J Braz Comput Soc (2013) 19:75–87
DOI 10.1007/s13173-012-0082-6

ORIGINAL PAPER

A semi-automated method for acquisition of common-sense and inferentialist knowledge

Vladia Pinheiro · Vasco Furtado · Tarcísio Pequeno · Wellington Franco

Received: 23 February 2012 / Accepted: 4 July 2012 / Published online: 11 August 2012

Abstract This paper presents a semi-automated method for the acquisition of common-sense and inferentialist concepts in Portuguese. Its innovative feature is a module of reasoning over the pre-existing knowledge that aims to offer original content to the user, helping in the expression of semantic relationships and validation of new concepts. This reasoning process is based on heuristics and syntactic analysis of noun phrases. A qualitative evaluation with users who interacted with the system built on the proposed method showed that the interactions made in the process of knowledge acquisition are more productive since the user is reminded about semantic relations and common-sense knowledge about the new concepts.

Keywords Knowledge acquisition method · Common-sense knowledge · Inferentialist knowledge

1 Introduction

Common-sense knowledge consists of spatial, physical, social, temporal and psychological facts, and knowledge, possessed by most people, which are fruits of daily life experiences [2,12]. Often, this knowledge is a set of implicit and basic assumptions that support and explain the reasoning necessary to carry out intelligent tasks by computers (e.g., understanding texts in natural language). For example, when someone says "I bought candy," it is implicit that they used money; that the effect of falling off a motorcycle is you get hurt; that objects roll down inclined surfaces; that politicians are involved in corruption and scandals.

Particularly in the area of Natural Language Processing (NLP), there is a consensus that the understanding of texts by computer systems depends not only on linguistic knowledge but also on world knowledge [14]. However, one of the challenges of research in this area is the continuous evolution of semantic–linguistic resources that express world knowledge to support NLP tasks, such as information extraction, information retrieval, question and answer systems, text summarization, semantic annotation of texts, among others. This challenge is even greater when we consider the Portuguese language [23].

In the search for semantic expression models, language philosophies lend some inspiration to the PLN researchers through their semantic theories, where the goal is to understand the nature of the content of concepts. In general, a *concept* refers to the semantic value expressed by a linguistic expression of a natural language, when used in a sentence. A concept can be named by simple terms that belong to the open classes of words—nouns, verbs, adjectives, adverbs (e.g., "crime", "death", "to reside"), or by expressions composed of more than one term, whether or not linked by closed classes of words—prepositions and conjunctions (e.g. "math test", "street mugging", "crime of passion"). In this work, we use the terms "concept", "content of a concept" or "conceptual content" to refer to the semantic value of a linguistic expression that named a concept. For example, when we refer to the concept "politician", named by the linguistic expression

This is a revised and extended version of a previous paper that appeared at STIL 2011, the 8th Brazilian Symposium in Information and Human Language Technology.

V. Pinheiro · V. Furtado · T. Pequeno · W. Franco (✉)
Universidade de Fortaleza (UNIFOR),
Av. Washington Soares 1321, Fortaleza, CE 60.811-905, Brazil
e-mail: jwellingtonfranco@lia.ufc.br

V. Pinheiro
e-mail: vladiacelia@unifor.br

V. Furtado
e-mail: vasco@unifor.br

T. Pequeno
e-mail: tarcisio@unifor.br

"politician", we are refering to the semantic relations that define the semantic value of this concept.

In this sense, the InferenceNet resource [27] was constructed containing common-sense and inferentialist semantic relations about concepts and sentences, which are expressed in Portuguese and English. The semantic bases of the InferenceNet resource were constructed according to the Semantic Inferentialist Model (SIM) [24] and express the pragmatic character of natural language through preconditions and consequences (post-conditions) of the use of concepts and sentences.

As occurs with other semantic bases, one of the difficulties is to assure continuous evolution of the InferenceNet linguistic resource effectively and with the timeliness required by the applications. Methods of automatic knowledge acquisition (KA), although widely used in NLP [5,10], are not shown to be useful for capturing tacit and common-sense knowledge, because this knowledge is not commonly derivable from grammatical and structural properties of texts available in linguistic corpora [14]. On the other hand, traditional semi-automatic KA methods (model-based)—for example, those adopted in the Open Mind Common Sense (OMCS) project [34] and OMCS-Br [38] face difficulties in capturing users pragmatic knowledge and common-sense knowledge. One difficulty stems from the fact that people have these types of knowledge (common-sense and pragmatic), but they do not know how to make it explicit; the knowledge is so ingrained in people's minds that it is difficult to remember and even more difficult to externalize it through structured semantic relations. Another difficulty is that even when people are able to make common-sense knowledge explicit, it is difficult to assure consistency with the existing conceptual content, avoiding duplication of content, and strengthening the connection of the semantic network.

In this work, we propose a semi-automated method for support the acquisition of common sense knowledge that characterizes a concept expressed in Portuguese. The differential of the method is its ability to reason over a preexisting knowledge-base and to infer relations that helps the user to build new relationship between concepts. The reasoning process is based on heuristics that, according to the grammatical structure of a linguistic expression, allows inferring common-sense relations that characterizes the concept dubbed by that expression. For example, to acquire the common-sense relations of the concept named in Portuguese "*crime passional*" [in English: "crime of passion"], one uses a specific strategy for the grammatical structure "<noun><adjective>", which utilizes the preexisting semantic knowledge of the concepts "*crime*" and "*passional*", or the concepts "*crime*" and "*paixão*". Moreover, the method provides for an interactive process that favors better accuracy in the capture and validation of semantic relations by the user. The KA method proposed in this paper was implemented and evaluated for the bilingual conceptual base InferenceNet and for the portuguese common-sense base OMCS-Br.

2 Noun phrases

According to [33], the smallest unit of meaning within a clause is known as a phrase. The phrase has a main element, called the head, which defines the nature of the phrase. In Portuguese, the following phrases are defined [15]:

- Noun Phrase (NP), when the head of the phrase is a noun;
- Adjectival Phrase (AdjP), when the head of the phrase is an adjective;
- Verb Phrase (VP), when the head of the phrase is a verb;
- Prepositional Phrase (PP), when the head of the phrase is a preposition;
- Adverbial Phrase (AdvP), when the head of the phrase is an adverb.

In the syntactic analysis of a sentence, the phrases are identified and appropriately qualified. This task is performed by constituent parsers or by dependency parsers. In Computational Linguistic, parsing, or more formally, syntactic analysis, is the process of analyzing a text, made of a sequence of part-of-speech (e.g., words), to determine its grammatical structure with respect to a given formal grammar [13]. Dependency grammar is a class of modern syntactic theories that are all based on the dependency relation between a word (a head) and its dependents. Constituent grammars, on the other hand, focus on identification of the phrases structure (noun phrases, verb phrases, etc.). There are several parsers available with their precision results in the dependency parsing task: PALAVRAS (precision = 99 %) [6], Brill TBL (precision = 97 %) [8], TreeTagger (precision = 96 %) [31], and FreeLing (precision between 97–98 %) [22]. For example, Fig. 1 shows the results of the dependence parsing performed by the parser PALAVRAS [6] in the sentence "*Os ladrões oportunistas agiram impunemente durante a greve da Polícia Militar do Ceará*" (which translates literally as "The opportunistic thieves acted with impunity during the strike by the Military Police of Ceará"). The following phrases were identified, whose nuclei are underlined: "*os ladrões oportunistas*" (NP); "*agiram*" (VP); "*impunemente*" (AdvP); "*durante*" (PP); "*a greve*" (NP); "*a Polícia Militar do Ceará*" (NP).

