Named-Entity Recognition for Modern Tibetan:

Tagset and Guidelines

Prepared for the project “Named-Entity Recognition in Tibetan and Mongolian Newspapers”
For the Mongolian and Inner Asian Studies Unit, Cambridge University
Funded by: Cambridge Language Sciences, Cambridge University
Project PI: Dr Hildegard Diemberger (Social Anthropology, Cambridge)
Coordinator and Lead Author: Dr Robert Barnett (East Asian Languages and Religions, SOAS)
Advisors: Dr Nathan Hill (SOAS), Dr Marieke Meelen (Cambidge), Dr Thomas White (Cambridge)
Research Assistant: Tsering Samdrup (East Asian Languages and Religions, SOAS)
Final revision: July 12, 2021.
Repository: http://doi.org/10.5281/zenodo.4536516

Introduction

In June 2019 Cambridge Language Sciences funded the Mongolian and Inner Asian Studies Unit at Cambridge University’s Department of Social Anthropology to carry out a six-month incubator project, “Named-Entity Recognition in Tibetan and Mongolian Newspapers”. The project’s purpose was to assess the feasibility of developing Named Entity Recognition (NER) for modern Tibetan and vertical Mongolian. NER is a basic tool in Natural Language Processing (NLP) that enables automated recognition of named entities, such as people, institutions, professions, titles and other entities, within a body of texts. This tool can significantly increase the ability of researchers to analyse large sets of data and is particularly useful for historians and other scholars and analysts studying open-source materials – texts produced and openly circulated by governments and media outlets. The analysis of open-source materials is especially important in the case of regions or states, such as Tibet, Xinjiang and certain other parts of the People’s Republic of China (PRC), where open-source materials generally represent the only information available to the public about conditions in those regions.

Although some forms of NER and other NLP procedures have reportedly been developed within China for modern Tibetan,[1] the data underlying those initiatives have not been made publicly available. The majority of NLP work carried out outside China on Tibetan has focused on classical Tibetan, primarily consisting of religious texts,[2] and at the time of this project had not been developed for modern Tibetan.

Data

The dataset for the project consisted initially of 3.11m syllables in Tibetan consisting of news articles varying from approximately 100 to 1,000 words in length downloaded from Chinese-language news aggregator sites within China that deal exclusively with news about Tibet, primarily tibet.cpc.people.com.cn and tibet.people.com.cn. These are government-run news-aggregator sites that republish official news reports in Tibetan about Tibet, almost all of which are translations from Chinese-language reports produced by government or Party outlets such as *Xizang Ribao* (Tibet Daily), Xinhua (New China News Agency), or *Renmin Ribao* (People’s Daily). To provide a wider lexical range, we added one set of literary texts (short stories) comprising 29,000 syllables.

[1] See Liu H, Nuo M, Ma L et al. (2011). Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Field. *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, 2011*, pp. 168–177.

[2] See Hill, Nathan W., & Garrett, Edward. (2017). A part-of-speech (POS) lexicon of Classical Tibetan for NLP [Data set]. Zenodo. http://doi.org/10.5281/zenodo.574876.

From this data we selected 280,000 syllables for annotation. In order to prepare the data for annotation, we divided the texts into 26,000 utterances, following basic Tibetan punctation rules for clause and sentence boundaries.

The Tags

Existing annotation guidelines for NER in English-language texts that we consulted[3] were often very complex and did not always distinguish official from unofficial titles and position-holders or identify political slogans and campaign names, all of which are important for historical-political researchers. We therefore developed a purpose-built tagset or schema for use by such researchers working on open-source materials. After testing seven different designs, we finally settled on a schema consisting of 17 tags.[4]

The 17-tag schema for modern Tibetan NER is designed to identify named entities in modern Tibetan texts, including named persons, places, organisations, position-titles, slogans, campaigns, and dates. It is intended primarily to facilitate the study and analysis of official texts in modern Tibetan, including newspapers, such as by historians and political scientists. Its chief feature is accordingly that it differentiates official or governmental positions, titles and entities from non-official positions, titles and entities. It has seven tags for types of *organisations* (administrative entities or other institutions), one for *places*, and three for types of *persons* (named individuals, groups or titles). Six other tags identify names of *ideologies*, *slogans*, *created works, weapons, time* or *date.*

This is the complete list of tags recommended for NER with modern Tibetan:

A. *Persons*

Persons refers to humans or other named individuals, such as animals and gods; generalised social categories consisting of persons; and non-governmental titles or positions given by which these individuals are known.

