Automatic semantic relation extraction from Portuguese texts

Leonardo Sameshima Taba, Helena de Medeiros Caseli
Federal University of São Carlos
Rod. Washington Luís Km. 235
leonardo_taba@dc.ufscar.br, helenacaseli@dc.ufscar.br

Abstract

Nowadays we are facing a growing demand for semantic knowledge in computational applications, particularly in Natural Language Processing (NLP). However, there aren't sufficient human resources to produce that knowledge at the same rate of its demand. Considering the Portuguese language, which has few resources in the semantic area, the situation is even more alarming. Aiming to solve that problem, this work investigates how some semantic relations can be automatically extracted from Portuguese texts. The two main approaches investigated here are based on (i) textual patterns and (ii) machine learning algorithms. Thus, this work investigates how and to which extent these two approaches can be applied to the automatic extraction of seven binary semantic relations (is-a, part-of, location-of, effect-of, property-of, made-of and used-for) in Portuguese texts. The results indicate that machine learning, in particular Support Vector Machines, is a promising technique for the task, although textual patterns presented better results for the used-for relation.

Keywords: Semantic relation extraction, Information extraction, Text mining

1. Introduction

The usage and importance of semantic information in Natural Language Processing (NLP) tasks is growing by the minute. However, the rate at which semantic information can be produced and analyzed by humans is much less than that which is needed by NLP applications. As one of the efforts that will hopefully help bridge that gap, this paper presents an automatic semantic relation extraction method using lexical-syntactic data.

Semantic relation extraction is the task of finding semantic relations between terms in texts. There's not a single formal definition for "semantic relation" and "term". Therefore, in this paper, "semantic relation" stands for any relation, explicit or implicit, between terms on a semantic level. A "term" is a contiguous sequence of tokens, which in turn are defined as any sequence of characters separated by spaces.

This work focuses on the Portuguese language, which still lacks high quality linguistic resources and tools, especially in the semantic level. Seven semantic relations are targeted: hyponymy (is-a), meronymy (part-of), locality (location-of), causality (effect-of), property-of (something has a certain property), made-of (something is made of some material) and used-for (something is used for a certain end). These relations are a subset of the ones used in the Open Mind Common Sense (OMCS) project[1] and were chosen motivated by the needs of the Brazilian branch of the OMCS project[2]. In order to extract these seven relations automatically a textual pattern strategy and two supervised machine learning algorithms, C4.5 decision trees (Quinlan, 1993) and Support Vector Machines (Vapnik, 1995), were evaluated.

2. Related work

There has been extensive work on the subject of semantic relation identification, mostly for the English language. The first researched approach was the textual patterns paradigm, pioneered by Hearst (1992). In her paper, Hearst describes six textual patterns that indicate the presence of a hyponymy relation between two noun phrases. She also proposed an algorithm to find patterns that imply a semantic relation R. Hearst applied her patterns on encyclopedic and journalistic corpora and found that 63% of the identified relations were of good quality.

Berland and Charniak (1999) follow Hearst's algorithm, but search for meronymy relations. Their results, obtained by applying the patterns on a 100 million words journalistic corpus, show that, on average, 55% of the relations found were correct. Girju and Moldovan (2002) also follow Hearst's algorithm, looking for causality relations on a journalistic corpus and reporting a 65% accuracy.

Freitas and Quental's (2007) work is one of the few that focuses on the Portuguese language. They adapted Hearst's patterns to Portuguese, creating 4 patterns that indicate hyponymy, and applied them to a corpus composed of around 2 million words of the public health domain. The results are compatible with Hearst's, showing that 73% of the relations found were of high quality.

Noticing the shortcomings of the textual patterns approach – namely, high precision but low recall – and encouraged by the increasing abundance of available textual data, researchers turned to machine learning (ML) techniques which leverage large quantities of text in order to try to find semantic relations.