In this paper, the focus will be on the noun phrases (NP), because these are usually used to describe the "things" in the world, and therefore, are primarily used to name concepts of the natural language.

The head of the noun phrase (NP) may consist of a noun (proper or common) or a pronoun (personal, demonstrative, indefinite, interrogative, possessive, or relative). When the head is a pronoun, this pronoun per se will represent the NP.

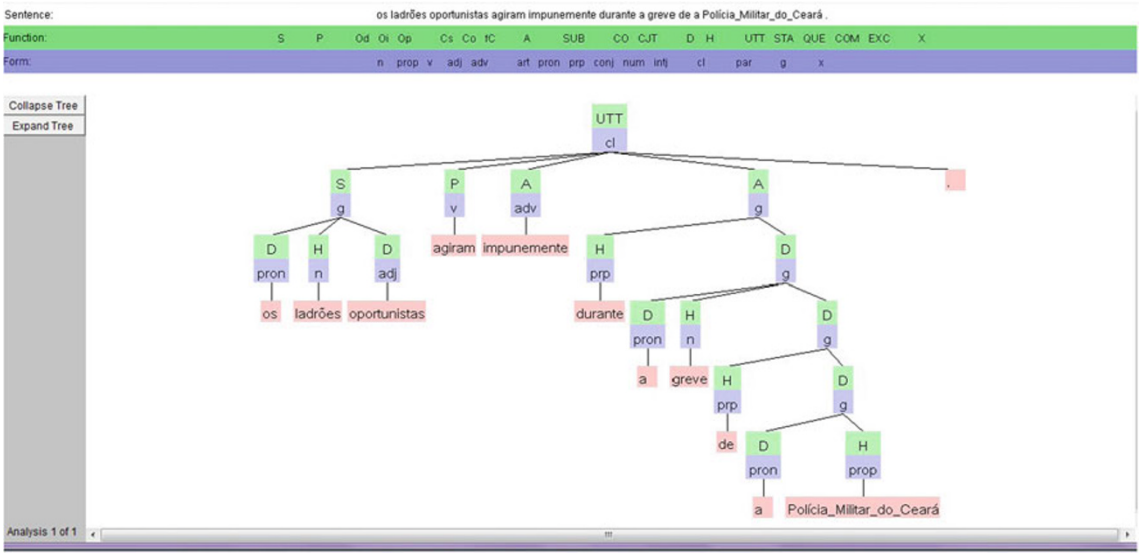

Fig. 1 Syntactic analysis of the sentence "Os ladrões oportunistas agiram impunemente durante a greve da Polícia Militar do Ceará", *highlighted* for an analysis of the phrases

In addition to the head (H), the NP may also have determinant(s) (DET) and/or modifier(s) (MOD). In Portuguese, determinants precede the head and modifiers follow the head. The determinants of noun phrase are represented by articles, pronouns, and numerals. We can cite the following as examples of noun phrases with DET + H: ***a luz*** [the light]; ***o sol*** [the sun]; ***um jornal*** [a newspaper]; ***certas tardes*** [certain afternoons].

Modifiers, in turn, are represented by adjectives and adjectival phrases. They serve to characterize or express an evaluation on the nouns. To characterize the noun, adjectival phrases are most often used. The following noun phrases, of the form H + MOD, are examples in which an adjectival phrase (underlined) characterizes the head: ***bola de futebol*** ["soccer ball," lit.: "ball of soccer"]; ***panela de arroz*** ["pan of rice"]; ***pista de corrida*** ["race track," lit.: "track of race"]; ***amor de mãe*** ["mother's love," lit.: "love of mother"]. The modifiers are also used to express an evaluation of the noun. In this case, simple adjectives are most often used. The following noun phrases, of the form H + MOD, are examples in which a simple adjective (underlined) expresses an evaluation of the noun: ***bola estragada*** ["ruined ball"]; ***juiz ladrão*** ["crooked judge"]; ***criminoso cruel*** ["cruel criminal"]; ***crime passional*** ["crime of passion"]; ***amor materno*** ["maternal love"].

Table 1 shows the main structures of noun phrases and respective examples.

Table 1 Main structures of noun phrases and examples

Structure of the NP	Example
DET + H + MOD	*os aguaceiros de verão*
H + MOD	*chuva grossa*
DET + DET + H	*uma certa crença*
DET + H + DET + H + MOD	*a terra e a areia assentadas*
MOD + H + MOD	*grande movimentação de bichos*
DET + DET + H + MOD	*uma certa alegria despropositada*

3 A knowledge common sense and inferentialist semantic base—InferenceNet

The motivation and construction process of InferenceNet's website are described in [27]. In the context of this work, InferenceNet's Conceptual Base is the most important since it contains the inferential and common-sense content of concepts of the Portuguese and English languages, defined and agreed upon in a community or area of knowledge. Moreover, InferenceNet is linked to the Linked Open Data cloud (LOD) [26], which permits the retrieval of related semantic content in other bases, such as DBPedia [4], WikiPedia, Yago [37], WordNet [19]. According to the inferentialist view [7], the content of a concept must be expressed, becoming explicit, through the use of it (the concept) in inferences, as premises or conclusions of reasoning. Moreover, what determines the use of a concept in inferences or potential inferences in which this concept may participate are: (i) its pre-conditions or premises of use what gives someone the right to use the concept and what could exclude such a right, serving as premises for utterances and reasoning; and (ii) its post-conditions or conclusions of use what follows or what are the consequences of using the concept, which let one know what someone is committed to by using a particular concept, serving as conclusions from the utterance per se and as premises for future utterances and reasoning. Formally, this base is represented

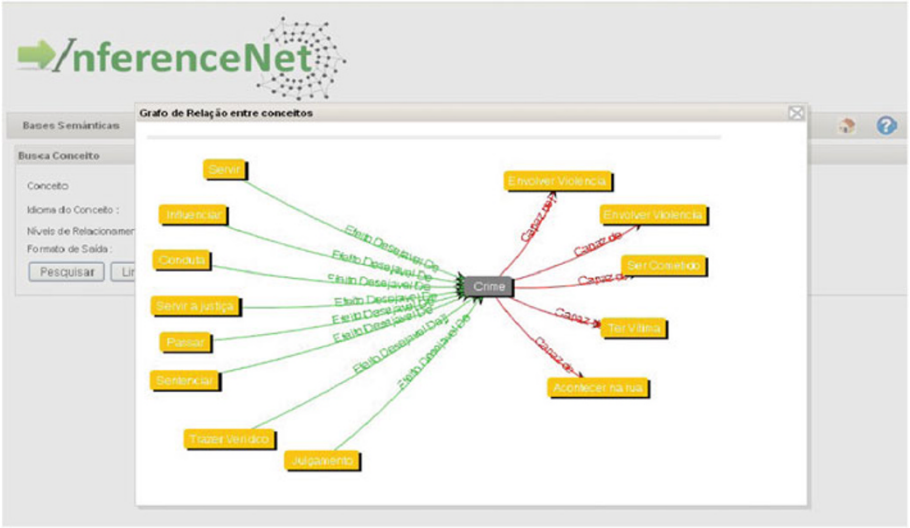

Fig. 2 Part of the conceptual base for the concept **crime** with some pre-conditions (*incoming arrows*) and post-conditions (*outgoing arrows*)

in a directed graph $G_c(C, R_c)$. Each inferential relationship $r_{cj} \in R_c$ (set of inferential relations of the concept $c \in C$) is represented by a tuple $(rel\_Name, c_i, c_k, type)$, where rel_Name is the name of an InferenceNet semantic relation (Capableof, PropertyOf, EffectOf etc.), c_i and c_k are concepts of a natural language, and $type =$ "Pre" or "Pos" (pre-condition or post-condition for using the concept c_i). Figure 2 presents part of the conceptual base for the concept **crime**.

