,

Extração de Informação sobre Personagens Literários em Português

Extraction of Literary Character Information in Portuguese

Eckhard Bick
University of Southern Denmark

Abstract

Este capítulo descreve o PALAVRAS-DIP, um sistema para a identificação automática de personagens e dos seus perfis sociais na literatura portuguesa e brasileira. O sistema foi concebido como um módulo adicional para um analisador morfossintáctico e semântico. Etiquetamos as entidades nomeadas (NE) humanas para profissão e posição social, e usamos as etiquetas relacionais do formalismo Constraint Grammar (Gramática de Restrições, CG) para estabelecer co-referências (por exemplo, anáfora de pronomes, verbos com sujeito zero) assim como relações familiares entre as personagens. A anotação de base resultante permite a extração de redes de personagens. O programa de extração reconhece e agrupa as variantes de nomes de personagens e distingue entre nomes que têm função narrativa e nomes contextuais de referência cultural. O desenvolvimento do sistema foi motivado pelo DIP, uma avaliação conjunta sobre 100 romances históricos, evento em que uma versão protótipo do sistema obteve medidas F razoáveis para as tarefas de identificação de personagens (63,4%) e de unificação/co-identificação de nomes (68,1%), mas teve problemas com as relações familiares (15,5%).

Keywords

leitura distante, extração de informação, reconhecimento de entidades nomeadas, constraint grammar, resolução de anáforas

Abstract

This chapter describes PALAVRAS-DIP, a system for the automatic identification of characters and their social profiles in Portuguese and Brazilian literature. The system has been designed as an add-on module for a morphosyntactic and semantic parser. We tag human named entities (NE) for profession and social position, and use Constraint Grammar (CG relational tags to keep track of co-reference (e.g. pronoun anaphora, zero-subject verbs) and family reations between the characters. The resulting base annotation allows the extraction of character networks. The extraction program recognizes and bundles character name variants and distinguishes between names with a narrative function and simple cultural references. System development was motivated by DIP, a shared-task evaluation on 100 historical novels, where a prototype version achieved reasonable F-scores for character identification (63.4%) and alias resolution (68.1%), but underperformed for family relations (15.5%).

Keywords

Distant reading, IE, NER, Constraint Grammar, anaphora resolution

1. Introduction

At first glance, the task of automatically extracting characters and their social relations from literature seems like an extension of named entity recognition (NER). However, while classical NER does identify candidate tokens for characters, the method needs to be adapted to the literary genre (e.g. Bornet & Kaplan (2017), for French), and as characters may be identified by many different variants of their name (e.g. first or second name, with or without a honorific, title, middle name, nickname etc.), name instances need to be unified. Also, the basic NER tag of "person" does not make the distinction between narrative, "functional" names and cultural reference names referring to gods, poets and historically important people. Finally, characters form social networks that go beyond simple recognition. In order to extract these networks, characters' social attributes and their mutual relations must be extracted too — information that may change chronologically throughout a book. In a wider context early literary analysis, other narrative character information may added, such their actions, plot event participation and affect states (Goyal et al., 2010). Much of the previous work in the field has been done on English (e.g. Labatut & Bost (2019); Valls-Vargas et al. (2014), often using classical, older texts. The work described here has a similar focus on classical literature, but addresses Portuguese, an under-represented language where previous research had targeted the

DOI: 10.21814/lm.15.1.397

less complex topic of children's stories Mamede & Chaleira (2004). Our research was carried out in the context of the DIP shared task (Desafio de identificação de personagens), a Portuguese-language character identification challenge organized by Linguateca[1], NuPILL, UEMA and UiO (Santos et al., 2022), and described in detail in this volume. The system uses a morphosyntactic and semantic parser, PALAVRAS (Bick, 2014) to provide a grammatical base annotation and named entity (NE) mark-up. The new DIP extension unifies name instances, verifies name gender at the text level and adds tags for title, profession or social standing for those names that it deems characters rather than cultural references. It also adds a new type of relational tags for family relations and extends PALAVRAS' experimental co-reference annotation using longer spans, text variables and explicit referent tags. The DIP extension is a rule-based system based on the same formalism as PALAVRAS itself, Constraint Grammar (CG3). Specifically, we use the CG3 variant (Bick & Didriksen, 2014), which supports the use of long-distance relational tags, as well as the capture, use and unification of both tag-level and text-level variables.

