Whole-Part Relations Rule-Based Automatic Identification: Issues from Fine-Grained Error Analysis

Ilia Markov[1], Nuno Mamede[2,3], and Jorge Baptista[3,4]

[1] Centro de Investigación en Computación (CIC),
Instituto Politécnico Nacional (IPN)
México D.F., Mexico
markovilya@yahoo.com

[2] Universidade do Algarve/FCHS and CECL
Faro, Portugal
jbaptis@ualg.pt

[3] INESC-ID Lisboa/L2F – Spoken Language Lab
Lisboa, Portugal
{Nuno.Mamede,jbaptis}l2f.inesc-id.pt

[4] Universidade de Lisboa/IST
Lisboa, Portugal
Nuno.Mamede@ist.utl.pt

Abstract. In this paper, we focus on the most frequent errors that occurred during the implementation of a rule-based module for semantic relations extraction, which has been integrated in STRING, a hybrid statistical and rule-based Natural Language Processing chain for Portuguese. We focus on *whole-part* relations (*meronymy*), that is, a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. In this case, we target the type of meronymy involving human entities and *body-part nouns*. We describe with some detail the decisions that were made in order to overcome the errors produced by the system and the solutions adopted to improve its performance.

Keywords: whole-part relation, meronymy, body-part noun, Portuguese, error analysis.

1 Introduction

Automatic identification of semantic relations contributes to cohesion and coherence of a text and can be useful in several other Natural Language Processing (NLP) tasks such as opinion mining, question answering, text summarization, machine translation, information extraction, information retrieval, and others [10].

The goal of this work is to improve the extraction of semantic relations between textual elements in STRING, a hybrid statistical and rule-based NLP

A. Gelbukh et al. (Eds.): MICAI 2014, Part I, LNAI 8856, pp. 37–50, 2014.

chain for Portuguese[1] [17], by targeting the most frequent errors that occured during the implementation of the whole-part relations extraction module. Whole-part relations (*meronymy*) is a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. In this case, we focus on the type of meronymy involving human entities and *body-part nouns* (henceforward, *Nbp*) when they co-occur in texts.

This paper is structured as follows: Section 2 briefly describes related work on whole-part dependencies extraction, while Section 3 explains how this task was implemented in STRING; Section 4 presents the evaluation procedure; Section 5 describes with some detail how the error analysis was carried out; Section 6 illustrates the results of the performance of the system after the error analysis; and Section 7 draws the conclusions.

2 Related Work

Meronymy is a complex relation that "should be treated as a collection of relations, not as a single relation" [15]. In NLP, various information extraction techniques have been developed in order to capture whole-part relations from texts.

Hearst [13] tried to find lexical correlates to the *hyponymic* relations (type-of relations) by searching in unrestricted, domain-independent text for cases where known hyponyms appear in proximity. The author proposed six lexico-syntactic patterns; he then tested the patterns for validity and used them to extract relations from a corpus. To validate his acquisition method, the author compared the results of the algorithm with information found in WordNet [4]. The author reports that when the set of 152 relations that fit the restrictions of the experiment (both the hyponyms and the hypernyms are unmodified) was looked up in WordNet: "180 out of the 226 unique words involved in the relations actually existed in the hierarchy, and 61 out of the 106 feasible relations (*i.e.*, relations in which both terms were already registered in WordNet) were found." [13, p. 544]. The author claims that he tried applying the same technique to meronymy, but without great success.

Girju *et al.* [10,11] present a supervised, domain independent approach for the automatic detection of whole-part relations in text. The algorithm identifies lexico-syntactic patterns that encode whole-part relations. The authors report an overall average precision of 80.95% and recall of 75.91%. The authors also state that they came across a large number of difficulties due to the highly ambiguous nature of syntactic constructions.

Van Hage *et al.* [12] developed a method for learning whole-part relations from vocabularies and text sources. The authors reported that they were able to acquire 503 whole-part pairs from the AGROVOC Thesaurus[2] to learn 91 reliable whole-part patterns. They changed the patterns' part arguments with known entities to

[1] `https://string.l2f.inesc-id.pt/` [last access: 22/09/2014].

