Evaluating Techniques for Learning Non-Taxonomic Relationships of Ontologies from Text

Ivo Serra[1], Rosario Girardi[1], Paulo Novais[2]

[1] Computer Science Departament – Federal University of Maranhão (UFMA) - São Luiz, Ma – Brazil.

[2] Departament of Informatics – University of Minho - Braga – Portugal.

ivocserra@gmail.com, rosariogirardi@gmail.com, pjon@di.uminho.pt

Abstract. *Learning Non-Taxonomic Relationships is a sub-field of Ontology Learning that aims at automating the extraction of these relationships from text. Several techniques have been proposed based on Natural Language Processing and Machine Learning. However just like for other techniques for Ontology Learning, evaluating techniques for Learning Non-Taxonomic Relationships is an open problem. Three general proposals suggest that the learned ontologies can be evaluated in an executable application or by domain experts or even by a comparison with a predefined reference ontology. This article proposes two procedures to evaluate techniques for Learning Non-Taxonomic Relationships based on the comparison of the relationships obtained with those of a reference ontology. Also, these procedures are used in the evaluation of two state of the art techniques performing the extraction of relationships from two corpora in the domains of biology and Family Law.*

Keywords. *Learning non-taxonomic relationships; Ontology; Ontology learning; Natural language processing; Machine learning*

1. Introduction

Manual construction of ontologies by domain experts and knowledge engineers is a costly task, thus automatic and/or semi-automatic approaches for their development are needed. Ontology Learning (OL) (Buitelaar, Cimiano and Magnini 2006) (Cimiano, Volker and Studer 2006) (Girardi 2010) aims at identifying the constituent elements of an ontology, such as non-taxonomic relationships (Serra, Girardi and Novais 2012), from textual information sources.

Several techniques for learning non-taxonomic relationships have been proposed. Some of them use linguistic patterns (Girju, Badulescu and Moldovan 2003), while others use statistical solutions (Sanchez and Moreno 2008) (Serra, Girardi and Novais 2013) or even machine learning (ML) (Fader, Soderland and Etzioni 2011) (Maedche and Staab 2000) (Mohamed, Junior and Mitchell 2011) (Villaverde, Persson, Godoy and Amandi 2009). All of them compare their results with a reference ontology. However, there are few studies on the comparison of results from one technique to another and moreover, there is a lack of formalization of evaluation procedures.

According to Dellschaft and Staab (Dellschaft and Staab 2006) there are three ways to evaluate a learnt ontology: the resulting ontology can be evaluated in an executable application; by domain experts or even by comparing it with a predefined reference ontology (gold standard).

The use of an ontology in an executable application aims at measuring the effectiveness of a system that uses the ontologies being evaluated. A disadvantage of this approach is that other factors may impact the output of the system and sometimes the ontology is, in fact, a small part of the system with little interference in its results. The manual evaluation approach has its advantages, since it is expected that experts know the concepts and relationships of their domains of expertise, and therefore they are supposedly able to tell whether a given domain ontology is good or not. Disadvantages of these two proposals are their subjectivity and delay. Moreover, these methods are not feasible for large-scale evaluations. Thus, the comparison with a reference ontology is a plausible alternative since it permits the automation of the evaluation process. Proposals based on the comparison with reference ontologies are shown in Maedche and Staab (Maedche and Staab 2000) and Dellschaft and Staab (Dellschaft and Staab 2006). The main disadvantage of this approach is that a reference ontology is a handmade artifact and if it presents modeling problems, the evaluation method rewards ontologies with similar problems and penalizes ontologies with concepts or relationships that do not appear in the reference ontology.

This paper formally defines two procedures for evaluating techniques for Learning Non-Taxonomic Relationships of Ontologies (LNTRO) with respect to a reference ontology and uses them to comparatively evaluate two state of the art LNTRO techniques: Technique for Learning Non-taxonomic Relationships (TLN) (Serra, Girardi and Novais 2013) and Learning relationships based on the Extraction of Association Rules (LEAR) (Villaverde, Persson, Godoy and Amandi 2009), to extract relationships from the corpus Genia (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004) and Family Law doctrine (FindLaw 2013).

The paper is organized as follows: Section 2 introduces a general process for LNTRO (Serra, Girardi and Novais 2012). Section 3 presents a discussion about related work. In section 4, two procedures for the evaluation of LNTRO techniques according to the generic process for LNTRO are presented. In Section 5, two LNTRO techniques that are used to illustrate the application of the evaluation procedures are briefly described. In Section 6, the results of the application of the two evaluation procedures to perform benchmarking of these LNTRO techniques are presented and discussed. Section 7 presents the conclusions and points out future lines of research for this work.

2. A General process for Learning Non-Taxonomic Relationships of Ontologies

Based on the analysis of some techniques of the state of art (Fader, Soderland and Etzioni 2011) (Girju, Badulescu and Moldovan 2003) (Maedche and Staab 2000) (Mohamed, Junior and Mitchell 2011) (Sanchez and Moreno 2008) (Villaverde, Persson, Godoy and Amandi 2009) we have developed a generic process for LNTRO (figure 1) (Serra, Girardi and Novais 2012). The objectives were to have a guideline to suggest new LNTRO techniques and to facilitate comparative evaluations between techniques regarding the solutions they adopt for each one of its phases.

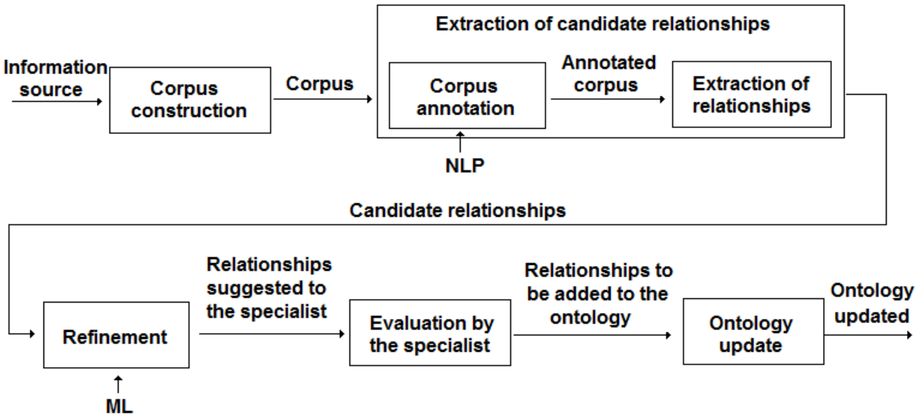

Information source
Corpus construction
Corpus
Extraction of candidate relationships
Corpus annotation
Annotated corpus
Extraction of relationships
NLP
Candidate relationships
Refinement
ML
Relationships suggested to the specialist
Evaluation by the specialist
Relationships to be added to the ontology
Ontology update
Ontology updated

Figure 1. A generic process for LNTRO.

The corpus construction task selects documents of the domain from which relationships can be extracted. This is usually a costly task and the outcome of any LNTRO technique depends on the quality of the used corpus.

The extraction of candidate relationships task identifies a set of possible relationships. It has the corpus built in the previous phase as input and candidate relationships as its product. It is composed of two sub-activities: corpus annotation and extraction of relationships. The corpus annotation task tags the corpus using Natural Language Processing (NLP) techniques that are necessary for the next steps of LNTRO. In the extraction of relationships activity, the annotated corpus is searched for evidence suggesting the existence of relationships. For example, Maedche and Staab (2000) consider the existence of two instances of ontology concepts in a sentence as evidence that they are non-taxonomically related. For Villaverde, Persson, Godoy and Amandi (2009), a relationship is identified by the presence of two ontology concepts in the same sentence with a verb between them.

The relationships obtained from the previous task should not be recommended to the specialist since there is usually a substantial amount of them that do not correspond to good suggestions. For this reason, in the refinement phase, machine learning (ML) techniques could be used to deliver the best suggestion to the specialist.

In the evaluation by the specialist task, he/she selects and possibly edits the relationships to be added to the ontology from those outputted from the previous phase. Finally, in the ontology update activity, the ontology is updated with the relationships that were chosen by the specialist.

One aspect of particular interest regarding LNTRO techniques is the type of representation adopted for the learned relationships. In the following we present some of the most common. The first is the one used by techniques that receive ontology concepts as input. There are two subtypes for this representation, depending if labels (typically verb phrases) are recommended. For the first subtype, the representation is $<c_1, vp, c_2>$ where c_1 and c_2 are ontology concepts and vp is a verb phrase. For example, considering the sentence "The court decree protects the property rights of the parties and provides support for the children" and "decree" and "property" as two ontology concepts, the relationship <*decree, protect, property*> would be extracted. Examples of techniques that use this representation are LEAR (Villaverde, Persson, Godoy and Amandi 2009) and TLN (Serra, Girardi and Novais 2012). For the second subtype, the representation is $<c_1, c_2>$, where c_1 and c_2 are two concepts. For example, considering "court" and "decree" as ontology concepts and the sentence "The court decree protects the property rights of the parties and provides support for the children", the relationship <*court, decree*> would be extracted. An example of a technique that use this representation is the LNTRO based on the extraction of generalized association rules (Maedche and Staab 2000).

