Internship – Impact of language evolution in historical texts on NLP models

ENP-China Team
jeremy.auguste@lis-lab.fr, baptiste.blouin@lis-lab.fr, benoit.favre@lis-lab.fr

1 Internship context

The ENP-China project falls within the "Digital Humanities" field and aims to study the mutation of the elites in China between the 19th and 20th centuries on the basis of a large amount of data, especially textual data, in Chinese and English from this period. Therefore, we need to conduct experiments and reflection on the use of NLP tools in the context of historical research. In this perspective, analyzing the influence of language changes during the considered period is essential in order to correctly apply NLP models that match the historian's needs (named entity recognition, entity relations, event extraction, ...).

2 Scientific context

Recent NLP advancements show impressive results on many tasks, however most models are only trained and most importantly only evaluated on contemporary texts, and of specific domains. Yet, in many fields of social sciences, studied texts are from very varied sources and, in the case of History, have been written in different time periods. However, languages change over time, and in many ways: syntax [6, 5], semantics [4, 3, 9], style [10] or even typography (e.g., classical Chinese doesn't have punctuation), meaning that is not obvious that methods that work on today's version of a language will also work on older versions.

Since most state-of-the-art NLP models are trained on contemporary texts, their ability to deal with historical texts can be impaired, which greatly affects performances on tasks such as named entity recognition [1, 2] or part-of-speech tagging [7].

Multiple approaches can be used in order to deal with this issue, for example by adapting existing models to the target time period. However, in order to use the correct models, the issue of finding what are the different time periods remains. [8] shows that between 1872 and 1949, Chinese can be split into multiple periods of the language, transitioning from classical Chinese to modern Chinese with clear language cuts in 1904, 1911 and 1937.

3 Goal of the internship

In this internship, we will try to give some answers to the broad question of how to deal with major and minor language changes in Chinese, and possibly in English, while still using state-of-the-art models for a wide range of NLP tasks. Thus, the main goal of the internship is to compare and propose multiple approaches on multiple time periods in order to know what methodology is the most appropriate when working with historical documents. Some possible approaches are: learning models only from data from a given time period, adapting existing models on a given time period, learning period-agnostic models (using for example GANs), finetuning BERT-like models on big quantities of data from a given time period. In order to evaluate and train these approaches, several tasks may be relied on across multiple periods of time: word segmentation (which is a non-trivial task in Chinese), article segmentation, named entity recognition, entity relations extraction and event extraction.

A complementary goal is to propose ways of identifying time periods where language shifts (syntactic or semantic) occur. Some shifts are well documented (e.g., the massive change from traditional to simplified Chinese) and broad language periods are known (e.g., ancient, old, middle and modern Chinese). However, small language changes — especially when transitioning between two major periods — would be interesting to document in order to be able to correctly apply NLP models on these texts (e.g., applying period-specific models, applying preprocessing in order to normalize texts, etc.). A first idea is to train models for each annotated time period on a task (word segmentation, NER, etc.) and evaluate them on each time period. This could then allow identifying models that produce similar predictions, and thus would give an idea of what periods have similar language phenomena.

4 Available data and other resources

In order to carry out these experiments, a corpus of historical texts in Chinese is being annotated with word segmentation, sentence/article segmentation, named entities, relations between entities and possibly events and will be available for the internship. In addition to these annotations, the corpus has been collected by extracting a set of documents every ten years between 1872 and 1949. It is also possible to use the huge quantity of non-annotated historical data in Chinese (2.3m documents) and English (10.5m). In addition to these project specific corpora, other historical corpora such as the HIPE [2] corpus or the different Academia Sinica datasets could also be of use.

The intern will work with members of IrAsia and LIS, allowing supervision from experts in natural language processing, but also in the historical context, the content, and the language of the corpora.

The national Jean-Zay computing cluster, which grants access to a huge number of V100 GPUs, will be accessible to the intern.

5 Prerequisites

Python will be used as a programming language and the intern must be able to set up a full NLP pipeline (preprocessing, training/finetuning, evaluation). Knowledge of the pytorch or tensorflow libraries is necessary, and (basic) knowledge of the Transformers[1] library is preferable. Ability to speak/read Chinese is **not** required. ENP-China team meetings and seminars are held in English.

[1] https://huggingface.co/transformers/

References

[1] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic evaluation of NER systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, number CONF, pages 97–107. Bochumer Linguistische Arbeitsberichte, 2016.

[2] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, *CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, page 38, Thessaloniki, Greece, 2020. CEUR-WS.

[3] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

[4] Adam Jatowt and Kevin Duh. A framework for analyzing semantic change of words across time. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 229–238, London, United Kingdom, September 2014. IEEE.

[5] Tom S. Juzek, Marie-Pauline Krielke, and Elke Teich. Exploring diachronic syntactic shifts with dependency length: The case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119, 2020.

[6] Anthony S. Kroch. Syntactic change. *The Handbook of Contemporary Syntactic Theory*, pages 698–729, 2001.

[7] Bai Li. Evolution of Part-of-Speech in Classical Chinese. *arXiv:2009.11144 [cs]*, September 2020.

[8] Pierre Magistry. Languages(s) of the SHUN-PAO, a Computational Linguistics account. In *10th International Conference of Digital Archives and Digital Humanities*, Taipei, Taiwan, December 2019.

[9] Syrielle Montariol. *Models of Diachronic Semantic Change Using Word Embeddings.* PhD thesis, Université Paris-Saclay, February 2021.

[10] Sanja Štajner and Ruslan Mitkov. Diachronic stylistic changes in British and American varieties of 20th century written English language. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 78–85, 2011.