[2] `http://www.fao.org/agrovoc` [last access: 12.08.2014].

introduce web-search queries. Corresponding whole entities were then extracted from documents in the query results, with a precision of 74%.

The Espresso algorithm [24] was developed in order to harvest semantic relations in a text. The algorithm extracts surface patterns by connecting the seeds (tuples) in a given corpus. The algorithm obtains a precision of 80% in learning whole-part relations from the Acquaint (TREC-9) newswire text collection, with almost 6 million words.

Some work has already been done on building *knowledge bases* for Portuguese, most of which include the concept of whole-part relations. These knowledge bases are often referred to as *lexical ontologies*, because they have properties of a lexicon as well as properties of an ontology [14,26]. Well-known, existing lexical ontologies for Portuguese are Portuguese WordNet.PT [19,20], later extended to WordNet.PT Global (Rede Léxico-Conceptual das Variedades do Português) [21]; MWN.PT-MultiWordNet of Portuguese [25]; PAPEL (Palavras Associadas Porto Editora Linguateca) [23]; and Onto.PT [22]. Some of these ontologies are not freely available for the general public, while others just provide the definitions associated to each lexical entry without the information on whole-part relations. Furthermore, the type of whole-part relation targeted in this work, involving any human entity and its related *Nbp*, can not be adequately captured using those resources (or, at least, only those resources)[3].

Attention was also paid to two well-known parsers of Portuguese, in order to discern how do they handle the whole-part relations extraction: the PALAVRAS parser [2], consulted using the Visual Interactive Syntax Learning (*VISL*) environment, and LX Semantic Role Labeller [3]. Judging from the available online versions/demos of these systems, apparently, none of these parsers extracts whole-part relations, at least explicitly.

3 Whole-Part Dependency Extraction Module in STRING

3.1 STRING Overview

STRING is a fully-fledged NLP chain that performs all the basic steps of natural language processing (tokenization, sentence splitting, POS-tagging, POS-disambiguation and parsing) for Portuguese texts. The architecture of STRING is given in Fig. 1.

STRING has a modular, pipe-line structure, where: (i) the preprocessing stage (tokenization, sentence splitting, text normalization) and lexical analysis are performed by LexMan; (ii) followed by RuDriCo, which applies disambiguation rules, handles contractions and several special types of compound words; (iii) the MARv module then performs POS-disambiguation, using HMM and the Viterbi

[3] At the later stages of our research (May, 2014), we came to know the work of Cláudia Freitas [7]; however, since all the lexicon, grammar rules and evaluation procedures had been already accomplished by then, we decided not to take it into consideration at this moment but to use it in future work.

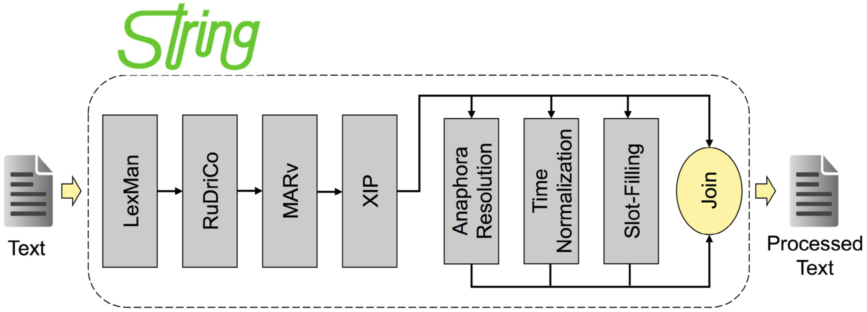

Fig. 1. STRING Architecture

algorithm; and, finally, (iv) the XIP parser (Xerox Incremental Parser) [1] segments sentences into chunks (or elementary sentence constituents: `NP`, `PP`, etc.) and extracts dependency relations among chunks' heads (`SUBJ`ect, `MOD`ifier, etc.). XIP also performs named entities recognition (NER). A set of post-parser modules have also been developed to handle certain NLP tasks such as anaphora resolution, temporal expressions' normalization and slot-filling. As part of the parsing process, XIP executes *dependency rules*. Dependency rules extract different types of dependencies between nodes of the sentence chunking tree, namely, the chunks' heads. Dependencies can thus be viewed as equivalent to (or representing) the syntactic relations holding between different elements in a sentence. Some of the dependencies extracted by XIP represent rather complex relations, such as the notion of *subject* (`SUBJ`) or *direct object* (`CDIR`), which imply a higher level of analysis of a given sentence. Other dependencies are much simpler and sometimes quite straightforward, like the determinative dependency `DETD`, holding between an article and the noun it determines, *e.g.*, *o livro* 'the book' – `DETD(livro,o)`. Some dependencies can also be seen as auxiliary dependencies, and are required to build the more complex ones.