The second type of representation is used when ontology concepts are not given as input to the LNTRO technique. In this case, noun phrases extracted from the corpus are used as ontology concepts. Here again there are two subtypes depending if labels are recommended. For the first subtype, the representation is $<np_1, vp, np_2>$ where: np_1 and np_2 are noun phrases and vp is a verb phrase. For example, from the sentence "The judge granted the custody of the child to his grandmother." the relationship <*the judge, granted, the custody*> would be extracted. Examples of techniques that use this representation are: LNTRO based on Web queries (Sanchez and Moreno 2008) and LNTRO based on logistic regression (Fader, Soderland and Etzioni 2011). The second subtype is $<np_1, np_2>$.

The procedures and evaluation measures (recall, precision and F-measure) used in the case studies presented in section 6 are suitable for use with LNTRO techniques that adopt relationships of the types $<c_1, c_2>$ or $<c_1, vp, c_2>$ and reference relationships defined for the same set of concepts, because the match of the concepts of the learned relationships and those of the reference ones is exact and therefore there is no need for more tolerant measures like the ones presented in section 3 (Maedech and Staab 2002) (Brewster, Alani, Dasmahapatra and Wilks 2004).

3. Related work

Various approaches for the evaluation of ontologies have been considered in the literature (Tartir, Arpinar and Sheth 2010). In this section we discuss some relevant ones and explain the motivations for our proposal.

Maedche and Staab (2002) propose several similarity measures for comparing different constituent elements of a learned ontology with a gold standard. Although the need to define a gold standard may be considered a drawback, an important positive aspect is that once it is defined the comparison of two ontologies can proceed entirely automatically. Maedech and Staab (2002) use a two level definition of ontology: the lexical level, which convey the terms that represent ontological structures and that is defined by a lexicon and the conceptual level formed by ontological structures like concepts and their relationships. In the following we discuss two evaluation approaches they propose for each of these layers.

The string matching (*SM*) is a measure that evaluate the similarity between two lexical entries. The *SM* returns a degree of similarity between 0 and 1, where 1 stands for perfect match and 0 for bad match. It considers the number of changes that must be made to transform one string into another (ed – edit distance) and weighs the number of these changes against the length of the shortest string of these two. For example, *SM*("TopHotel";"Top_Hotel") = 7/8. The *SM* between two lexical entries l_i and l_j is formally defined by the equation 1:

$$SM(l_i, l_j) := \max\left(0, \frac{\min(|l_i|, |l_j|) - \mathrm{ed}(l_i, l_j)}{\min\,(|l_i|, |l_j|)}\right) \in [0,1] \tag{1}$$

In order to provide a summarizing figure for the lexicons (L_1 and L_2) of two ontologies O_1 and O_2, Maedech and Staab (2002) also define the averaged String Matching $\overline{SM}(L_1; L_2)$ (equation 2):

$$\overline{SM}(L_1 + L_2) := \frac{1}{|L_1|} \sum_{l_i \in L_1} \overset{max}{l_j \in L_2} \; SM(l_i, l_j) \tag{2}$$

SM diminishes the influence of string pseudo-differences in different ontologies, such as use vs. not use of underscores or hyphens, use of singular vs. plural, or use of additional markup characters. However, sometimes *SM* may be deceptive, when two strings resemble each other though they there is no meaningful relationship between them, e.g. "power" and "tower". However, experimental results (Maedech and Staab 2002), show that *SM* may be very helpful for proposing good matches of strings.

Maedech and Staab (2002) also propose a similarity measure for non-taxonomic relationships (*RO*) formally specified by a pair of ontology concepts (c_i, c_j) corresponding the domain and range of the relationship. This measure verifies the similarity of relationships based on how similar their domain and range concepts are. The similarity between two concepts (the concept match *CM*) is computed by the upwards cotopy (*UC*) as defined in equation 3. $H(c_i, c_j)$ correspond to the ancestors of c_i and H is a taxonomy.

$$UC(c_i, H) := \{c_j \in C \mid H(c_i, c_j)\} \qquad (3)$$

Based on the definition of the upwards cotopy (*UC*) the concept match (*CM*) is then defined by the equation 4. F_1 and F_2 are functions that map the concepts of the correspondent ontologies O_1 and O_2 to their lexical entries in the ontology lexicons.

$$CM(c_1, O_1, c_2, O_2) := \frac{|F_1(UC(c_1, H_1)) \cap \ F_2\ (UC(c_2, H_2))|}{|F_1(UC(c_1, H_1)) \cup \ F_2\ (UC(c_2, H_2))|} \qquad (4)$$

Then, *RO* (equation 5) of relations r_1 and r_2 is defined by the geometric means of the similarity of their concepts. The geometric mean reflects the intuition that if either domain ($D(r_i)$) or range ($R(r_i)$) concepts fail to match, the matching accuracy converge to 0.

$$RO(r_1, O_1, r_2, O_2) := \sqrt{CM(D(r_1), O_1, D(r_2), O_2) * CM(R(r_1), O_1, R(r_2), O_1)} \qquad (5)$$

Differently from Maedech and Staab (2002), Brewster, Alani, Dasmahapatra and Wilks (2004) have proposed methods to evaluate the congruence of an ontology with a given corpus in order to determine how appropriate it is for the representation of the knowledge of the domain represented by the texts instead of a gold standard ontology. In general, the method performs automated term extraction on the corpus and count the number of terms that overlap between the ontology and the corpus. The ontology is penalized for terms present in the corpus and absent in ontology, and for terms present in the ontology but absent in the corpus. Another approach is to use a vector space representation of the terms in both the corpus and the ontology under evaluation. This permits an overall measure of the "fit" between the ontologies and the corpus. This approach has been tested in the evaluation of the similarity between a corpus and ontologies in the domain of art (Brewster, Alani, Dasmahapatra and Wilks 2004).

4. Procedures to Evaluate LNTRO Techniques

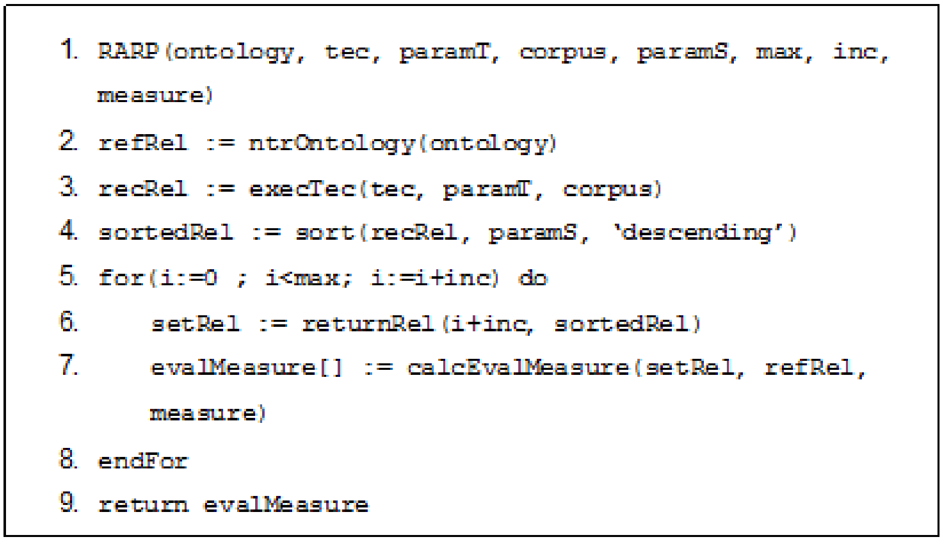
The evaluation approaches proposed in this paper differ from the previous mentioned (section 3) in the following aspects. First, we consider the scenario where we want to evaluate the result of LNTRO techniques that adopt relationships of the types <c_1, c_2> or <c_1, *vp*, c_2> against reference relationships defined for the same set of ontology concepts. In this case, measures like recall, precision and F-measure, that check the exact match, are adequate because instead of what happen for *SM* (Maedech and Staab 2002), the lexical representation of the concepts of the learned relationships coincide with those of the reference ones. Second, the *RO* approach (Maedech and Staab 2002), that evaluate the similarity between non-taxonomic relationships of ontologies take into account the hierarchical position of the domain and range concepts in the ontology taxonomy. This is not the case of the proportions of this paper, because the relationships learned by the LNTRO techniques being evaluated are compared against a set of non-taxonomic relationships of either types <c_1, c_2> or <c_1, *vp*, c_2>, without regard to the hierarchical relationships of their concepts. Third, differently of Brewster, Alani, Dasmahapatra and Wilks (2004) but in the same way of Maedech and Staab (2002) we evaluate the learned relations against reference ones. Defining reference relations is a job that is not more laborious then building a corpus of the domain.

In sections 4.1 and 4.2 the two procedures to evaluate LNTRO techniques, Recommendation of relationships with the Annulment of the Refinement Parameters (RARP) and Recommendation of relationships with the Maximization of the Evaluation Measure (RMEM) are presented. Both of them are based on the principle of the comparison with a reference ontology, as proposed by Dellschaft and Staab (Dellschaft and Staab 2006).