The work of Girju et al. (2003) uses C4.5 decision trees (Quinlan, 1993) to extract part-whole relations from journalistic corpora. Using the same idea from Hearst's algorithm, some of the meronym pairs from WordNet were searched for in these corpora and some patterns that may indicate the part-whole relation were derived, such as "NP (noun phrase) of PP (prepositional phrase)", "NP's PP" and "NP verb PP". However, these patterns are very ambiguous as they can indicate relations other than meronymy. In order to solve that problem, Girju et al. (2003) propose learning semantic restrictions over the participants in the relation. The researchers reported 83% precision and 72% recall, re-

[1] http://openmind.media.mit.edu/
[2] http://www.sensocomum.ufscar.br/

sulting in a F-measure of 77.1%.

Snow et al. (2005) is based on Hearst's work, but uses logistic regression and naive Bayes classifiers to try to automatically find new patterns that indicate hyponymy. The resources used by them are a journalistic corpus processed by a dependency parser and WordNet (Fellbaum, 1998). Basically, Snow et al. transform the dependency paths between nouns into features. The classifiers were trained over a set of annotated dependence paths and the reported results, in terms of F-measure, were 34.80% for logistic regression and 31.75% for naive Bayes.

Zelenko et al. (2003) use Support Vector Machine (SVM) (Cortes and Vapnik, 1995) classifiers. One of the useful characteristics of SVMs is that they can use different *kernels*, which are functions that calculate the similarity between two objects. Zelenko et al. (2003) search for person-filiation and organization-place (a type of location-of) relations in a journalistic corpus. The authors used a parse tree kernel which was tested in two classifiers, an SVM and a voted perceptron. A training set of shallow parse trees was manually tagged with positive and negative instances. The best classifier was the SVM, which reported a F-measure of 86.80% for the person-filiation relation and 83.30% for the organization-place relation.

Among the state-of-the-art in semantic relation extraction with ML we can cite (Girju et al., 2010). Girju et al.'s approach uses features from different linguistic levels and different resources such as a tokenizer, POS tagger, parser, hand-crafted dictionaries, WordNet (Fellbaum, 1998) and others. Girju et al. (2010) focuses on the seven SemEval Task 4 (Girju et al., 2007) relations. Seven binary SVM classifiers were trained, one for each relation. The final average F-measure of the classifiers was 72.4%.

Based on these related works and others, this study investigates the automatic extraction of seven semantic relations in Portuguese texts through the use of textual patterns, decision trees and SVMs.

3. Resources and tools

Two corpora were used in this work: the first one is CETENFolha[3], a journalistic corpus composed of around 24 million words from articles of the Brazilian newspaper *Folha de São Paulo*. This corpus was morphologically tagged by the PALAVRAS parser (Bick, 2000). The second corpus used was composed of 646 articles (around 870,000 words) from *Pesquisa FAPESP*[4] (Aziz and Specia, 2011), a scientific divulgation magazine. This corpus was also morphologically tagged by the PALAVRAS parser.

In order to apply supervised learning methods to extract semantic relations from texts, a sample of both corpora was manually annotated with the terms of interest and relations between them. Roughly 3,800 sentences were initially annotated by two annotators; each one marked around 2,000 sentences, of which around 200 were annotated by both in order to evaluate the agreement rate between them. This rate was calculated over these 200 common sentences as the number of relations marked in the same way by both divided by the number of total distinct marked relations. The concordance rate in this first stage of annotation was 69.15%.[5] A subset of the FAPESP corpus was also annotated, this time by only one annotator, which marked around 500 sentences.

Table 1: Number of instances for each relation (annotated manually) and the negative class (generated automatically) in both corpora

Relation	CETENFolha	FAPESP
property-of	5114	853
is-a	2950	367
part-of	2105	306
location-of	1742	248
effect-of	169	84
used-for	138	68
made-of	82	60
none	112794	23985
Total	125094	25971

Table 1 shows the number of annotated training instances for each relation in each corpus. Additionally, a negative class "none" was created automatically, composed of all pairs of terms that didn't have any relation annotated between them.

Another resource used in the first experiment (with textual patterns) was the OMCS-Br common sense facts database, comprised of about 115 thousand instances distributed between the seven relations of interest of this work.