3.2 A Whole-Part Extraction Module in STRING

Next, we describe the way a whole-part dependency involving *Nbp* is extracted in the Portuguese grammar for XIP. To this end, a new module of the rule-based grammar was built, which contains most of the rules required for this work. Example (1) is a simple case where there is a determinative `PP`, complement *de* 'of' N of the *Nbp*, so that the meronymy is overtly expressed in the text:

(1) *O Pedro partiu o braço do João* 'Pedro broke the arm of João'

The next rule captures the meronymy relation between *João* and *braço* 'arm':

```
IF( MOD[POST](#2[UMB-Anatomical-human],#1[human]) & PREPD(#1,?[lemma:de]) &
    CDIR[POST](#3,#2) & ~WHOLE-PART(#1,#2) )
    WHOLE-PART(#1,#2)
```

This rule is built using the XIP dependency rules' syntax, and it reads as follows: first, the parser determines the existence of a `[MOD]`ifier dependency, already calculated, between an *Nbp* (variable #2) and a human noun (variable #1); notice that, according to XIP conventions, the governor of the dependency is its first argument, hence *João* is said to be a modifier of *braço* ‘arm’; this modifier must also be introduced by preposition *de* ‘of’, which is expressed by the dependency PREPD; then, a constraint is defined that the *Nbp* must be a direct object (CDIR) of a given verb (variable #3); and, finally, that there is still no previously calculated WHOLE-PART dependency between the *Nbp* and the human noun (variable #1); this last constraint is meant to ensure that there is only one meronymy relation between each *Nbp* and a given noun; if all these conditions are met, then, the parser builds the WHOLE-PART relation between the human determinative complement and the *Nbp*.

The meronymy extraction module contains 29 general rules addressing the most relevant syntactic constructions triggering this type of meronymic relations, and a set of 87 rules for the 29 *disease nouns* (*Nsick*), in order to capture the underlying *Nbp* (*e.g.*, *gastrite-estômago* ’gastritis-stomach’). A set of around 400 rules has also been devised to prevent the whole-part relations being extracted in the case the *Nbp* are elements of idiomatic expressions (*e.g., O Pedro partiu o coração à Ana* ‘Pedro broke the heart to Ana’). This work also addresses the cases where a whole-part relation holds between two *Nbp* in the same sentence (*e.g., A Ana pinta as unhas dos pés* (lit: Ana paints the nails of the feet) ‘Ana paints her toes’ nails’) and the case of determinative nouns that designate parts of an *Nbp*, though they are not themselves *Nbp* (*e.g., O Pedro encostou a ponta da língua ao gelado da Ana* ‘Pedro touched with the tip of the tongue the ice cream of Ana’). Each one of these cases triggers different sets of dependencies. 54 rules were built to associate the *Nbp* with their respective parts, to handle the cases where there is an *Nbp* and a noun that designates a part of that same *Nbp*.

4 Evaluation

For the evaluation of the work the first fragment of the CETEMPúblico corpus [27] (14,7 million tokens and 6,25 million words) was used in order to extract sentences that involve *Nbp* and *Nsick*. Using the *Nbp* (151 lemmas) and the *Nsick* (29 lemmas) dictionaries, specifically built for STRING lexicon, 16,746 *Nbp* and 79 *Nsick* instances were extracted from the corpus. In order to produce a golden standard for the evaluation, a random stratified sample of 1,000 sentences was selected, keeping the proportion of the total frequency of *Nbp* in the source corpus. This sample also includes a small number of *Nsick* (6 lemmas, 17 sentences). The 1,000 output sentences were divided into 4 subsets of 225 sentences each. Each subset was then given to a different annotator (native Portuguese speaker), and a common set of 100 sentences was added to each subset in order to assess inter-annotator agreement. The annotators were asked to append the whole-part dependency, as it was previously defined in a set of guidelines, using the XIP format. To assess inter-annotator agreement we used ReCal3: Reliability Calculator [6], for 3 or more annotators. The results showed that the Average Pairwise Percent Agreement equals 0.85, the Fleiss’ Kappa inter-annotator agreement is 0.62, and the Average Pairwise Cohen's Kappa 0.63. According to Landis and Koch [16] these figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [5], these results correspond to an inter-annotator agreement halfway between "fair" and "good". In view of these results, we assumed that the remaining, independent and non-overlapping annotation of the corpus by the four annotators is sufficiently consistent, and can be used as a golden standard for the evaluation of the system output.