4.1. Procedure RARP

The aim of the procedure RARP (Recommendation of relationships with the Annulment of the Refinement Parameters) is to evaluate a LNTRO technique by comparing the reference relationships with a group of learnt relationships ordered by a pruning parameter of the refinement solution. This group is divided into subgroups with the same quantity of relationships. Then, for each of them considered cumulatively from the first, the evaluation measure is calculated. The pruning parameters of the solution to the refinement phase should be annulled. For example, in the case of the algorithm for the extraction of association rules (Srikant and Agrawal 1995), this corresponds to the adjustment of the values of minimum support and minimum confidence to zero, and in the case of the Bag of labels approach (section 5.1.4), it corresponds to setting to zero the value of the minimum frequency.

The evaluation procedure is formalized in Figure 2. In line one, the eight arguments: reference ontology ("ontology"), the technique being evaluated ("tec"), the values of the parameters with which the technique should be executed ("paramT"), the corpus from which the relationships are extracted ("corpus"), the parameter for sorting the refined relationships ("paramS"), the number of relationships to be considered in the result ("max"), the size of the subgroups of relationships for which the evaluation measure is calculated ("inc") and an evaluation measure ("measure") are informed to the procedure RAPR. In line two, "ntrOntology" receives as argument a reference ontology ("ontology") and returns its non-taxonomic relationships. These relationships, and those extracted by the LNTRO technique will be used to calculate the values of the evaluation measures. In line three, "execTec" takes as arguments the technique to be executed ("tec"), the values to its parameters ("paramT") and the corpus ("corpus") and returns the relationships recommended by this technique. In line four, "sort" returns the relationships recommended by the technique ("recRel") ordered in decreasing order ("descending") by the parameter of the refinement solution "paramS". In the next line there is a loop ranging from zero to a value below the number of relationships that should be considered for the calculations ("max"). The variable "inc" used in the increment, as well as "max", is informed by the user and corresponds to the size of the subgroups of relationships used to calculate the evaluation measure. For example, in the experiments conducted in section 6, "max" and "inc" received 100 and 5 respectively. In line six, the variable "setRel" receives the first "i + inc" non-taxonomic relationships recommended by the technique ("sortedRel"). In the next line, the vector "evalMeasure" receives, for each of its positions, the value for the evaluation measure ("measure") calculated for groups of relationships of size "inc" considered cumulatively from the first ("setRel"). To perform the calculation, the reference relationships obtained in step two ("refRel") are also informed. For example, if we consider max = 100 and inc = 5, then the position "0" of the vector contains the value of the evaluation measure calculated for the first five recommendations and the position "1" contains the value of the evaluation measure calculated for the first 10 recommendations. Finally, the vector "evalMeasure" containing the values for the evaluation measure is returned.

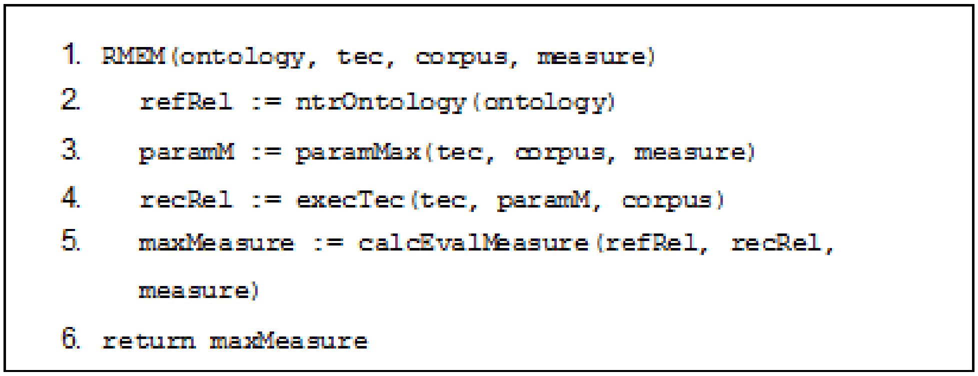


```
RARP(ontology, tec, paramT, corpus, paramS, max, inc,
measure)
refRel := ntrOntology(ontology)
recRel := execTec(tec, paramT, corpus)
sortedRel := sort(recRel, paramS, 'descending')
for(i:=0 ; i<max; i:=i+inc) do
    setRel := returnRel(i+inc, sortedRel)
    evalMeasure[] := calcEvalMeasure(setRel, refRel,
    measure)
endFor
return evalMeasure
```

Figure 2. The evaluation procedure RARP.

4.2. Procedure RMEM

The aim of the RMEM (Recommendation of relationships with the Maximization of the Evaluation Measure) procedure is to evaluate LNTRO techniques in terms of an evaluation measure to be maximized. Thus, the technique must be executed with a configuration that allows it to get the highest value for the evaluation measure considered. The evaluation procedure is formalized in the code of Figure 3.



```
1. RMEM(ontology, tec, corpus, measure)
2.    refRel := ntrOntology(ontology)
3.    paramM := paramMax(tec, corpus, measure)
4.    recRel := execTec(tec, paramM, corpus)
5.    maxMeasure := calcEvalMeasure(refRel, recRel,
      measure)
6. return maxMeasure
```

Figure 3. The evaluation procedure RMEM.

In the first line, the four arguments, the reference ontology ("ontology"), the technique to be evaluated ("tec"), the corpus from which relationships are extracted ("corpus") and the evaluation measure ("measure") are informed to the procedure RMEM. In line two, "ntrOntology" receives as argument the reference ontology ("ontology") and returns its non-taxonomic relationships, which are assigned to "refRel". In line three, "paramMax" takes as arguments the LNTRO technique ("tec"), the corpus from which relationships should be obtained ("corpus") and the evaluation measure to be maximized ("measure") and returns the values for the parameters of the technique ("tec") that maximize the value of the evaluation measure ("paramM"). In the fourth step, "execTec" execute the LNTRO technique ("tec") with the parameter values obtained in step three ("paramM") on the corpus informed as its third argument ("corpus") and returns the set of relationships recommended by the technique ("recRel"). In step five, the reference relationships obtained in step two ("refRel") and those recommended by the technique in step four ("recRel") are used to calculate the evaluation measure ("measure") informed as the third argument of the function "calcEvalMeasure". Finally, the maximized value of the evaluation measure ("maxMeasure") is returned.

4.3. Discussion

The RMEM approach allows the evaluation of LNTRO techniques based on their capacity to obtain the maximum value for an evaluation measure via the adjustment of their pruning parameters that is how specialists set the technique. However, it does not take into account the absolute number of valid non-taxonomic relationships recommended. For example, consider two LNTRO techniques *A* and *B* that obtained respectively 28% and 42,8% as the maximum values for recall, which indicate according to RMEM the superiority of *B*. However, *B* got 3 valid relationships in a set of 7 recommended relationships. Beside, *A* obtained a greater number of valid relationships (7 in 24 recommendations), which could be considered a more valuable result than that of the technique *B*.

The RARP approach performs the evaluation of LNTRO techniques considering the absolute number of valid relationships obtained. For example, consider a LNTRO technique *A* that obtained for the first 30 recommendations, considered in groups of 10 from the first, the cumulative values for precision: 50%, 35% and 30%. It means that 5, 2 and 2 relationships were present in each of the three groups of ten recommendations. Also, consider a technique *B* that obtained 40%, 30% and 30% for the same groups of recommendations. It means that 4, 2 and 4 relationships were present in each of the three groups of ten recommendations. In this case the technique *B* is considered superior, because despite not having obtained the highest value for the evaluation measure, it had an absolute number of valid relationships greater than that of the technique *A*. A drawback of RARP is that by canceling the pruning parameters, in the case of LNTRO techniques that have more than one parameter, some relationships within the observed group of recommendations that would be excluded will be not and then the value for the evaluation measure presented by the LNTRO technique can be different from that presented by

RARP. For example, consider a group of 30 recommended relationships ordered by the frequency of occurrence in the corpus. If the LNTRO technique also uses TF, IDF or TF-IDF (Salton and Buckley 1987) to prune the recommendations, and if it is not canceled as stated by RARP, some relationships within the list of the 30 recommendations could be excluded, resulting in a different value for the evaluation measure in relation to that calculated by RARP.

5. Evaluated LNTRO Techniques

To illustrate the application of RARP and RMEM, two LNTRO techniques presented in the next sections are used: Technique for Learning Non-taxonomic Relationships (TLN) (Serra, Girardi and Novais 2013) and Learning relationships based on the Extraction of Association Rules (LEAR) (Villaverde, Persson, Godoy and Amandi 2009). These techniques were chosen for the case studies (section 6) because, considering the same set of ontology concepts for both learned and reference relationships (type <c_1, *vp*, c_2>), they allow exact match between these two and therefore permit the use of the evaluation measures recall, precision and F-measure.

5.1. TLN

TLN (Serra, Girardi and Novais 2013) is a semi-automatic and parameterized LNTRO technique that uses NLP and statistical solutions to extract non-taxonomic relationships of predefined ontology concepts from an English corpus. The solutions adopted by TLN for each phase of the generic process of LNTRO (section 3) are summarized in Table 1 and described in sections 5.1.1 to 5.1.5.

Phases		TLN solutions
Corpus construction		Not approached
Extraction of candidate relationships	Corpus annotation	Chunk and lemmatization
	Extraction of relationships	Sentence rule (SR), sentence rule with verb phrase (SRVP) and apostrophe rule (AR)
Refinement		Frequency of co-occurrence and Bag of labels
Evaluation by the specialist		Manual selection and edition of the relationships
Ontology update		Execution of the procedure to update the ontology file in owl format with non-taxonomic relationships.

