010. *Language, Technology, and Society*. Oxford: Oxford University Press.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E. M.; Chang, K.; and Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 1630–1640. URL https://www.aclweb.org/anthology/P19-1159/.

Touretzky, D.; Gardner-McCune, C.; Martin, F.; and Seehorn, D. 2019. Envisioning AI for K-12: What Should Every Child Know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1): 9795–9799. doi:10.1609/aaai.v33i01.33019795. URL https://aaai.org/ojs/index.php/AAAI/article/view/5053.

Wikipedia Contributors. 2017. Wikimedia Downloads. URL https://dumps.wikimedia.org/. [Online; accessed November 2018].

Wollowski, M. 2016. Training Watson - A Cognitive Systems Course. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 4136–4137. AAAI Press. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12462.