2. Grammatical base annotation

PALAVRAS' annotation scheme comprises information from various linguistic levels, including lemma, morphology, syntactic function and dependency structure. At the semantic level, in addition to the afore-mentioned NER, the parser provides a (disambiguated) noun ontology, as well as framenet structures and semantic roles (Bick, 2022). This linguistically high level of pre-annotation is an important prerequisite for the extraction of character networks, as pointed out by Chaturvedi et al. (2017), who use linguistic information such as frame semantics to complement the simpler bag-of-words approach in their feature vectors when tracking character relationships.[2] Both PALAVRAS itself and the add-on DIP module are rule-based systems. This makes for great transparency and allows fairly straightforward error correction and genre adaptation, but as a rule-based set-up cannot simply copy its tokenization and category distinction from a body of training data, some adaptations have to be made to meet a given annotation standard — in this case the one dictated by the DIP conventions. Of course, the problem is limited, as it does not concern internal tagging, but only what is visible in the final output. Specifically, changes had to be made regarding the inclusion (or non-inclusion) of honorifics, titles, family terms and title-like profession terms in names. As a rule of thumb, title-like words were included in character names, if they can co-occur with a name in the vocative.

- part-of-name: *Com(p)adre, Dama, Dom, Don(a), Doutor, Dr., Frau, Fräulein, Frei, Herr, Lady, Lord, Madame, Mademoiselle, Maestro, Mano, Miss, Mister, Monsenhor, Monseor, Monsior, Monsenhor, Monsieur, Mlle, Nhá, Nhô, Padre, Prima, Primo, Prof(a)., Senhor(a), Senhorita, Sô, S(n)r., S(n)r.ª, Tia, Tio*

- not part-of-name: *Conde(sa), Duque(sa), Imperador(a), Príncipe, Rei, Rainha, Vinsconde(sa), coronel, juiz, mãe, pai, neto, avó etc.*

Another necessary adaption concerned the textual input itself, as most DIP texts were historical in nature and contained lexical and orthographial variation not seen in modern texts, often compounded by what might be photo reproduction, "de-pdf'ing" and OCR posing problems. The resulting unrecognizable wordforms pose a problem to both the base parser and later tasks — inparticular, name unification. As PALAVRAS has been used on both historical data, transcribed speech and social network input, it contains some non-standard lexical extensions, and it does perform a certain amount of normalization itself, including some automatic spell checking useful for the task at hand. In the output, both the original and the normalized wordforms are provided, but the lemma as well as morphosyntactic and semantic annotation will be based on the normalized form. The parser does not normalize names, though, as this will produce many false positive "corrections." Therefore, the DIP module implements its own, "relaxed" name unification method, where a Levenshtein (spelling) distance[3] of 1 is tolerated for names between 4 and 9 letters (e.g. *Luíza, Luiza, Luísa, Luisa* or *Hamlet, Hamleto*), and a Levenshtein distance of 1 or 2 for longer words (e.g. *Christovam, Chrystovam, Christovão*), provided there also is a gender match (i.e. not *Francisco, Francisca*).

[1] https://www.linguateca.pt/DIP/

[2] In this work, relationships are seen as (evolving) latent states. The family relations addressed in DIP can be seen as a more stable subset of overall relations.

[3] meaning 1 exchanged, added or removed letter

3. Co-reference resolution

Though a name or its variant may occur many times in a given literary text, crucial information about the character, such as profession, marital status and descent (parents' names), may be provided explicitly only once, and henceforth assumed to be known to the reader. The information may also be implicit-only, for instance providing a work place or typical tool instead of a profession, or hinting at family relations through the form of address in dialogue. In any case, all (or as many as possible) mentions of a given character need to be kept track of in order to make such connections where and when they occur. The importance of this task is illustrated by the fact that most character occurrences are not names, but pronouns[4]. For our Portuguese data, zero-subject finite verbs with pronoun ellipsis (50.3% in the first 100 books of the DIP collection) are more frequent than finite verbs with personal pronoun subjects (9.7%) and need to be lumped with the latter. Together, pronoun references and zero-subject verbs account for over half (52.6%) of all established character references in the data when excluding reflexives, and 57.5% including np (noun phrase) mentions. The percentage of indirect mentions rises to 64.8% when also including +HUM personal pronouns without an established character reference — the metrics used for English literature by Bamman et al. (2014), who also report a high prevalence (74%) of pronominal mentions. Finally, we might add participle and infinitive clauses, which usually make do with an implicit subject, linked to a preceding finite verb. In any case, given the high prevalence of pronominal and zero-subject character references, it is of prime importance to resolve anaphora relations.