Table 1. TLN solutions for LNTRO.

5.1.1. Corpus construction

TLN does not define a specific solution to be adopted in this phase and the specialist is the one responsible for choosing the one that best suits the needs for that situation. Some helpful references are (Baroni and Bernardini 2004) (Fletcher 2004) (Sinclair 1989).

5.1.2. Corpus annotation

This phase aims at adding annotations to the corpus. These annotations are needed for the application of the extraction rules selected by the expert in the extraction of relationships phase. TLN applies five techniques of NLP executed in the order by which they are described in the following paragraphs.

Tokenization is a basic NLP task and its execution is a prerequisite for the application of any other NLP technique. Sentence splitter is necessary because the sentence is the linguistic unit from which non-taxonomic relationships are extracted by applying the rules selected in the extraction of relationships phase. Lemmatization is used to improve the recall of the search for ontology concepts in the corpus. For example, the match between the ontology concept "lawyer" and the term "lawyers" would not occur if the corpus was not lemmatized.

Morphological analysis classifies words in grammatical categories and is used in conjunction with verb phrase chunking to find verb phrases suggested as labels of the relationships. For example, the verb phrases "violates" and "can draw up" are labels for the relationship between the concepts "party" and "agreement" extracted from the following two sentences respectively: "If one party violates a settlement agreement the other may bring a lawsuit alleging a breach of contract" and "Although parties can draw up a separation agreement without the assistance of lawyers, it is often risky to do so". These two NLP techniques are executed only if the SRVP rule (section 5.1.3) is used in the extraction of candidate relationships.

5.1.3. Extraction of candidate relationships

In this phase, a set of extraction rules selected by the specialist are used to extract candidate relationships from the previously annotated corpus. TLN provides three types of extraction rules: the sentence rule (SR), the sentence rule with verb phrase (SRVP) and the apostrophe rule (AR). To illustrate their application, sentences from Genia (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004), a corpus in the domain of biology and the concepts "gene regulation", "morphogenesis", "amino acid" and "protein" from its corresponding ontology will be used.

The SR extraction rule is based on the intuition that two consecutive concepts in the same sentence are probably non-taxonomically related. Considering the c_1 and c_2 ontology concepts, this rule can be formalized in the regular expression in PCRE (Perl Compatible Regular Expressions): *(Lc)* $(?:(?!L_c|'s?).)*(?=(L_c))$ The L_c parameter is a string that corresponds to the concepts of the ontology on which we want to learn the non-taxonomic relationships separated by the disjunction operator of PCRE. The parameter L_c can be formally defined as the following concatenation: $L_c = c_1$ "|" c_2 "|" ... "c_n; $\forall c_i \in C$. The sub expressions (L_c) and $(?=(L_c))$ perform the match and extraction of the ontology concepts respectively in the left and right ends of the text that matches the entire regular expression. The sub expression $(?:(?!L_c|'s?).)$ checks if there is not a concept or one of the strings *'s* or *'* between the two concepts returned by (L_c) and *(?=(Lc))*. For each match for the entire regular expression in the corpus, a candidate relationship $<c_1, c_2>$ is generated. Considering the sentence, "Among the most important cellular processes, gene regulation controls morphogenesis and the versatility and adaptability of most living organisms", the tuple <*gene regulation*, *morphogenesis*> would be extracted.

The SRVP extraction rule considers that two consecutive concepts in the same sentence with a verb phrase (*vp*) between them are probably non-taxonomically related and can be formalized in the regular expression in PCRE: (L_c) $(?:(?!L_c|'s?).)*(V)(?:(?!L_c's?).)*(?=(L_c))$. The parameter V can be formally defined by the following concatenation: $V = vp_1$ "|" vp_2 "|" ... "|" vp_n; $\forall v_i \in F$. F correspond to the set of verbal phrases present in the document being considered. For each match of the entire regular expression in the corpus, a candidate relationship $<c_1, vp, c_2>$ is generated. This rule tends to provide lower recall than the Sentence Rule (SR) because, for example, it cannot extract the tuple <*amino acid, protein*> from the sentence "The DNA sequence of a gene encodes the amino acid sequence of a protein.", which corresponds to a valid relationship in this domain. However, SRVP tends to offer higher precision than the SR.

The AR extraction rule is based on the intuition that two consecutive concepts with either strings "’s" or "’" between them have a high probability of being non-taxonomically related. The apostrophe rule can be formalized with the regular expression in PCRE: $(L_c)'s?(L_c)$. The sub expression *'s?* checks if between the two concepts returned by (L_c) there is one of the strings *'s* or *'*. For each match of the entire regular expression in the corpus, a candidate relationship $<c_1, c_2>$ is generated.

5.1.4. Refinement

TLN provides two statistical solutions for this phase: *Frequency of co-occurrence* and *Bag of labels*. The general idea of *Bag of labels* is to calculate the frequency of pairs of ontology concepts (<c_1, c_2>), independently of their order, and to store the corresponding verb phrases in its bag of labels. The result is presented to the specialist that chooses the most appropriate verbal label for that relationship. This solution is used to filter the relationships extracted with the SRVP extraction rule. The *Frequency of co-occurrence* calculates the frequency of pairs of ontology concepts (<c_1, c_2>), independently of their order. This solution is used to filter the relationships extracted with AR or SR rules. For both solutions, the specialist can experimentally set the pruning parameter minimum frequency.

5.1.5. Evaluation by the Specialist and ontology update

No technique of NLP, ML or Statistics is capable of replacing the expert decision in an environment of ambiguous nature, as is the learning from natural language sources. Therefore, the goal of this phase is to make the best possible suggestions to the user and give him/her the control over the final decision. Thus, the result of the technique should be evaluated by a specialist before the relationships can be definitely added to the ontology. Issues such as the scope of the knowledge to be represented, the level of generalization, the real need of adding a relationship, its direction and label must ultimately be evaluated, selected, and possibly adjusted by an expert. Then, a procedure to update the owl file of the ontology with these non-taxonomic relationships is executed.

5.2. LNTRO based on the extraction of association rules

LNTRO based on the Extraction of Association Rules (LEAR) (Villaverde, Persson, Godoy and Amandi 2009) has two phases: "Identification of occurrences of relationships" and "Mining associations". The first phase receives a corpus and a set of concepts of an ontology as input and outputs a set of tuples of the type <c_1, *vp*, c_2>. Initially, to increase the recall of the search, each ontology concept is extended with its synonyms using Wordnet (Fellbaum 1998). Then, in order to identify the verbs, Chunk is performed. For sentences that have exactly two concepts and a verb phrase between them, a tuple <c_1, *vp*, c_2> is generated. For example, for the sentence "The court judged the custody in three days", the tuple <*court, judge, custody*> is generated.

Once a set of tuples outputted from the previous phase (candidate relationships) is obtained, the "Mining associations" task can be performed aiming at refining the results of the previous phase before suggesting relationships to the specialist. For this purpose, an algorithm for mining association rules

(Srikant and Agrawal 1995) is used. This algorithm extracts rules of the form $X \rightarrow Y$, which means that the occurrence of X implies the occurrence of Y. A typical application is to extract from a database of sales transactions, rules representing the purchasing behavior of customers. For example, <*coffee, bread*> → <*butter*> means that who purchases coffee and bread generally purchases butter.

In the context of the present LNTRO technique, the extracted rules have the form ($<c_1, c_2> \rightarrow vp$), where $<c_1, c_2>$ denotes two concepts and *vp* the associating verb phrase. Two thresholds are used to prune the rules: minimum support and minimum confidence. Support is the percentage of transactions containing all items that appear in the rule and is given by the formula: Support($<c_1, c_2> \rightarrow vp$) = $|\{\mathrm{t} \in \mathrm{T}| \{c_1, c_2, vp\} \subseteq \mathrm{t}\}| / |\mathrm{T}|$, where "T" correspond to the set of transactions in the form $<c_1, vp, c_2>$ from which the rules are extracted. Confidence measures how one can trust the rule and is given by the formula: confidence($<c_1, c_2> \rightarrow vp$) = support($<c_1, c_2> \rightarrow vp$) / support($<c_1, c_2>$).

The product of this phase are non-taxonomic relationships represented by association rules in the form $<c_1, c_2> \rightarrow vp$, having values of support and confidence greater than the minimum defined experimentally by the specialist. For example, in the sentence "Among the most important cellular processes, gene regulation controls morphogenesis and the versatility and adaptability of most living organisms", "gene regulation" and "morphogenesis" are concepts and "controls" is a verb phrase. In the first phase, the tuple <*gene regulation, control, morphogenesis*> is generated representing the fact that the extraction condition described previously was satisfied. In the second phase, if the rule <*gene regulation, morphogenesis*> → *control* has values of support and confidence greater than or equal to the minimum support and confidence, it is recommended to the specialist.

Table 2 shows which solutions have been adopted for each one of the generic phases for LNTRO as defined in section 3.

Phases		**LEAR solutions**
Corpus construction		Not approached
Extraction of candidate relationships	Corpus annotation	Chunk and lemmatization
	Extraction of relationships	Extract from sentences candidate relationships in the form of tuples ($<c_1, vp, c_2>$)

Phases	LEAR solutions
Refinement	Uses an algorithm for the extraction of association rules to suggest non-taxonomic relationships in the form of rules (<c_1, c_2> → *vp*)
Evaluation by the specialist	Not approached
Ontology update	Not approached

Table 2. LEAR solutions for LNTRO.