The results of the system performance are showed in Table 1, where `TP`=*true-positives*; `TN`=*true-negatives*; `FP`=*false positives*; `FN`=*false negatives*; and the first line correspond to the 100 sentences that were subject to multiple annotators' classification, while the 900 sentences are the remainder instances of the sample taken form the corpus. The number of instances is higher than the number of sentences, as one sentence may involve several instances, and we count 5 partial `TP` as 0.5. The relative percentages of the `TP`, `TN`, `FP` and `FN` instances are similar between the 100 and the 900 set of sentences. This explains the similarity of the evaluation results and seems to confirm our decision to use the remaining 900 sentences' set as a golden standard for the evaluation of the system's output with enough confidence.

Table 1. System's performance for *Nbp*

Number of sentences	TP	TN	FP	FN	Precision	Recall	F-measure	Accuracy
100	8	73	7	14	0.53	0.36	0.43	0.79
900	73.5	673	55	118	0.57	0.38	0.46	0.81
Total:	81.5	746	62	132	0.57	0.38	0.46	0.81

5 Error Analysis

The results of the evaluation of the task showed that there were 62 false positive cases and 132 false negatives. We begin this section by the analysis of some false positives cases and then move on to the false negatives.

5.1 False Positives

Disambiguation of *Nbp* in Context. To begin with, we tackled a number of cases with the ambiguous noun *língua* 'tongue/language'. In order to preclude the building of whole-part relation in cases such as *língua portuguesa* 'Portuguese language', *a língua de Camões* 'the language of Camões', *professor de língua* (lit: teacher of language) 'language teacher', etc., where the noun *língua* 'language' is not used in the meanining of an anatomical part, we adopted one of the following strategies: we removed the *Nbp* (`sem-anmov`) feature from the nouns lexical set of features. This is carried out by the following rules, which are applied before the chunking stage:

— in the case of gentilic adjectives, one rule had to be done for each one of this type of adjectives:

```
2> noun[lemma:língua,sem-anmov=~], adj[gentcontinent=+].
2> noun[lemma:língua,sem-anmov=~], adj[gentregion=+].
2> noun[lemma:língua,sem-anmov=~], adj[gentcountry=+].
2> noun[lemma:língua,sem-anmov=~], adj[gentcity=+].
```

— in the case of combinations of *língua* 'tongue/language' with renowned authors of a given language, a PP structure has to be spelled out; so far, we built rules for over a dozen authors, epitomes of their national languages, which occurred with some frequency in the CETEMPúblico corpus:

```
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Camões].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Shakespeare].
```

— a similar rule is necessary for PP complements with country names (*a língua de Portugal* 'Portugal's language'):

```
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[country=+].
```

Besides, there could also be a determiner for examples such as *a língua do Brasil é o Português* 'the language of the Brazil is the Portuguese' Thus, a second rule is necessary:

```
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], art[lemma:o], noun[country=+].
```

This second rule is required because some country names are obligatorily preceded by a definite article (*o Brasil* 'the Brazil', *os Estados Unidos* 'the United States', *etc.*)

Difficult Cases. A certain number of cases were found where the use of the *Nbp* is clearly figurative, but it is not neither an idiom nor a compound word, so we were unable to devise any strategy to avoid capturing the whole-part relation:

(2) *À farta ementa associou-se um acontecimento a que certamente não foi alheio o* ***dedo*** *organizativo de José Perdigão, que no filho encontrou precioso instrumento...*
'To the abundant menu, an event was associated, which was certainly not unconnected with the organizational finger of José Perdigão, who found in [his] son a [precious=] most valuable tool...'