To this end, we expanded a set of existing, experimental anaphora rules in PALAVRAS' Constraint Grammar pipe with additional CG rules, improving its coverage, scope and accuracy and adapting it to the task at hand. Here, we use the RELATIONS[5] operator to establish referent links between pronouns and underspecified noun phrases and a target referent, optimally a named entity (NE) of the PERSON category. The equivalent solution for subject-less verbs establishes links to a preceding surface subject, or — as a fallback — another subject-elliptic verb. In both cases, name targets will also be mapped, as `<REF:name>` tags, on the anaphorical element itself. This is useful for "promoting" the antecedent information, if link targets are themselves anaphorical (e.g. chains of pronouns or subject-elliptical verbs), in which case the ultimate name referent may be outside the rolling CG focus window. For syntax, this window would be just one sentence at a time, but for co-reference resolution, we opted for a ±6 sentences as a compromise between reach and recall on the one hand, and precision on the other. The basic co-reference rule algorithm is based on recency and similarity of antecedents, as suggested by Elson et al. (2010)), weighting features such as +HUM, top-level subject-hood, definiteness and topic or focus function[6]. As most personal pronouns in Portuguese are marked for gender and number, and verbs are inflected for (subject) person and number, this morphological information can be used to impose tag conditions on the name or np antecedent. This is true not only of clause-level pronouns, but also of possesives, where a reference link will allow us correctly assing family relations mentioned in the possessive's head (e.g. his father/mother/son/daughter). The relative pronoun que presents a special case, as it is underspecified with regard to gender and number. However, dependency syntax makes it relatively easy to recover an antecedent that can then be associated with information provided in the dependent relative clause (e.g. Pedro, que se casou com Júlia em tenra idade [Peter, who had married Julia at an early age] or seu amigo, que trabalhava como porteiro no turno da noite [his friend, who worked night shifts as a porter]).

For pronouns in a +HUM semantic frame slot, or for subject-less verbs with a human verb frame, the co-reference antecedent should be a human name. But if none can be found, without interfering blocking material (e.g. a non-human top-level subject), in the context window, or if there is more than one candidate, it may be necessary (or safest) to settle for an intermediate antecedents, even if it is a pronoun or zero-subject verb itself. We call such underspecified references "stepping stones". If the stepping stone itself can be assigned a reference link to a name antecedent, this link can later be recovered by a meta rule, and raised to the originnal pronoun that has been "stranded" with a stepping-stone reference. In real-life narrative text, there will

[4] Sometimes, +HUM noun phrases, e.g. o vigário [the vicar], are also used to refer to names, but with a much lower frequency, according to PALAVRAS' anaphora annotation.

[5] The RELATIONS operator allows the assignment of bi-directionally named, non-unique relations between tokens.

[6] Bamman et al. (2014) used gender and linear word distance. The equivalent to the latter, in our CG rules, are so-called barriers (e.g. non-matching top-level subjects or paragraph breaks), blocking further search left.

often be multiple stepping stones, for instance in a chain of action statements with pronominal subjects all referring to the same human agent introduced in the beginning of a paragraph. In the example, a subject complement (@SC, id-6), médico (physician), has been assigned an attribute relation (R:n-attr:5) to a pronoun subject (id-5). This pronoun is itself linked to a top-level subject antecedent, a character named XXX (id-1), four sentences to the left, through three stepping stones — first a subject-less finite verb (id-4), then two pronouns (id-2 and id-3). The verb carries an elliptic-subject relation (R:e-subj:3) to the closest of the pronouns (id-3). All pronouns carry referent relations (R:ref:id) linking them to either a stepping stone (R:ref:4) or the first, full subject (R:ref:1). Ultimately, this links the profession information ('doctor' id-6) to the full subject (id-1), even across text (. . .) spanning multiple sentences. Secondary annotation rules can propagate these links (e.g. R:ref:3 on id-5 or R:n-attr:1 on id-6) and assign explicit attribute tags to names, here the profession tag NA:Hprof/médico (on *XXX*, id-1).