This work also used some computational tools, namely the parser PALAVRAS (Bick, 2000), the machine learning suite WEKA (Hall et al., 2009) and SVM Light (Joachims, 1998).

4. Textual patterns strategy

The textual pattern strategy is simpler than the ML approach, so it was the first to be investigated in this research. Particularly, the works of Hearst (1992) and Freitas and Quental (2007) were the main references in this work. As said in Sect. 2., (Hearst, 1992) was one of the first works to use patterns in order to find semantic relations (in that case, hyponymy). Moreover, Hearst also defined an iterative algorithm (Fig. 1) to discover new patterns that indicate a certain semantic relation. Freitas and Quental (2007), based on Hearst's work, translated her patterns to Portuguese.

4.1. Experiment 1

Using these works as base, the first experiment consisted in the application of the 4 hyponymy patterns defined in (Freitas and Quental, 2007), plus the manual construction of textual patterns for the 6 remaining semantic relations of interest. Hearst's algorithm was also applied in order to find new patterns for all 7 relations. In total, 17 patterns were

[3] http://www.linguateca.pt/cetenfolha/

[4] http://revistapesquisa.fapesp.br

[5] Since the terms of the training instances were not fixed, they can be different between human annotators. As a consequence, it was not possible to measure the inter-annotator agreement using the kappa coefficient (Carletta, 1996).

Figure 1: Hearst's algorithm (Hearst, 1992)

1. First, a semantic relation of interest is chosen (e.g. hyponymy, meronymy, etc.);
2. A list of pairs of terms for which it is known that the relation is valid is constructed (e.g. "Brazil-country" and "dog-animal" for hyponymy). That list can be obtained through the search of manually defined patterns or from pre-existing lexical or knowledge bases;
3. The corpus is then searched for sentences in which these terms occur close to each other and the context (words around the terms or the whole sentence) is stored (e.g. "*Brazil* is a developing *country*");
4. Next, these stored contexts are analyzed and common contexts are hypothesized as patterns that indicate the relation of interest;
5. When a pattern is defined, it is used to find more instances of the targeted relation. Return to step 2.

defined by hand (Table 2)[6]. The 115 thousand instances from the OMCS-Br database were used as seed instances in the application of the algorithm. 7 new patterns were found with the execution of one iteration of the algorithm (Table 3).

Table 2: Manually defined patterns for the 7 semantic relations plus Freitas and Quental's (Freitas and Quental, 2007) hyponymy patterns

Relation	#	Pattern
is-a	1	T1 (tais como|como) T2 {, T3}* (e|ou) TN
	2	T2 {, T3}* ,? (e|ou) outros T1
	3	tipos de T1: T2 {, T3}* (e|ou) TN
	4	T1 chamad(o|a|os|as) de? T2
property-of	1	T1_N T2_ADJ
	2	T2_ADJ T1_N
	3	T1_N " T2_ADJ "
part-of	1	T1 com T2
	2	T1 {verbo fazer} parte de T2
	3	T1 {verbo ser} parte de T2
made-of	1	T1_N de T2_N
	2	T1 (é|são)? feit(o|a|os|as) de T2
location-of	1	T1 entrou em T2
	2	T1 ,? localizad(a|o) em T2
effect-of	1	T2_V .* devido=a T1
	2	T2_V por=causa=de (a|o|as|os)? T1
used-for	1	T1 (que podem ser)? usadas? para T2_V

[6]In the patterns, shown as regular expressions, T1 represents the first term in the relation and T2 the second. Terms T3, T4, ..., TN, if present, are always related to T1, e.g. R(T1, T3), R(T1, T4) (where R stands for one of the semantic relations). The notations "_N", "_ADJ" and "_V" indicate that a term must be a noun, an adjective or a verb, respectively.

