Verb Clustering for Brazilian Portuguese

Carolina Scarton[1,4], Lin Sun[2], Karin Kipper-Schuler[3], Magali Sanches Duran[4], Martha Palmer[3], and Anna Korhonen[2]

[1] Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
c.scarton@sheffield.ac.uk
[2] Computer Laboratory, University of Cambridge
William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
linsun84@gmail.com, alk23@cam.ac.uk
[3] Department of Linguistics, University of Colorado at Boulder
295 UCB Boulder, Colorado 80309-0295
karin_schuler@yahoo.com, martha.palmer@colorado.edu
[4] Interistitutional Center for Computational Linguistics, ICMC, University of São Paulo
Avenida Trabalhador são-carlense, 400 - Centro, 13566-590, São Carlos - SP
magali.duran@uol.com.br

Abstract. Levin-style classes which capture the shared syntax and semantics of verbs have proven useful for many Natural Language Processing (NLP) tasks and applications. However, lexical resources which provide information about such classes are only available for a handful of worlds languages. Because manual development of such resources is extremely time consuming and cannot reliably capture domain variation in classification, methods for automatic induction of verb classes from texts have gained popularity. However, to date such methods have been applied to English and a handful of other, mainly resource-rich languages. In this paper, we apply the methods to Brazilian Portuguese - a language for which no VerbNet or automatic class induction work exists yet. Since Levin-style classification is said to have a strong cross-linguistic component, we use unsupervised clustering techniques similar to those developed for English without language-specific feature engineering. This yields interesting results which line up well with those obtained for other languages, demonstrating the cross-linguistic nature of this type of classification. However, we also discover and discuss issues which require specific consideration when aiming to optimise the performance of verb clustering for Brazilian Portuguese and other less-resourced languages.

1 Introduction

Verbs are central to many Natural Language Processing (NLP) tasks. Typically the main predicates of sentences, they provide the key syntactic and semantic information required for language understanding. Information regarding verbs is traditionally obtained from lexical resources such as WordNet [1], FrameNet [2], PropBank [3] and VerbNet [4].

In this paper we are particularly concerned with Levin-style verb classification. Beth Levin [5] has classified verbs according to their participation in diathesis alternations.

A. Gelbukh (Ed.): CICLing 2014, Part I, LNCS 8403, pp. 25–39, 2014.

These are alternations in the syntactic realization of a verb that may sometimes also result in a slight change of meaning. For example, sentences (1) and (2) (from [5], p. 2) illustrate the locative alternation for verbs ”spray” and ”load”:

1. (a) Sharon sprayed water on the plants.
   (b) Sharon sprayed the plants with water.
2. (a) The farmer loaded apples into the cart.
   (b) The farmer loaded the cart with apples.

Levin classes capture the regularity in verb meaning and behaviour at syntax-semantics interface. VerbNet [4], an extensive computational verb lexicon for English, extends Levin's original classification with additional classes and member verbs, and provides detailed syntactic-semantic information for the classes that span across the entire verb lexicon.

Because Levin classes capture useful generalisations about verb behaviour and meaning, they have proven useful for many NLP tasks and applications. Examples include information retrieval [6], semantic role labelling [7, 8], semantic parsing [9], word sense disambiguation [10–13], among others. Since many Levin-style classes are applicable across languages [14], they can also be useful for cross-linguistic tasks. However, their effective exploitation in the multi-lingual context has been limited to date because resources like VerbNet are currently only available for English and a handful of other languages, such as Spanish [15], Chinese [16] and Arabic [17].

Although manual development of VerbNets is under way for many languages, it is extremely time-consuming and cannot reliably capture domain variation. Therefore techniques that can automatically induce or update verb classifications from texts have gained popularity. Particularly attractive are unsupervised techniques such as clustering because they do not require manual annotations for training and are therefore easier port across NLP tasks.

For English, verb clustering approaches have been developed by Kingsbury and Kipper-Schuler [18], Sun and Korhonen [19] and Reichart and Korhonen [20], among others. The best techniques have been applied successfully to domains such as biomedicine [21] and they have produced promising results with demonstrated improvement on application tasks such as argumentative zoning [22] and metaphor identification [23].

For languages other than English, only a small number of clustering works exist that focus on Levin style syntactic-semantic classification, e.g. [24–26]. Interestingly, the recent experiment performed by Sun et al. demonstrates that it is possible to take an unsupervised clustering method developed for English [27] and apply it successfully to French [25], using French NLP tools for feature extraction, but without language specific feature engineering. If this approach was applicable to a wider range of languages, it could greatly support the development of VerbNets across languages and language domains.

In this paper, we explore this approach for a another language: Brazilian (Br.) Portuguese. Although Portuguese is a major world language (around 215 million speakers worldwide), no manually developed VerbNet or automatic verb clustering approach exists for it yet. While the language belongs to the family of Romance languages like

French, it is lexically more distant to English and is also less-resourced in terms of basic corpora and NLP tools than French, and therefore likely to be more challenging for clustering.