- "XXX" main referent: top level @subject (PROP, +HUM, np-def)
  → R:be:6
  → <NA:Hprof/médico>
  . . .
- subject pronoun <REF:XXX> **R:ref:1**
  . . .
- subject pronoun <REF:XXX> **R:ref1** R:subj:4
  . . .
- subject-less VFIN R:e-subj:3
  . . .
- subject pronoun **R:ref:4** R:be:6
  → R:ref:3 → R:ref:1
  → <REF:XXX>

- "médico" <Hprof> @SC §ATR <R:n-attr:5>
  → R:n-attr:1

4. Quoted speech

In many literary works, an important part of character-related information is to be found in quoted speech. The density of quotes is quite text dependent. Thus, for English, Elson et al. (2010) found a spread of 19-71% for text included in quotes. For the DIP data (100 non-pdf texts), the average was 30.5% for direct speech. This

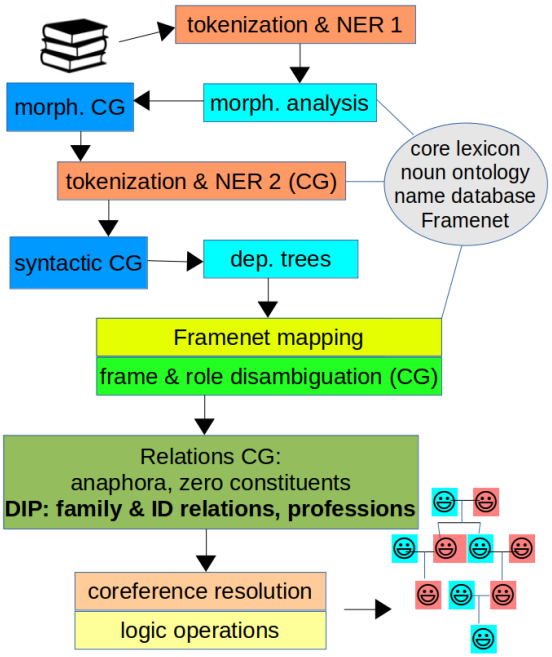

Figure 1: Relation chains

fact is relevant for the parser, as direct speech often manifests as an independent utterance without an externally linked dependency head (which would be typical of indirect speech), making it more difficult to keep track of who is who. To tackle this problem, and to facilitate the identification of speaker and addressee in dialogue turn-taking, we perform a quote mark-up as an early annotation step. The relevant tags are <quote-edge> (quote opener), <quote-end> and <quote-ana> (quote continuation after a quoting verb). In the literary data, dashes were used rather than quotation marks, leaving quote closures unmarked, unless they were followed by a quoting verb. To compensate for the missing dependency relations, we also mark the syntactic top node(s) in a quoted speech (<quote>), as well as the quoting verbs (<v-quote>), even if occurring in a separate sentence. In-quote name mentions are distinguished from body text name mentions by tagging the forme <quo>, and the latter <nquo>. In its newest version, our grammar also keeps track of turn-taking, marking alternate turns as <turn-1> and <turn-2>, and setting a speaker variable for each turn and <quote> mark.

This annotation not only makes it possible to extract and correctly name-link information from inside direct speech, but also helps establishing what is a character and what is not. Thus, narrative characters are more likely to occur in direct speech, or to produce it. A direct syntactic clue are vocatives (surface addressees), which almost always refer to characters. Co-reference rules will

link np vocatives (typically family terms or titles) to the speaker (quoting subject) in an adjacent turn, or — conversely — link named vocatives to noun speakers (or the antecedents of speaker pronouns). In both cases, the link can be used to infer attributive information from an np to a name. Even without extracting further information, speaker-discourse links can be used to establish relations between turn-taking characters based on the quantity of one-on-one dialogue. Thus, Elson et al. (2010) define and analyze character relationships as networks of social conversations. Similarly, dialogue relations could be used alongside family relations and cooccurrence-strength to the extraction and visualisation of social networks (Section 6). Linking discourse turns to literary characters is not a big discipline in NLP, but there is prior work on Portuguese, who present a rule-based system with trained decision trees for children's stories, reporting a 89% success rate for discourse separation, while recall and precision for speaker character identification were 10.6% and 65.7%, respectively. Our own speech annotation method was experimental and incomplete at the time of the shared task, but now recognizes 98.1% of all direct speech utterances, with an overall F-score for speaker identification of 92.0% (Bick, 2023).

