https://www.linguateca.pt/DIP/

Identifying Literary Characters in Portuguese: Challenges of an International Shared Task

Diana Santos, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes,
Cristina Mota, Emanoel Pires, Rebeca Schumacher Fuão

Introduction

We introduce the problem of identifying characters in literary text, and mention some specific issues that are special for Portuguese in the context of presenting DIP (*Desafio de Identificação de Personagens* - Character identification challenge). DIP is a shared task to foster work in the area and produce resources for computational literature studies in Portuguese. We see this as a natural first step of distant reading in Portuguese, given the importance of characters for literary studies.

Motivation from literary studies

A literary character is important in fiction, as it sustains the plot and moves it in a particular direction, and may also organize the discourse. Since they are created by the author in an attempt to revive or project experiences, they are ideological products. If we can get information on characters from thousands of works, we may be able to read the (character) landscape by epoch, literary genre, and/or author, expanding our base with many works outside the literary canon, which may provide interesting opportunities for postcolonial, gender and queer studies. Finally, the form of the names themselves is relevant, not only because address forms reflect different social status, but because some epithets have relevant interpretations.

Motivation from CS studies

From a Computer Science (CS) perspective, one can see the problem as a standard information extraction task, that from literary works must populate a knowledge base with characters, their attributes, and relations to other characters.

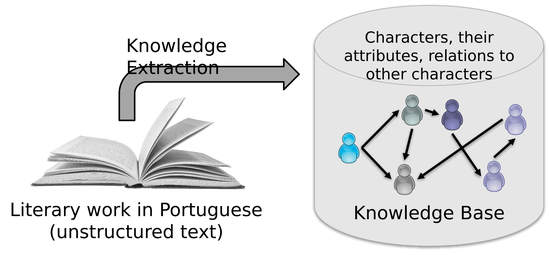

Fig. 1: The task of DIP in a nutshell.

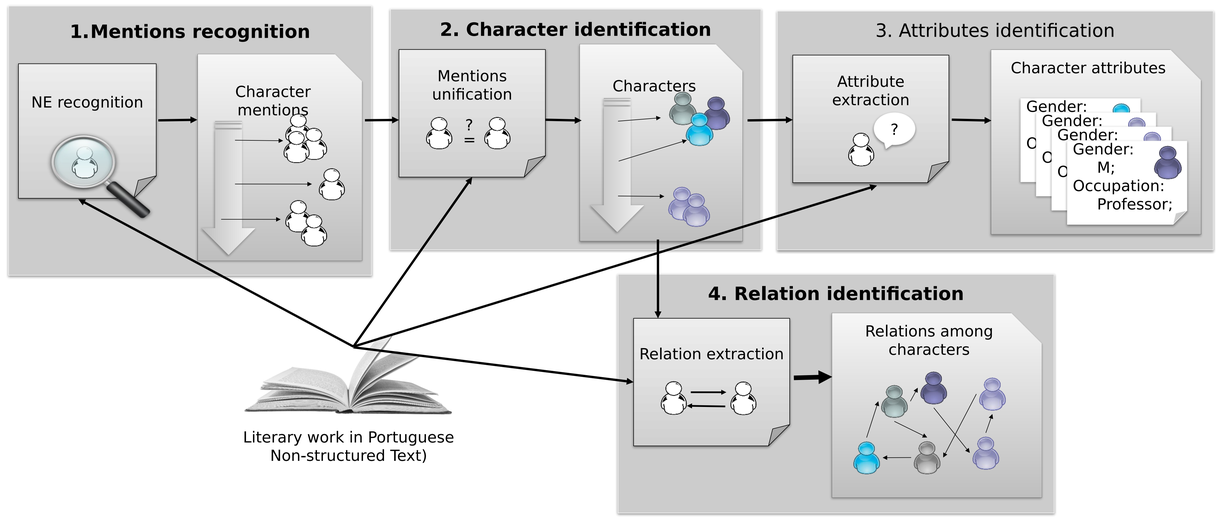

Fig. 2: Different subtasks in DIP (figure adapted from [1]).

In order to perform this extraction task, one may have to use various NLP techniques, which make the challenge more related to understanding Portuguese text.

Figure 2 presents the main steps in DIP. The first is named entity recognition of the person names present in the narrative. Note that DIP is not interested in all person names, only plot characters. In general, historical people, or characters from other literature should not be flagged as characters.

As one character is rarely always called the same way throughout a book, we have the widely known coreference problem. This is what the unification step is supposed to solve, identifying all mentions that refer to the same character. Mentions are often depending on the context: depending on who is talking correferring to her, the name may be radically different. So, a serious challenge is to identify the set of denominations by which a novel character is mentioned in a work.

The two last steps presented in Figure 2, attributes and relation identifications, aim at recognizing the gender, profession/occupation/social status of the detected characters, and the relationship between them.

DIP allows the creation of character networks, currently a hot task, as can be seen in the overview by [1]. These can in turn be used for genre prediction, (visual) text summarization, comparing fiction with (social) reality, comparing different literatures, and deciding who are the main characters.

The expected results of DIP are: (i) a list of characters, each character represented by a list of possible mentions; (ii) the gender of the character (M, F, or M and F); and (iii) the profession, or occupation, or social status of the character (can be more than one, or none) the family relations among any characters.

The DIP setup

DIP organization will provide 200 books in digital form to the community, which has 48 hours to return the results. This effectively prevents a close reading of the 200 works, ensuring that the analysis is done automatically.

In order to have large numbers of works to process, distant reading of literary collections necessarily includes works from several time periods – after all, one of the goals of distant reading is to address trends and changes in time, see [2]. This means that, specifically for Portuguese, systems will have to process several different orthographies and styles, including different ways to describe professions and relations. For example, boticário or cacaolista are not exactly modern words to refer to a pharmacist. The historical novel subgenre, quite frequent in Portuguese, brings a set of additional problems [3], such as old names, jobs and address forms.

After the submission period is over, the golden collection (containing the right information for 40 out of the 200 books) is made publicly available, and the evaluation results are computed. A workshop presenting the results and the different approaches of the participants will then be organized, followed by the publication of a journal volume. All data amassed about the literary works will also be released.

The systems will be evaluated separately on the five tasks, with the final score per book being the sum of the five measures. The ranking among the systems is done by macro-averaging over the golden collection.

References

1. Labatut, V., Bost, X.: Extraction and analysis of fictional character networks: A survey. ACM Computing Surveys 52(5) (2019)
2. Underwood, T.: Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press (2019)
3. Santos, D., Bick, E., Wlodek, M.: Avaliando entidades mencionadas na coleção ELTeC-por. Linguamática 12(2), 29–49 (dezembro 2020)