Table 3: Patterns defined after one iteration of Hearst's algorithm (Hearst, 1992)

Relation	#	Pattern
is-a	5	T2 {, T3}* ,? (e|ou) (qualquer|quaisquer) outro{s}? T1
	6	T2 é (o|a|um|uma) T1
	7	T2 são T1
property-of	4	de T1_ADJ T2_N
part-of		–
made-of		–
location-of	3	T1 chega a o T2
	4	T1 em (o|a|os|as) T2
effect-of		–
used-for	2	T1 para (o|a|os|as) T2_V

All 24 patterns (Tables 2 and 3) were then applied over the manually annotated sentences from CETENFolha. The relation instances found through these patterns were compared to the manually annotated relations. The results, presented in the next section, show that textual patterns have good precision but low recall.

The low recall in relation extraction using textual patterns and the high cost of manual analysis (of the corpus, for the construction of patterns, or of the contexts, after the application of Hearst's algorithm) favored the investigation of machine learning approaches.

5. Machine learning strategy

5.1. Classifiers

Two ML classifiers were investigated in this work, C4.5 decision trees (Quinlan, 1993) and Support Vector Machines (SVMs) (Vapnik, 1995). SVMs are binary classifiers but our task involves the discrimination between 7 different classes (plus one negative class) so the one-vs-all strategy was adopted.

5.2. Features

In this work, a training instance is defined as a pair of terms in a sentence. Therefore, given a pair of terms in a sentence, the goal of a classifier is to decide whether one of the seven relations of interest exists between the terms or if there is no relation. To make the training of decision trees and SVMs possible, these instances have to be featurized.

Thus, different features of the superficial, morphological and syntactic levels were defined in order to featurize the training data. Some examples of features are the distance between terms, number of commas between terms, morphological classes, among others. Their full description can be found in (Taba, 2013).

5.3. Experiment 2

Experiment 2 consisted in the training of a C4.5 decision tree and SVM classifiers on the annotated data from corpus CETENFolha (Table 1) and their evaluation through 10-fold cross-validation. The purporse of this experiment was to find out the effectiveness of using decision trees and SVMs to extract semantic relations from Portuguese texts.

The J48 algorithm, an open source Java implementation of the C4.5 decision tree, found in the Weka (Hall et al., 2009) machine learning software collection, was used to perform the decision tree experiments. The C parameter[7] used was 0.25, chosen after empirical tests done with varying C values in the range between 0.05 and 1, in intervals of 0.05.
The chosen SVM implementation was SVM Light[8] (Joachims, 1998), with a third degree polynomial kernel and parameter[9] $C = 0.01$. These parameters were empirically selected after tests were made with all the combinations of degrees 1 to 4 and C from 10^2 to 10^{-4} (in intervals of degrees of 10).

5.4. Experiment 3

Experiment 3 was conducted to verify the impact of each subset of features on the classifiers. Therefore, 7 subsets (Table 4) consisting of different combinations of superficial, morphological and syntactic features were used. Then, 7 decision tree and 7 SVM classifiers were trained on each of these subsets and then evaluated through 10-fold cross-validation.

Table 4: Subsets of features used in experiment 3

	Features		
#	Superficial	Morphological	Syntactic
1	X		
2		X	
3			X
4	X	X	
5	X		X
6		X	X
7	X	X	X

5.5. Experiment 4

The final experiment consisted in the training of a decision tree and an SVM classifiers with all the features and all annotated data from CETENFolha (Table 1). The trained classifiers were then tested over the annotated sentences of the FAPESP corpus. The purpose of this experiment was to verify whether the methods and features described in this paper are useful to finding relation instances in new data and also to evaluate the adequacy of the training corpus, which is journalistic, when confronted with a distinct genre corpus (FAPESP is of the scientific dissemination genre).
The results for each experiment are described in the next section.

6. Results and Discussion

6.1. Experiment 1

Experiment 1's results[10] are summarized in Table 5. As expected, the recall values were, in general, quite low, confirming what was said in Sect. 2.. This can be observed especially in relation is-a, that obtained a precision of 61% but only 1% of recall.
These results show that some relations are simpler to be extracted, such as the is-a and property-of relations, while others, like part-of and made-of, are harder. One of the reasons for the differences in difficulty between relations is the ambiguity and plasticity of natural languages. The ML approach, which involves deeper linguistic knowledge, attenuates that problem and yields better recall.

