NER for Tibetan and Mongolian Newspapers

Robert Barnett, Christian Faggionato, Marieke Meelen, Sargai Yunshaab,
Tsering Samdrup, Nathan Hill and Hildegard Diemberger

rb25rb25@gmail.com - cf36@soas.ac.uk - mm986@cam.ac.uk

Challenge 1: Modern Tibetan (20th, 21st cs.) and Vertical Mongolian, used by about 11m people within the People's Republic of China (PRC), are extremely low-resourced and under-researched outside the PRC.

Challenge 2: Tibetan has no word or sentence boundaries. Could we develop a tool for NER, the most important NLP tool for historical and political researchers, without word-level segmentation?

Challenge 3: No NER has been developed for Modern Tibetan. Could we produce an NER tagset that would meet the needs of historians and policy analysts working on Tibetan newspapers?

Stage 1 - Surveying the State of NLP for Vertical Mongolian & Modern Tibetan

NLP for Vertical Mongolian (VM)

Vertical Mongolian (VM) is the script used in Inner Mongolia, a region within the PRC. This script, also known as Traditional Mongolian, is based on Old Uyghur (8th to 17th cs.). Written from the top down and from left to right, it is used by c. 4m people within the PRC.

Fig. 1. Masthead of *Xinjiang ribao* (Xinjiang Daily News), Mongolian edition, Nov. 11, 2021

One of our researchers went to Hohhot (the capital of Inner Mongolia) to meet with NLP developers at Inner Mongolia University (IMU). From this we learnt that:

- They and others have developed NLP tools for VM, but these are not publicly available [1]
- Very few open-source tools have been developed for VM outside China
- Many NLP tools have been developed for Cyrillic Mongolian (CM), the script used in Mongolia (outside the PRC)
- CM > VM conversion tools are not yet reliable, partly because of coding problems with the Unicode set [2]
- To develop a toolkit for VM would mean starting from scratch

NLP for Modern Tibetan (MT)

- NLP tools developed within the PRC for MT have not been made publicly available
- NLP tools developed outside the PRC have been trained on Classical Tibetan, but this stage of the language (pre-20th c.) has different morphosyntactic features from MT
- We needed to see how the Classical Tibetan tools would work with MT, so we tested the segmenter and POS taggers developed by Meelen, Roux & Hill. [3]

Testing a Classical Tibetan Segmenter on MT

- We used **website-capture software** Sitesucker and Webcopy to scrape 3.11m syllables of data from news sites in MT. For geoblocked sites we used Transocks, mimicking China-located IPs.
- We manually **segmented sentences** by adding utterance boundaries after clausal punctuation markers in Tibetan called *shad* and after final *-g+*<space>, where Tibetans omit the *shad*.
- We **applied the existing word segmentation tool**. [4]

Fig. 2. Raw data from a Modern Tibetan news site, the "Chinese Communist Party News"

Stage 2 - Developing NER for Modern Tibetan, a Non-segmented, Low-resourced Language

Testing Classical Tibetan Segmentation

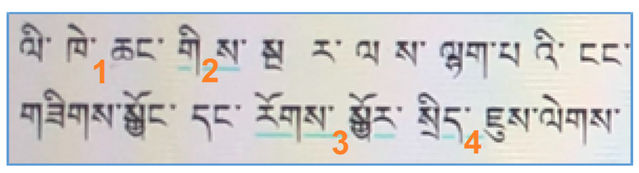

Fig. 3. An MT sentence segmented with a Classical Tibetan segmenter. 12 of the 17 syllables are incorrectly segmented.

We found that the Classical Tibetan word segmenter **produced errors** when applied to MT:

- multisyllabic names, such as Li Keqiang (li.khe.chang), were segmented into two or three separate words (Fig. 3: 1)
- words with the final consonant *-s* or *-r* were segmented as if these were the agentive or dative-locative particles: *gis* (agentive particle) was changed to *gi sa*, 'of earth' (Fig. 3: 2)
- 2-syllable words became 2 words: *rogs.skyor* 'help' became *rogs* and *skyor*, and *srid jus* 'policy' became *srid* and *jus* (Fig. 3: 3, 4).

We tried to develop NER training data by tagging **data without word boundaries** or Part-of-Speech (POS) tags. This was more convenient for the taggers (native speakers used to reading unsegmented Tibetan).