One motivation for a new linguistic resource is the low existence of linguistic resources with large-scale inferentialist semantic knowledge for the Portuguese language. Lexical-semantic bases in Portuguese, for example, WordNet.Pt [17], WordNet.Br [32], VerbNet-Br [30] and Propbank-Br [11] are already available, and a common-sense base for Brazilian Portuguese—OMCS-Br—contains 250,000 common-sense relations. InferenceNet represents an evolution, because in addition to expressing around 700,000 common-sense relationships about concepts, these relations are qualified in terms of conditions of use of the concepts, allowing better and richer inferences from texts [25,28].

4 Method of common-sense and inferentialist knowledge acquisition

Figure 3 presents the phases of our method for acquisition of common-sense and inferentialist knowledge. First, the user enters with a linguistic expression <EXP>, used for dubbing the new concept to be acquired. If there is a concept for that expression, the common sense and inferentialist relations, already existing in the knowledge base, are retrieved to be validated by the user. Otherwise, a new concept should be acquired and then the method is executed according to the following steps:

1. Syntatic analysis of EXP, in order to define the gramatical structure of the NP;
2. Generation of conceptual content through the execution of the heuristics to acquire a new concept by the reasoner, from a preexisting Conceptual Base. In this step, the reasoner infers common-sense and inferentialist semantic relations, serving as a baseline for the new concept;
3. Validation of the baseline by the user and definition of the content of the new concept.

Primarily, the proposed KA method consists of a heuristic reasoning applied to pre-existing conceptual and semantically related content, which generates a baseline of knowledge for the new concept. Furthermore, the method enables an interactive process with the user, which can include new inferential and common sense relations and can exclude proposed relationships, closing a mechanism of validation of the conceptual content to be acquired. It is noteworthy that this method is independent of the semantic knowledge base used as the baseline. The more resources of semantic knowledge that are available and interconnected, the more the heuristics will generate a richer baseline for the new concept. This is in line with the current view that knowledge intensive methods for NLP will reason better if they consider several knowledge bases as a joint semantic resource. The current design of ConceptNet, version 5.0, emphasizes this idea and suggests the combination of semantic knowledge of various bases, such as WordNet, Open Mind Common Sense (OMCS), Wikipedia, DBPedia, Wiktionary, ReVerb, and other bases. In this sense, the semantic resource InferenceNet was linked to the LOD cloud (Linked Open Data) through DBPedia, Yago, and WordNet, as described in [26]. In the next subsection, the heuristics are detailed.

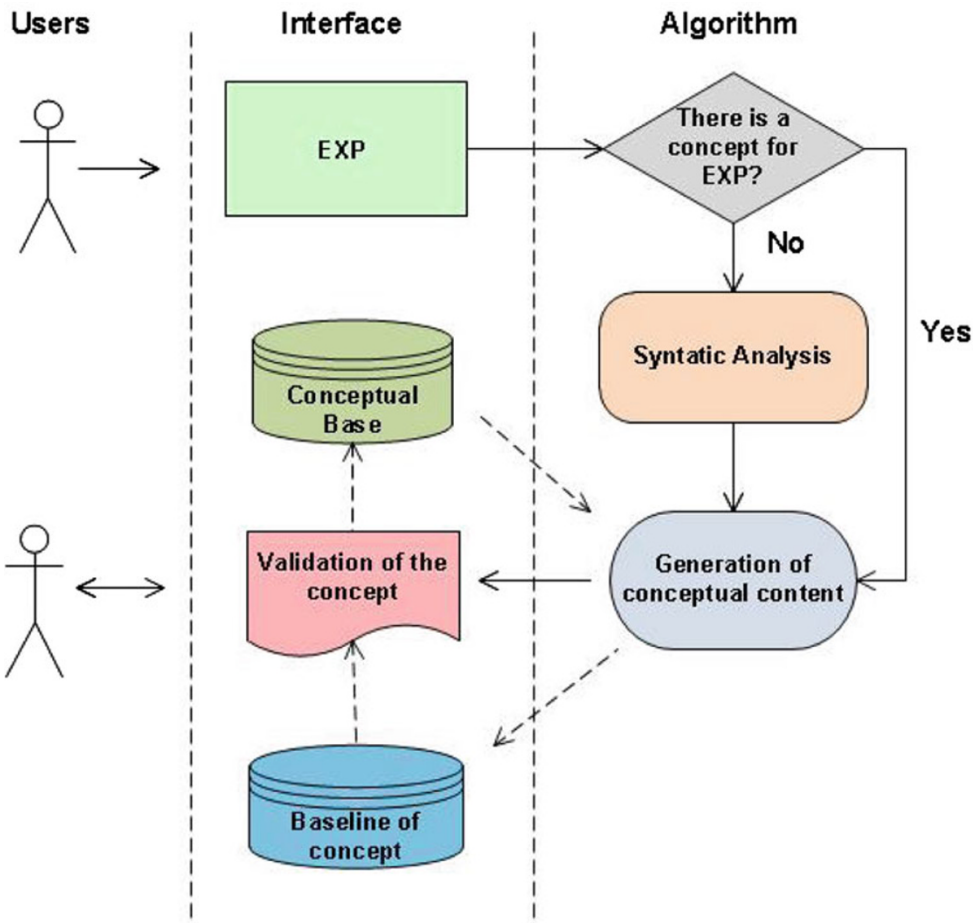

Fig. 3 The method of common-sense and inferentialist KA for concepts in natural language

4.1 Heuristics for common-sense and inferentialist knowledge acquisition

Heuristics are responsible for the generation and proposition of the semantic content for the input linguistic expression EXP, which names the new concept to be acquired. According to the grammatical structure of EXP, a set of heuristics searches in a semantic content related to a pre-existing knowledge base (e.g. InferenceNet, OMCS-Br, etc.) and generates new semantic relations, which are the basis for validation by the user and, then, for definition of the content of the new concept. As stated previoulsy, the proposed heuristics include only noun phrases, because these are usually used to describe the "things" in the world. Table 2 shows the grammatical structures of noun phrases considered by the heuristics.