6. Applying RARP and RMEN in the evaluation of LNTRO techniques

To illustrate the application of the evaluation procedures RARP and RMEM, four experiments were conducted (sections 6.1 to 6.4). They consisted in applying RARP and RMEM to comparatively evaluate the LNTRO techniques TLN (Serra, Girardi and Novais 2013) and LEAR (Villaverde, Persson, Godoy and Amandi 2009) in the extraction of non-taxonomic relationships.

The Genia (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004) and Family Law doctrine (FindLaw 2013) corpora and corresponding ontologies were used as input. The corpus Genia (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004) has 2000 documents and 18545 sentences, whereas the 38 non-taxonomic relationships extracted from the ontology were used as reference to calculate the evaluation measures recall, precision and F-measure. The corpus Family Law doctrine (FindLaw 2013) describe a set of rules that regulate the marriage, its validity, effects, dissolution and the relationships between parents and children, among others. It is composed of 926 documents and 8.334 sentences. The corresponding ontology represents the knowledge of this domain and its 42 non-taxonomic relationships were used as reference for the calculation of the evaluation measures.

6.1. Using RARP to Evaluate TLN and LEAR with the corpus Genia

TLN (Serra, Girardi and Novais 2013) was configured with SRVP for the phase "Extraction of Relationships" and Bag of labels for the phase "Refinement" with zero to the minimum frequency whereas LEAR (Villaverde, Persson, Godoy and Amandi 2009) was configured with zero for both minimum support and minimum confidence. We considered that for LEAR a match between a relationship recommended by the technique and a reference one occurred whenever the three elements <c_1, *vp*, c_2> of a reference relationship coincided with the corresponding elements of a relationship recommended by the technique in the form of an association rule (<c_1, c_2> → *vp*). For example, the relationship <*cell, virus*> → *host* recommended by LEAR matches the reference relationship <*cell, host, virus*>. For TLN, a match occurred whenever a pair of concepts and their corresponding verb phrase, in the case of a reference relationship, coincided respectively with a pair of concepts recommended by TLN and a verb phrase in its bag of labels. Table 3 shows the number of valid relationships for each group of five recommendations and also the recall, precision and F-measure for these groups considered cumulatively from the first one. Figures 4, 5 and 6 show the recall, precision and F-measure graphs for both techniques corresponding to Table 3.

In the observed range of recommendations (first hundred), TLN obtained values for recall greater than or equal to those obtained by LEAR, being of 0,0842 (approximately 8,4%) the average difference. This result can be explained by the fact that the same amount of reference relationships identified by both techniques (37) were distributed over a larger number of relationships recommended by LEAR (524) than that recommended by TLN (134). The number of recommendations made by LEAR was 3,9 times greater than that of TLN. Generally the trend is that the values of recall for TLN when configured with SRVP in the phase "Extraction of relationships" are equal or greater than those presented by LEAR considering the same corpus and ontology concepts as input, since the set of tuples extracted by LEAR (candidate relationships) will be greater than or equal to that of TLN.

	TLN				**LEAR**			
Nº of recommen-dations	**Nº of valid relations**	**Recall**	**Precision**	**F-measure**	**F-measure**	**Precision**	**Recall**	**No. of valid relations**
5	3	0,0789	0,6000	0,1395	0,0465	0,2000	0,0263	1
10	3	0,1578	0,6000	0,2500	0,1250	0,3000	0,0789	2
15	2	0,2105	0,5333	0,3018	0,1886	0,3333	0,1315	2
20	1	0,2368	0,4500	0,3103	0,2758	0,4000	0,2105	3
25	1	0,2631	0,4000	0,3174	0,3174	0,4000	0,2631	2
30	3	0,3421	0,4333	0,3823	0,3529	0,4000	0,3157	2
35	1	0,3684	0,4000	0,3835	0,3561	0,3714	0,3421	1
40	1	0,3948	0,3760	0,3852	0,3589	0,3500	0,3684	1
45	1	0,4370	0,3655	0,4180	0,3373	0,3111	0,3684	0
50	1	0,4473	0,3400	0,3863	0,3409	0,3000	0,3947	1
55	1	0,4736	0,3272	0,3870	0,3440	0,2909	0,4210	1
60	0	0,4736	0,3000	0,3673	0,3469	0,2833	0,4473	1
65	2	0,5263	0,3076	0,3883	0,3300	0,2615	0,4473	0
70	1	0,5526	0,3000	0,3888	0,3148	0,2428	0,4473	0
75	0	0,5526	0,2800	0,3716	0,3008	0,2266	0,4473	0
80	2	0,6052	0,2875	0,3898	0,3050	0,2250	0,4736	1
85	1	0,6315	0,2823	0,3902	0,2926	0,2117	0,4736	0
90	1	0,6578	0,2777	0,3906	0,2968	0,2111	0,5000	1
95	2	0,7105	0,2842	0,4060	0,2857	0,2000	0,5000	0
100	2	0,7630	0,2800	0,4096	0,2898	0,2000	0,5263	1

Table 3. Recall of TLN and LEAR for the first 100 recommendations from the corpus Genia.

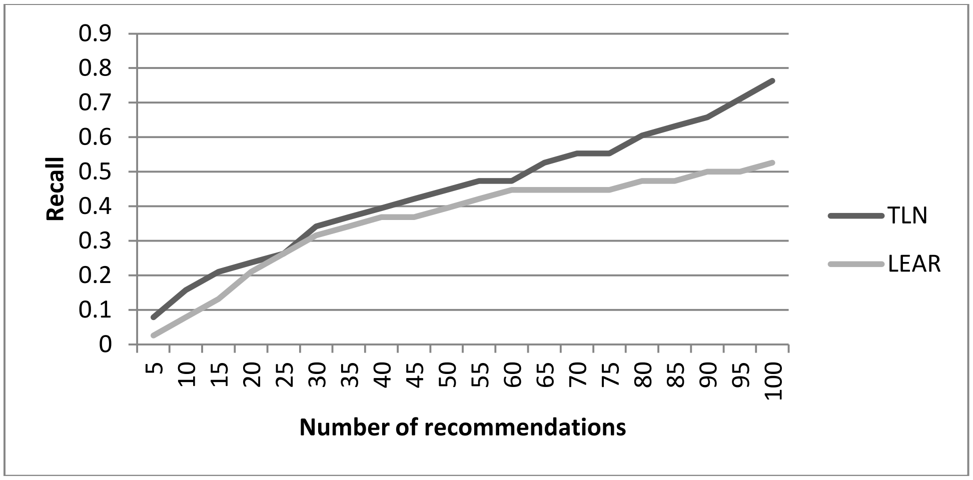


Figure 4. Recall graph of TLN and LEAR for the 100 recommendations from the corpus Genia.

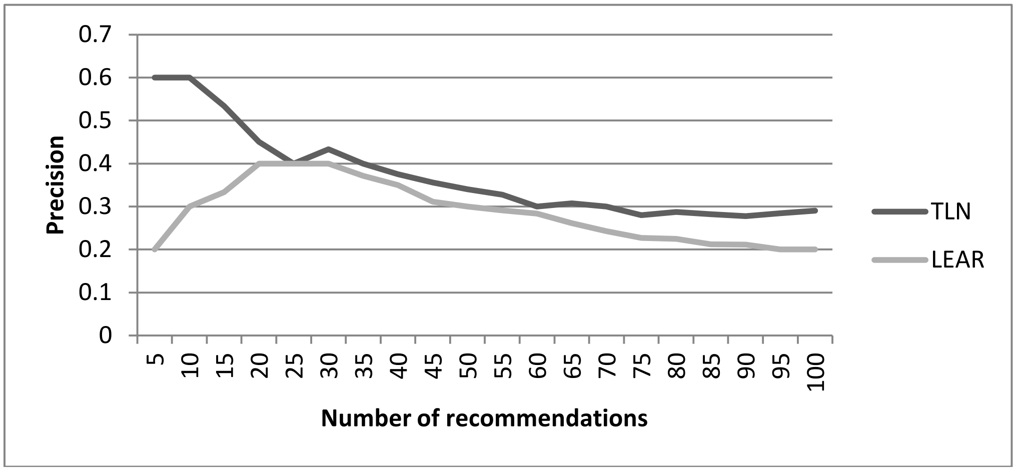


Figure 5. Precision graph of TLN and LEAR for the 100 recommendations from the corpus Genia.