Developing an NER Tagset for MT

We began with 6 tags: PERSON, ORGANISATION, PLACE, TIME and DATE, but to meet the needs of historians and political analysts, we needed to distinguish governmental entities ("State Council") from non-governmental ones ("Temple"), official positions ("Party Secretary") from social titles ("Lama", "Rinpoche"), and administrative place terms ("Lhokha Municipality") from natural place names ("Lhasa").

Fig. 4. The final NER tagset for Modern Tibetan we produced using LightTag. It has 17 tags for NER and one for wrong segmentation.

Using the the ACE guidelines [5] as a basis, we sub-divided PERSONS into 3 sub-tags and ORGANISATIONS into 7:
PERSONS - Group/Individual/Title
ORGANISATIONS - Government/Position/NGO/Commercial/Media/Education/Religion
After consulting with Tibet scholars, we added further tags needed for political and historical analysis: SLOGANS, IDEOLOGIES, TITLES of Created Works, and WEAPONS, a total of 17 tags.

NER Annotation Without Segmentation

We chose **LightTag** [6] as our online tagger, because (a) to provide automatic suggestions, it uses character-based modelling (which is essential for our unsegmented data) [7] and (b) it scored highly with Neves and Ševa 2021 in their comparison of online taggers. [8]

We found that the human annotators' acceptance rate of LightTag's automated suggestions improved from 64 per cent in the first six sessions to 81 per cent in six of the project's final seven sessions. [9]

Using the 17-tag scheme on unsegmented, untagged data within LightTag, we **annotated 9,884 terms**, creating a unique new set of training data.

Creating Annotation Guidelines: The Metonymy Rule & Tag Hierarchies to Solve Ambiguous Terms

Annotation Philosopy & Principles

Applying the NER tags required careful assessment of context, because many named entities are polysemous. Bearing in mind the research aims of Tibetan historians and policy analysts, we developed a set of annotation guidelines. These included:

Guideline 1. The Metonymy Rule: A PLACE is not a PLACE if it is an agent ("China passed a law"). In such cases, that place-name is tagged as ORGANISATION-Government.

Guideline 2. Multi-word expressions: Use a single tag for each expression or string that names an entity even if that expression or string combines different types of named entities, such as name+place, name+organisation, or title+name, as in *brag phyi grong tso*, 'Dragchi village', or *bla ma tshe ring*, 'Lama Tsering'. Ideally, we would tag both the entire expression and its component terms.

Guideline 3. Non-specific entities: a term that denotes an organisation or entity was tagged as ORGANISATION even if its specific name or identifier is not given: e.g. both "prefecture" and "Shigatse Prefecture" are treated as named entities.

Guideline 4. Tag boundaries: Do not include the *tseg* (raised dot, a syllable boundary marker) or quotation marks in a tagged string unless they are internal to the string.

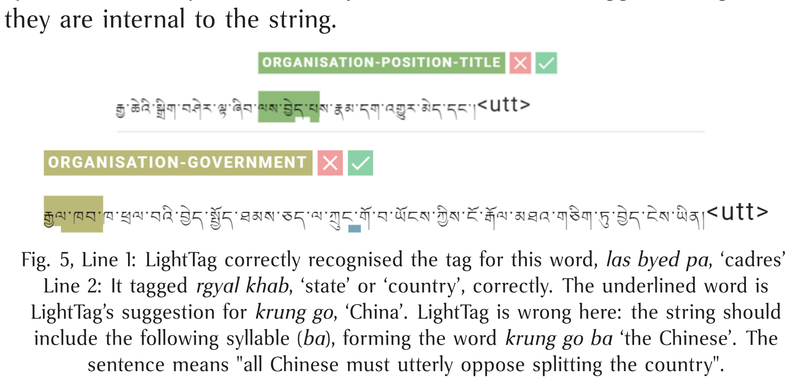

Fig. 5, Line 1: LightTag correctly recognised the tag for this word, *las byed pa*, 'cadres'. Line 2: It tagged *rgyal khab*, 'state' or 'country', correctly. The underlined word is LightTag's suggestion for *krung go*, 'China'. LightTag is wrong here: the string should include the following syllable (*ba*), forming the word *krung go ba* 'the Chinese'. The sentence means "all Chinese must utterly oppose splitting the country".

