Subject: Re: German dialogue dataset translation
From: Tianhao Shen <thshen@tju.edu.cn>
Date: 06.06.23, 14:05
To: Lena Schwertmann <Lena.Schwertmann@hpi.de>
CC: Mehrad Moradshahi <mehrad@stanford.edu>

Hi Lena,

Really sorry for the late reply. I tested positive for COVID again and caught a high fever for several days. For the questions you ask:

1. I think you can first focus on the word forms that appear in the translations since we did this in other languages.
2. You should make sure all of the candidate values of a slot have the same part of speech (e.g., all nouns or adjectives), and you can decide the most suitable form of values for your language once they have consistent POS.
3. Not exactly. The canonicalized value is not always the shortest version, and this depends on the slot. For example, when we talk about cuisines, we usually use adjectives in English (e.g., Japanese cuisine). And for product countries/regions of a movie, we use nouns (e.g., America).

Hope these answers are helpful for you. Please let me know if you have any other questions.

Best,
Tianhao

Subject: Re: German dialogue dataset translation
From: Mehrad Moradshahi <mehrad@stanford.edu>
Date: 31.05.23, 01:53
To: Lena Schwertmann <Lena.Schwertmann@hpi.de>
CC: Tianhao Shen <thshen@tju.edu.cn>

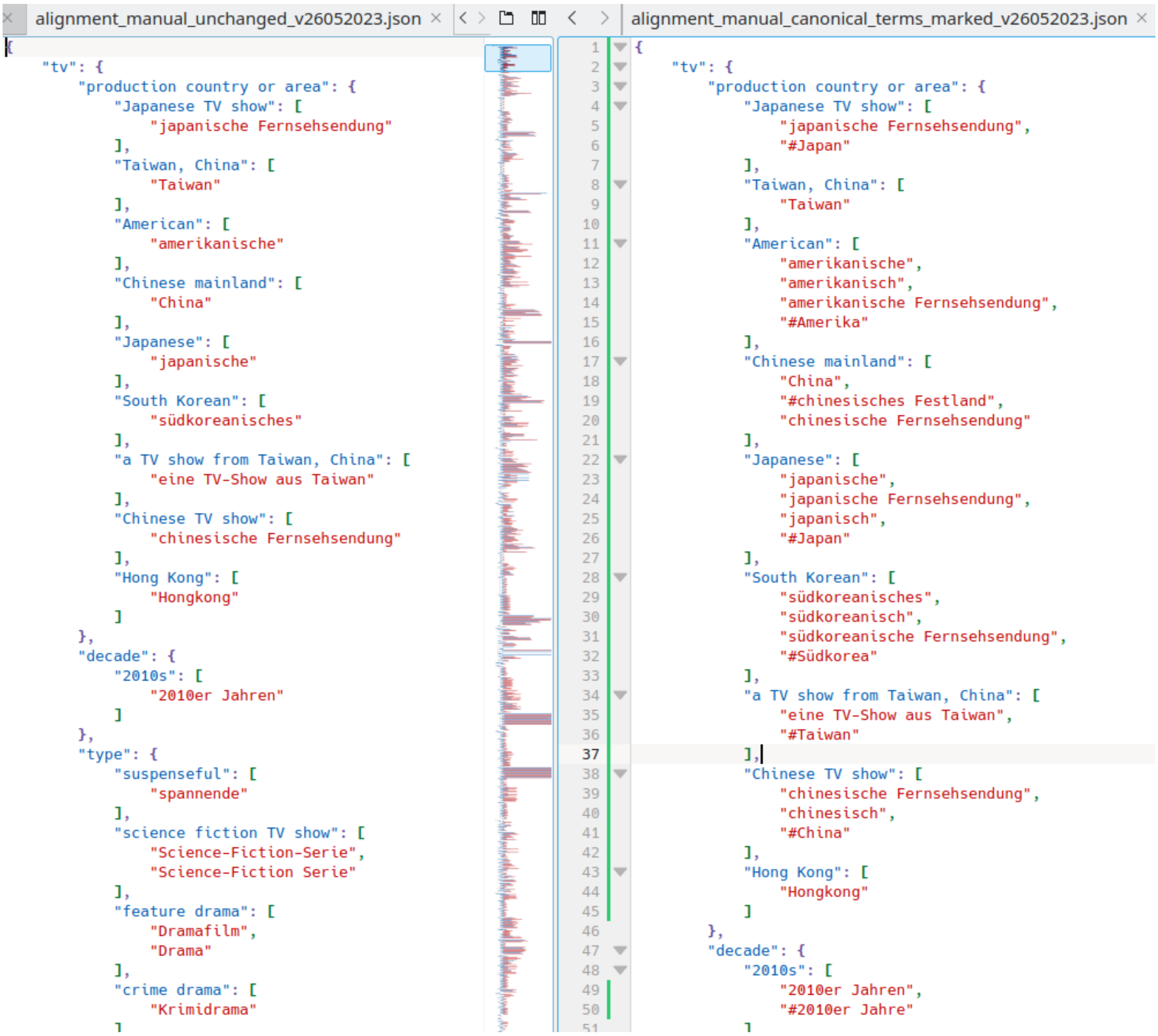
Hi Lena,

Apologies for my late reply. I'm a bit busy these days as I'm preparing to defend my thesis within a week. I'm looping in Tianhao who have been working on this project and is more familiar with the translation process. He should be able to help you with your questions. Thanks!

Best,
Mehrad

> On May 26, 2023, at 3:04 AM, Lena Schwertmann <Lena.Schwertmann@hpi.de> wrote:

> Hi Mehrad,
>
> My main question when working on step 7 is: Is it intended that I add additional translations here and select them as the canonical translation? In my case I think that that is very often necessary to adhere to the "POS consistency" that you request.
>
> In the guideline you write "
>
> In German, a lot of compound words and grammatical inflections exist, which influences the translations. For instance, compared to English and French, German has more cases which influence the specific word form, e.g. "American" can be translated as "amerikanisch, amerikanischer, amerikanisches, amerikanischen, amerikanischem", depending on the case :D
>
> So I have a couple questions:
>
> 1. Is it desirable that I add all these possible inflections? Or is this "completeness" not so important? In that case only the word shapes actually appearing in the translations would appear.
>
> 2. Should I - if possible - select the word stem that all these versions of the same word have in common as the canoncial translation? For instance, then I would select the noun "Amerika" for American instead of "amerikanisch" which is the actual case-neutral translation of the adjective American.
>
> 3. In general, do I understand correctly that the canonical translation should be the shortest version of the word, e.g. that it makes sense to always add the "basic" version of an adjective, e.g. "günstig" for cheap, when only the inflection "günstiges" appears in the original alignment file?