Table 5: Experiment 1 – Results of the application of all 24 textual patterns on the CETENFolha corpus in terms of precision, recall and F-measure.

Relation	**#**	**Precision**	**Recall**	**F-measure**
is-a	1	0,0%	0,0%	0,0%
	2	42,8%	0,1%	0,2%
	3	0,0%	0,0%	0,0%
	4	80,0%	0,1%	0,2%
	5	0,0%	0,0%	0,0%
	6	59,4%	0,7%	1,4%
	7	70,0%	0,2%	0,4%
Total		61,1%	1,2%	2,3%
property-of	1	50,0%	10,7%	17,6%
	2	61,9%	28,9%	39,4%
	3	100,0%	0,2%	0,4%
	4	50,8%	0,6%	1,2%
Total		57,5%	39,8%	47,0%
part-of	1	4,2%	0,1%	0,2%
	2	60,0%	0,1%	0,2%
	3	100,0%	0,1%	0,2%
Total		12,3%	0,3%	0,6%
location-of	1	83,3%	0,3%	0,6%
	2	100,0%	0,1%	0,2%
	3	100,0%	0,2%	0,4%
	4	7,3%	2,5%	3,7%
Total		9,0%	3,2%	4,7%
effect-of	1	71,4%	3,1%	5,9%
	2	66,7%	2,4%	4,6%
Total		69,2%	5,5%	10,2%
made-of	1	1,3%	28,4%	2,5%
	2	100,0%	1,3%	2,6%
Total		1,3%	29,7%	2,5%
used-for	1	100,0%	0,7%	1,4%
	2	42,7%	25,7%	32,1%
Total		42,3%	26,5%	32,6%
Average		36,5%	18,2%	24,3%

6.2. Experiment 2

The results of experiment 2 are shown in Tables 6 and 7. Experiment 2 shows a significant improvement on recall when compared to those of experiment 1. These results show the effectiveness of the use of ML methods in the automatic extraction of semantic relations.
However, it's important to note that the used-for relation had a higher F-measure using patterns (32.6%, against 8.0% for decision trees and 26.2% with SVMs), showing that, at least for that relation, the textual patterns approach yields better results. One possible explanation for that re-

[7] Confidence factor in the tree pruning. The lower, the more pruning is done.

[8] `http://svmlight.joachims.org`

[9] Compensation factor between training errors and the margin of the support vectors.

[10] The precision, recall and F-measure values for the total of a certain semantic relation are based on the sum of found and correct instances of all patterns of that relation

sult can be that the features used by the ML classifiers can't capture relevant information for the extraction of that relation. The low number of training examples annotated for that relation also aggravates the weak performance of the classifiers.

Table 6: Experiment 2 – Results of the evaluation of decision trees through *10-fold cross-validation* in terms of precision, recall and F-measure

Relation	Precision	Recall	F-measure
property-of	90.5%	80.0%	84.9%
is-a	76.9%	56.3%	65.0%
part-of	66.4%	37.2%	47.7%
location-of	62.2%	21.4%	31.8%
effect-of	34.0%	10.4%	16.0%
made-of	33.3%	2.8%	5.2%
used-for	17.5%	5.1%	8.0%
Average	54.4%	30.4%	39.0%

Table 7: Experiment 2 – Results of the evaluation of SVMs through *10-fold cross-validation* in terms of precision, recall and F-measure

Relation	Precision	Recall	F-measure
property-of	89.6%	81.6%	85.4%
is-a	78.2%	65.1%	71.0%
part-of	56.9%	41.4%	47.9%
location-of	51.8%	27.8%	36.2%
effect-of	45.5%	16.9%	24.6%
made-of	58.2%	24.3%	34.3%
used-for	50.8%	17.7%	26.2%
Average	61.6%	39.2%	47.9%

6.3. Experiment 3

The performance of the classifiers trained with each subset of features and evaluated through 10-fold cross-validation is shown in Table 8. From the values presented it is possible to note that the ML methods have better results when all features are used in the training of the classifiers (subset 7).

6.4. Experiment 4

The last experiment's results are presented on Tables 9 and 10. These results show that the algorithms and data used in the training had a good generalization capacity when confronted with a corpus of a different genre.