We develop and release the first gold standard Levin classification for Br. Portuguese and apply the state-of-the-art verb clustering approach developed for English [27] to this language. Using the NLP tools developed for Br. Portuguese for feature extraction, we experiment with the same basic features and the same clustering method as for English. The results are encouraging and support the hypothesis that Levin-style classes can indeed be cross-linguistically applicable: it was possible to obtain a gold standard largely via translation from English to Br. Portuguese, and the best performing features and clustering techniques matched with those for English and French. The level of clustering performance for Br. Portuguese lags behind that of resource-richer languages. We investigate reasons for this and discuss future work in this area.

2 Related Work

Several approaches have been developed for classifying English verbs into Levin-style classes in both supervised [28–30] and unsupervised [18–20, 27] manner. These approaches have employed a variety of different features for classification, ranging from shallow features (e.g. co-occurrences of verbs with other words) extracted from raw or part-of-speech (POS) tagged text to deeper features (e.g. grammatical dependencies) extracted from manually or automatically parsed data. Also lexical(-semantic) features which correspond more closely with the features Levin used for her manual classification have been used, such as verb subcategorization frames (SCFs), selectional preferences (SPs) and recently also diathesis alternation approximations [31]. These more sophisticated features have been learned from parsed data. Clustering approaches have ranged from the simple k-means [18] to more sophisticated techniques such as spectral clustering [27], hierarchical clustering using graph factorization [19] and determinantal point processes [20], among others.

For example, the state-of-the-art approach of Sun and Korhonen [27] which we plan to use as a starting point in our work, uses a variety of features based on co-occurrences, verb SCFs and lexical as well as selectional preferences of verbs on their argument heads, and yields promising performance when used with spectral clustering. When this approach was evaluated against gold standards based on Verbnet [30, 32], both containing hundreds of verbs in 15-20 classes, it achieved the highest performance (at around 80 F-measure) with deep linguistic features: SCFs refined with selectional preferences.

For languages other than English, few works exist. The most substantial related work focuses on German [33] where verb clustering has yielded promising results, but this work has emphasis on semantic rather than VerbNet style syntactic-semantic classification. Although the two classification types share many properties, the mapping between the two is only partial and many to many due to fine-grained nature of semantic classes based purely on synonymy [11].

The prior works most related to ours are those by Ferrer [24] for Spanish and Sun et al. [25] and Falk et al. [26] for French. Ferrer applied a simple hierarchical clustering

approach developed for English to Spanish, and evaluated it against a manual classification of Vazquez [34] which is similar in nature (but not identical) to that of Levin's. The experiment included 514 verbs in 31 classes and produced results only slightly better than the random baseline.

Sun et al. [25] used the same features as Sun and Korhonen [27] for English to cluster 171 French verbs to 16 classes. The gold standard was obtained by translating the Levin-based gold standard of Sun et al. [32] from English to French, and a good correspondence was reported between the two gold standards. The authors reported the best results (64.5 F-measure) on high frequency verbs with the same combination of features (SCFs and selectional preferences) and the same clustering method (spectral clustering) as for English. Falk et al. [26] employed a neural clustering method for French verbs. They achieved 70 F-measure when evaluating on a slightly modified version of the Sun et al. 2010 gold standard for French. However, the method is not fully comparable to other works mentioned here because it uses features from lexical resources rather than those obtained solely by NLP.

The work reported on manual development of VerbNets for different languages [15–17] seems to support the hypothesis that Levin-style classes can be, to a considerable degree, cross-linguistically applicable. The experiment reported by Sun et al. [25] provides further evidence for this because it shows that it is possible to take an unsupervised technique developed for one language and apply it to another language without language specific tuning (other than use of language-specific corpora and basic NLP tools for feature extraction) and get promising results.

However, this experiment focused on French only. French shares some of its vocabulary with English, and like English, has large corpora, POS-taggers, parsers and lexical acquisition tools available. If this approach proves more widely applicable so that it can be successfully employed for other, including also less-resourced languages, verb clustering could offer a useful tool for hypothesizing Levin classes for other languages. We will take this line of research further and investigate whether the approach could be applied to Br. Portuguese which, like the majority of worlds languages, is less-resourced in terms of NLP than French and is thus likely to be a more challenging test case.

Portuguese is a Romance language which has its origins in Latin and is currently the seventh most spoken language in the world. From the 215 million people speaking Portuguese, 85% speak Br. Portuguese. Br. Portuguese differs from the European Portuguese largely in terms of lexicon. As we are dealing with the verb lexicon, we differentiate between the two variations of the language and focus on Br. Portuguese only. However, our work could be easily extended to accommodate European Portuguese as well.

Some major lexical resources are currently under development for both variants of Portuguese. The ones related to verbs include PropBank-Br [35] (based on PropBank), FrameNet Brasil [36] and FrameCorp [37] (based on FrameNet), WordNet.Br [38], WordNet.Pt [39, 40], one of the Wordnets in the MultiWordNet Project [41] (based on WordNet) as well as VerbNet.Br [42] based on English VerbNet. The latter project, which is most closely related to our work, provides alignments between English VerbNet, WordNet and WordNet.Br, and enables semi-automatically inferring Levin classes

for Br. Portuguese from the alignment data. The classification created using this method is noisy and has not been manually validated.