>
> Below in the picture I give some examples of how I could treat this and would appreciate a brief feedback whether the translations selected as canoncial with # correspond to your idea of POS consistency.
>
> On the left you can see the unchanged alignment_manual.json file that I get as an output from Step 6, on the right you can see how I would modify it manually according to how I understand the instructions in step 7:



```
{
    "tv": {
        "production country or area": {
            "Japanese TV show": [
                "japanische Fernsehsendung"
            ],
            "Taiwan, China": [
                "Taiwan"
            ],
            "American": [
                "amerikanische"
            ],
            "Chinese mainland": [
                "China"
            ],
            "Japanese": [
                "japanische"
            ],
            "South Korean": [
                "südkoreanisches"
            ],
            "a TV show from Taiwan, China": [
                "eine TV-Show aus Taiwan"
            ],
            "Chinese TV show": [
                "chinesische Fernsehsendung"
            ],
            "Hong Kong": [
                "Hongkong"
            ]
        },
        "decade": {
            "2010s": [
                "2010er Jahren"
            ]
        },
        "type": {
            "suspenseful": [
                "spannende"
            ],
            "science fiction TV show": [
                "Science-Fiction-Serie",
                "Science-Fiction Serie"
            ],
            "feature drama": [
                "Dramafilm",
                "Drama"
            ],
            "crime drama": [
                "Krimidrama"
            ]
```

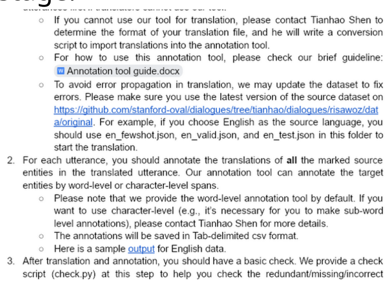
```
{
    "tv": {
        "production country or area": {
            "Japanese TV show": [
                "japanische Fernsehsendung",
                "#Japan"
            ],
            "Taiwan, China": [
                "Taiwan"
            ],
            "American": [
                "amerikanische",
                "amerikanisch",
                "amerikanische Fernsehsendung",
                "#Amerika"
            ],
            "Chinese mainland": [
                "China",
                "#chinesisches Festland",
                "chinesische Fernsehsendung"
            ],
            "Japanese": [
                "japanische",
                "japanische Fernsehsendung",
                "japanisch",
                "#Japan"
            ],
            "South Korean": [
                "südkoreanisches",
                "südkoreanisch",
                "südkoreanische Fernsehsendung",
                "#Südkorea"
            ],
            "a TV show from Taiwan, China": [
                "eine TV-Show aus Taiwan",
                "#Taiwan"
            ],
            "Chinese TV show": [
                "chinesische Fernsehsendung",
                "chinesisch",
                "#China"
            ],
            "Hong Kong": [
                "Hongkong"
            ]
        },
        "decade": {
            "2010s": [
                "2010er Jahren",
                "#2010er Jahre"
            ]
```