1. <noun> or <adjective>—When EXP is not found in the conceptual basis, the user is shown a set of pre-existing concepts in the base, which are: (i) semantically related (e.g., synonyms); (ii) named with the same root of <noun> or <adjective>; (iii) named with the primitive form of <noun>; (iv) nouns related to <adjective>. For this, lexical resources, such as TEP [18], Onto.PT [21], and others can be used. For example, for the linguistic expression "*torcedor*" [sports fan], the heuristic would present the concepts "*fã*" ["fan"], "*torcida*" ["group of fans"] and "*torcer*" ["to cheer for"]. For the word "*passional*", the heuristic would present the concept "*paixão*"["passion"]. Then the user selects which of the concepts presented can be used as the basis for the acquisition of the new concept. The heuristic returns a list of semantic relations of the selected concept, previously contained in the database.
2. <noun><adjective> or <adjective><noun>—In these cases, <noun> is characterized by <adjective>, giving it an attribute, property, status, mode of being, or aspect. One can therefore perceive a case of specialization, in

Table 2 Main grammatical structures of noun phrases

Structure of noun phrase EXP	Examples
<noun>	*vingança, pistolagem*
<noun> <adjective>	*crime passional, impunidade penal*
<adjective> <noun>	*má urbanização*
$<adjective_1><noun><adjective_2>$	*má iluminação pública*
$<noun_1>$ <"DE"> $<noun_2>$	*aula de português, bola de plástico*

Table 3 Types of semantic relationships of InferenceNet that will be inherited from <adjective> to <noun><adjective> or <adjective><noun>

Nature of the relationship	Type of semantic relationship	Type of inferential relationship
RELATIVE TO PROPERTY	PropertyOf	Pre-condition
RELATIVE TO EVENT	LastSubEventOf, PreRequirementEventOf, FirstSubEventOf, SubEventOf	Pre-condition
CAUSAL	EffectOf, DesirousEffectOf	Post-condition
MOTIVATIONAL	DesireOf, MotivationOf	Pre-condition
FUNCTIONAL	UsedFor;	Pre-condition

which "<noun><adjective>" or "<adjective><noun>" expresses a particular situation or a type of $< noun >$. For example, in the case of the expression "*crime passional*", the adjective "*passional*" is characterizing the noun "*crime*", attributing properties relative to "*paixão*" ["passion"] and a type of "*crime*". This heuristic defines the following steps:

(a) Recursive call to the heuristic (1) for EXP1 = <noun> and EXP2 = <adjective>, returning a list of semantic relations of concepts associated with EXP1 and EXP2;
(b) Inheritance of the content of <noun> to the new concept "<noun><adjective>" or "<adjective> <noun>", since in both there is the expression of a particular case or type of $< noun >$ and therefore, the entire content of $< noun >$ can be transcribed (or inherited) to "<noun><adjective>" or "<adjective><noun>". For example, the semantic relationship "<***crime***> <*capableOf*> <***have victim***>" is transcribed to a new semantic relationship "<***crime passional***> <*capableOf*> <***have victim***>";
(c) Partial transcription of the content of <adjective> to the new concept "<noun><adjective>" or "<adjective> <noun>". In this case, <adjective> is characterizing <noun> and some semantic relations of <adjective> must be transcribed to <noun> so as to give it characteristics or qualities.
The following metarule is used in this step:

$$\frac{<A> is\ characterized\ by\ <B>, <B>\ <rel\_name>\ C}{\rightarrow <A\ characterized\ by\ B>\ <rel\_name>\ C}$$

To define which make this inference valid, each *rel_name* from the semantic base should be analyzed according to the nature of the semantic relationship. In general, structural semantic relations (e.g., isA, madeOf, partOf) usually should not be inherited because they express content restricted to <adjective>. For example, the fact that "<***paixão***> <isA> <***feeling***>" does not imply that "<***crime passional***><isA><***feeling***>".

Pragmatic semantic relationships as functional, causal, incidental or motivational relationships commonly give rise to characteristics that are attributed from <adjective> to <noun>. For example, the fact that "<***paixão***> <*effectOf*> <***jealousy***>" authorizes the generation of the content "<***crime passional***> <*effectOf*> <***jealousy***>". As an example, Table 3 shows the types of semantic relationships of InferenceNet defined for the application of the metarule above. However, other semantic knowledge bases can be used, simply by analyzing the nature of relations expressed and which ones can be inherited from <adjective> to "<noun> <adjective>" or "<adjective><noun>". At the end of the process, the heuristic returns the list of semantic relations generated, which were associated with "<noun> <adjective>" or "<adjective><noun>".

3. $< adjective_1 >$ <noun> $< adjective_2 >$—In this case, the user is asked if $< adjective_1 >$ is qualifying "<noun> $< adjective_2 >$", for example, as occurs in the phrase "*má iluminação pública*". If the user confirms, the heuristic (2) is called for EXP="<noun> $< adjective_2 >$" and then for EXP="$< adjective_1 >< np_2 >$" with $< np_2 >$ = "<noun> $< adjective_2 >$". Otherwise, the heuristic (2) is called for EXP="$< adjective_1 >$ <noun>" and for EXP="<noun> $< adjective_2 >$". At the end, the heuristic returns the list of semantic relations selected recursively.
4. $< noun_1 >$ <"DE"> $< noun_2 >$—In this case, the user is asked if $< noun_2 >$ is characterizing "$< noun_1 >$", for example, as occurs in the phrase "*aula de português*". If the user confirms, the heuristic (2) is called for EXP="$< noun_1 >< noun_2 >$" with $< noun_2 >$ being an adjective phrase that is expressing a characterization of $< noun_1 >$. Otherwise, the heuristic (1) is called for EXP="$< noun_1 >$" and for EXP="$< noun_2 >$". At the end, the heuristic returns the list of semantic relations selected recursively.

Figure 4 shows the algorithm that implements the proposed heuristics, exemplifying for the new concept "*crime passional*" from InferenceNet.

5 Evaluation

The evaluation is aimed at analyzing two aspects: (i) how the heuristics facilitate the acquisition of conceptual common-sense and pragmatic knowledge for the Portuguese language; and (ii) the quality of conceptual content generated by the heuristics, i.e., whether the proposed content actually expresses the semantic value of the concept desired by the user. In this evaluation, the algorithm was implemented for KA of concepts for the InferenceNet base, and the dependency parser PALAVRAS [6] was used. For retrieval of synonyms and related words (according to heuristic 1, in Sect. 4.1), we used a service available on the web http://www.dicionarioinformal.com.br, merely for the sake of technical simplicity. However, the method can be applied to acquire new content for other common-sense knowledge bases, such as OMCS-Br, and other dependency parsers for the Portuguese language can be used, such as MaltParser [20].

Below, we present the two evaluation experiments that were performed.

5.1 An empirical evaluation of the quality of the interactive knowledge acquisiton

The evaluation methodology of the first experiment we have proposed to evaluate the quality of our proposal had the steps outlined as follows.

