First and Last names for dialogue on Swedish NER data

In our experience using Named Entity Recognition (NER) models for supporting dialogue assistants, we have observed that, when speaking about certain entities in certain contexts, it is necessary to have some more fine-grained distinction between entities. For instance, a dialogue system that needs to take data of the user may need to distinguish between first and last names or different location entities (street, postal code, city, country, etc).

That has motivated us to perform a study that takes the data of this project further than annotation and evaluation inside the same data and domain, but also to the development of a dialogue application in Swedish.

To that end, we have added more granularity in the person names entity (PER) in an extract of the data. We have done a semi-automatic reannotation on an extract of the Språkdatalabb data to include distinction between first and last person names.

Following that, we have trained an NER model with the spaCy NER pipeline.

To check that there is not any loss performance after splitting the person name into entities and to understand the benefit of having this new granularity in person names, we have evaluated the performance of the different NER models using two different methods: (i) numerically through some testing in written data from the same project and (ii) in a dialogue assistant demonstration performed during the Språkdatalabb Reference Group Workshop (05/12/2019). We will talk briefly about those two evaluations.

Data

The data used for this reannotation process is an extract of the annotations made by Recorded Future for the Språkdatalabb project. We received this data in June 2019, so it contains the part of the data that was annotated by that date. The data consists of 45036 sentences which contain the following annotated entities:

- PER (person name).
- LOC (location).
- ORG (organization).
- TIT (title). E.g. “pressekreterare”, “programledaren”, “Åklagaren”.
- REL (religion). E.g. “judarna”, “muslimsk”, “hinduiska”.
- NAT (nationality). E.g. “brasilianaren”, “nordiska”, “västsvenskt”.

- PRO (product). E.g. “The late late show with James Corden”, “En natt på jorden”, “Duracell Ultra”. nt.

Moving towards first and last names

In this study we have performed the reannotation of the person name entities so they reflect more granularity. That is, we have split the person name (PER) entities in the data into first (PER-first) and last name (PER-last).

The original annotated data presents the NER tags with IOB format (Inside, Outside, Beginning), which contains information about the boundaries of multiple word entities. This extra information allowed us to create a semi-automatic process to reannotate the PER entities as following.

Programledaren TIT Lotta B-PER Bouvin-Sundberg I-PER frågar 0 : 0 Fredrik B-PER Reinfeldt I-PER , 0 är 0 du 0 [...] Bouvin-Sundberg PER ställer 0 om 0 frågan 0 ett 0 par 0 gånger 0 , 0 varpå 0 Reinfeldt PER [...] Men 0 enligt 0 president TIT Juan B-PER Manuel I-PER Santos I-PER saknas 0 fortfarande 0 314 0 människor 0	Programledaren TIT Lotta PER-first Bouvin-Sundberg PER-last frågar 0 : 0 Fredrik PER-first Reinfeldt PER-last , 0 är 0 du 0 [...] Bouvin-Sundberg PER-last ställer 0 om 0 frågan 0 ett 0 par 0 gånger 0 , 0 varpå 0 Reinfeldt PER-last [...] Men 0 enligt 0 president TIT Juan PER-first Manuel PER-first Santos PER-last saknas 0 fortfarande 0 314 0 människor 0

We wanted to make this conversion of the entities from PER/B-PER/I-PER to PER-first/PER-last without having to revisit manually all the data, which would have been a very time-consuming task. Therefore, we studied the data by extracting all the PER entities and ended up with some assumptions:

- All the person names that appear in the data as B-PER (and maybe also as PER) but *not* as I-PER are first names. E.g.: Fredrik, Lotta, Juan... Therefore, we can generalise and annotate all those cases as **PER-first**.
- In the same way, all the person names that appear in the data as I-PER (and maybe also as PER) but *not* as B-PER are last names. E.g.: Andersson, Eneroth, Bush... Therefore, we can generalise and annotate all those cases as **PER-last**.

We have observed some very rare exceptions for the two assumptions, but we can generalise them and still be certain that they will not have any impact for this evaluation.

For the names that appear in the data both as B-PER and I-PER, we had to perform some manual disambiguation work, which took only a few days. After disambiguating these cases, we were able to change all of them automatically into PER-first or PER-last.

Evaluation

1. Plain NER evaluation

We have done some evaluation on models trained with the **spaCy 2.1.8** NER pipeline. The evaluation of these models has been performed in a split of the same Svenskt Språkdatalabb data that was not used to train the models. We used each of the trained models to run named entity recognition on the test data and obtain some results from that.