Thanks for your help!

Best regards,
Lena

On 14.03.23 06:23, Mehrad Moradshahi wrote:

> Hi Lena,
>
> Thanks for your email and welcome to the team! Great to hear about the progress.
>
> Regarding your question:
>
> Please first make sure you’re running the code in the “main" branch (or if you’re working on your own branch, it’s synced with upstream main branch).
> I haven’t ran the code recently but I remember we do still see errors when running those scripts on data splits (I don’t remember the exact number). It’s fine to stay within ~100 failed examples - some

> of these stem from annotation errors in the original dataset.
> To check, you can compare the failed examples between English data and German data to see what the issue might be.
>
> Also, we now have a document which provides step-by-step instructions on what needs to be done in each stage:
>
> 
>
> utterances first if translators cannot use our tool.
>
> - If you cannot use our tool for translation, please contact Tianhao Shen to determine the format of your translation file, and he will write a conversion script to import translations into the annotation tool.
> - For how to use this annotation tool, please check our brief guideline: Annotation tool guide.docx
> - To avoid error propagation in translation, we may update the dataset to fix errors. Please make sure you use the latest version of the source dataset on https://github.com/stanford-oval/dialogues/tree/tianhao/dialogues/risawoz/data/original. For example, if you choose English as the source language, you should use en_fewshot.json, en_valid.json, and en_test.json in this folder to start the translation.
>
> 2. For each utterance, you should annotate the translations of **all** the marked source entities in the translated utterance. Our annotation tool can annotate the target entities by word-level or character-level spans.
>   - Please note that we provide the word-level annotation tool by default. If you want to use character-level (e.g., it's necessary for you to make sub-word level annotations), please contact Tianhao Shen for more details.
>   - The annotations will be saved in Tab-delimited csv format.
>   - Here is a sample output for English data.
> 3. After translation and annotation, you should have a basic check. We provide a check script (check.py) at this step to help you check the redundant/missing/incorrect
>
> Dataset Translation Guideline
> docs.google.com
>
> This should help identify and address some of the issues you're facing during data translation. Please let me know if this is helpful.
>
> Best,
> Mehrad
>
>> On Mar 13, 2023, at 7:07 AM, Lena Schwertmann <Lena.Schwertmann@hpi.de> wrote:
>>
>> Dear Mehrad,
>>
>> I am a graduate research assistant with Prof. de Melo and I took over the project from Jim to finalize the German version of the fewshot file. I had a conversation with Jim about his work, but having now spent some time with the files, I still have a couple of open questions.
>>
>> This is the current state of the project: The database files are translated, and I successfully ran the convert.py and preprocess.py scripts on the de_fewshot.json file I got from Jim. The next step you describe in your dialogues repo, the check_entity.py script, produces 276 errors, i.e. instances where entities from the "output_text" are not contained in the "input_text" of the DA and DST parts of the dialogue.
>>
>> While I see that some of these instances surely stem from us needing to improve individual translations and entity alignments, I think that many of the errors are also inherited from the original en_fewshot.json file: When I run the same scripts on en_fewshot.json and fr_fewshot.json (as they are contained in your repo), I get output files with 158 and 467 errors, respectively. I attach these 3 output files here.
>>
>> Can you help understand me why this is the case? As I see similar errors in the other files, I am also generally unsure whether all the errors need to be resolved.
>>
>> How did you handle this with the other language teams? Maybe a meeting with you or someone from the French team could be fruitful?
>>
>> Thanks and best regards,
>> Lena
>>
>> -------- Forwarded Message --------
>> **Subject:** Fw: German dialogue dataset translation
>> **Date:** Fri, 24 Feb 2023 11:37:25 +0100
>> **From:** Maar, Jim <Jim.Maar@student.hpi.uni-potsdam.de>
>> **To:** Schwertmann, Lena <Lena.Schwertmann@hpi.de>

From: Mehrad Moradshahi <mehrad@stanford.edu>
Sent: 24 November 2022 09:01
To: Maar, Jim
Subject: Re: German dialogue dataset translation

Hi Jim,

I'm not sure what the issue is either. I tried adding you again with Editor access. Please check if you can upload the file now.