The graph on Figure 2 summarizes the results obtained by textual patterns (applied on the CETENFolha corpus) and the decistion tree and SVM classifiers (trained and tested on CETENFolha) for each one of the 7 semantic relations.

Table 11 shows some examples of correctly identified relation instances using both textual patterns and ML strategies.

6.5. Comparison with related works

Considering the is-a relation, we can compare this work with those of (Hearst, 1992) (for English) and (Freitas and Quental, 2007) (for Portuguese). Hearst obtained a precision of 63% and Freitas and Quental got 73.4%, while experiment 1 (Table 5) resulted in an average precision of 61.1%. The application of decision trees and SVM classifiers (Tables 6 and 7) resulted in 76.9% and 78.2% precisions, respectively. The difference between these results can be attributed to the different methods and corpora used in each work. Recall can't be compared as it wasn't calculated by the related works.

Table 9: Experiment 4 – Results of the decision tree classifier tested over the FAPESP corpus in terms of precision, recall and F-measure

Relation	Precision	Recall	F-measure
property-of	89.1%	83.3%	86.1%
is-a	60.0%	26.2%	36.4%
part-of	59.5%	35.0%	44.1%
location-of	39.1%	10.9%	17.0%
effect-of	40.0%	7.1%	12.1%
made-of	0.0%	0.0%	0.0%
used-for	38.5%	7.4%	12.3%
Average	23.3%	22.8%	23.0%

Table 10: Experiment 4 – Results of the SVM classifier tested over the FAPESP corpus in terms of precision, recall and F-measure

Relation	Precision	Recall	F-measure
property-of	91.0%	84.3%	87.5%
is-a	57.6%	32.1%	41.2%
part-of	52.1%	37.1%	43.3%
location-of	39.6%	17.7%	24.5%
effect-of	50.0%	8.3%	14.3%
made-of	36.4%	7.3%	12.1%
used-for	40.9%	13.2%	20.0%
Average	52.5%	28.6%	37.0%

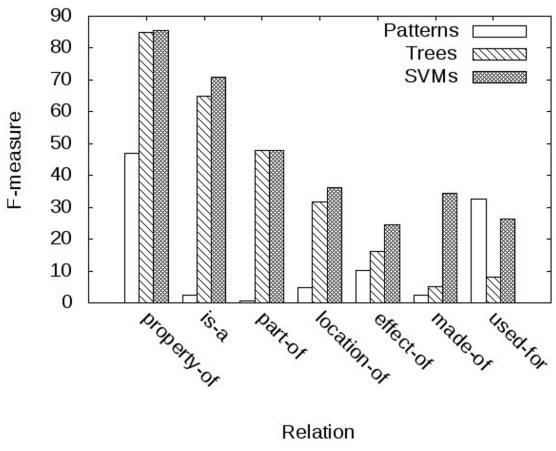

Figure 2: F-measure obtained by textual patterns (experiment 1), decision trees and SVMs (experiment 2) for each semantic relation

Table 8: Experiment 3 – Average results of 10-fold cross-validation for the classifiers trained with different subsets of features

	Decision tree			SVM		
Subset	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
1	46.6%	23.3%	31.7%	0.0%	0.0%	0.0%
2	37.4%	16.1%	22.5%	34.4%	45.6%	39.2%
3	50.0%	7.2%	12.6%	0.2%	14.3%	0.4%
4	56.2%	27.9%	37.3%	64.7%	35.6%	45.9%
5	52.0%	29.0%	37.2%	48.0%	13.7%	21.3%
6	47.1%	18.8%	26.9%	60.2%	35.0%	44.2%
7	54.4%	30.4%	39.0%	61.6%	39.2%	47.9%

Table 11: Examples of correctly identified relation instances, the context in which they occurred and the method that identified them

