CS4248 Natural Language Processing
Staff Proposed Projects

v210226 – Initial commit.

This document describes the staff proposed projects for NUS CS4248 Natural Language Processing, as proposed for Semester 2020 (Academic Year 2020/2021, Semester 2). Most of these projects are of particular interest to the lecturer's research group (the Web IR and NLP Group; WING, `http://wing.comp.nus.edu.sg`), and could be extended into publishable work or a longer-term project. The listing of references is purposefully incomplete and limited to work done at NUS; it is not meant to be representative of all of the relevant work, rather as an entryway to discovering the appropriate references. Most references can be found open access through a web search of the title.

Many of these tasks have been done in the pre-deep learning NLP era, and with sufficient data, some may be suitable to be re-examined with deep learning methodologies for classification and/or sequence labeling.
You will note that all of the projects are still open ended as specified. With respect to project proposals, if your team decides to take up any of the below projects, you will need to further customize your project to better define the dimensions of your investigation.

1 Webpage classification without the webpage

Webpage classification is the multiclass classification task of assigning a webpage to a certain topic category. The search engine Yahoo! built a taxonomy of categories and in its early days, employed humans to classify websites to categories. Typical information used to classify webpages to categories use the link structure (hyperlinks) in the page as well as those pointing to it, as well as the semi-structured content on the page itself.

In this project, we limit ourselves to using only the uniform resource locator (URL) string of a webpage for classification. As the URL does not have spaces, this requires good tokenization and a good understanding of how URL tokens give rise to certain topical associations.

This is a standard, supervised multiclass classification task, as we consider taking an input URL u of a webpage (may not even exist), and assigning it to a class c, of potentially a large set of classes. This is a good project for those interested in a smaller project that can be easily completed with simple methods, yet has a very extensible set of means of improvement using suitably trained word and subword embeddings. Catalogues of URLs are easily obtained from web sources and both gold-standard ground truth (from web catalogs and hierarchical reference sites (e.g., Wikipedia) as well as obtainable self-supervised bronze data can be used as side information to potentially improve classification. This task is easily scalable, given the small size of input data.

Extensions could include also investigating the improvements using other information (content of the webpage or both in- and out- citations), aligning multiple categories of ground truth, or hierarchical classification.

1.1 References

- Hashemi, M. (2020) *Web page classification: a survey of perspectives, gaps, and future directions*. Multimed Tools Appl 79, 11921–11945 `https://doi.org/10.1007/s11042-019-08373-8`
- Min-Yen Kan and Hoang Oanh Nguyen Thi (2005) *Fast webpage classification using URL features*. In Proc. of Conf. on Info and Knowledge Management (CIKM '05). Bremen, Germany, November 2005. Poster Paper. pp. 325-236.
- Min-Yen Kan (2004) *Web Page Classification Without the Web Page*. In Proceedings of the 13th International World Wide Web Conference (WWW2004), May 2004. New York, New York, USA. Poster Paper.

2 Instructor intervention prediction in online forums

Discussion forums, such as those found on Slack, Piazza or LumiNUS, are a primary means for helping students and teaching staff communicate with each other. However, most discussion forums threads are visualized simply using a *most-recently-updated* rule. If there are important or timely threads that are worth paying attention but whose placement in the visualization have been superseded by less important threads, these can be missed by both teaching staff and students alike. As such, when forums get large, such discussion forums become difficult to manage.

In this project, we take the viewpoint of an instructor of a large class (i.e., a massive online open class; MOOC). We build a binary classifier of whether an instructor will intervene on a thread in a particular state – having X posts with its particular content. We can do this by using historically completed courses, and observing when an instructing staff decides to reply to a student post. This assumes that what instructors have actually intervened on are actually important conversations.

This project is a bit more open-ended as there are a lot of potential issues to examine in both the natural language, social network and forum structure. It is more suitable for teams that have different interests. Aside from the classification, teams may find it useful to model the sequential information in subsequent posts (sequence labeling; see Project 3). There is also more flexibility and the prerequisite in the engineering of crawlers for additional data. There is also currently a student in my research group, who is also our CS4248 teaching assistant, ZHANG Tianyang, who may be able to provide some simple pointers on your team's work.