Thanks for sharing the file. We're writing up some notes to document the process of translations/verification we've done for the English data.
This is our codebase: https://github.com/stanford-oval/dialogues

GitHub - stanford-oval/dialogues: A unified interface for dialogue datasets

github.com

A unified interface for dialogue datasets. Contribute to stanford-oval/dialogues development by creating an account on GitHub.

There are some instructions in the README file on how you can turn your data into the right format and verify dataset and database are correct.
Please give it a try and let me know if you have any questions. I will share with you the other document once we finish it.

Best,
Mehrad

> On Nov 23, 2022, at 3:22 AM, Maar, Jim <Jim.Maar@student.hpi.uni-potsdam.de> wrote:
>
> Hi Mehrad,
>
> As a visitor, I'm not authorized to upload or create files in the shared folder. I don't know if there's an easy fix because many other people seem to have the same Problem https://support.google.com/drive/thread/83982595/add-files-in-visitor-session?hl=en.
>
> We have the translated fewshot data as a csv file in the format, that is used in the UI Tool. We also have the translated database files. Is there a straightforward way to get the dataset into the right format?
>
> Best,
> Jim
>
> **From:** Mehrad Moradshahi <mehrad@stanford.edu>
> **Sent:** 01 November 2022 18:15:09
> **To:** de Melo, Gerard

Cc: Tianhao Shen; Maar, Jim; Monica Lam
Subject: Re: German dialogue dataset translation

Hi Gerard,

That sounds great! Would you be able to share the data with us? You can upload it to this shared folder: https://drive.google.com/drive/u/1/folders/13CO3KzQSi1z4q3qS0zeDBeIdKI36S6VZ

There are multiple steps we need to take before the dataset is usable for training.
We need to first process the dataset into the right format. This is the codebase we're using to do that: https://github.com/stanford-oval/dialogues
There are instruction in the README file on how to add a new dataset and validate its contents. Besides dataset, you also need to add database files containing entities for each domain.

Please let us know if you have any questions.

Best,
Mehrad

> On Oct 31, 2022, at 6:59 AM, Gerard de Melo <gerard.demelo@hpi.de> wrote:
>
> Hi Mehrad and Tianhao,
>
> We have the few-shot data ready now in German. Is there any code you have that we could already use to play around with different models?
>
> Thanks,
> Gerard
>
> Gerard de Melo, Professor at HPI / University of Potsdam
> http://gerard.demelo.org/
>
> ===
> Chair for Artificial Intelligence and Intelligence Systems
> http://deepdata.demelo.org/
>
> Hasso Plattner Institute
> and
> University of Potsdam – Digital Engineering Faculty
>
> Legal Information:
> Hasso-Plattner-Institut für Digital Engineering gGmbH
> Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany
> Registergericht: AG Potsdam, HRB 12184 P
> Geschäftsführung: Christoph Meinel, Marcus Kölling
> ===
>
> On 2022-10-04 21:09, Tianhao Shen wrote:
>
>> Hi Gerard,
>> Thank you so much for your interest in this project!
>> For your question, the answer is yes. It is possible to annotate using our tool after translation is done. However, the output format of German translation should be determined in advance (e.g., a csv format with the following columns: dialogue id, turn id, role, English utterance and corresponding German translation), then we can write a script to convert this file into the format which can be read by our annotation tool. Please let me know if you have any questions about this process.
>> Best,
>> Tianhao
>> On Oct 1, 2022, at 06:53, Mehrad Moradshahi wrote:
>>
>>> Hi Gerard,
>>>
>>> Yes I believe you can do that. It should be possible to feed the translated sentences to the annotation tool.
>>> I'm CCing Tianhao (from the Chinese team) that developed the annotation tool to provide more details. Thanks!
>>>
>>> Best,
>>> Mehrad
>>>
>>>> On Sep 27, 2022, at 8:49 PM, Gerard de Melo <gerard.demelo@hpi.de> wrote:
>>>> Hi Mehrad,
>>>> Thanks a lot for the quick response and the detailed explanations!
>>>> The translation is indeed a bit more involved than expected. Professional translation services normally are not able to use custom UI tools or do span annotations. Are some other teams doing this in two steps, i.e. first having translators create the translation, then having other linguistic annotators use the UI to create the alignment?