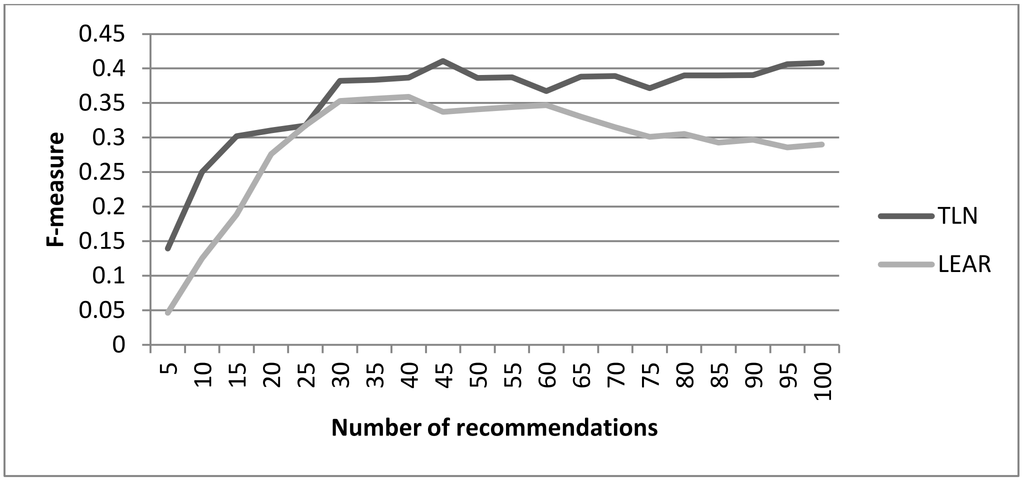


Figure 6. F-measure graph of TLN and LEAR for the 100 recommendations from the corpus Genia.

The difference between the recall values obtained by TLN and LEAR increases with the growing in the amount of relationships recommended. The difference in the group of the 5 first recommendations was 0,0526; in the group of the 10 first was 0,0789 and in the groups of 80 and 100 first were 0,1315 and 0,2368 respectively. This means that LEAR was gradually identifying less valid relationships than TLN with the growing number of recommendations and that TLN was more effective in performing the separation between true and false non-taxonomic relationships than LEAR.

With respect to the precision, just like for the recall, for the first hundred recommendations, TLN had values equal or greater than those obtained by LEAR, being of 0,0852 (approximately 8,5%) the average difference. This observation can be explained by the fact that the same amount of reference relationships identified by both techniques (37) are distributed over a larger number of relationships recommended by LEAR (524) than those recommended by TLN (134).

Generally, the trend is that the values of precision of TLN when configured with any of its extraction rules in the phase "Extraction of relationships" are greater than those presented by LEAR for the same observed range of recommendations, considering the same corpus and ontology concepts as input. This occurs because the reference relationships tend to be more dispersed in LEAR than in TLN.

TLN achieved the highest precision in the 10 first recommendations being it equal to 0,6. For the same interval, LEAR obtained lower values, since it corresponded to the range of recommendations with higher confidence values. In this range, reference relationships tend to be more rare. In the first 20, 25 and 30 recommendations, LEAR obtained 0,4; its highest value of precision.

From the first 30 recommendations, both TLN and LEAR show a downward trend in precision. However, LEAR presents a greater decrease. The difference between the values of precision between TLN and LEAR in the first 30, 70 and 90 recommendations were 0,0333; 0,0572 and 0,0666. This observation suggests that in the case of TLN the reference relationships are more concentrated in the beginning of its recommendations (5 to 30 first). Moreover, the loss of precision of TLN, with the increase in the number of recommendations, tends to be smaller, since the remaining reference relationships are distributed over a smaller number of candidate relationships. With respect to the F-measure, TLN had values equal or greater than those obtained by LEAR, being of 0,0678 (approximately 6,7%) the average difference. Finally, considering the evaluation procedure adopted (RARP) and the values obtained for the evaluation measures, we consider that TLN was more effective than LEAR in learning non-taxonomic relationships from the Genia corpus (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004) under the conditions described in this experiment.

6.2. Using RMEM to Evaluate TLN and LEAR with the corpus Genia

In the experiment conducted in section 6.1 the evaluation measures were calculated for groups of five recommendations considered cumulatively for the first hundred recommendations of the evaluated techniques when the pruning parameters of the refinement solutions were annulled. To allow working with values that do not annul the pruning parameters we developed the evaluation procedure RMEM (Recommendation of relationships with Maximization of the Evaluation Measure).

This experiment uses RMEM to comparatively evaluate the effectiveness of TLN (Serra, Girardi and Novais 2013) and LEAR (Villaverde, Persson, Godoy and Amandi 2009) on the extraction of relationships in the situation where we want to maximize a measure via adjusting their pruning parameters. The match between a relationship recommended by a technique and a reference one was considered as already described in section 6.1. To maximize the recall, TLN was configured with SRVP for the phase of "Extraction of Relationships" and Bag of labels for the phase of "Refinement", with 0,0146 for the minimum frequency, whereas LEAR was set with 0,0019 for the minimum support and 0,6667 for the minimum confidence. To maximize the precision, TLN was configured with 0,0146 for the minimum frequency and LEAR with 0,0019 and 0,6667 for the minimum support and minimum confidence respectively. With respect to F-measure, TLN was set with 0,0097 for the minimum frequency and LEAR with 0,0019 and 0,4286 for the minimum support and minimum confidence respectively. The parameter values for the maximization of the evaluation measures were found manually by inspecting the lists of non-taxonomic relationships recommended by the LNTRO techniques when their pruning parameters were annulled. Table 4 presents the maximum values of recall, precision and F-measure for TLN and LEAR and their corresponding numbers of recommended and valid relationships.

	TLN			LEAR			A – B (%)
	Nº of recommendations	Nº of valid relations	maximum value (A)	Nº of recommendations	Nº of valid relations	maximum value (B)	
Recall	134	37	0,9736	524	37	0,9736	0%
Precision	13	8	0,6153	26	11	0,4230	19,2%
F-mesure	41	18	0,4390	32	12	0,3750	6,4%

Table 4. Maximum recall, precision and F-measure for TLN and LEAR from the corpus Genia.

TLN and LEAR obtained the same value for the maximum recall. This fact was expected since, despite being spread over a larger number of recommendations in the case of LEAR, the same reference relationships are all present in the recommendations made by both techniques. Generalizing, when TLN is configured with SR it obtains the same recall than LEAR, when configured with SRVP, it obtains a recall equal or greater than LEAR and presents a recall lower or equal when configured with AR.

For the maximum precision, TLN obtained approximately 0,61; a value 0,19 higher than that obtained by LEAR which was approximately 0,42. It happened because TLN made a better separation between true and false relationships since it tends to concentrate them in the beginning of the list of recommendations. For the maximum F-measure, TLN obtained approximately 0,43; a value 0,06 higher than that obtained by LEAR which was approximately 0,37. Finally, considering the evaluation procedure RMEM and the values obtained for the evaluation measures, we consider that TLN was more effective in learning non-taxonomic relationships from the Genia corpus (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004).

6.3. Using RARP to evaluate TLN and LEAR with the corpus Family Law doctrine

In this experiment the evaluation procedure RARP was applied to LEAR and TLN configured with the sentence rule with verb phrase for the phase of "Extraction of relationships" and Bag of labels and "Extraction of association rules" for the refinement phase respectively.

The Family Law doctrine corpus and ontology (FindLaw 2013) were used as the source for the extraction of relationships and as the reference ontology respectively. It was considered that for LEAR a match between a relationship recommended by the technique and a reference one occurred whenever the three elements $<c_1, vp, c_2>$ of a reference non-taxonomic relationship coincided with the corresponding elements of a relationship recommended by the technique in the form of an association rule ($<c_1, c_2> \rightarrow vp$).

For TLN a match occurred whenever a pair of concepts and their corresponding verb phrase, in the case of a reference relationship, coincided with a pair of concepts recommended by TLN and a verb phrase in its corresponding bag of labels respectively. For example, the relationship <*court, divorce*> → *grant* recommended by LEAR matches the reference relationship <*court, grant, divorce*>. There was also the match of this reference relationship with a recommendation made by TLN, since the concepts "court" and "divorce" coincided and the verb phrase "grant" is present in their respective bag of labels.

Table 5 shows the number of valid relationships for each group of five recommendations and also the recall, precision and F-measure for these groups considered cumulatively from the first one.

Figures 7, 8 and 9 show the recall, precision and F-measure graphs for both techniques corresponding to Table 5.

In the observed range of recommendations, the first hundred, TLN obtained values for recall equal to or greater than those obtained by LEAR being of 0,2035 (approximately 20%) their average difference. This observation can be explained by the fact that the same amount of reference relationships, identified by both approaches (34), are spread over a larger number of recommendations in the case of LEAR (551) than in the case of TLN (108). The number of recommendations made by LEAR was 5,1 times higher than that of TLN.