1. PERSON-individual. This is for named humans, eg. “Tsering”; “Nathan”; “Dalai”; “Panchen”; “Xi Jinping”; “Buddha”. This includes named individual animals, deities and ghosts.

   Notes: (1) Mythical, fictional and idealized peoples such as James Bond and Aku Tonpa are tagged as person-individual if they purport to be or are represented as real persons (dreamlike or conceptual figures would be tagged as ideology-concept); (2) Names of species and professions are tagged as person-group; (3) Where a string includes multiple elements that could be tagged separately, such as title+name, we

[3] Linguistic Data Consortium (2005). ACE (Automatic Content Extraction) English annotation guidelines for entities. Version, 5(6), pp. 2005-08, https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf; Marrero M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J.M. (2013). Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5), pp. 482-489. https://julian-urbano.info/files/publications/003-named-entity-recognition-fallacies-challenges-opportunities.pdf; and Schön, S., Mironova, V., Gabryszak, A. and Hennig, L. (2020). A Corpus Study and Annotation Schema for Named Entity Recognition and Relation Extraction of Business Products. arXiv preprint arXiv:2004.03287, https://www.aclweb.org/anthology/L18-1704.

[4] When working on texts that had been segmented, we included an additional tag to mark instances of wrong segmentation.

tag the entire string as a single unit (see Multiword Expressions, see Section #3.2 below).

2. PERSON-group. This is for a social or cultural category, including those formed by professions and ethnicities, or for a single member of that category, eg., “Tibetans”; “Tibetan”: “monks”; “monk”; “lawyers”; “nomad”; “children”; “elder”; “Buddhist”; “the youth”; “[community] leader”, “veterans”. It is also used for a named species or race, both for the name of that group and for a member of that category. It is not used for formal, organised groups.

   Notes: The person-group tag includes (1) informal, descriptive terms for those affiliated to a government or Party entity, such as “the delegates”, “the members”, “retired officials”, or “workers”, if these are general descriptions, not formal titles of positions; (2) ideological groupings or constructs such as *mi rig* (“nation” or “nationality”) and ཀྲུང་གོ་མི་དམངས (*krung go mi dmangs*, “the Chinese people” or “the Chinese nationality”), since they are used as if they describe actual groups; (3) terms such as “international society” or “international community”, since they do not consist only of governmental entities; (4) informal political associations such as *spun zla* (“brethren”, in the sense of fellow-members), since that is not a formal title.

   The person-group tag is *not* used for (1) groups that are named after a professional or social title, such as “lamas”, “geshes”, “doctors”, “company directors”, which would be tagged as person-title; (2) organised groups, such as “Young Pioneers” or “dance group”, which are tagged as organisation-government or organisation-non-governmental, depending on their affiliation; (3) groups that are named after a governmental position or political category, such as “officials”, “cadres”, “officials and workers”, “comrades”, “soldiers”, or “Aid Tibet Cadres”, which would be tagged as organisation-position-title. In other words, the fact that a term refers to a social title or a governmental affiliation takes precedence over the fact that it also sometimes refers to a group.

3. PERSON-title. This tag is used for titles, positions, and professions of persons that are awarded by society, not by a government, eg. *Dr.; Madame; Rinpoche; Lama;* and *Geshe*. It is also used for pluralities or groups of people holding formal titles, such as “lamas”, “geshes”, “rinpoches”, “doctors”, “university chancellors”, or “company directors” – ie., a title takes precedence over affiliation to a group.

   Notes: The tag person-title includes terms that refer to a general group or category of title-holders, such as “trulkus”, rather than to specific holders of that position.

