Using Spreadsheets to Review Annotations Offline

Report prepared for the project "Named-Entity Recognition in Tibetan and Mongolian Newspapers" hosted by the Mongolian and Inner Asian Studies Unit, Cambridge University (PI: Dr Hildegard Diemberger).

Author: Robert Barnett (East Asian Languages and Cultures, SOAS)

February 2021

Introduction

In June 2019 Cambridge Language Sciences awarded an incubator grant to the Mongolian and Inner Asian Studies Unit of the Department of Social Anthropology at Cambridge University to develop Named Entity Recognition (NER) for modern Tibetan and to assess the current availability of NLP tools for vertical Mongolian. The project was designed in particular to facilitate reading and analysis of contemporary newspapers and other open-source materials issued by Chinese authorities in Tibetan and vertical Mongolian.

The project compiled 3.11m syllables of data in Tibetan, consisting of news articles and open source materials, by downloading from or scraping official media websites within Tibet. From this data, it selected, processed and uploaded texts containing 280,000 syllables in Tibetan (grouped in 26,000 utterances/sentences) to Lighttag, an online annotation site. Using Lighttag, the project team (Annotator: Tsering Samdrup) annotated 74.3% of the utterances, or approximately 186,000 syllables leading to 9,884 annotations (for raw data, see DOI:10.5281/zenodo.4536516). Of these, 6,736 annotations (68%) had been produced using our final tagset (see DOI:10.5281/zenodo.4536516) and were thus suitable for review.

For the review process, we assessed the effectiveness of Lighttag's review function. It allows one easily to revisit each annotation in context, but at present only allows two options – either to accept or reject an annotation. This can create false positives or false negatives if a reviewer does not want to make a final decision regarding an annotation on the spot, as often can happen while he checks other instances, consults the guidelines, or simply needs more information about the term in question. In other words, a reviewer needs a third option, allowing him or her to be able to skip an annotation and return to it later.

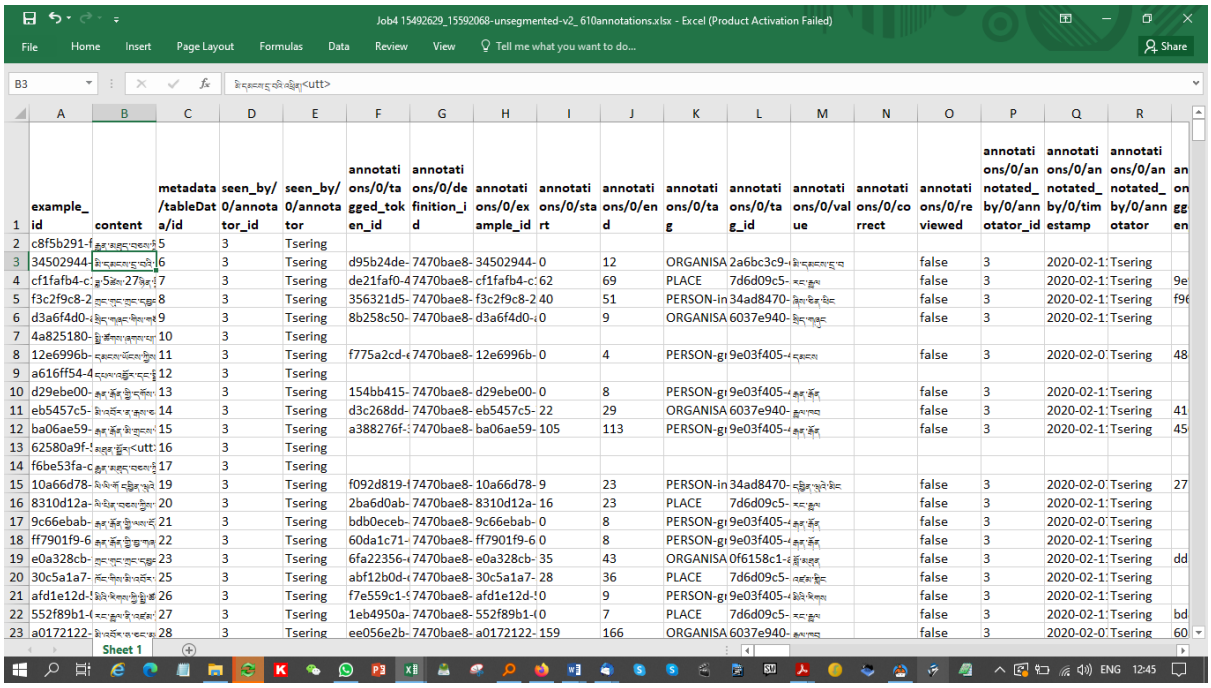
We also found that the process of reviewing is far more efficient if the reviewer can see all the annotations for each given term, together with the context. Viewing all annotations for each term allows the reviewer to identify immediately any inconsistences in the annotations for a particular term, to consult the guidelines applicable to that term, and to apply them consistently to any given term.