5. Character annotation

The second layer of our CG annotation rules exploits existing relational links of the base annotation and the co-reference module, making relevant implicit information explicit on name tokens that refer to characters, or — as an intermediate step — on +HUM nouns or pronouns that are referent-linked to such a character token. First of all, this means mapping explicit name referent tags (red in Figure 1), e.g. `<REF:Pedro=da=Silva>`, but it also involves tags for the characters' social attributes and relations.

5.1. Social attributes

Social attributes are harvested from profession and title nouns (green annotation in Figure 1), and tagged on name tokens with an 'NA' (noun attribute) prefix, e.g. `<NA:Hprof:ministro>`. The process can make use of existing framenet information in PALAVRAS' base annotation, such as R:attr relational tags, the syntactic functions of subject and object complement, apposition or noun predicate, as well as the semantic roles of §ATR (attribute), with a name head, and §ID (identity, with a name dependent. These syntactic or semantic links are exploited to relate names to profession nouns (e.g. 'carpenter') and titles (e.g. 'chairman'), or to family nouns (e.g. 'father' or 'daughter'). For the latter, additional rules are needed to identify the argument of a given family noun, which may be "hidden" in a postnominal pp (e.g. 'X, mother of Y). Of course, as discussed in the co-reference Section, the immediate framenet or dependency links may not lead to a name, but rather to a noun or pronoun, in which case the link needs to be propagated to a name antecedent, following anaphora links and possibly bypassing one or more of the aforementioned "stepping stones". In addition to existing attributive relations, rules can also make use of a variety of specific clues and semantic reasoning, exploiting, for instance, profession-specific verbs (such as teaching for teachers), or nouns denoting profession-related tasks, products or institutions.

5.2. Family relations

The annotation of family relations is more complex than that of social attributes, as it involves two targets rather than one. We want to know not just that somebody is a daughter, but also whose daughter. Family relations are marked at both ends of a relation and may either be symmetrical (siblings, spouses) or asymmetrical (parent–child). A CG rule establishing such a relation, will add both tags at the same time, e.g. R:parent:id at one end and R:child:id at the other. Especially if one of the two names is "out of reach" (outside the window of analysis, or not mentioned as a proper noun), the name in question may also be recovered from a co-reference link or tag (including stepping stones), and the information tagged on the other name with an RI prefix, e.g. `<RI:parent_of:Maria>`. With few exceptions (heuristic proper nouns without morphological clues or attributes), gender is already tagged in PALAVRAS input. Therefore, family relation tags can be kept gender-neutral, reducing tag set size and grammar complexity. All in all, the grammar covers eight basic family relations. Four of these are symmetrical (parent, sibling, spouse, cousin, "gbfriend" [girl friend or boy friend]), four are paired/asymmetrical. Apart from the parent–child pair, the latter group includes the portmanteau categories of "auncle" [aunt or uncle] and "nephie" [nephew or niece]). Where relevant, the set of relations can be expanded with prefixes for 'great-' (g-), 'great-great-' (gg-), 'in-law' (i-) and 'god' (god-), e.g 'gparent', 'isibling' or 'godchild'. Finally, there is a non-directional relation 'widow' and one non-family relation, 'friend'. The various combinations correspond to about 40 Portuguese words.

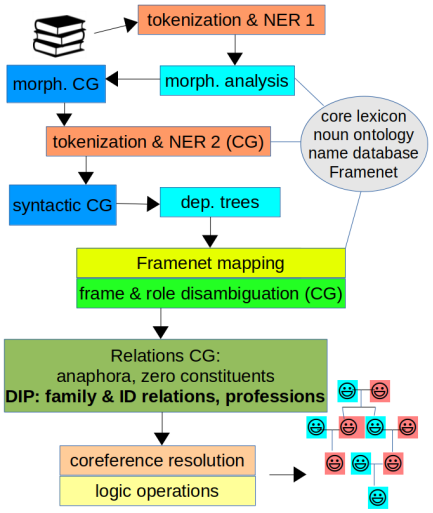

Figure 2: System architecture

5.3. Non-characters

Because PALAVRAS does not itself distinguish between characters and "cultural" name references (saints, poets, emperors etc.), because neither can be covered by closed lists, and because some names may function as either, we also need rules to help making this distinction. These rules exploit morphological, semantic and contextual clues[7] to assign a `<cult>` (cultural) or `<noncult>` (fictional) tag. For instance, names with an affection suffix (-inh[oa], -zinh[oa]) or a family relation are deemed to be characters, as are names that are part of speaking or movement frames. Conversely, surname-only names[8], or names in a royal or religious context will be tagged as `<cult>`. The final cast extraction also employs text-level statistics as a further means of distinction.