1. Selection of 20 adults with experience in interactive systems of the Internet and with no knowledge of the KA method proposed in this study. The individuals were randomly assigned to 2 (two) groups of 10 people—one group for each test scenario;
2. Selection of concepts used in Portuguese that did not previously exist in the InferenceNet base: "*crime passional*", "*violência policial*", "*má iluminação pública*", "*bom juiz honesto*", "*aula de português*", "*bola de plástico*". These concepts were selected because they cover all the heuristics proposed in this work.
3. Definition of test scenarios:
   - **Scenario 1** The users, with no time limit, will include semantic relations for the chosen concepts, through InferenceNet's web site, which has an interactive interface that allows users to enter common-sense and inferentialist relations in the InferenceNet base.
   - **Scenario 2** On InferenceNet's web site, the user enters the linguistic expression EXP corresponding to the concept and interacts with the portal to validate the conceptual content generated by the algorithm implemented. The users were asked to modify and exclude semantic relations if they disagreed with them, and to include new relationships if deemed necessary, also with no time limit.
4. Generation of a baseline, where an adult human evaluator validated the semantic relationships generated by the algorithm and defined a baseline for the concepts of this evaluation. This human evaluator is not a linguist, since the semantic relations were common-sense relations, which require the evaluator to have only a level of proficiency in natural language (e.g. in the Portuguese language) and daily experience. The baseline was used for qualitative analysis of the semantic content at the end of the KA process experienced by the 10 users in Scenario 2.

In each scenario, the time to perform the activity was measured, as well as and how many semantic relations were included or excluded for each concept selected. Tables 4, 5, and 6 show the average results. In Scenario 2, the algorithm generated the following 1,082 relations for the concepts in question, distributed as follows: ***crime passional***—45 pre-conditions and 17 post-conditions; ***violência policial***—67 pre-conditions and 1 post-condition; ***má iluminação publica***—13 pre-conditions; ***bom juiz honesto***—69 pre-conditions and 7 post-conditions; ***aula de português***—53 pre-conditions and 1 post-conditions; ***bola de plástico***—808 pre-conditions and 1 post-conditions. Figure 5 presents the screenshot of the InferenceNet's web site used in this scenario.

Based on the results collected, we showed that the proposed method enables more productive interactions for KA: in Scenario 1 users took, on average, 3 min 22 s to include 7 semantic relations (average of pre-conditions and post-conditions included for the 6 concepts), while in Scenario 2 users performed 9.55 exclusions and inclusions of semantic relationships in 3 min 49 s (average time). We noted that in Scenario 1, users found it difficult to express common-sense semantic relations about the concept and, in some cases, even to remember what semantically characterizes that concept. In Scenario 2, the user is prompted to interact with the semantic relationships generated, resulting in better inclusion— exclusion ratio per minute (2.0 in Scenario 1, against 2.5 in Scenario 2). Another interesting fact is that the number of semantic relations generated by the method is much larger than the inclusions made by users in Scenario 1, even considering the exclusions made in Scenario 2: 1,026 semantic relations generated by the methods against 69 semantic relations included by the users in Scenario 1. Regarding the quality of conceptual content generated by

Fig. 4 Algorithm that implements heuristics for generating conceptual conceptual content

```
SemanticRelations [ ] generateContent (exp)

// 1st iteration:  exp=" crime passional"
// 2nd  iteration: exp = "crime"
// 3rd iteration: exp = "passional"

If knowledgeBaseExists (exp) then
     return retrieveContent(exp);
     // 2nd iteration: for exp = "crime"
     // retrieved semantic relations of "crime"
     // Examples: (capableOf, "crime","envolver violência", Pre); (effectOf, "crime","sofrimento", Pos)
else {
     if structure(exp) = "<noun>" or "<adjective>":   // HEURISTIC (1)
          relatedConcepts [ ] = retrieveRelatedConcepts (exp);
          relatedConcept = selectionUser(relatedConcepts [ ]);
          return retrieveContent(relatedConcept);

          // 3rd iteration: for exp = "passional"
          // retrieved semantic relations of "paixão" – relatedConcept selected by user
          // Examples: (eventPreRequisitOf,"paixão","amante", Pre);
          //              (effectOf,"paixão","sofrimento", Pos)
          //              (usedFor ,"paixão","romance", Pre); (isA, "paixão","sentimento", Pos)

     If structure(exp) = "<noun><adjective>" or "<adjective><noun>": // HEURISTIC (2)
                    // 1st iteration:  exp="crime passional"

          exp1 = firstTerm(exp);                                   // exp1 = "crime"
          exp2 = secondTerm(exp);                                  // exp2 = "passional"
                    content1 = generateContent(exp1)      ;    // recursive  call for exp1= "crime"
                    content2 = generateContent(exp2);          // recursive  call for exp2= "passional"
          if structure(exp1) = "<adjective>" then {
               content1 = selectContent(content1);
          } else {
               content2 = selectContent(content2);
               // selects relations of content2 (referring to exp2="passional") per step 2.c
               //  Examples: (eventPreRequisitOf,"paixão", "amante", Pre);
               //              (effectOf, "paixão", "sofrimento", Pos)
               //              (usedFor, "paixão", "romance", Pre);
               //            Note: the relation "isA("paixão", "sentimento", Pos)" was not selected
          }
          return content1+content2;
     if structure(exp) = "<adjective1><noun><adjective2>" :  // HEURISTIC (3)
          if <adjective1> qualifies "<noun><adjective2>" then {
               exp1 = <adjective1>;
               exp2 = <noun> <adjective2>;
               content1 = generateContent(exp1);
               content2 = generateContent(exp2);
               content1 = selectContent(content1);
          } else {
               exp1 = <adjective1><noun>;
               exp2 = <noun><adjective2>;
               content1 = generateContent(exp1);
               content2 = generateContent(exp2);
          }
          return content1+content2;
     if structure(exp) = "<noun1><DE><noun2>" :  // HEURISTIC (4)
          if <noun2> characterizes "<noun1>" then {
               exp1 = <noun1>;
               exp2 = <noun2>;
               content1 = generateContent(exp1);
               content2 = generateContent(exp2);
               content2 = selectContent(content2);
          } else {
               exp1 = <noun1>;
               exp2 = <noun2>;
               content1 = generateContent(exp1);
               content2 = generateContent(exp2);
          }
          return content1+content2;

}
```

Table 4 Results collected in the two evaluation scenarios and the baseline

Scenarios	"crime passional"			"violência policial"		
	Incl	Excl	Time	Incl	Excl	Time
Scenario 1	2.9 pre	–	00:02:55	4.1 pre	–	00:02:31
	1.6 pos	–		2 pos	–	
Scenario 2	0 pre	8.2 pre	00:03:46	0 pre	17.9 pre	00:03:52
	0 pos	0.7 pos		0.1 pos	0.4 pos	
Baseline	4 pre	10 pre	n/a	5 pre	23 pre	n/a
	1 pos	5 pos		1 pos	1 pos	

Table 5 Results collected in the two evaluation scenarios and the baseline

Scenarios	"bom juiz honesto"			"má iluminação pública"		
	Incl	Excl	Time	Incl	Excl	Time
Scenario 1	5.7 pre	–	00:03:49	2.9 pre	–	00:02:25
	2.2 pos	–		1.2 pos	–	
Scenario 2	0.9 pre	5.8 pre	00:02:16	0 pre	6.3 pre	00:01:42
	0.1 pos	0.6 pos		0.1 pos	0 pos	
Baseline	–	15 pre	n/a	4 pre	7 pre	n/a
	–	2 pos		3 pos	0 pos	