Taken together, these considerations led us to try carrying out the review process offline rather than on Lighttag. To do this, we converted the data to forms that could be viewed on Excel, using standard software rather than writing special code. The initial download and conversion process was time-consuming because of unexpected errors in the conversion, and because texts using non-Latin scripts can easily be lost during conversion to a program like Excel. Once these errors had been recognised, the process became relatively simple and reviewing became far faster and simpler than when performed online.

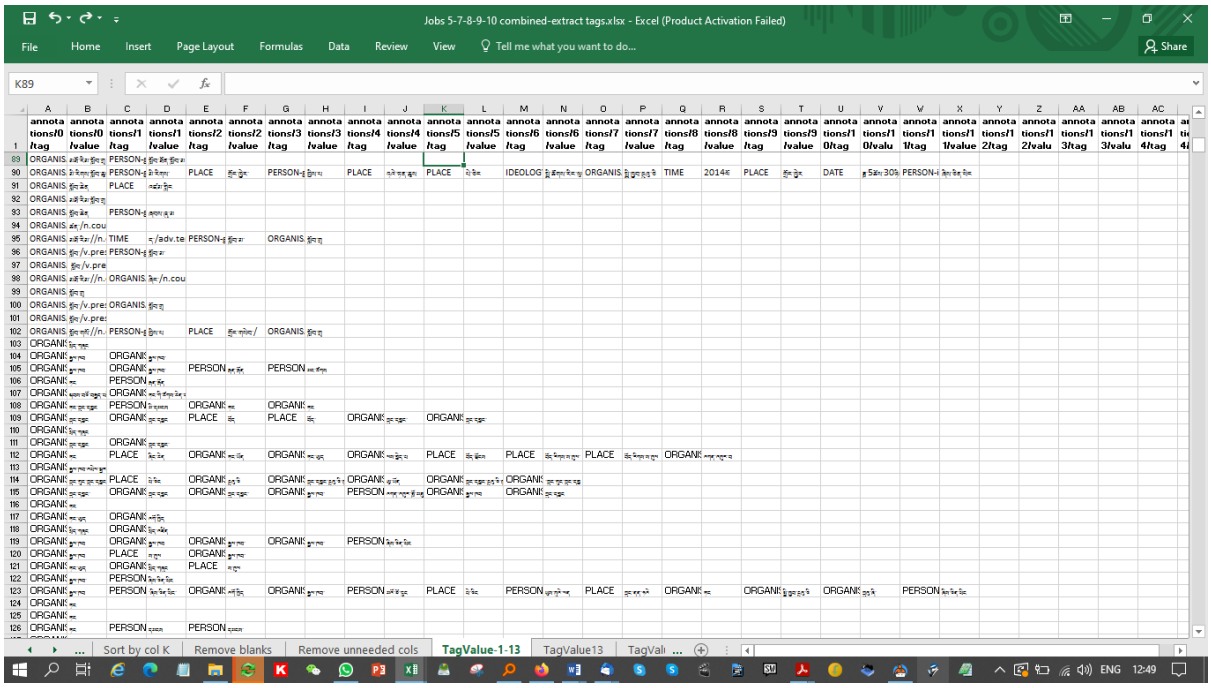
Download and conversion process

The method we used for preparing materials for reviewing was as follows:

(a) Download the completed job or task from Lighttag as a .JSON file.
(b) Convert the file to .CSV or .XLS format using http://www.convertcsv.com/json-to-csv.htm or a similar site.
(c) At this stage, you need to be very careful if your data includes text in non-Latin script. If it does, do *not* open any .CSV file in Excel, or it will lose and permanently erase the non-Latin content. Instead, *import* the file, using the menu items Data-Import-From Text, selecting the file, and choosing Comma delimited in the delimitations dialogue box. Once imported, the file will look like this:



(d) Concatenate the files for each job or task into a single file
(e) Check that each file has the same headers for the same column and adjust accordingly. (Note that if any classification was added in one job or task, it will have additional columns compared to jobs where there was no classification. Note also that some downloaded files had metadata or serial numbers for each example added as column 3, whereas others did not.)
(f) Simplify the data by removing all columns except Example_id, Content, Annotation/0/tag, Annotation/0/value, Annotation/1/tag, Annotation/1/tag…. etc. You may have 10 or 15 tags in a single example. (Note: the deleted columns give information about the time and identity number of each annotation, which are not needed for training). This will show each example with all its tags:



(g) To remove examples which had no tags, select the Annotation/0/value column, click on Data-Filter, and select “Blanks”. Then remove all the rows that remain, which will be blanks, ie. those that have no annotations. This will remove all rows that show examples (utterances) but which had no tags.

(h) Now select in turn all the tags which were the second, third, fourth or fifth, etc., in each example or row. The first tag is numbered Annotation/0/tag. So to do select the second tag, select column Annotation/1/tag, click on Data-Filter, check “Select All” and uncheck “Blanks”. Then copy columns example_id (an identifier added by Lighgttag for each example), content, Annotation/1/tag. and Annotation/1/value to your clipboard.

(i) Paste the copied list of Annotation/1/tags, content and values from the clipboard below the list of Annotation/0/tags, content and value.