[7]With only a "proper noun" tag as input, i.e. from a parser without the NE category of PERSON, the distinction between character and non-character would have to subsume the ±PERSON distinction. Paul & Das (2017), for instance, also using grammatical context features, describes a neural network-version of such a classifier for an English version of Mahabharata, achieving F-scores of 76% and 88% for proper nouns and proper noun-containing NP's, respectively.

[8]This surprisingly safe heuristics was originally motivated by inspection of DIP's first example novels, but can be corroborated on the rest of the collection: typically, only famous people (scientists, poets etc.) are referred to by surname alone. Character names exhibit more variation and are normally introduced at least once with a more complete name. Also, character surnames are usually used with honorifics. First names, on the other hand, do occur on their own, especially for children and servants.

6. Cast extraction

The third module in our system pipeline, after primary (general) and secondary (task-driven annotation), extracts and structures the (now) explicit character information encoded in the annotation (Figure 2). The extractor first builds a data structure for all name IDs and name relations, storing, for each person name token:

- gender labels
- social attribute labels (NA tags)
- relative_of:name tags (RI tags)

As a fall back, in the absence of an explicit RI tag, the extractor will follow family relation links and retrieve the target name, either directly (for name lemmas), or from a `<REF:name>` tag, or by following stepping stone links.[9] If this process leads to circular ID references or self-relations,[10] the extractor will ignore the information in question.

6.1. Character name unification

For each name in the data structure, the extractor loops through all other names and decides which are aliases ("synonyms") of the same name, creating named synsets based on:

- maximum number of shared core name elements (first names and surnames)
- unification of attributes for gender and social "role" (profession or family role)
- ratio of occurrence inside/outside quotes

For this process, titles/honorifics (e.g. Sra — 'Mrs') and morphological variation (-inho — 'dear') are ignored, and titles in isolation are not regarded as names. Synset names should be unambiguous, and will therefore typically consist of a multi-part version of a given name. Theoretically, isolated instances of ambiguous first names can then be attributed to different synsets based on social role.

In conjunction with coreference resolution, the cast extractor weeds out `<cult>`-marked names,[11] unless they can be assigned to an ex-

[9]Since the extractor has built a data structure for the whole text, it is not limited by the ±6 sentences analysis window.

[10]This is rare, but can theoretically happen due do errors in the CG annotation, in particular dependency errors or frame link errors.

[11]In the case of conflict, i.e. if there are both `<cult>` and `<noncult>` tags for a given name, `<noncult>` wins with a simple majority, while `<cult>` is valid only if it is tagged on more than half of the occurrences of that name.

isting name synset. 1-part names that are rare in text in both absolute and relative terms, and that are not part of a synset, are also discarded — unless they have been specifically tagged as <noncult> (cp. Section 3). The same goes for multi-part names if they consist only of honorifics and/or start with an article (e.g. o=Santo=Padre — 'the Holy Father'). In unclear cases, occurrence in direct speech is regarded as an indicator of characterhood. A first-person narrator is regarded as a special case character. Thus, if there are 1st-person verbs or pronouns, outside of quotes, these are flagged and synset-linked to possible 3rd-person, named mentions of the narrator. In other words, a 1st-person narrator will be regarded as part of the cast.

6.2. Logic operations

In a second stage, the cast extractor expands the family tree using logic:

- Propagation: If X is a child of Y, and Z is a parent of Y, then X is a grandchild of Z;
- Symmetry: If X is a sibling of Y, then Y is a sibling of X

Also, professions (or other, in-context relatively unique, noun attributes) may be used for unification: If X is a doctor, and Y the child of a doctor, then Y is a child of X. Because human NPs, just like names, may be given relation-tags, this method, albeit a bit risky, may even be applied where the profession-providing NP has not been name-resolved: If X is the spouse of a (nameless) doctor, and Y the child of (said nameless) doctor, then Y is a child of X.

6.3. Output formats

The native output of the cast extractor is an alphabetical list of name synsets with their members and gender, followed by profession attributes and a list of family relations. A condensed format with numbered name synsets (nns) is available for evaluation, consisting of two .csv files for each text, one for character synsets, gender and profession, one for family relations.