`WHOLE-PART(José Perdigão,dedo) 'WHOLE-PART(José Perdigão,finger)'`
In this case, the whole-part relation is correctly extracted, but the *Nbp dedo* 'finger' is not to be interpreted literally, but figuratively, and can be connoted with several idiomatic expressions such as *meter o dedo/a mão em* 'sb. put [one's] finger/hand in sth.' 'to have a role in / to interfere with'.

5.2 False Negatives

Noun or `NP` Modifiers (not involving verbs). The rules that have been developed only involve verb arguments (subject or complements) and did not consider the situations where an *Nbp* is a modifier of a noun or an adjective. Therefore, in several situations, the whole-part relations have not been captured. For example:

(3) *Um mágico de carapuço na cabeça*
'A magician with a hood over the head'

In this case, there is only a complex `NP`, with all the `PP` depending on the head noun *mágico* 'magician'. It is also possible to consider that in these cases an adjective or a verb past participle has been zeroed (*e.g.*, *Um mágico de carapuço enfiado/posto/colocado na cabeça* 'A magician with a hood stucked/placed over the head'). The meronymy module did not contemplate these complex `NP`s, including those with a zeroed adjective/past-participle, as most of the rules always involved a verb argument. This will have to be taken into consideration in future work.

Missing Features. One of the main reasons why the whole-part relation has not been captured derived from the fact that many human nouns are still unmarked with the human feature (or any of its subsumed features). For example, in the sentence:

(4) *Numa espécie de altar, um transexual padece com uma coroa de agulhas espetadas na cabeça, apoiado a umas muletas, provavelmente a sua cruz, nesta paródia à crucificação*
'In a kind of altar, a transsexual suffers with a crown of needles stuck in his head, supported by crutches, probably his cross, in this parody of the crucifixion'

In this case, the whole-part relation between the subject of *padecer* 'suffer' and the body-part *cabeça* 'head' was not captured just because the noun *transexual* (*id.*) had not been attributed the feature human.

In some cases, the rules were not triggered because the human entity is expressed by a personal pronoun and this category is not marked with the human feature. In the next sentence, the system failed to establish the whole-part relation because it can not ascribe the human feature to the relative pronoun *que* 'who' that is the subject of the relative clause.

(5) *Segundo o responsável do hospital, o doente –* ***que*** *também sofreu graves ferimentos na cabeça – poderia ser ainda sujeito a uma segunda intervenção cirúrgica*
'According to the head of the hospital, the patient - **who** also suffered serious head injuries - could still be subjected to a second surgical intervention'

However, the antecedent of the pronoun has been correctly extracted:
`ANTECEDENT_RELAT(doente,que) 'ANTECEDENT_RELAT(patient,who)'`

According to [18], relative pronouns are among the most successful cases of anaphora resolution in STRING. Therefore, it is possible that after this module comes into play, the features of the antecedent are inherited by the pronoun and the whole-part module be allowed to process the sentence again.

An opposite situation occurs when some features associated to the *Nbp* preclude the correct extraction of the whole-part dependency. The noun *corpo* ‘body’ is one of that cases and a very complex one. It is an element of several compound nouns, which are identified during lexical analysis and do not interfere in the dependency extraction step. Furthermore, it can be considered as an *Nbp* (*e.g.*, *o corpo da vítima* ‘the body of the victim’) and also a collective noun, functioning as a type of determiner, as in

(6) *O corpo (=conjunto) dos docentes da faculdade*
‘The staff of the (= set) of the teachers of the faculty’

Because of this a `QUANTD` (quantifying) dependency is extracted between *corpo* ‘body’ and the immediately following `PP`, which prevents the extraction of whole-part relation; therefore, rules were build to partially disambiguate this particular noun by removing the features associated to its collective noun interpretation.

```
3> noun[lemma:corpo,sem-anmov=+,sem-sign=~,sem-cc=~, sem-ac=~,sem-hh=~,sem-group-of-things=~],
prep[lemma:de], (art[lemma:o]), noun[lastname=+].
3> noun[lemma:corpo,sem-anmov=+,sem-sign=~,sem-cc=~, sem-ac=~,sem-hh=~,sem-group-of-things=~],
prep[lemma:de], (art[lemma:o]), noun[firstname=+].
```

These rules read as follows: if the noun *corpo* ‘body’ is followed by preposition *de* ‘of’ and a first or a last proper name, then we remove all the other features of *corpo* ‘body’ except the one that marks it as an *Nbp*.