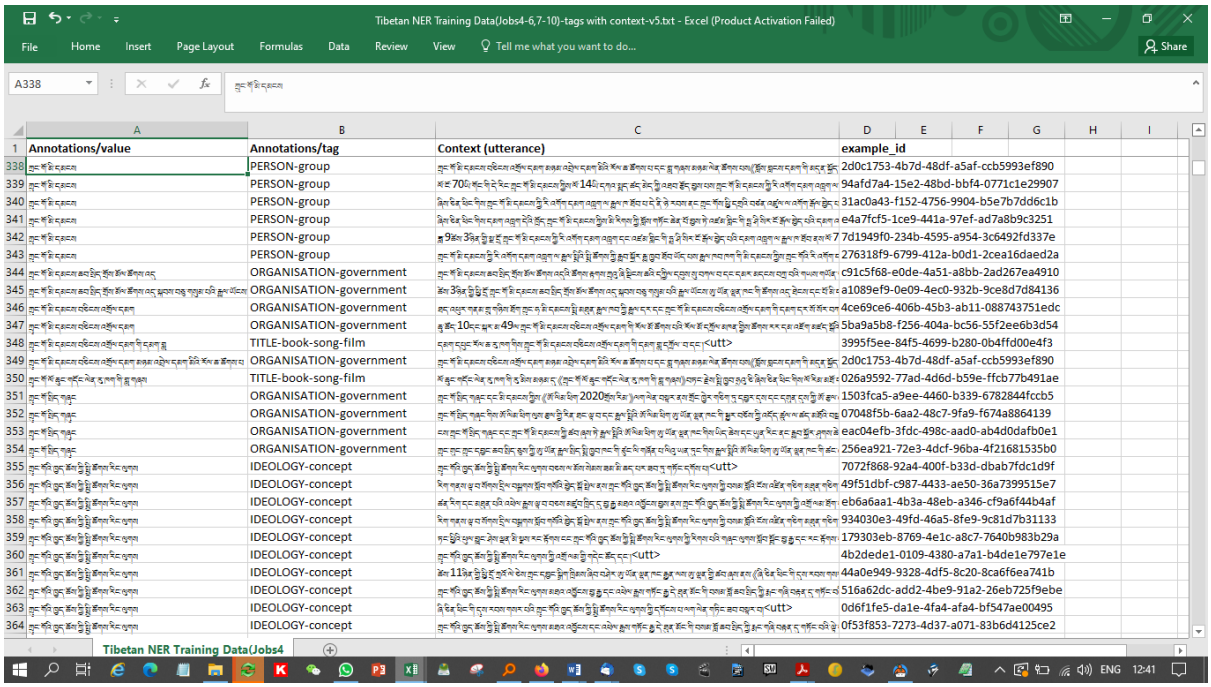
(j) Repeat with Annotation/2/tags, value and content, etc.

(k) Sort the data alphabetically A-Z by Annotation/value (first level) and Annotation/tag (second level). Rearrange the columns as needed.

If you are using Excel, it is best to save the file either as an xls file or as a Unicode .txt file, not as a .csv file, in case you forget to import rather than to open the file when you next view it in Excel: if you open a .csv file in Excel, the non-Latin script will be permanently lost. The best way to avoid this problem is to save/convert xls files to UTF-8 csv format instead of plain csv format.

Reviewing

You will now be able to view and review all tags for each word, organised according alphabetically, together with the context in which the word or term was found:



This will quickly show any inconsistencies and many of the tagging errors, if any. Our review showed approximately 244 out of 6,700 annotations needed re-assessment or correction, a review rate of 3.62%.

In the case of our data, which involved tagging named entities, most inconsistencies were due to confusion over certain guidelines. For example, words for places often had different tags depending on their context because of the metonymy rule – i.e., a place name is sometimes used for a government or institution. So we would tag China as place in "We go to China" but as organisation-government in "China voted for reform". Similarly, we found some confusion about the tag person-group: despite its name, this tag refers not to organised groups, but to social or cultural *categories* of people; organised groups of people would be organisation-government or organisation-non-governmental. Another common source of confusion was over which social or cultural categories of people should be tagged as person-group, and which should be tagged as person-title. We decided that general descriptive terms such as "youth", "the elderly", "farmers", "Buddhists", "monks", should be annotated with the former tag, while categories defined with by a formal or organisational title, such as "director", "lama", "geshe", or "doctor", should be annotated with the latter tag, even in the case of pluralities. Terms like "comrade", "official", "soldiers", "workers" were sometimes tagged as person-group and sometimes as organisation-government, a decision which depends largely on what type of society or ideological context one is dealing with (in a Leninist system, these terms are probably best viewed as political and organisational terms rather than as social categories or groups) Finally, many place-names, such as "region", "district", "county", "autonomous prefecture", and "city", were treated inconsistently because they both as places and as administrative terms used by governments.

By comparing all the annotation decisions made for each term and checking the context of that term, we were able to settle on which principles and guidelines to apply during the tagging process.

Conclusion

Carrying out the review process offline using standard spreadsheet software made the process of reviewing annotations far simpler than online reviewing, and allowed for much greater consistency when applying guidelines to the annotation process.