   It does *not* include (1) titles and positions in government or CCP organisations, such as “comrade”, “cadre”, “Party secretary”, “official”, “cadre”, “staff”, or “soldier”, which are tagged as organisation-position-title, because the fact that a term refers to a governmental position takes precedence over the fact that it includes a title; (2) words denoting membership of a social, non-governmental category like “monk”, “retired worker”, “[community] leader”, or “child soldier” when used in a general sense and not as a specific title; these are tagged as person-group, because titles take precedence over general terms; (3) compound expressions which include a title plus an individual’s name (“Dr Tenzin”, “Lama Topgyal”), which are tagged as person-individual – ie., an individual’s name takes priority over a title where found together; (4) titles used in society (not government) to identify a particular individual, such as

"Panchen Erdeni" and "the Dalai Lama", or a number of specific holders of that title, such as "Retring Rinpoches" or "the Dalai Lamas", even if the individuals' names are not given – these are tagged as person-individual, since they stand for a specific holder of that title.

B. Organisations

Organisational entities are governments, the CCP, institutions, administrative entities, corporations, agencies, and other groups of people which have an established organisational structure.

4. ORGANISATION-government. This is for government or CCP entities, groups and associations, including military entities belonging to a government. It includes governments and governmental entities from any country, not just China (eg., "Japan", "the UK government", "the French Foreign Ministry"). It includes "mass organisations" in China like the Women's Federation, since they are run by the Party. Examples include "the Foreign Ministry", "the authorities", "the centre", "the State Council", "the county government", "the Party Committee", "the Women's Federation", "the Communist Youth League", "the army", "the PLA", and "Bank of China", as well as power-stations, government banks, government hospitals and performance troupes. It includes temporary organisations, like a Party Congress, and political movements if those are organised by a government or the CCP.

   Notes: The tag organisation-government includes (1) generalised terms for military or police groupings such as *dmag dpung* (military troop, force) and *dpung ru* (military unit), even though they are not specific units, and even if belonging to governments other than China; (2) informal groupings of officials, such as བོད་སྐྱོར་རུ་ཆེན (*Bod skyor ru chen,* "the Aid Tibet contingent"), because their membership depends on holding government positions – ie., the fact that a term refers to a governmental position takes precedence over the fact that it sometimes refers to a group. (3) government or Party groupings or associations even if they are not permanent or are not location-specific, such as a temporary "propaganda team" or "dance troupe".

   This tag is *not* used for (1) military or police groupings run by the UN or by rebel armies, which would be tagged as organisation-non-governmental; (2) organised groups or political movements that are non-governmental, such as an NGO or a private singing group, which would be organisation-non-governmental; (3) private banks and private hospitals, which are relatively rare in China, which are tagged as organisation-non-governmental; (4) terms such as "international society" or "international community", which are tagged as person-group since they are not limited to governments.

5. ORGANISATION-non-governmental. This tag is used for non-governmental organisations, eg., the UN, NATO, the World Bank, political parties apart from the CCP, charities, NGOs, and performance groups not organised by the state.

6. ORGANISATION-position-title. This is for governmental or political positions or titles such as "Party Secretary" (*shuji*), "chairman", "committee member", or "lieutenant". It includes general, descriptive terms for positions in a government or similar institution, such as "official", "cadre", "worker", "grassroots official" or "soldier", when these are used to refer

to an official or political category rather than in a loose or general sense. It is used for pluralities or groups of people with such positions, such as “officials”, “cadres”, “workers”, “grassroots officials”, “soldiers”, “officials and workers”, “Aid Tibet Cadres”, and *blo mthun* (“comrade”), except when used in a non-official sense. It includes positions in foreign governments, the UN, and in non-governmental organisations or institutions.

7. ORGANISATION-Media. Use this for media and news organisations.

8. ORGANISATION-Religion. Use this for religious organisations, institutions, events and festivals. Religious adherents are tagged as a social group, so “Tibetan Buddhism” is tagged as organisation-religion, but Tibetan Buddhists are a person-group.

9. ORGANISATION-Commercial. This is for commercial institutions, factories, shops, cinemas, and enterprises including SOEs and other government-run commercial operations. However, private banks and private hospitals, which are rare in China, are tagged as organisation-non-governmental.

10. ORGANISATION-Educational. This is for educational institutions.

C. Place

Place refers to entities that can be geographically located.

11. PLACE. This includes geographical entities, features and landmasses including boundaries, planets, rivers, towns, countries and addresses, eg. “Lhasa”, “Tibet”, “the centre of the country”, “the Barkor”.