Tag Hierarchies. The greatest difficulty for a tagger is deciding which guideline to follow for polysemous terms. We therefore established a set of disambiguation and priority rules. These give precedence, for example, to:

- the **individual identity** of a person or location over their administrative or organisational status ("President Xi" > PERSON not POSITION)
- the **geographic identity** of a place-term even if it can also be used as an administrative term, ("town" > PLACE not ORGANISATION, unless used metonymically)
- the **geographic identity of an ideological place-name** over its ideological function ("Motherland" > PLACE not IDEOLOGY)
- the **governmental, administrative or political function** of a term even if it also serves as a social or cultural title ("Comrade", "Chairman" > ORGANISATION-position-title, not PERSON-title
- **political/governmental roles** in the PRC context even if a term also refers to a social group ("Soldier", "Worker" > ORGANISATION-position-title not PERSON-group.

Stage 3 - Post-processing: Manual Correction, Evaluation & Results

Manual Correction & Evaluation

An **advantage of LightTag** is that its character-based modelling means that word segmentation is not necessary. A **disadvantage of LightTag** is that it generally presents a randomised sequence of utterances rather than showing all tagged instances for each term, making reviewing and correction more difficult. The final round of manual correction was therefore done offline, with terms arranged alphabetically to show tagging inconsistencies.

Developing NER for this purpose requires **extensive collaboration with native speakers and historians** to finalise complex decisions about tagset design and hierarchies. We found that NER for Tibetan historians and policy analysts needs a tagset with at least 17 tags, with guidelines highlighting political and social hierarchies to facilitate their research.

Results

- To create a basic NLP toolkit for VM, we need to start from scratch
- For MT, we need to adapt tools trained on Classical Tibetan data
- Using LightTag, we created **an NER model for MT with an accuracy of 94.6%**, available at Zenodo [9]
- In collaboration with Tibetan historians and policy analysts, we created **a detailed NER annotation manual**, also at Zenodo [9]
- With a native speaker and a Tibetologist working together part-time for six months, we developed **a gold-standard data set consisting of 6,624 NER annotations**, deposited at Zenodo [9], which can be used as future training data
- As a result of this Incubator Project, we successfully applied for a follow-up grant with which we developed similar resources for Uyghur, the language of c. 12m people in Xinjiang, also within the PRC. [10]

Future Tasks

In follow-up projects, we will focus on these remaining issues:

- We did not tag **ethnicity, citizenship, and gender** because they are not always known, LightTag does not allow double-tagging, and we needed to avoid too many tags
- We did not tag **loanwords** from Chinese, etc. We hope to develop an automated process for detecting loanwords in MT
- In future we would **add 4 more tags**: historical periods ("Middle Ages"), named events ("5th Plenary"), money, or foreign (in this case, non-PRC) governmental entities
- All our training data so far is from contemporary news sites. We plan to add training data from **Tibetan newspapers published in the 1930s-40s**, before Tibet became part of the PRC
- Choices over **tag hierarchies** and **multi-word expressions** are difficult and deserve further testing and discussion.

Acknowledgments: This research was funded by the Isaac Newton Trust, Cambridge University Press and the Cambridge Language Sciences Incubator Fund. It was hosted by the Mongolia and Inner Asia Studies Unit, University of Cambridge. **References:** [1] Yunshaab, S. (2019), "Survey of Natural Language Processing for Vertical Mongolian: Current Situation," *Report for MIASU, University of Cambridge*. Zenodo, http://doi.org/10.5281/zenodo.5103499. [2] Faggionato, C. (2019), "Testing Available Mongolian NER tools and Future Perspectives," *Report for MIASU, University of Cambridge*. Zenodo, http://doi.org/10.5281/zenodo.5103499. [3] Meelen, Roux and Hill (2021), 'Improving the ACTib: combining a neural segmenter and POS-tagger with rule-based error correction' in *TALLIP*, see code on https://github.com/lothelanor/actib. [4] Meelen & Hill (2017), 'Segmenting and POS tagging Classical Tibetan' in *Himalayan Linguistics*, 16(2), 64-86. [5] Linguistic Data Consortium (2005), ACE (Automatic Content Extraction) English annotation guidelines for entities. Version, 5(6), pp. 2005-08, https://www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf. [6] LightTag, https://actib.lighttag.io/. [7] Perry, T. (n.d) Character Level NLP, https://www.lighttag.io/blog/character-level-NLP/. [8] Neves & Seva (2021), An extensive review of tools for manual annotation of documents, *Briefings in Bioinformatics*, (22:1), pp. 146–163, https://doi.org/10.1093/bib/bbz130. [9] Barnett et al. (2021), Named-Entity Recognition for Modern Tibetan Newspapers: Tagset, Guidelines and Training Data, Zenodo, https://doi.org/10.5281/zenodo.453651. [10] https://github.com/HKBUproject/historical-uyghur-chinese-corpus.