	TLN				**LEAR**			
Nº of recommendations	**Nº of valid relations**	**Recall**	**Precision**	**F-measure**	**F-measure**	**Precision**	**Recall**	**Nº of valid relations**
5	2	0,0476	0,4000	0,0851	0,0426	0,2000	0,0238	1
10	3	0,1190	0,5000	0,1923	0,1154	0,3000	0,0714	2
15	3	0,1905	0,5333	0,2807	0,1404	0,2667	0,0952	1
20	2	0,2381	0,5000	0,3226	0,2258	0,3500	0,1667	3
25	3	0,3095	0,5200	0,3881	0,2687	0,3600	0,2143	2
30	3	0,3810	0,5333	0,4444	0,3056	0,3667	0,2619	2
35	2	0,4286	0,5143	0,4675	0,3117	0,3429	0,2857	1
40	2	0,4762	0,5000	0,4878	0,3171	0,3250	0,3095	1
45	1	0,5000	0,4667	0,4828	0,2989	0,2889	0,3095	0
50	2	0,5476	0,4600	0,5000	0,3043	0,2800	0,3333	1
55	1	0,5714	0,4364	0,4948	0,2887	0,2545	0,3333	0
60	0	0,5714	0,4000	0,4706	0,2941	0,2500	0,3571	1
65	1	0,5952	0,3846	0,4673	0,2804	0,2308	0,3571	0
70	2	0,6429	0,3857	0,4821	0,2679	0,2143	0,3571	0
75	0	0,6429	0,3600	0,4615	0,2564	0,2000	0,3571	0
80	1	0,6667	0,3500	0,4590	0,2623	0,2000	0,3810	1
85	1	0,6905	0,3412	0,4567	0,2520	0,1882	0,3810	0
90	1	0,7143	0,3333	0,4545	0,2576	0,1889	0,4048	1
95	2	0,7619	0,3368	0,4672	0,2482	0,1789	0,4048	0
100	1	0,7857	0,3300	0,4648	0,2394	0,1700	0,4048	0

Table 5. Recall of TLN and LEAR for the first 100 recommendations from the corpus Family Law doctrine.

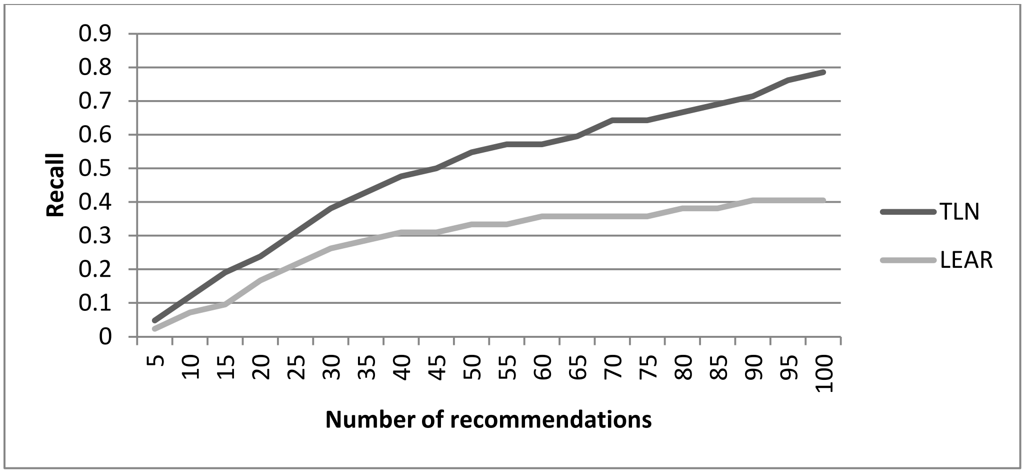


Figure 7. Recall graph of TLN and LEAR for the 100 recommendations from the corpus Family Law doctrine.

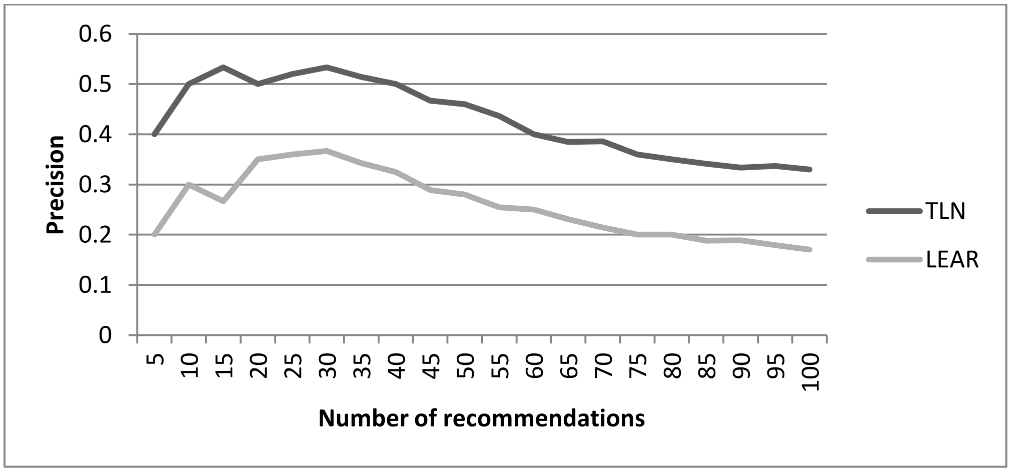


Figure 8. Precision graph of TLN and LEAR for the 100 recommendations from the corpus Family Law doctrine.

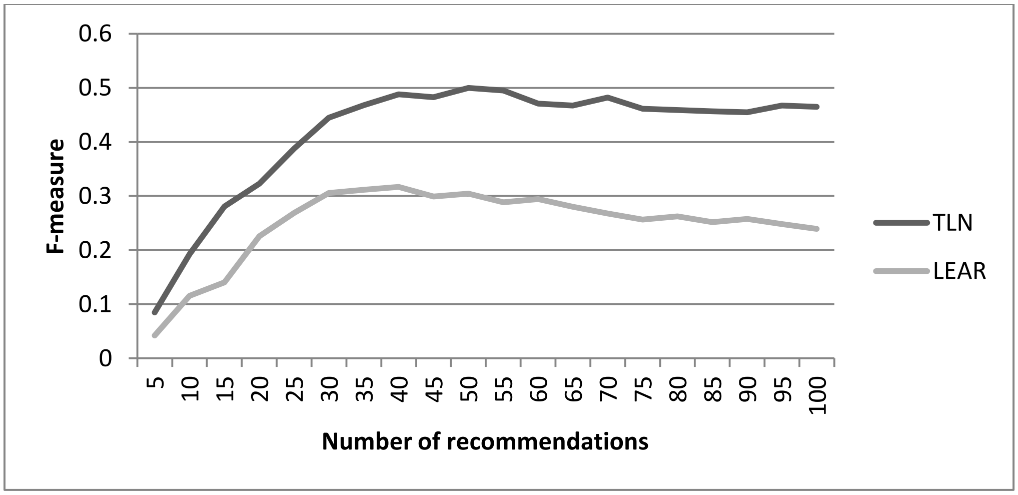


Figure 9. F-measure graph of TLN and LEAR for the 100 recommendations from the corpus Family Law doctrine.

The difference between the values for recall obtained by both approaches increased with the growth in the amount of relationships recommended. The difference in the group of the first five recommendations was of 0,0238, in the first group of 10 was of 0,0476 and in groups of 80 and 100 it was of 0,2857 and 0,3810 respectively. This means that LEAR was gradually identifying less valid relationships than TLN with the increase in the number of recommendations. This suggests that the algorithm Bag of labels of TLN was more effective in performing the separation between true and false non-taxonomic relationships and that consequently TLN is able to identify in a lower number of recommendations a greater number of relationships in relation to the algorithm of Extraction of association rules (Srikant and Agrawal 1995) of LEAR.

With regard to precision, in the same way as for the recall, TLN, throughout the range observed recommendations (the first hundred) had values equal or higher than those obtained by LEAR being of 0,1715 (approximately 17%) the average difference. This observation can be explained by the fact that the same amount of the reference relationships identified by both techniques (34) are distributed over a larger number of recommended relationships in the case of LEAR (551) then in the case of TLN (108). With regard to the F-measure, TLN had values equal or greater than those obtained by LEAR, being of 0,1676 (approximately 16%) the average difference.

Finally, considering the evaluation procedure adopted (RARP) and the values obtained for the evaluation measures, recall and precision, we consider that TLN was more effective than LEAR in learning relationships from the Family Law doctrine corpus (FindLaw 2013) under the conditions described in this experiment.

6.4. Using RMEM to evaluate TLN and LEAR with the corpus Family Law doctrine

This experiment uses the evaluation procedure RMEM to comparatively measure the effectiveness of TLN and LEAR. We consider that a match between a reference relationships and a recommendation made by the LNTRO techniques was as described in section 6.3.

To maximize the recall, both approaches, TLN and LEAR were configured with the same extraction rule (sentence rule with verb phrase) for the phase of "Extraction of Relationships" and had their pruning parameters annulled. To maximize the precision, TLN had its pruning parameter set to 0,0203 whereas LEAR was set with 0,0054 for the minimum support and 0,5000 for the minimum confidence. With respect to F-measure, TLN was set with 0,0122 for the minimum frequency and LEAR with 0,0036 and 0,4286 for the minimum support and minimum confidence respectively. Table 6 presents the maximum values of recall, precision and F-measure for both techniques and their corresponding numbers of recommended and valid relationships.

	TLN			LEAR			A – B (%)
	No. of recommendations	No. of valid relations	maximum value (A)	No. of recommendations	No. of valid relations	maximum value (B)	
Recall	108	34	0,8095	551	34	0,8095	0%
Precision	9	5	0,5555	13	5	0,3846	17%
F-measure	39	21	0,5384	33	11	0,3333	20,5%

Table 6. Maximum recall and precision for TLN and LEAR using RMEM and the corpus Family law doctrine.

TLN and LEAR obtained the same value for the maximum recall, since despite being spread over a larger number of recommendations in the case of LEAR, the same reference relationships are all present in the recommendations made by both approaches.