Thanks,
Gerard
Gerard de Melo, Professor at HPI / University of Potsdam
http://gerard.demelo.org/

===

Chair for Artificial Intelligence and Intelligence Systems
http://deepdata.demelo.org/
Hasso Plattner Institute
and
University of Potsdam – Digital Engineering Faculty
Legal Information:
Hasso-Plattner-Institut für Digital Engineering gGmbH
Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany
Registergericht: AG Potsdam, HRB 12184 P
Geschäftsführung: Christoph Meinel, Marcus Kölling

===

On 2022-09-28 03:25, Mehrad Moradshahi wrote:

> Hi Gerard and Maar,
> I'm excited that you're interested in working on the project!
> The translation process is quite involved and we've discussed it over email with the other teams. We have written down a summary of the steps in the paper we're writing (we're aiming to submit it by December 1st).
> Please look at Section 3:https://www.overleaf.com/read/djhmkwggscsf <https://www.overleaf.com/read/djhmkwggscsf>
> The main challenge is during translation, then entities are gonna change based on translator preferences. The Chinese team has shared with us an annotation UI tool that translators can use to translate and also annotate entity spans during translation. Please see here for a guide and the tool:https://drive.google.com/drive/folders/1f3_WhW6YYBVmGSP3zBlIPvY6JYnn74X2<https://drive.google.com/drive/folders/1f3_WhW6YYBVmGSP3zBlIPvY6JYnn74X2>
> Github repo for the tool if you wanna build it for Linux or MacOS system: https://github.com/danovw/annotate_translation<https://github.com/danovw/annotate_translation>
> I'm also sharing with you the original dataset in Chinese and the translated data in English:https://drive.google.com/drive/folders/13CO3KzQSi1z4q3qS0zeDBeIdKI36S6VZ?usp=sharing<https://drive.google.com/drive/folders/13CO3KzQSi1z4q3qS0zeDBeIdKI36S6VZ?usp=sharing>
> It shows which parts of the data need to be translated and what information are kept during translation to process the final dataset.
>
>> The paper also mentions that you have a framework for doing automated translations while preserving entity names. To what extent is this relevant? Would it be possible for you to run that framework with German as the target language such that we can then focus on post-editing the translations?
>
> Yes we can generate dataset in German using our translation + alignment tools. In fact we'll be using that to generate the training data. However, the final format of that dataset is different than what the annotation tool I shared above expects. I can definitely share with you a sample of the translated data in German (the entities would be in English). Then you would need the human post-editing to both correct the bad translations, replace the slot values with correct translation in German, and keep track of .entity spans in the source and target sentences.
> Please let me know if you have any other questions.
> Best,
> Mehrad
>
>> On Sep 27, 2022, at 11:00 AM, Gerard de Melo <gerard.demelo@hpi.de <mailto:gerard.demelo@hpi.de>> wrote:
>> Hi Mehrad,
>> We are ready to begin the translations (I have CC'ing our collaborator Jim Maar) and wanted to double-check how to proceed here. Obviously, we need to translate the user_utterance and system_utterance sentences.
>> Are the annotations supposed to remain in English only?
>> The paper also mentions that you have a framework for doing automated translations while preserving entity names. To what extent is this relevant? Would it be possible for you to run that framework with German as the target language such that we can then focus on post-editing the translations?
>> Thanks,
>> Gerard
>> Gerard de Melo, Professor at HPI / University of Potsdam
>> http://gerard.demelo.org/ <http://gerard.demelo.org/>
>>
>> ===
>>
>> Chair for Artificial Intelligence and Intelligence Systems
>> http://deepdata.demelo.org/
>> Hasso Plattner Institute
>> and
>> University of Potsdam – Digital Engineering Faculty
>> Legal Information:
>> Hasso-Plattner-Institut für Digital Engineering gGmbH
>> Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany
>> Registergericht: AG Potsdam, HRB 12184 P
>> Geschäftsführung: Christoph Meinel, Marcus Kölling
>>
>> ===
>>
>> On 2022-09-10 01:33, Mehrad Moradshahi wrote:
>>
>>> Yes. Only those three splits need to be translated/ post-edited by humans.

> > > > > > > > Best,
> > > > > > > > Mehrad
> > > >
> > > > <en_fewshot_output.csv>
> > >
> > > <fewshot_entity_check_1_fr_13032023.tsv><fewshot_entity_check_1_en_13032023.tsv>
> > > <fewshot_entity_check_1_de_13032023.tsv>