Relation	**Original context**	**Method**
is-a(Itália, país) is-a(Italy, country)	*...ir para Itália ou qualquer outro país para...* ...go to Italy or any other country for...	is-a pattern #5
property-of(mito, velho) property-of(myth, old)	*...enterra em=parte o velho mito explicitado por...* ...buries in part the old myth explicited by...	property-of pattern #1
part-of(painéis, campanha) part-of(panels, campaign)	*...que os painéis fizessem parte de a campanha de o...* ...that the panels made part of the campaign of the...	part-of pattern #2
made-of(medalhas, ouro) made-of(medals, gold)	*...Vegard=Ulvang , que ganhou três medalhas de ouro em o...* ...Vegard Ulvang, that won three gold medals in the...	made-of pattern #1
effect-of(úlcera, morreu) effect-of(ulcer, died)	*...Jandira morreu devido=a a úlcera perfurada .* ...Jandira died due to the perforated ulcer...	effect-of pattern #1
location-of(carro de FHC, Colégio Alberto=Levy) location-of(FHC's car, Colégio Alberto Levy)	*O carro de FHC chega a o Colégio Alberto=Levy , em...* FHC's car arrives at the Colégio Alberto Levy, in...	location-of pattern #3
used-for(recurso, alterar a foto) used-for(resource, modify the photo)	*...sobre o recurso usado para alterar a foto* ...about the resource used to modify the photo	used-for pattern #1
is-a(Jovem=Pesquisador, programa) is-a(Young Researcher, program)	*...participou de o programa Jovem=Pesquisador com...* ...participated in the Young Researcher program with...	Decision tree
property-of(ação, predatória) property-of(action, predatory)	*...a ação predatória de o homem ...* ...the predatory action of man is...	SVM
part-of(USP, Equipe) part-of(USP, Team)	*Equipe de a USP detalha os mecanismos...* Team from USP details the mechanisms...	Decision tree
made-of(jatos, gases) made-of(jets, gases)	*...evolução de jatos de gases com...* ...evolution of jets of gases with...	SVM
effect-of(começou a chover e a ventar forte, desistir) effect-of(started raining and gusting strongly, stop)	*...a volta começou a chover e a ventar forte , perto=de Ribeirão=Preto , e tivemos de desistir...* ...it started raining and gusting strongly, close to Ribeirão Preto, and we had to stop...	Decision tree
used-for(terapia celular, tratar) used-for(cellular therapy, treat)	*...usam terapia celular para tratar experimentalmente...* ...use cellular therapy to experimentally treat...	Decision tree
location-of(sangue, corpo) location-of(blood, body)	*...o sangue que corre por o corpo contém...* ...the blood that runs through the body contains...	SVM

Concerning the remaining relations, we can cite (Girju et al., 2003) that extracts the part-of relation using C4.5 decision trees. The researchers reported a precision of 83% and a recall of 72%, resulting in a F-measure of 77.1%. That score is considerably higher than the one obtained in experiment 2 (47.7% with decision trees and 47.9% with SVMs). Zelenko et al. (2003) focuses on the location-of relation and uses an SVM classifier with a parse tree kernel. The F-measure obtained was 83.3%, easily surpassing the results found in experiment 2 (31.8% for decision trees and 36.2% for SVMs). It's possible that the use of a specific kernel influenced the performance of the algorithm.

Finally, (Girju et al., 2010) use SVM classifiers to search for 7 semantic relations, 2 of which – effect-of and part-of – are also focused in this work. The resulting F-measure obtained for each relation was 82% (effect-of) and 68% (part-of), both higher than the respective values obtained in experiment 2. It must be taken into account that (Girju et al., 2010) is the related work that employs the greatest number of resources and linguistic tools.

Concluding, even if the results presented in this paper don't surpass all the ones presented in the related works, it's important to note that each investigated work uses different corpora and semantic relation extraction methods, in addition to focusing on different kinds of relations and using varied quantities of training data, factors that must be taken in account when making comparisons. It's also worth mentioning that this is the first work that investigates ML techniques to extract semantic relations with the Portuguese language in focus and, compared with some of the related work for English (Snow et al., 2005; Girju et al., 2010), uses few tools and linguistic resources. Table 12 summarizes the comparison with related work in terms of F-measure (with exception of the first line, presented with precision values).