2.1 References

- Muthu Kumar Chandrasekaran and Min-Yen Kan (2018) *Countering Position Bias in Instructor Interventions in MOOC Discussion Forums*, In The 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia: July 2018.
- Muthu Kumar Chandrasekaran, Carrie Demmans Epp, Min-Yen Kan and Diane Litman (2017). *Using Discourse Signals for Robust Instructor Intervention Prediction.* In Proceedings of the Thirty-First AAAI conference on Artificial Intelligence (AAAI-17), San Francisco, USA, 3415-3421, AAAI.
- Muthu Kumar Chandrasekaran, Kiruthika Ragupathi, Min-Yen Kan, and Bernard C. Y. Tan (2015). *Towards Feasible Instructor Intervention in MOOC Discussion Forums.* In Proceedings of 36th International Conference on Information Systems, (ICIS '15), Forth Worth, Texas. Research in Progress Paper. Muthu Kumar Chandrasekaran, Min-Yen Kan, Kiruthika Ragupathi and Bernard C. Y. Tan (2015). *Learning instructor intervention from MOOC forums: Early Results and Issues.* In Proceedings of Education Data Mining (EDM '15), Madrid, Spain.
- The NUS MOOC Transacts Corpus: `https://github.com/WING-NUS/NUS-MOOC-Transacts-Corpus`

3 Reference string parsing in scholarly documents

In the bibliography section of a scholarly work are sets of natural language strings, in the form of references. Reference string parsing is the sequence labeling task (similar to part-of-speech tagging) of associating a multi-class label with individual tokens. Similar to the first project, there is a component of this project that involves tokenization (tokens like "(2002)" and "4(3)" need to be properly tokenized), as well as token representation (through embeddings or other work). As a sequence labeling task, the input is a sequence of tokens $t_1, t_2, ..., t_n$ and the output task is to associate one of a closed set of labels $c \in \mathcal{C}$ to ascribe to each token (e.g., *author*, *title*, *year*, etc.). Key in this work is the challenge of handling certain conventionalized styles (e.g., Harvard, APA) and out-of-vocabulary tokens (there are always new names, places, scholarly terminology that may be encountered).

The basic task has been well-studied. This is a bonus in the sense that standardized evaluation metrics and datasets have been defined for this task. Like the other tasks, this is a base project with lots of possible extensions, which can be distributed to subteams of the group. These include:

- Handling subword embeddings. Names and technical terminology often have internal structure. Paying attention to such regularities may allow your labeler to achieve better accuracy on those individual tokens.
- Regularities among references within a set. In a particular scholarly document, reference usually conform to a particular standard style of reference. This can lend stronger evidence for a regular sequence structure among all references within a set. Note: most evaluation datasets aren't defined on a set (reference section) basis, so you'll need to create your own data for testing.
- Using domain-specific word embeddings. Word embedding systems like BERT have been re-trained on large corpora of scholarly documents and could be fine tuned on reference data; see SciBERT for starting leads here.
- Entity Linking. Many words within a reference string come from a (semi-)closed set of words. Cross referencing lexicons or online statistics (e.g., Google n-Grams) may assist your parsing accuracy, especially on strings that your labeler is less certain about.
- Implementing hierarchical labeling. In particular, in the author list the names have structure (first, middle, last) and trying to infer the correct parsing of names may be done more effectively. Current systems use a simple list of rules to parse names which is largely effective. Trying to improve parsing accuracy here may be a challenge topic.

3.1 References

- Animesh Prasad, Manpreet Kaur and Min-Yen Kan (2018) *Neural ParsCit: a deep learning-based reference string parser*. International Journal on Digital Libraries. May 2018.
- Markus Hänse, Min-Yen Kan and Achim P. Karduck (2010) *Kairos: Proactive Harvesting of Research Paper Metadata from Scientific Conference Web Sites.* Proceedings of the International Conference on Asia-Pacific Digital Libraries (ICADL '10), Brisbane, Australia, June. pp. 226-235.
- Isaac G. Councill, C. Lee Giles and Min-Yen Kan (2008) *ParsCit: An open-source CRF reference string parsing package*. In Language Resources and Evaluation Conference (LREC 08). Marrakesh, Morocco, May.