- characters.csv:
  `nns,syn-1|syn-2|...,gender,prof-1|prof-2|...`
- relation.csv:
  `nns-1,relation,nns-2`

Task	Average F-score	F-score spread
character identif.	63.4	40–80
name unification	68.1	(20–) 40–90
gender	89.5	60-100
prof./occupation	24.6	(0–) 10–55
family relations	15.5	0–60

Table 1: System performance (shared task)

7. Evaluation

As already explained elsewhere in this volume, the DIP shared task addressed Portuguese and Brazilian literary works, mostly historical novels from the 19th and early 20th century, and comprised five subtasks: (a) character identification, (b) co-reference resolution, (c) character gender, (d) profession/occupation or other position in society, (e) family relations. Two novels were provided as examples by the organizers, with manually extracted character information. The test run had to be performed in 48 hours on a collection of 100 books (mostly older novels), 20 of which had manually extracted gold casts for the evaluation. The literary period constraint, motivated by public domain availability, caused some annotation problems for the base parser, as texts contained a certain amount of orthographical variation (e.g. *cavallo/cavalo* — 'horse', *sabbados/sabados* - 'Saturdays') not found in modern Portuguese, as well as errors introduced by OCR scanning, or combinations of both (e.g. *offerecimento* — 'offering'). Our cast extractor was the only participating system that solved the task for the provided data (historical literature) and within the given time frame. It achieved reasonable F-scores for character identification (63.4%), co-reference resolution (68.1%) and gender assignment (89.5%), but did not perform well for professions (F=24.6%) and family relations (F=15.5%).

Results differed a great deal between works, not least for the difficult subtasks (d) and (e). Thus, the best books had F-scores of 80-90% for the identification subtasks (a) and (b), 100% for gender, and 50-60% for the social information subtasks (d) and (e). On the flipside, several books scored 0 for relations, as did one for professions. Disregarding one outlier, identification was more robust, with the lowest-scoring books at around 40% for (a) and (b). The very pronounced text dependence of the task was also noted by Dekker et al. (2019), who evaluated the performance of different (English) NER tools in the co-occurrence-based extraction of social net-

works. Here, systems worked well for e.g. *Huckleberry Finn* and *Game of Thrones*, with the best ones achieving F-scores of 80-90% for the isolated NER task of person recognition, while many had single-digit results for *Brave New World*. Also, all systems performed worse for classical novels and better for modern novels. For the full task of character detection, Vala et al. (2015), when evaluating their 8-stage system on three different English literary data sets, also report a substantial spread in accuracy (F-scores of 44.8, 54.0 and 75.8), similar to our own spread found for Portuguese (F=40-80). When interpreting our results, it should be born in mind that character identification is more than named entity recognition. Thus, errors were caused mainly by the (sometimes unclear) distinction between "real" characters and cultural background names, not by the underlying NER, which was much more robust, as it conflates the two categories into "person". Second, the unification of names with each other (b) and with pronouns and non-name np's (both here only evaluated indirectly) makes for additional complexity compared to the underlying NER task. Also, when developing the system, the distinction between titles and occupation or social position was not always entirely clear from the examples. Therefore, some errors in (d) resulted from fuzzy definitions and not the rules or algorithm of the software. Finally, historical spelling variation and non-standard uppercasing created false name candidates caused by POS errors.

8. Network analyses

Automatic cast extraction provides a quantitative-comparative[12] angle to literary analysis difficult to achieve by other means. Thus, it is possible to compare the works of different authors or across different periods og genres, focusing on features such as cast complexity, gender distribution, and societal representativeness. The same data can also be used for the vizualisation of character networks, where quantitative network parameters allow e.g. the distinction between central and peripheral characters. The example in figure 3 shows a close-up of one such network, generated with Cytoscape[13], for the Brazilian novel "Quincas Borba", by Machado de Assis (published 1891).