Table 6 Results collected in the two evaluation scenarios and the baseline

Scenarios	"aula de português"			"bola de plástico"		
	Incl	Excl	Time	Incl	Excl	Time
Scenario 1	6.1 pre	–	00:04:54	6.9 pre	–	00:04:22
	2 pos	–		1.9 pos	–	
Scenario 2	1 pre	9.2 pre	00:04:20	0.7 pre	15.7 pre	00:07:49
	0 pos	0 pos		0.2 pos	0.7 pos	
Baseline	–	12 pre	n/a	–	240 pre	n/a
	–	–		–	–	

the heuristics, we compared the conceptual graphs of the six concepts, after the inclusions and exclusions made by the users, and the baseline conceptual graph. As the main result, we found that 72 % of the relations generated were validated by humans, i.e., 783 of the 1,082 semantic relations generated by the algorithm were confirmed by the 10 users of scenario 2 and were contained in the baseline, constructed as explained in item 4 above (considering the average of exclusions of distinct semantic relations from the baseline and by the 10 users who participated in Scenario 2). It is important to note that in Scenario 2, users were limited to excluding only those relations that seemed invalid for the concept, and included practically no new relationships. Although we did not question the users as to the reason for this behavior, mainly because we perceived this characteristic only during analysis of the results, we believe that the users considered the content presented as sufficient and that a validation of the inappropriate and/or incorrect relations would be enough.

5.2 A quantitative evaluation of the heuristics used by the KA method

In another experiment, we sought to measure the coverage of the proposed heuristics in the InferenceNet base, i.e., how capable the heuristics are of retrieving concepts similar to new concepts. The list of new concepts was formed with 500 nouns of markers from the base of collaborative maps created through the WikiMapps tool http://www.wikimapps.com [29].

In the first scenario, we apply the heuristics for commonsense and inferentialist KA (detailed in Sect. 4.1), with no user interaction, to retrieve concepts related to the markers. In the second scenario, we applied the LUCENE indexing

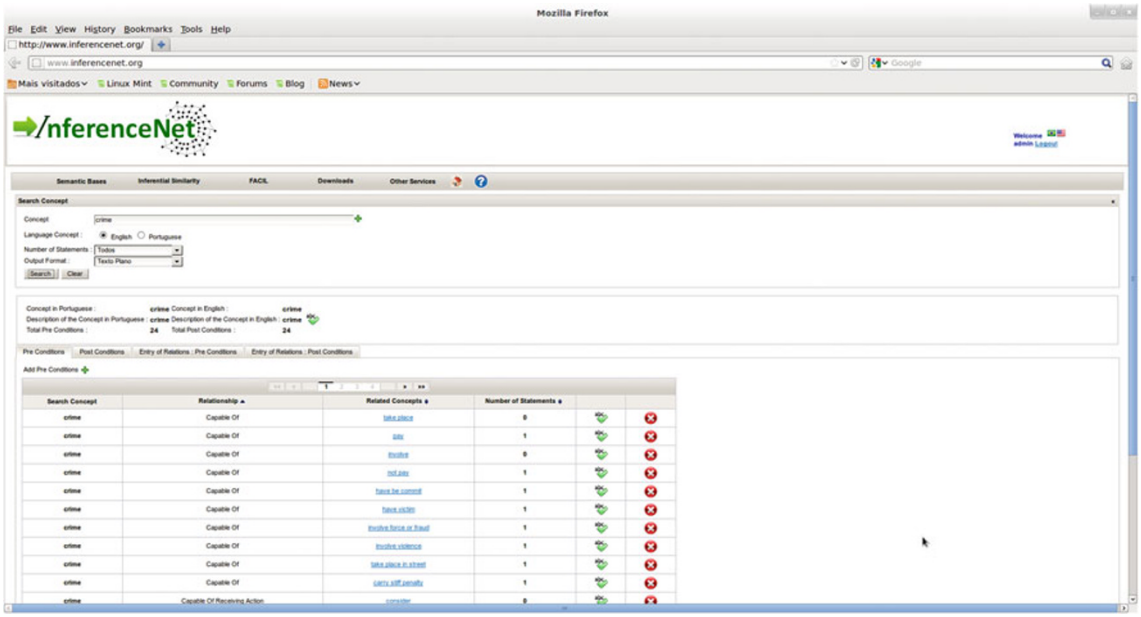

Fig. 5 Screenshot of the InferenceNet's web site used in Scenario 2

tool to syntactically retrieve concepts related to the names of the markers. The LUCENE tool was used for its syntactic indexing capacity. We wanted to measure how capable of retrieving related concepts the heuristics would be in relation to simple searches that considered lexical and syntactic variations. In both scenarios, the InferenceNet Conceptual Base was used as a repository of concepts. Finally, a human analyzed the quality of the concepts retrieved in both scenarios, discarding non-conformities. This human was an adult specialist of the WikiMapps project. Note that again we did not need any special expertise of the evaluator because the task was only to know if a concept retrieved by the heuristics were related to the concept associated to the marker.

As a result, we found that the heuristics were able to retrieve concepts for 81 % of the markers, and the LUCENE syntactic indexing tool retrieved concepts for 62 % of the markers. It is noteworthy that in the latter case, the concepts retrieved were only those syntactically similar to the marker name. For example, for the "*educação*" ["education"] marker, the concept "*educar*" ["to educate"] was retrieved by LUCENE. In the case of the heuristic process proposed in this work, semantically related concepts were retrieved. For example, for the marker "*favela*" ["shantytown"], the concept "*gueto*" ["ghetto"] was retrieved, and for the marker "*desastre ambiental*" ["environmental disaster"], the concepts "*ambiente*" ["environment"] and "*floresta*" ["forest"] were retrieved.

In the third and fourth evaluation scenarios, the first two scenarios were adapted to use the common-sense base OMCS-Br [3]. Our goal was to ascertain how dependent the heuristics were on a conceptual base. As expected, the use of the OMCS-Br base in combination with proposed heuristics resulted in 71 % coverage (scenario 3) and with the syntactic search alone, we had 37 % coverage. The results, compared to the scenarios in which InferenceNet was used, show a reduction in the coverage for both methods because OMCS-Br have fewer concepts and relations. However, the use of the heuristics show an evident gain against the search exclusively synthatic.

6 Related work

Research in KA has focused on the acquisition of common-sense knowledge based on the collaborative effort of specialists as well as various web users. The CYC project [16] is one of the oldest examples of a common-sense base constructed based on specialists. In the early years of the project, CYC already had 1.6 million rules and 180,000 concepts. The initial effort of knowledge acquisition was carried out by a group of specialists who were paid to perform this task. A problem with this approach is that the knowledge gained is dependent upon specialists. As an improvement, [39] propose a KA system by interactive dialogue.The methodology is similar to the method proposed here, in that the user, prior to including a concept, chooses a similar concept that belongs to the CYC base and, in an interactive manner, the user can accept or reject a set of assertions of the similar concept to be acquired for the new concept. For example, if the concept that the user wishes to include is "computer" and if there is the concept "notebook" in the CYC base, the user can select the latter and be guided through a process

of questions and answers aimed to acquiring common-sense facts for "computer" based on what is already known about "notebook". [39] does not present an assessment of this KA method.