    Notes: The place tag includes (1) terms for locations that have no specific name or identifier, such as simply “town”, “city”, “river”, “lake”; (2) terms for locations even if they also indicate an administrative category, such as “county”, “town”, “prefecture”, “municipality”, “township”, “province”, and “region”, since we give precedence to location over administrative status; (3) political, ideological and conceptual place-terms, such as *mes rgyal* (“ancestral realm”, “motherland”), *rgyal yongs* (“entire nation”), *rgyal khab* (“state” in the sense of a country or its administrative apparatus, as in མོང་གོལ་རྒྱལ་ཁབ *mong gol rgyal khab*, “the Mongol state”) and ཀྲུང་ཧྭ་མི་དམངས་སྤྱི་མཐུན་རྒྱལ་ཁབ་ (*krung hwa mi dmangs spyi mthun rgyal khab,* the People’s Republic of China), since we give precedence to their location over their ideological function; (4) organized and administrative locations, such as nature reserves, since we give precedence to their location over their administrative status; (5) mythical, fictional and idealized places such as “Shambhala” and “Shangrila”, if they purport to have a real location on the globe (not just in a dream or vision, in which case they are tagged as ideology-concepts).

    The place tag does *not* include (1) place-terms that are used in a metonymic sense, such as “the county passed a law”, in which case the term would be tagged as organisation-government or organisation-non-governmental depending on affiliation; (2) place-terms used within standardised collocations, such as *labs chen mes rgyal* (“the Great Motherland”), which are tagged as slogans; (3) idealized, ideological, or conceptual place-terms like "The Silk Road", "Belt and Road", "Maritime Silk Road" and *nub ljongs bde ba chen* ("western paradise") which have no fixed or clear location and are therefore tagged as ideology-concepts.

D. Other types of named entities

This includes names and titles of ideologies, slogans, political campaigns, created works, weapons, and indications of date or time.

12. IDEOLOGY – concepts. This is for names of philosophies, theories, and concepts, eg., socialism, capitalism. It includes *splittism* and *separatism*.

13. SLOGAN-campaigns-policy names. This is for slogans (frequently repeated statements, usually political), such as "Power grows out of a gun", and policy names, campaign titles, and movement titles, such as "the Three Represents", "the Cultural Revolution," "Democratic Reform".

14. TITLE of book. This is for the title of a book, film, song or similar work.

15. WEAPON. This is for devices, equipment and technology primarily used as instruments for physically harming or controlling other entities, eg., *missile, gun, radar, cameras, surveillance equipment.*

16. TIME: an hour, or hour+minutes, or period of the day, eg., *2pm, early afternoon, evening, noon. Early* and *late* are tagged as time even though they are not specific.

17. DATE: a specific day of the week, month or year, or a combination of these, eg., *Monday* or *February* or *1836* or a specific day plus month, eg., *Feb 24* or *Feb 24, 2016*.

Guidelines and Principles

As a result of the annotation process, we settled on a series of principles guiding our tagging decisions. The most important of these consisted of which order of precedence to apply when facing competing factors in a tagging decision. These largely reflected our focus on political history and analysis. However, we also had to make decisions about technical questions, such as how to deal with multiword expressions and tag boundaries. These were the main principles we applied in the annotation process:

Precedence: In determining which tag to use where there are options, firstly we gave precedence to the individual *identity* of a person or location, where specified, over their administrative or organisational status. This means that if an individual or place is identifiable from a term (because it includes the name), we tagged it as person-individual or place even if it also includes a title, administrative or ideological function except for place-terms used metonymically. Secondly, we gave precedence to the geographic nature of a place-term (such as "town" or "county") and tagged it as place rather than as an organisation even if that place-term can also be used as an administrative term, unless it is used metonymically. Thirdly, where an organisational term or title does not include an identifiable name, we gave precedence to the *governmental, administrative* or *political* functions of that term even if it includes a title, and tag it as organisation-position-title, not as person-title. Fourthly, the fact that a term refers to a social *title* or a governmental *affiliation* takes precedence over the fact that it also sometimes refers to a group, so we tag it as person-title or organisation-position-title, rather than as person-group. Fifthly, *specific titles* take precedence over general terms for positions or professions. Our decisions concerning precedence reflected the project's focus on the study and analysis of official, public documents and thus prioritised identification of governmental and social agents, appointments, policies, and ideologies.