- **NER model without first and last names**

  General NER scores:
  Precision 83.3122 Recall 81.2469 F-score 82.2666
  PER:
  Precision 91.6497 Recall 91.2470 F-score 91.4479

- **NER model with first and last names**

  General NER scores:
  Precision 83.3355 Recall 81.5118 F-score 82.4135
  PER-first:
  Precision 91.6226 Recall 93.5285 F-score 92.5658

PER-last:
Precision 89.6887 Recall 93.3828 F-score 91.4985

Without entering into explaining what these metrics mean specifically, they do not really show any loss in performance when having the new granularity in person names. This evaluation is also useful to check if the model is effectively separating between first and last names, which is doing.

However, this plain numerical evaluation is not suitable for understanding the benefit of distinguishing between first and last names in a dialogue context. For that, we need to understand how that affects the dialogue itself.

2. Dialogue evaluation

During our presentation in the Svenskt Språkadalabb Reference Group Workshop (05/12/2019), we explain why there is a need of having more granular entities from the dialogue perspective, mentioned briefly our reannotation process and made a demo in a dialogue application that benefits from that distinction between first and lasts names.

The application is a virtual receptionist which goal is to determine if it has to grant access to a building (e.g. a company office) to a visitor who wants to visit some host in the building. The virtual receptionist needs to know the first and the last name(s) of both host and visitor to know for sure who they are and if they are allowed to enter the building, and then it needs to distinguish between them. An interaction example of this would be:

System (S)> Vem ska du besöka?
Human user (U)> Jag ska träffa Krona.
S> Vad heter den du ska besöka i förnamn : Maria eller Samuel?
U> Maria.
S> Vem kan jag hälsa ifrån?
U> Lisa.
S> Vad heter du i efternamn?
U> Jonsson.
S> Hej Lisa Jonsson. Jag öppnar dörren och meddelar Maria Krona att du är här. Välkommen.

There are two things to remark in this interaction. Firstly, the system has recognised "Krona" as a last name. This receptionist system knows the names of all the hosts (it's reading a database) and it has found that there are two with the last name "Krona". Therefore, it asks specifically for the first name since it knows that "Krona" is a last name.

Secondly, the visitor has only said their first name "Lisa". Since the system has recognised that as a first name through the NER model, then it can know that the

information it lacks is the last name of the visitor. Without this distinction between first and last names, it gets more complicated for the system to know if the visitor has said a first name or a last name, and therefore the system would have to ask for the full name instead of asking only for either the first and the last name. This would lead to more frustration for the user and therefore a worse user experience.

A few examples of how the user could talk to the virtual receptionist and how the distinction between first and last names would work can be seen in the slides of that presentation provided along with this documentation.

As explained above, NER models which are capable of distinguishing between both first and last names would be clearly beneficial in scenarios like the example receptionist. These models reflect better a reality in which person names are not just a uniform block: they are composed by several pieces which, in this case, are first and last names.

Suggestions for future work

Lately we have been moving into using the new DIET[1] classifier implemented in Rasa for both the NLU (Natural Language Understanding) and NER tasks of our dialogue systems. We have observed in other projects that this provides better entity recognition in certain contexts and the possibility of having entity roles in utterances, which makes possible to have more advanced dialogue features. Such entity roles allow, for instance, to train models that distinguish between two different types of person names in the virtual receptionist when said in the same utterance:

> System (S)> Vem ska du besöka?
> Human user (U) > Jag heter Lisa Jonsson och jag ska träffa Samuel Krona.
> S> Hej Lisa Jonsson. Jag öppnar dörren och meddelar Samuel Krona att du är här.

More recent pre-trained models such as BERT[2] or XLM[3] could also be investigated and tested for this NER task. Using such models was a suggestion made by Peltarion in the Svenskt Språkadalabb Reference Group Workshop (05/12/2019). We have been following their progress in using those two models for Swedish data in the project Language Models for Swedish Authorities and the results seem very promising.

A comparison study for our data using different architectures (e.g. DIET vs. BERT vs. XLM vs. ?) for NER would be relevant as a follow-up and evaluation of this

[1] https://blog.rasa.com/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/

[2] https://huggingface.co/transformers/model_doc/bert.html

[3] https://huggingface.co/transformers/model_doc/xlm.html

project data and in addition also consider entity roles as an important future addition complementing the value that was targeted within the scope of this project. This could be suggested as a topic in future projects with existing or new project partners.