They do not solve all the cases, naturally, since the distinction between the determiner and the *Nbp* can not yet be done, as it would require a previous word sense disambiguation module.

Ambiguous `FIXED` Expressions, Incorrectly Captured. In some cases, the `FIXED` expressions have been incorrectly captured instead of the whole-part relations, because they are ambiguous and have been used in the literal sense. For example:

(7) *Ele arrancava-me os cabelos todos* ‘He pulled out all my hair’

```
FIXED(arrancava,cabelos) 'FIXED(pulled out,hair)'
```

In the idiom *arrancar os cabelos* ‘to despair’, there is obligatory correference between the subject and the *Nbp*, so there is no way the sentence could be interpreted figuratively. The problem, thus, relies in the incorrect representation of the constraints of the idiom, not of the grammar.

In this case, the correct relation should be:

```
WHOLE-PART(me,cabelos) 'WHOLE-PART(my,hair)'
```

No Syntactic Relation Between *Whole* and *Part*. In some cases, the *whole* and the *part* are not syntactically related (and can be far away from each other in a sentence):

(8) *O facto do corpo ter sido encontrado na cozinha, leva os bombeiros a suspeitar que a vítima, com graves problemas de saúde, tenha desmaiado e caído à lareira, o que poderá ter estado na origem do incêndio*
'The fact that the body was found in the kitchen, makes the firefighters to suspect that the victim, with serious health problems, had fainted and fallen into the hearth, which may have been the origin of the fire'

In this example, the *part corpo* 'body' is the subject of the *ter sido encontrado* 'have been found', while the *whole vítima* 'victim' is the subject of *tenha desmaiado* 'had fainted'; each noun is in a different subclause, and there is no syntactic dependency between the two nouns. However, the annotator was able to identify this meronymic relation `WHOLE-PART(vítima,corpo)` `'WHOLE-PART(victim,body)'`, which is beyond the scope of our current parser. Eventually, a bag-of-words machine learning approach could overcome this difficulty, which can not be done by this rule-based approach.

Difficult Cases. In spite of our best efforts, some *Nbp* were still missing from the lexicon, as in the case of *defesas imunitárias* 'immune defenses':

(9) *O que se pensa que acontece na artrite reumatóide é que a cartilagem é atacada pelas defesas imunitárias do doente, como se ela fosse um autêntico "corpo estranho"*
'What we think happens in rheumatoid arthritis is that the cartilage is attacked by the immune defenses of the patient as if it was an authentic "foreign body"'

In such cases, we have completed the dictionary, naturally.

In the next example, there is also a problem with the compound noun *cabelo(s) branco(s)* 'white hair(s)':

(10) *Um deles, de óculos e cabelo branco, olha para o relógio e depois perscruta com alguma inquietação as bancadas a meia nau*
'One of them, wearing glasses and with white hair, looks at his watch and then peers restlessly to the seats at midship'

For the moment, even though *cabelo(s) branco(s)* 'white hair(s)' is already tokenized as a compound noun, it has not been given the *Nbp* feature; therefore, the system did not capture any meronymic relation for this element. Even so, the problem is in the missed apposition relation of the two `PP`s with the subject complex `NP`, whose head is a pronoun (namely, *um deles* 'one of them'). Since no dependency exists between the subject (*um* ''one') and the apposition and also because the subject is a pronoun, no feature is there to trigger the meronymy rules.

6 Evaluation after Error Analysis

Ones all the corrections were taken into consideration, we ran the system again in order to carry out a second evaluation of the system's performance. The results are shown in Table 2 (the abbreviations and the legend are explained in Table 1).

