Pseudonymization of Language Learner Data

Beáta Megyesi and Elena Volodina

We present de-identification and pseudonymization of a learner corpus within the ongoing research infrastructure project SweLL[1]. The main project aim is to make available a linguistically annotated corpus of essays written by second language (L2) learners of Swedish. To ensure that the data collected in the project can be used openly in research protecting the subjects' integrity, we developed data handling flow, a set of metadata about the learners, pseudonymization principles of learner texts, and tools in support of pseudonymization. During data collection and storage, the data needs to be handled in a secure way, and the participating subjects must be de-identified in the corpus, where common personal identifiers such as names, age, geographic places, dates must be identified, masked and eventually replaced. These identifiers might occur in metadata about the learner, and in the learners' text(s).

The SweLL project adopted a rather restrictive approach to metadata describing important aspects about each produced text and learner so that learners are de-identified while still providing important information for research purposes about the learner's gender, age given in 5-year interval spans, total time in Sweden, education level, mother tongue, and languages spoken in various communicative situations. The metadata does not provide exact date of birth, arrival date to Sweden, the country of origin or nationality of the learner, and no information is given about the educational establishment, where the essays have been collected.

De-identification through metadata might not be solely satisfactory, since the texts written by a learner may, and in fact often contain personal information about the learner. Pseudonymization involves the identification of personal information that can relate to the subject (e.g. My name is *Ali*), and the classification of that information, masked into certain predefined types (e.g. My name is *first_name*). As the first step, we manually mark-up text segments that reveal personal information in the corpus data. The identified segments are categorized as personal names, institutions (referring to schools, work place, sport teams), geographic data (such as country, city, region, areas, street name, numbers), transportation types and line names/numbers, age, date, phone number, email address, personal web page, social security number, account number, certificate/license number, profession and education, and sensitive information revealing physical or mental disabilities, political views, unique family relations, and any other items not covered by the previous categories.

Each marked text string with a category is then replaced in a systematic way to reproduce a "natural" text to increase reading flow. This step includes assigning unique id-numbers to each entity within a certain category type so if the particular entity is repeated in the text, the same running number is assigned to it and can be replaced by the same word. We also add morphological information to each masked entity to be able to replace it in the same morphological form as the original.

There are several ways to mask the sensitive information through substitution, either by rendering, or by replacement with another pre-defined token of the same category. Rendering is applied to information that can be collected from general resource lists, such as personal names and surnames; city and country names, nationalities and languages; geographic names; street names; names of schools, institutions, work places; etc. Replacement applies to strings containing information with certain formatting where general resource lists cannot suffice. Such cases include middle names or initials, numerical information such as phone numbers or dates. In some cases, when the annotator does not know how to categorize a certain text string, the original text is kept but marked by a placeholder. Distinction is made between objects that need to be replaced because of sensitivity, and objects that might be sensitive but can be replaced later, or to be removed later.

The pseudonymized corpus is under development, as are the tools supporting the pseudonymization process. We expect the corpus and the tools to be released as open source by the end of 2020.

References:

[1] https://spraakbanken.gu.se/eng/swell_infra

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C-J., Sundberg, G., Wirén, M., Volodina, E. (2018) Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In *Proceedings of the 7th NLP4CALL*, SLTC workshop, Stockholm, Sweden.

Volodina E., Granstedt L., Megyesi B., Prentice J., Rosén D., Schenström C-J., Sundberg G. & Wirén M. (2018). Annotation of learner corpora: first SweLL insights. Proceedings of SLTC-2018.