In 2000, the Open Mind Common Sense (OMCS) project [34] was launched with the aim of collecting—from the Internet and from volunteer collaborators—sentences expressing facts of ordinary life. The OMCS corpus gave rise to the triples of common-sense knowledge in ConceptNet [12]. The new version of the OMCS [35] already provides functionalities that help the user to refine and validate the knowledge collected. Version 3.0 of this project is distinguished by expanding the project to other languages and by the expression of common-sense relations of a negative nature. For example, "dogs cannot fly." Currently, ConceptNet is the largest common-sense base, containing 35,854,766 relations and is currently in version 5.0, which is characterized by the combination of knowledge acquired from other bases and corpora such as Wikipedia.

Speer et al. [36] proposes the interactive game called 20 Questions with the dual purpose of motivating voluntary contributions to the OMCS project and increasing the rate of new knowledge acquisition. This game uses a hierarchical cluster model to define a set of 20 questions that will be used to motivate the user and to define a cluster of concepts. For example, for acquisition of the concept "apple", the game asks the following questions:

- Is it an example of a place? Answer: No
- Is it an example of food? A: Yes
- Can you find it in a store? A: Yes ...

Based on these responses, the clustering algorithm can define that the new concept "apple" belongs to the same cluster of concepts "cheese", "bread", "meat", etc. This method of KA was evaluated in two ways. The first evaluation consisted of a questionnaire for users of the game to compare the proposed method with the traditional way of including common-sense relationships in the OMCS project. For example, questions about how much more amusing the game is, and about how intuitive the game is. On average, 80 % of users evaluated that the 20 Questions game is more amusing than the traditional method. However, 56 % of users did not consider it intuitive. In the second evaluation, the authors measured the time it took to include a concept by using the game and by not using of the game. Users who used the game took 50 % less time than users who do not use the game. There was no assessment on the quality of the content acquired.

The Verbosity project [1] is also an interactive game for KA of common-sense knowledge. Just as [36], the main idea of Verbosity is to transform the process of common-sense KA into something amusing and interesting. It consists of a guessing game between a Narrator and a Guesser. The Narrator chooses a word and gives tips for the Guesser to discover the related concept. The tips are formulated by a template with a set of types of predetermined common-sense relations (contain, is a type of, is about, is the opposite of, is used for, is within, etc.). At the end of the process, if the Guesser is able to discover the concept that the Narrator is thinking of, the set of relations on the concept is acquired for a common-sense base. For example, the Narrator chooses the concept "computer" and formulates tips, such as "It contains a Keyboard." The Narrator keeps formulating other tips until the Guesser discovers the concept chosen by the Narrator. At the end of the process, the formulated and answered tips will be expressed as common-sense knowledge. In the example, the relation "computer contains keyboard" will be expressed in the knowledge base. The evaluation of this method concluded that the average number of inclusions was 29.58 common-sense relations, in an average usage time estimated at 23.58 min.

ReVerb [12] is a system for extracting open (non-domain specific) common-sense relationships that uses a set of syntactic and lexical constraints. In general, it uses regular expressions to recognize sentences and morphological modifications, such as converting verbs to the infinitive form. The lexical constraint is intended to discard sentences with poorly formed or complex relationships. For example, the sentence "The Obama administration is offering only modest targets for reducing greenhouse gases at the conference", ReVerb extracts the relation "X is offering only modest targets for reducing greenhouse gases at Y" with the arguments X = "Obama" and Y = "conference". This relationship does not meet the lexical constraints because the relationship is very specific. It also has a sorting algorithm to exclude possible meaningless or incomplete relationships, i.e., relationships that have no relevant information. This model is specific to the English language. To evaluate this system, 500 relations extracted by ReVerb from texts on the web were chosen at random, which were reviewed by two evaluators. As a result, 86 % of the relations extracted by ReVerb were corroborated by human evaluators.

In [9], the authors propose an automatic method to generate new triples of knowledge based on common-sense metarules. The proposed algorithm automatically searches an extended WordNet,[1] base for the concepts that have a given property, and generates new axioms using common-sense facts. As an example, we can cite the acquisition of new relations for the concept "glass". If "glass" has the property of "transparent", and "see through" is a characteristic of "transparent", then we can conclude that "see through" is also characteristic of "glass". The method was evaluated through human validation. About 50 axioms generated by the method were randomly chosen and the users were asked

[1] http://xwn.hlt.utdallas.edu/, accessed on February 10, 2011.

which seemed correct and which didn't make sense. Overall, we had a little more than 98 % accuracy for the proposed method.

For the Portuguese language, there are two important common-sense KA projects. The Open Mind Common Sense Brazil (OMCS-Br) project collects common-sense knowledge in Portuguese by collaborators on the web [3], such as the traditional KA strategy of the OMCS. Currently it has around 255,000 common-sense relationships. The InferenceNet Conceptual Base was initially translated from ConceptNet 2.0 by expert human translators and heuristics were applied to generate new common-sense and inferentialist knowledge relations [27]. It currently has 700,000 common-sense and inferentialist relations.

Our method is not intended to supplant the progress of these and other methods for acquiring common-sense knowledge. Instead, it is a complementary solution to leverage the process of KA. In this sense, the differential of the method proposed in this paper is the retrieval of similar content from the previous knowledge base, which facilitates more productive interactions for the acquisition of common-sense and inferentialist knowledge for new concepts. The interactions are more productive because the process helps the user to remember common-sense relations based on the content of related concepts. For example, for acquisition of the new concept *crime passional*, the algorithm proposes semantic relations retrieved from conceptual content of "***paixão***" ["passion"], namely: (eventPreRequisitOf, "*paixão*", "amante", Pre); (effectOf, "*paixão*", "*sofrimento*", Pos); (usedFor, "*paixão*", "*romance*", Pre), enriching process of KA of the new concept.

A comparative analysis with state of the art, presented here, allowed us to position our proposal in relation to the work of [12,36], which bear some resemblance to our proposal; all of them, in one way or another, use a base of previous relations and concepts as the baseline for KA. None of these presented an evaluation that would enable a comparison regarding the quality of knowledge acquired. Our process of KA uses heuristics based on the grammatical structures of the concepts, and thereby augments the possibility of related concepts that will serve as a baseline for KA that will elicit, for the user, ideas about common-sense relations. The projects Verbosity [1] and ReVerb [12] are different from the proposal of this work because they do not use a conceptual base to support the process of KA.

7 Conclusion

In this paper, we propose a semi-automated method for common-sense and inferentialist KA. The differential of the method in relation to the state of the art is an automatic reasoning process that generates new common-sense and pragmatic facts for concepts of the Portuguese language, based on content of other similar concepts and according to the grammatical structure of noun phrases. Moreover, the proposed method enabled more productive and richer interactions for KA because, with a baseline for validation, the user is prompted about common-sense semantic relations concerning the new concept. The interactive process with the end user allows a validation of the common sense semantic relations generated and, consequently, better quality in the acquisition of knowledge of this nature.

The method was implemented and evaluated for the common-sense and inferentialist base of the Portuguese language—InferenceNet—and obtained 72 % validation by human users. As future work, we can cite the further development of the algorithm with new heuristics for the generation of inferential content that contemplate other prepositions in the grammatical structures "<noun> <preposition> <noun>", for example, heuristics that consider the prepositions "in" and "with". In addition, we intend to explore other levels of the semantic network in order to discover and raise more and more implicit semantic relations to the user. The current version only explores the semantic relationships of the first level of a concept.