The necessary .csv tables were created by exporting standardized, unique character id names (u-names) as network nodes and as-is "surface names" (s-names) as their attributes. Family relations were exported as directed arcs ("edges") between source (SN) and target (TN) node (e.g. SN child_of TN). In order to further populate the network, an un-named relation was assigned to a given pair of character tokens, if their in-text difference was less than 3 "semantic" tokens (defined as carrying a semantic role or frame tag). Relation frequencies were then used as a strength attribute for the relation in question. We end up with 6 columns in the *Cytoscape* .csv table: SN u-name, SN s-name, relation, relation strength, TN u-name, TN s-name. The table allows the computation of various network parameters, such as edge count, stress, closeness and clustering, which can be used to evaluate and describe the narrative importance and connectedness of the characters. Figure 3 shows a close-up of the central portion of one possible vizualisation of the character network, using edge-weighted spring-embedded layout, with family relation edges in red, and unnamed relations in gray. Though based on a mathematical model, the graphic is immediately interpretable by a human reader, singling out a handful of main characters and providing an overview of their mutual connections.

9. Conclusion

We have discussed the implementation of a Constraint Grammar-based system for the extraction of character information from Portuguese text. The method harnesses existing mark-up from a morphosyntactic and semantic parser to assign relational tagging for name co-reference and family relations, as well as social attributes. Character names are unified and distinguished from non-characters using clues like dialogue participation and network centrality. In the DIP shared task, a first version of our system achieved reasonable results for character identification and unification. However, information about family relations and social position proved to be much more difficult to extract, as it is often provided through indirect clues, or through attribution in direct speech subject to long turn-taking sequences. Future versions of the cast extractor should address these problems, for instance by developing a full-fledged speaker- and adressee attribution module and by the use of text-level coreference variables.

[12] Quantitative analysis is particularly robust in the face of a certain error rate, since distributional patterns are likely to be visible despite errors in individual characters.

[13] https://cytoscape.org/

Figure 3: Character network (Cytoscape)

Acknowledgments

We are grateful to the DIP team at Linguateca, NuPILL, UEMA and UiO for preparing and organizing the shared task, and appreciate the work that has gone into the manual compilation of gold-standard evaluation data.

References

Bamman, David, Ted Underwood & Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *52nd Annual Meeting of the Association for Computational Linguistics*, 370–379. doi 10.3115/v1/p14-1035.

Bick, Eckhard. 2014. PALAVRAS, a constraint grammar-based parsing system for Portuguese. In *Working with Portuguese Corpora*, 279–302. Bloomsburry Academic.

Bick, Eckhard. 2022. PFN-PT: A framenet annotator for Portuguese. *Domínios de Lingu@gem* 16(4). 1401–1435. doi 10.14393/dl52-v16n4a2022-7.

Bick, Eckhard. 2023. Attribution of quoted speech in Portuguese text. In *Contraint Grammar: Methods, Tools and Applications (NoDaLiDa 2023 Workshop)*, forthcoming.

Bick, Eckhard & Tino Didriksen. 2014. CG-3 — beyond classical constraint grammar. In *Nordic Conference of Computational Linguistics (NoDaLiDa)*, 31–39.

Bornet, Cyril & Frédéric Kaplan. 2017. A simple set of rules for characters and place recognition in french novels. *Frontiers in Digital Humanities* 4. n/p. doi 10.3389/fdigh.2017.00006.

Chaturvedi, Snigdha, Mohit Iyyer & Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *AAAI Conference on Artificial Intelligence*, 3159–3165. doi 10.1609/aaai.v31i1.10982.

Dekker, Niels, Tobias Kuhn & Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science* 5. e189. doi 10.7717/peerj-cs.189.

Elson, David, Nicholas Dames & Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *48th Annual Meeting of the Association for Computational Linguistics*, 138–147.

Goyal, Amit, Ellen Riloff & Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 77–86.

Labatut, Vincent & Xavier Bost. 2019. Extraction and analysis of fictional character networks. *ACM Computing Surveys* 52(5). 1–40. 10.1145/3344548.

Mamede, Nuno & Pedro Chaleira. 2004. Character identification in children stories. In *Advances in Natural Language Processing*, 82–90. 10.1007/978-3-540-30228-5_8.

Paul, Apurba & Dipankar Das. 2017. A deep dive into identification of characters from Mahabharata. In *14th International Conference on Natural Language Processing (ICON)*, 447–455.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese. In *15th International Conference on Computational Processing of Portuguese (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39.

Vala, Hardik, David Jurgens, Andrew Piper & Derek Ruths. 2015. Mr. Bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 769–774. 10.18653/v1/D15-1088.

Valls-Vargas, Josep, Santiago Ontañón & Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. In *7th Intelligent Narrative Technologies Workshop*, 38–44.