For the maximum precision, TLN obtained 0,5555; a value approximately 17% higher than that obtained by LEAR which was 0,3846. It happens because TLN made a better separation between true and false relationships. For the maximum F-measure, TLN obtained 0,5384; a value approximately 20,5% higher than that obtained by LEAR which was 0,3333. Finally, considering the evaluation procedure RMEM and the values obtained for the evaluation measures, we consider that TLN was more effective in learning non-taxonomic relationships from the corpus Family law doctrine (FindLaw 2013).

7. Concluding Remarks

The evaluation of LNTRO techniques is not a trivial task and despite its relevance to the area of LNTRO there is still little research in this direction. Three general propositions about how to conduct this task are (Dellschaft and Staab 2006): comparison of the ontology learnt with a reference one, manual evaluation of the ontology learnt by domain experts and its use as the knowledge base in a software system.

Two proposals based on the comparison with a reference ontology (gold standard) are the *SM* (String Matching) and *RO* (Relations of Ontology) (Maedech and Staab 2002). *SM* is a measure to evaluate the similarity between ontology lexicons that considers the number of changes that must be made to transform one string into another. The *RO* verify the similarity of non-taxonomic relationships based on how similar their domain and range concepts are. Despite their relevance, these proposals do not address the needs of this work for two reasons: First, we consider the scenario where we want to evaluate the effectiveness of LNTRO techniques by comparing the learned relationships against reference ones, despite of the position of the concepts in a taxonomy, what makes *RO* inappropriate. Second, because the concepts of the relationships learned by the considered LNTRO techniques (TLN and LEAR) have exact match with those of the reference relationships, once both are of the type $<c_1, vp, c_2>$, more tolerant measures like *SM* are not suitable.

This paper presented two procedures, based on the comparison of the ontology learnt with a reference one, for the evaluation of LNTRO techniques: Recommendation of relationships with the Annulment of the Refinement Parameters (RARP) and Recommendation of relationships with the Maximization of the Evaluation Measure (RMEM). The procedure RARP aims at evaluating a technique in terms of an evaluation measure calculated for a range of recommendations. This range is divided into subsets of fixed size and the values of the evaluation measure are calculated for these subgroups taken cumulatively starting from the first one. The aim of the procedure RMEM is to evaluate LNTRO techniques in terms of an evaluation measure to be maximized. Thus, the technique must be executed with the configuration that allows it to get the highest value for the evaluation measure considered.

The main positive aspect of RMEM approach is that it evaluates LNTRO techniques based on their capacity to obtain the maximum value for an evaluation measure via the setting of their refinement parameters that is in practice the way specialists have to prune the results of a technique (recommended relationships). Its main disadvantage is that it does not take into account the absolute number of valid non-taxonomic relationships learned.

The RARP approach has as its main positive aspect the fact that it verifies which LNTRO technique is able to recommend the maximum amount of non-taxonomic relationships, even if it has not obtained the highest value for the evaluation measure considered. This aspect is relevant because in the final analysis what the experts want is to get the greatest amount of relationships and not the highest value for an evaluation measure. Its main disadvantage is that by canceling the pruning parameters, in the case of LNTRO techniques that have more than one parameter, some relationships within the observed group of recommendations that would be excluded will be not and then the value for the evaluation measure presented by the LNTRO technique can be different from that presented by RARP.

In the case of the experiments of sections 6, there were no conflicting results of RARP and RMEM. Both procedures indicated the superiority of TLN against LEAR in learning non-taxonomic relationships from text, what can be formally verified by the values obtained for the evaluation measures recall, precision and F-measure for both LNTRO techniques evaluated. However, as discussed in section 4, there is no guarantee that this fact will happen for other experiments made with different LNTRO techniques, corpora and corresponding ontologies. In this case a more careful interpretation of the results is needed.

For the corpus Genia (Rinaldi, Schneider, Kaljurand, Dowdal, Andronis, Persidis and Konstanti 2004), the results obtained by RARP showed that for the first hundred recommendations, TLN (Serra, Girardi and Novais 2013) obtained 8,4%, 8,5% and 6,7% as the average difference for recall, precision and F-measure respectively. RMEM showed that both techniques obtained the same value for the maximum recall. For the maximum precision and F-measure, TLN obtained values 19,2% and 6,4% higher than those obtained by LEAR.

For the corpus Family Law doctrine (FindLaw 2013), the results obtained by RARP showed that for the first hundred recommendations, TLN obtained 20,3%, 17,1% and 16,7% as the average differences for the evaluation measures recall, precision and F-measure respectively. RMEM showed that both techniques obtained the same value for the maximum recall. For the maximum precision and F-measure, TLN obtained values 17% and 20,5% higher than that obtained by LEAR.

Although RARP and RMEM were useful to conduct the evaluation of TLN (Serra, Girardi and Novais 2013) and LEAR (Villaverde, Persson, Godoy and Amandi 2009), both procedures are just partially automated. Thus a work yet to be done is to fully implement these two procedures to completely automate the evaluation process, which is one of the main advantages of using a reference ontology. Also, we intend to use RARP and RMEM in the evaluation of LNTRO techniques that adopt noun phrases extracted from text as ontology concepts (types <np_1, np_2> and <np_1, vp, np_2>) like the LNTRO based on Web queries (Sanchez and Moreno 2008) and LNTRO based on logistic regression (Fader, Soderland and Etzioni 2011). However, since noun phrases do not correspond to names commonly used for ontology classes, the match of the learned relationships and the reference ones will generally not be exact and therefore more tolerant measures like *SM* (Maedech and Staab 2002) are needed.

Acknowledgments

This work is supported by CNPq, CAPES and FAPEMA, research funding agencies of the Brazilian government.

References

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC). Lisbon: ELRA, pp. 1313-1316.

Brewster, C., Harith, A., Dasmahapatra, S., & Yorick, W. (2004). Data Driven Ontology Evaluation. In, International Conference on Language Resources and Evaluation, Lisbon, Portugal, pp. 24 - 30.

Buitelaar, P., Cimiano, P., & Magnini, P. (2006). Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands.

Cimiano, P., Volker, J., & Studer R. (2006) Ontologies on Demand? – A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. In: Information, Wissenschaft und Praxis 57 (6-7), pp. 315-320.

Dellschaft, K., & Staab, S. (2006). On how to perform a gold standard based evaluation of ontology learning. In: International Semantic Web Conference, 5., Athens. Proceedings. Springer, 2006. pp. 228 - 241.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying Relations for Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge: MIT Press. pp. 23-24.

Find Law - for legal professionals. Resources and links for both state and federal laws. http://www.findlaw.com/casecode/. Accessed 16 April 2013.

Fletcher, W. H. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton (eds.) Corpus Linguistics in North America. Amsterdam: Rodopi. pp.191-205.

Girardi, R. (2010). Guiding Ontology Learning and Population by Knowledge System Goals. In: Proceedings of International Conference on Knowledge Engineering and Ontology Development, Ed. INSTIIC, Valence, pp. 480 – 484.

Girju, R., Badulescu, A., & Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, pp. 1-8.

Maedche, A., & Staab, S. (2002). Measuring Similarity between Ontologies. In: Proc. of the European Conference on Knowledge Acquisition and Management–EKAW. Madrid, Spain, LNCS/LNAI 2473. Springer. pp. 251–263.

Maedche, A., & Staab, S. (2000). Mining non-taxonomic conceptual relations from text. In Knowledge Engineering and Knowledge Management. Methods, Models and Tools: 12th International Conference. Proceedings. pp. 189-202.

Mohamed, T. P., Junior, E. R. H. & Mitchell, T. M. (2011). Discovering Relations between Noun Categories. In Proceedings of the conference on empirical methods in natural language processing (EMNLP 2011), Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1447-1455.

Rinaldi, F., Schneider, G., Kaljurand, K., Dowdal, J., Andronis, C., Persidis, A., & Konstanti, O. (2004). Mining relations in the GENIA corpus. Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, pp. 61 - 68.

Salton, G., & Buckley, C. (1987). Term Weighting Approaches in Automatic Text Retrieval. Cornell University.

Sanchez, D., & Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. Data and Knowledge Engineering, 64(3), pp. 600-623.

Serra, I., Girardi, R., & Novais, P. (2012). Reviewing the Problem of Learning Non-Taxonomic Relationships of Ontologies from Text, International Journal of Semantic Computing. Vol. 6-4, pp. 491-507.

Serra, I., Girardi, R., & Novais, P. (2013). PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text, In Proceedings of the 10th International Conference on Information Technology: New Generations.

Sinclair, J. (1989). Corpus creation. In Language, learning and community, eds. C Candlin and T McNamara, NCELTR Macquarie University, pp. 25-33.

Smith, B. (2003). Ontology. In: FLORIDI, L. (Ed.). The Blackwell guide to philosophy of computing and information. Malden: Blackwell,. pp. 155-166.

Srikant, R., & Agrawal, R. (1995). Mining generalized association rules. In Proceedings of the International Confeference on Very Large Databases (VLDB 19 95), pp. 407-419.

Tartir, S., Arpinar, I., Sheth, A. (2010). Ontological Evaluation and Validation. Theory and Applications of Ontology: Computer Applications, Springer, pp. 115-130.

Villaverde, J., Persson, A., Godoy, D., & Amandi, A. (2009). Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Expert Syst. Appl. 36(7), pp. 10288-10294.