References

1. Von Ahn L, Kedia M, Blum M (2006) Verbosity: a game for collecting common-sense facts. In: Proceedings of the SIGCHI conference on Human Factors in computing systems, pp 75–78. ACM, New York
2. Anacleto J, de Souza Godoi M, de Carvalho A, Lieberman H (2007) A common sense-based on-line assistant for training employees. Human Computer Interaction-INTERACT 2007, pp 243–254
3. Anacleto J, Lieberman H, Tsutsumi M, Neris V, Carvalho A, Espinosa J, Godoi M, Zem-Mascarenhas S (2006) Can common sense uncover cultural differences in computer applications? Artificial intelligence in theory and practice, pp 1–10
4. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: a nucleus for a web of open data. The semantic web, pp 722–735
5. Baker C, Ellsworth M, Erk K (2007) SemEval'07 task 19: frame semantic structure extraction. In: Proceedings of the 4th international workshop on semantic evaluations. Association for computational linguistics, pp 99–104
6. Bick E (2000) The parsing system "Palavras": automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus University Press, Aarhus
7. Brandom R (2001) Articulating reasons: an introduction to inferentialism. Harvard University Press, Cambridge
8. Brill E (1995) Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput Linguist 21(4):543–565
9. Cankaya H, Moldovan D (2009) Method for extracting common-sense knowledge. In: Proceedings of the fifth international conference on knowledge capture. ACM, New York, pp 57–64
10. Che W, Li Z, Li Y, Guo Y, Qin B, Liu T (2009) Multilingual dependency-based syntactic and semantic parsing. In: Proceedings of the thirteenth conference on computational natural language learning: shared task. Association for Computational Linguistics, pp 49–54

11. Duran M, Aluísio S (2011) Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. In: 8th Brazilian symposium in information and human language technology, pp 164–168
12. Havasi C, Speer R, Alonso J (2007) Conceptnet: a lexical resource for common sense knowledge. In: Recent advances in natural language processing V: selected papers from RANLP, vol 309, p 269
13. Jurafsky D, Martin J, Kehler A, Vander Linden K, Ward N (2000) Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall, New Jersey, p 427
14. Kay M (2003) Introduction to computational linguistics. In: Mitkov R (ed) The Oxford handbook of computational linguistics, vol 30, no 1, pp 17–22
15. Lemle M (1984) Análise sintática: teoria geral e descriÇão do português, vol 106. Editora Atica
16. Lenat D (1995) CYC: a large-scale investment in knowledge infrastructure. Commun ACM 38(11):33–38
17. Marrafa P, Amaro R, Chaves R, Lourosa S, Martins C, Mendes S (2005) WordNet.Pt—Uma Rede Léxico-conceptual do português on-line. In: XXI Encontro da Associação Portuguesa de Linguística, Porto, Portugal
18. Maziero E, Pardo T, Di Felippo A, Dias-da Silva B (2008) A Base de Dados Lexical e a Interface Web do TeP 2.0: Thesaurus Eletrônico para o Português do Brasil. In: Companion proceedings of the XIV Brazilian symposium on multimedia and the web. ACM, New York, pp 390–392
19. Miller G (1995) WordNet: a lexical database for english. Commun ACM 38(11):39–41
20. Nivre J, Hall J, Nilsson J (2006) Maltparser: a data-driven parser-generator for dependency parsing. In: Proceedings of LREC, vol 6, pp 2216–2219
21. Oliveira H, Gomes P (2010) Onto.PT: automatic construction of a lexical ontology for Portuguese. University of Coimbra
22. Padró L, Stanilovsky E (2012) Freeling 3.0: towards wider multilinguality. In: Proceedings of language resources and evaluation (LREC). European Language Resources Association
23. Pardo T, Caseli H, Nunes M (2009) Mapeamento da Comunidade Brasileira de Processamento de Línguas Naturais. In: The proceedings of the 7th Brazilian symposium in information and human language technology-STIL, pp 1–21
24. Pinheiro V (2010) SIM: Um Modelo Semântico Inferencialista para Expressão e Raciocínio em Sistemas de Linguagem Natural. Phd Thesis, Universidade Federal do Ceará
25. Pinheiro V, Furtado V, Pequeno T, Nogueira D (2010) Natural language processing based on semantic inferentialism for extracting crime information from text. In: IEEE international conference on intelligence and security informatics (ISI). IEEE (2010), pp 19–24
26. Pinheiro V, Furtado V, Pequeno THC, Ferreira C (2012) Towards a common sense base in Portuguese for the linked open data cloud. In: PROPOR. Springer, Berlin
27. Pinheiro V, Pequeno T, Furtado V, Franco W (2010) InferenceNet.Br: expression of inferentialist semantic content of the Portuguese language. In: PROPOR, Lecture notes in computer science. Springer, Berlin, vol 6001, pp 90–99. http://dblp.uni-trier.de/db/conf/propor/propor2010.html#PinheiroPFF10
28. Pinheiro V, Pequeno T, Furtado V, Nogueira D (2009) Information extraction from text based on semantic inferentialism. In: Proceedings of the 8th international conference on flexible query answering systems, FQAS '09. Springer, Berlin, pp 333–344
29. Santos H, Furtado V (2012) SeMaps: enabling semantics on crowd maps. In: 21st Brazilian symposium on artificial intelligence—SBIA. Springer, Curitiba
30. Scarton C (2011) Verbnet.Br: ConstruÇão Semiautomática de um Léxico Computacional de Verbos para o Português do Brasil. In: The proceedings of the eighth Brazilian symposium in information and human language technology (STIL 2011), Cuiabá, Brazil
31. Schmid H (1995) Treetagger: a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, p 43
32. Dias da Silva B, Di Felippo A, Hasegawa R (2006) Methods and tools for encoding the WordNet.Br sentences, concept glosses, and conceptual-semantic relations. In: PROPOR, vol 3960, pp 120–130
33. Silva M, Koch I (1989) Lingüística aplicada ao português: sintaxe. Editora Cortez
34. Singh P, Lin T, Mueller ET, Lim G, Perkins T, Zhu WL (2002) Open mind common sense: knowledge acquisition from the general public. In: CoopIS/DOA/ODBASE, pp 1223–1237
35. Speer R (2007) Open Mind commons: an inquisitive approach to learning common sense. In: Proceedings of workshop on common sense and intelligent user interfaces
36. Speer R, Krishnamurthy J, Havasi C, Smith D, Lieberman H, Arnold K (2009) An interface for targeted collection of common sense knowledge using a mixture model. In: Proceedings of the 14th international conference on intelligent user interfaces. ACM, New York, pp 137–146
37. Suchanek F, Kasneci G, Weikum G (2008) Yago: a large ontology from wikipedia and wordnet. Web Seman Sci Serv Agents World Wide Web 6(3):203–217
38. Tsutsumi M, Anacleto J, Carvalho A, Lieberman H, Neris V (2010) Can common sense uncover cultural differences in computer applications? Int Fed Inform Process Digital Lib 217(1)
39. Witbrock M, Baxter D, Curtis J, Schneider D, Kahlert R, Miraglia P, Wagner P, Panton K, Matthews G, Vizedom A (2003) An interactive dialogue system for knowledge acquisition in cyc. In: Proceedings of the workshop on mixed-initiative intelligent systems, pp 138–145