Table 2. Post-error analysis system's performance for *Nbp*

Number of sentences	TP	TN	FP	FN	Precision	Recall	F-measure	Accuracy
100	10	75	4	12	0.71	0.45	0.56	0.84
900	90	688	39	91	0.70	0.50	0.58	0.86
Total:	100	763	43	103	0.70	0.49	0.58	0.85

The precision improved by 0.13 (from 0.57 to 0.70), the recall by 0.11 (from 0.38 to 0.49), the F-measure by 0.12 (from 0.46 to 0.58), and the accuracy by 0.04 (from 0.81 to 0.85). Since only some of the errors detected, particularly the most frequent, were we able to correct at this stage, and some can still be improved by extending the current work to so far unaddressed situations (dependencies on nouns, anaphora resolution, to name a few) it is expectable that higher levels of performance will be achieved in future work.

7 Conclusions

In this paper, we present the most frequent errors in a rule-based module for whole-part relations extraction involving human entities and body-part nouns (*Nbp*) in Portuguese, which has been implemented and integrated in the STRING NLP system. Around 17 thousand sentences with *Nbp* and disease nouns were extracted from a corpus. 4 Portuguese native speakers annotated a stratified random sample of 1,000 sentences and produced a golden standard, which was confronted against the system's output. The results show 0.57 precision, 0.38 recall, 0.46 F-measure, and 0.81 accuracy. The recall is relatively small (0.38), which can be explained by the fact that in many sentences, the *whole* and the *part* are not syntactically related and are quite far away from each other; naturally, human annotators were able to overcome these difficulties. In some cases, the rules were not triggered because some human nouns and personal pronouns are unmarked with the human feature. Besides, as we focused on verb complements alone, the situations where an *Nbp* is a modifier of a noun or an adjective (and not a verb) have not been contemplated in this work, which produced a significant number of *false negatives*. Other, quantitatively less relevant, cases were also presented in the detailed error analysis made after the systems' first evaluation. The problem derived from pronouns (especially relative pronouns) not having the human feature raises the issue of the adequate placing of the meronymy module in the STRING pipeline architecture: if some part of this task could also be performed *after* anaphora resolution, it is likely that better results would be produced. The precision of the task is somewhat better (0.57). The accuracy is relatively high (0.81) since there is a large number of *true-negative* cases. A detailed error analysis was performed to determine the most relevant cases for these results, which led to some situations being implemented. A second evaluation of the system's performance was carried out, with the same golden standard, and the results showed that the precision improved by 0.13 (from 0.57 to 0.70), the recall by 0.11 (from 0.38 to 0.49), the F-measure by 0.12 (from 0.46 to 0.58), and the accuracy by 0.04 (from 0.81 to 0.85).

From the observations above, it is clear that most of the phenomena here described are not exclusive of the Portuguese language, for example, the *Nbp*-disease noun relations, though there may be language-specific lexical gaps. It is also obvious that the structural descriptions involved in the transformational processes (sentence alternations) here analyzed depend on the particulars of every language syntax (and morphology), which should be modeled independently from the meronymy extraction task. In this respect, this paper may hint both on similar and on different linguistic aspects of the meronymy here tackled, and the observations here made might be useful for other approaches to meronymy extraction, in other languages.

In future work, the extraction of other types of whole-part relations will be addressed such as component-integral object (*pedal* - *bicycle*), member-collection (*player* - *team*), place-area (*grove* - *forest*), and others [30]. We intend to target the extraction of these types of whole-part relations using *syntactic dependency-based n-grams* (the concept is introduced in detail in [28,29]) and other syntactic information, such as subcategorization frames [8,9] in a machine learning approach. Another line of future work will be to use the lists of *Nbp* and several *Nbp*-related words provided by Cláudia Freitas [7] in order to complete the existing *Nbp* lexicon in STRING and to improve the recall by focusing on the *false negative* cases already found, which have shown that several syntactic patterns have not been paid enough attention yet.

Acknowledgments. This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013; Erasmus Mundus Action 2 2011-2574 Triple I - Integration, Interaction and Institutions; and the Mexican Government (scholarship CONACYT).