Multiword expressions: We only used a single tag for each expression or string that names an entity. When an expression or string combines different types of named entities, such as name+place, name+organisation, title+name, or organisation-title+name, as in *brag phyi grong tso* ("Dragchi village"), *phu'u can zhing ud* ("Fujian Province Committee"), *Lha sa grong khyer* ("Lhasa city"), or *bla ma tshe ring* ("Lama Tsering"), we gave the whole expression one tag, instead of splitting the expression into subsidiary units each with their own tag. Both approaches are valuable and should be considered, but we chose to prioritise the full name of an entity.

Metonymy rule: When a place-name or ideological term refers to an action, thought or emotion by the entity represented by that term (as in "China banned all foreigners", "the county passed the law", or "the motherland was outraged"), that term was tagged as organisation-government or organisation-non-governmental according to affiliation, not as place or ideology-concept.

Non-specific organisations: An organisation or entity was tagged as organisation-government even if its specific name or identifier is not given: eg., "bureau", "department", "prefecture", "sub-police-station", "regiment", "committee", "discipline inspection committee".

Non-specific places: A place or location was tagged as place, not as an organisation, even if its specific name or identifier is not given and even if it can also be an administrative category, eg., "county", "town", "prefecture", "municipality", "township", "province", "region", "city", "village", except where the word has a metonymic function, when it would be tagged as organisation-government or organisation-non-governmental, depending on affiliation.

Mentions: We tagged only the name mentions (Joe Smith) and the title/position mentions ("Mr Smith") of a person or entity. We did not tag nominal mentions ("the guy wearing a blue shirt") or pronoun mentions ("he", "him", "her").

Tag boundaries: The *tseg* (raised dot, a syllable boundary marker) was be included in a tagged string only if it is internal, ie., it is between any two syllables that are within a tagged string. Any *tseg* at the start or end of a tagged string was not included in a tag, since it is not found in all instances (such as at the end of a clause or sentence). Quotation marks and other punctuation in the title of a song, film, book, or suchlike were not included in a tagged string if they are at the beginning and end of the string, since these are also not found in all instances. Quotation marks were included if they are internal to the string, ie., if part of the tagged string is before or after the punctuation marks.

We did not need to identify beginning, middle and end markers, since we used Lighttag to record our tags, and it automatically includes this information (using its colour-coding of each tag) in the results data.

Misspellings: If a name or title is misspelt, we did not tag it.

Patron: the term *byin bdag* ("patron") was not treated as a named entity and was not tagged.

Recommendations

We recommend that a future tagset for a study focusing on the analysis of public documents should include these five additional tags:

18. NAME-period: names of periods and dynasties (eg. “Middle Ages”, “Reform Era”, “Maoist era”, “Ming dynasty”)
19. EVENTS: names of events (“WWII”, “anti-Japanese War”, “Peaceful Liberation of Tibet”, “5th Plenary of 11th Congress”)
20. MONEY for currencies and currency amounts
21. MONUMENTS such as stellae (*do ring*) or statue
22. ORGANISATION-government-foreign for official entities that are not part of or affiliated to the domestic government, in this case China

Questions for Re-assessment

We made decisions on three technical issues that other users may want to revisit according to the technical options available to them:

(a) We chose to tag non-specific places that are also administrative entitles (“town”, “village”) as places rather than as organisations.

(b) We chose to tag multiword expressions (such as name+organisation, place+organisation, title+person) with one tag rather than two or three (eg., “China Communist Party” > “China” – place, “Communist Party” > organisation-government, “China Communist Party” > organisation-government).

(c) We did not include tags for four important features of named entities: the ethnicity of a person, the ethnicity of a term (loanword, etc), the nationality of a person or institution (their citizenship or whether they are domestic/foreign), and the gender of a person. The chief difficulty that arises with tagging these features is that these characteristics are not always known to a reader (a person with a Chinese name could be Tibetan, for example). Secondly, we wanted to avoid adding a second tag to a single word or string. In addition, we were reluctant to subdivide each tag into multiple subsidiary tags in order to identify these features. However, these constitute important information for historical and political research, and a practicable way should be developed to include this information in any NER procedure for research of this kind.

-end-