Transhistorical Resonance: Medieval Chinese Scholarship as Data

Budak, Nicholas Andrew
budak@stanford.edu
Stanford University, United States of America

Rominger, Gian Duri
gianr@princeton.edu
Princeton University, United States of America

One of the core issues for building Natural Language Processing (NLP) models for historical languages remains the lack of annotated datasets — often despite the long scholastic and exegetical traditions for these languages and their associated texts. [1] Specifically, annotating sufficient data can be a labor-intensive and error-prone process for researchers and students. Historical secondary sources offer an intriguing source of semi-structured data in the form of commentaries, dictionaries, and other collaborative scholarly work. This presentation focuses specifically on medieval Chinese annotation work for Old Chinese, a language that lost its last native speakers millennia ago.

Our project , a collaboration between a software engineer and a philologist, is a continuation of the work of Lu Deming (d. 630), author of the *Jingdian Shiwen* (c. 583), an exegetical work that provides phonological and semantic annotations on the ancient classics. [2] Providing some 55,000 annotations-in-context across a selection of 16 ancient classics, the*Jingdian Shiwen*exemplifies Lu's meticulous style of commentary. While the*Jingdian Shiwen*provides the invaluable context words that dictionaries lack, it keeps its length manageable by excerpting only a few characters around the annotation target, which can be interpreted as a very early form of textual compression.

As the*Jingdian Shiwen*incorporates material from 230 different sources of previous commentaries from the Han, Wei-Jin, and Six Dynasties periods (from 1stcentury to 6thcentury CE),[3]it further provides insights into the scholastic background available to and epistemological assumptions of a medieval annotator. This includes Lu Deming's focus on the canon of traditional learning, which features classical works stemming largely from the pre-imperial and early Han periods (before 1stcentury BCE). This source material ranges from the*Classic of Poetry* (*Shijing*) and historical chronicles (the*Chunqiu*with its commentaries) to dictionary entries (the*Erya*).

Lu's annotations focus on the correct reading for a given Chinese character, often including commentary on its meaning and appearance in other works. The meticulous, highly-structured annotation text is particularly well-suited to parsing using modern NLP methods. For this purpose, we assemble an NLP pipeline that can parse, segment, and label the various phonological, semantic, and paratextual components within each annotation. We utilize a novel combination of deterministic methods for segmenting the text based on “marker characters” followed by carefully-applied machine learning models to label spans and relationships within each annotation. The output is a dataset of glosses and pronunciations in context as well as a network of citations and references situated in the academic landscape of early medieval China.

Curating a dataset of attested pronunciations with context is particularly important to address a major hurdle for Chinese NLP: characters that signify different words depending on context (Liu, Lu, and Neubig 2017). This issue is more difficult for premodern Chinese, and especially for texts from before the stabilization of orthography, when multiple different characters could be used to write the same word.[4] All of the texts Lu annotated fall into this category, and active debates remain about meaning and pronunciation across the corpus.

By curating and publishing an open dataset derived from the*Jingdian Shiwen*, along with the code and models used to generate it, we invite collaboration from digital humanists and philologists alike. Our data is published in the JSON-lines format to ensure it is simple to consume for downstream applications that might add even more richness to the data, e.g. by transforming it into a fully-annotated TEI corpus. Our approach further highlights the debates around how the ancient classics were read, spoken, and experienced, as well as the tensions inherent in the annotation process itself, especially for texts that allow for a plurality of readings.

Notes

1. For problems of under-resourced languages and the overt focus on English in Digital Humanities more broadly, compare Dombrowski, and Burns forthcoming. At the same time, for possibilities of domain-specific NLP models, compare Bamman forthcoming.
2. See, for example, Lu 2013. We use the digitized version from the Kanseki Repository, available at Wittern 2015. Note that this material is available under a CC-BY-SA license. For an analysis of the phonological material presented in the *Jingdian Shiwen*, see, for example, Luo 2012; or Yue 2017.
3. For a larger summary of the historical context of the *Jingdian Shiwen*, see Mair 1986, 168; and Honey 2021, 215-221.
4. For overviews of the development of the Chinese writing system in antiquity, see Qiu, Mattos, and Norman 2000; Boltz 2003; and Galambos 2006.

Bibliography

Bamman, David. Forthcoming. “LitBank: Born-Literary Natural Language Processing.”Iin *Computational Humanities*, ed. Jessica Marie Johnson, David Mimno, and Lauren Tilton. Minneapolis: University of Minnesota Press.

Boltz, William G. 2003. *The Origin and the Development of the Chinese Writing System.* New Haven: American Oriental Society.

Dombrowski, Quinn, and Patrick Burns. Forthcoming. "Language is not a Default Setting: Countering Digital Humanities’ English Problem." In *Debates in the Digital Humanities 2022*, ed. Matthew Gold and Lauren Klein. Minneapolis: University of Minnesota Press.

Galambos, Imre. 2006. *Orthography of Early Chinese Writing: Evidence from Newly Excavated Manuscripts*. Budapest: Department of East Asian Studies, Eötvös Loránd University.

Honey, David B. 2021. *A History of Chinese Classical Scholarship, Volume III: Northern and Southern Dynasties, Sui, and Early Tang: The Decline of Factual Philology and the Rise of Speculative Hermeneutics*. Washington: Academica Press.

Liu, Frederick, Han Lu, and Graham Neubig. 2017. "Handling Homographs in Neural Machine Translation," *arXiv* (preprint) , arXiv:1708.06510.

Lu, Deming 陸德明. 2013. *Jingdian Shiwen* 經典釋文. Shanghai: Shanghai guji chubanshe 上海古籍出版社.

Luo, Changpei 羅常培. 2012. *Jingdian Shiwen yinqie kao* 經典釋文音切考. Beijing: Zhonghua shuju 中華書局.

Mair, Victor H. 1986. " *Tzu-shu* 字書 or *tzu-tien* 字典 (dictiona-ries)." In *The Indiana Companion to Traditional Chinese Litera-ture (Volume 2)*, ed. William H. Nienhauser, Jr., Charles Hartman, and Scott W. Galer, 165-171. Bloomington, Ind.: Indiana Univer-sity Press.

Qiu, Xigui, Gilbert Louis Mattos, and Jerry Norman. 2000. *Chinese Writing.* Berkeley: Society for the Study of Early China.

Wittern, Christian. 2015. "KR1g0003: 經典釋文 / SBCK." *Git-Hub*. Last modified August 22, 2022. https://github.com/kan-ripo/KR1g0003.

Yue, Limin 岳利民. 2017. '*Jingdian Shiwen' yinqie de yinyi pipei yanjiu* 《經典釋文》音切的音義匹配研究. Chengdu: Ba Shu shushe 巴蜀書社.