References

1. Ait-Mokhtar, S., Chanod, J., Roux, C.: Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering 8(2/3), 121–144 (2002)
2. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Ph.D. thesis, Aarhus Univ. Aarhus, Denmark: Aarhus Univ. Press (2000)

3. Costa, F., Branco, A.: LXGram: A Deep Linguistic Processing Grammar for Portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS, vol. 6001, pp. 86–89. Springer, Heidelberg (2010)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
5. Fleiss, J.L.: Statistical methods for rates and proportions, 2nd edn., pp. 38–46. John Wiley, New York (1981)
6. Freelon, D.: ReCal: Intercoder Reliability Calculation as a Web Service. Intl. J. of Internet Science 5(1), 20–33 (2010)
7. Freitas, C.: ESQUELETO - Anotaçã das palavras do corpo humano. Tech. Rep. Versão 5 (May 20, 2014), `http://www.linguateca.pt/acesso/Esqueleto.pdf`
8. Gelbukh, A.: Syntactic disambiguation with weighted extended subcategorization frames. In: Proceedings of PACLING-99, Pacific Association for Computational Linguistics, pp. 244–249. University of Waterloo, Canada (1999)
9. Gelbukh, A.: Unsupervised Learning for Syntactic Disambiguation. Computación y Sistemas 18(2), 329–344 (2014)
10. Girju, R., Badulescu, A., Moldovan, D.: Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In: Proceedings of HLT-NAACL, vol. 3, pp. 80–87 (2003)
11. Girju, R., Badulescu, A., Moldovan, D.: Automatic discovery of part-whole relations. Computational Linguistics 21(1), 83–135 (2006)
12. van Hage, W.R., Kolb, H., Schreiber, G.: A method for learning part-whole relations. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 723–735. Springer, Heidelberg (2006)
13. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conf. on Computational Linguistics, COLING 1992, vol. 2, pp. 539–545. ACL Morristown, NJ (1992)
14. Hirst, G.: Ontology and the lexicon. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 209–230. Springer (2004)
15. Iris, M., Litowitz, B., Evens, M.: Problems of the Part-Whole Relation. In: Evens, M. (ed.) Relational Models of the Lexicon: Representing Knowledge in Semantic Networks, pp. 261–288. Cambridge Univ. Press (1988)
16. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)
17. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In: Computational Processing of Portuguese, PROPOR 2012, vol. Demo Session (2012), `http://www.propor2012.org/demos/DemoSTRING.pdf`
18. Marques, J.: Anaphora Resolution. Master's thesis, Univ. of Lisbon/IST and INESC-ID Lisboa/L2F (2013)
19. Marrafa, P.: WordNet do Português: uma base de dados de conhecimento linguístico. Instituto Camões (2001)
20. Marrafa, P.: Portuguese WordNet: general architecture and internal semantic relations. DELTA 18, 131–146 (2002)
21. Marrafa, P., Amaro, R., Mendes, S.: WordNet.PT Global – extending WordNet.PT to Portuguese varieties. In: Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, pp. 70–74. ACL Press, Edinburgh (2011)

22. Oliveira, H.: Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese. Ph.D. thesis, Univ. of Coimbra/Faculty of Science and Technology (2012)
23. Oliveira, H.G., Santos, D., Gomes, P., Seco, N.: PAPEL: A Dictionary-Based Lexical Ontology for Portuguese. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 31–40. Springer, Heidelberg (2008)
24. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of Conf. on Computational Linguistics/ACL, COLING/ACL 2006, pp. 113–120. Sydney, Australia (2006)
25. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: developing an aligned multilingual database. In: 1st Intl. Conf. on Global WordNet, Mysore, India, pp. 293–302 (2002)
26. Prévot, L., Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A.: Ontology and the lexicon: a multi-disciplinary perspective (introduction). In: Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévot, L. (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective. Studies in Natural Language Processing, ch. 1, pp. 3–24. Cambridge Univ. Press (2010)
27. Rocha, P., Santos, D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M. (ed.) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000), pp. 131–140. São Paulo, ICMC/USP (2000)
28. Sidorov, G.: Non-continuous Syntactic N-grams. Polibits 48, 67–75 (2013)
29. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernandez, L.: Syntactic N-grams as Machine Learning Features for Natural Language Processing. Expert Systems with Applications 41(3), 853–860 (2013)
30. Winston, M., Chaffin, R., Herrmann, D.: A Taxonomy of Part-Whole Relations. Cognitive Science 11, 417–444 (1987)