Table 12: Summary of the comparison with related works in term of F-measure, except when noted otherwise

Relation	Best result obtained in this paper	Best result presented in related works
is-a	78,2% (precision) (SVM)	73,4% (precision) (Freitas and Quental, 2007)
property-of	85,4% (SVM)	–
part-of	47,9% (SVM)	77,1% (Girju et al., 2003)
location-of	36,2% (SVM)	83,3% (Zelenko et al., 2003)
effect-of	24,6% (SVM)	82,0% (Girju et al., 2010)
made-of	34,3% (SVM)	–
used-for	26,2% (patterns)	–

7. Conclusion and Future Work

Automatic semantic relation extraction is a task for which existing systems still don't have high performance (Girju et al., 2010), mainly due to its complexity. Considering the Portuguese language, the situation is even more dire, with few studies done about that subject (de Abreu et al., 2013). Also, according to (de Abreu et al., 2013), one of the reasons for the scarcity of Portuguese-based works is the lack of resources such as annotated data, lexical bases and high quality tools for that language. That scenario shows the importance of this work for the advancement of this complex and vast subject.

In that way, this work sought the study and comparison of the two main approaches – textual patterns and machine learning – for the automatic extraction of semantic relations, an underexplored area in the Portuguese language. The obtained results show that the machine learning approach brings better results than the one based on textual patterns, indicating that ML strategies are a promising direction for studies about this subject with the Portuguese language in focus. The only relation (among the 7 that were focused in this work) that got better results with textual patterns was used-for, possibly due to a low number of training examples.

The methods, features and presented results can be used and bring advancements to applications such as information retrieval and extraction, automatic translation and question & answer systems, and as support for the construction and enhancement of lexical resources such as ontologies, terminologies and dictonaries. The automatic relation extraction tool and the trained models used in this research are available at the Machine Translation Portal PorTAl.[11]

7.1. Future work

Among the possible improvements for this work, we can cite the definition of new features that help the ML methods with the classification of instances. Another possibility is the usage of different taggers for processing the training corpus. The annotation of more training examples can also bring better results for the classifiers, as supervised classifiers were used. It is interesting to verify if the annotation of more used-for relation examples will make the ML methods surpass the textual patterns approach for that relation.

8. Acknowledgements

The authors would like to thank the Brazilian support agencies *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* and *Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)* (#2011/04482-4, #2013/11811-0) for their financial support. This work is also part of the CAMELEON (CAPES-COFECUB #707-11) and AIM-WEST (FAPESP #2013/50757-0) projects.

9. References

Aziz, W. and Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*, Cuiabá, MT, October.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 57–64, College park, MD. ACL.

Bick, E. (2000). *The Parsing System" Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Carletta, J. (1996). Squibs and discussions assessing agreement on classification tasks: The kappa statistic. *Computational linguistics*, 22(2):249–254.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20:273–297, September.

de Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013). A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571, Nov.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press, Cambridge, MA.

Freitas, M. C. and Quental, V. (2007). Subsídios para a elaboração automática de taxonomias. In *Anais do XXVII Congresso da SBC*, V TIL, pages 1585–1594, Rio de Janeiro, Rio de Janeiro.

Girju, R. and Moldovan, D. (2002). Text mining for causal relations. In *Proceedings of the FLAIRS 2002*, pages 360–364, Pensacola, Florida. AAAI Press.

[11] http://www.lalic.dc.ufscar.br/portal/

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June. Association for Computational Linguistics.

Girju, R., Beamer, B., Rozovskaya, A., Fister, A., and Bhat, S. (2010). A knowledge-rich approach to identifying semantic relations between nominals. *Information Processing and Management*, 46(5):589–610.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, June.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational linguistics - Volume 2*, pages 539–545, Nantes, France. ACL.

Joachims, T. (1998). Making large-scale svm learning practical. LS8 24, Universitt Dortmund.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press.

Taba, L. S. (2013). Extração automática de relações semânticas a partir de textos escritos em português do brasil. Master's thesis, Universidade Federal de So Carlos.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer Verlag.

Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, March.