3 Gold Standard for Brazilian Portuguese

As no VerbNet exists which could provide gold standard classes for our experiments we created the first gold standard including Levin classes for Br. Portuguese[1]. We used an approach similar to that earlier employed by Sun et al. [25] for building a gold standard for French. They took a gold standard frequently used for evaluating verb clustering for English [32] and translated its 204 verbs and 17 classes to French. The majority of verbs and classes could be translated successfully. To cover for the ones that could not, Sun et al. considered synonyms of known member verbs and added these in, where possible. French subcategorization frames (SCFs) and alternations, defined manually for each class, were used as evaluation criteria. The final gold standard included 171 verbs in 16 classes.

We employed a similar approach because it had proved successful for French and we were interested in exploring the cross-linguistic potential of Levin classification. We used a native speaker of Portuguese with expertise in VerbNet to develop the first version of the gold standard. She performed the translation and defined syntactic-semantic criteria for each class. We ended up with 203 verbs in 16 classes (12.69 verbs per class). The majority of verbs (including their synonyms) got translated successfully. Only one class in the English gold standard was deemed incompatible with Portuguese (peer-30.3), showing a strong cross-lingual element between English and Portuguese classifications, similar to that earlier observed with English and French.

Because many of the verbs in the resulting gold standard were quite low in frequency in our corpus, we supplemented the resource with additional member verbs from VerbNet.Br – the resource recently developed by Scarton and Aluísio [42]. As the classifications in this resource are noisy, we used a native language expert to validate the class memberships according to the criteria we had developed during the translation of the first version of the gold standard. The resulting extended gold standard includes 540 verbs in 16 classes (c. 34 verbs per class).

Table 1 shows the resulting classes in the gold standard (indicated by original Levin class numbers) together with some example verbs.

4 Verb Clustering

4.1 Features

We employed a selection of syntactic and semantic feature sets that had proved promising for both English [27] and French [25]. To facilitate easy comparison of our results against earlier works we indicate each feature set using the same numbers as in [27] and [25]:

[1] We will release this gold standard together with a published version of this paper.

Table 1. Brazilian Portuguese gold standard classes with some example verbs

Number	Class	Portuguese Members
22.2	amalgamate	*alternar, contrastar, combinar, juntar, comparar,...*
31.1	amuse	*frustar, chatear, alegrar, decepcionar, encantar,...*
29.2	characterize	*diagnosticar, restabelecer, retratar, classificar,...*
36.1	correspond	*pechinchar, flertar, simpatizar, colidir, cooperar,...*
13.5.1	get	*arranjar, colher, reservar, adquirir, obter,...*
18.1	hit	*martelar, esmagar, espancar, bater,...*
43	light emission	*resplandecer, raiar, cintilar, piscar, brilhar,...*
37.3	manner of speaking	*cochichar, rosnar, sussurrar, berrar, ...*
47.3	modes of being with motion	*boiar, flutuar, vibrar, oscilar,...*
40.2	nonverbal expression	*bocejar, roncar, solućar, suspirar, sorrir,...*
45	other cos	*encurtar, afrouxar, alargar, estreitar, derreter,...*
9.1	put	*cravar, posicionar, mergulhar, situar, inserir,...*
10.1	remove	*erradicar, subtrair, descarregar, remover,...*
51.3.2	run	*marchar, nadar, passear, voar, correr,...*
37.7	say	*segredar, reportar, dizer, proclamar, exprimir,...*
11.1	send	*despachar, transportar, remeter, enviar,...*

- F1: SCFs and their relative frequencies with individual verbs (without parameterising for prepositions).
- F2: F1 with SCFs parameterized for the tense (i.e. POS tag) of the verb.
- F3: F1 with SCFs parameterized for specific prepositions.
- F7: Collocations (COs) extracted from the window of 6 words, with the relative word position recorded. We followed the work of Li and Brew [29] where COs were extracted from the window of words immediately preceding and following a POS-tagged and lemmatized verb (stop words were removed before the extraction).
- F13: All Lexical Preferences (LPs) in argument head positions: the type and frequency of words acting as prepositions (PREP), subjects (SUBJ), indirect objects (IOBJ) and direct objects (OBJ) in dependency-parsed data were considered.
- F16: F3 parameterized for LPs.
- F17: F3 refined with Selectional Preferences (SPs).

The extraction of these features requires POS-tagging and, with the exception of F7, also parsing data, and using additional technology to extract SCFs and SPs from the parsed data. We used the three publicly available corpora for Brazilian Portuguese to ensure that as much data as possible was available for clustering. These were (i) Lácio-Ref [43] which includes legal, news, scientific and literary texts - approximately 9 million words in total, (ii) PLN-BR-FULL [44] which provides 29M words of news texts and (iii) Revista Pesquisa FAPESP corpus [45] which contains 6M words of scientific text.

These corpora were POS-tagged and parsed using the rule-based PALAVRAS [46] parser which outputs grammatical relations. According to the evaluation performed by the authors, this rule-based parser achieves 99% of correctness for POS and 97% for syntax. We used the system of Zanette et al. [47] to extract SCFs from the resulting

parsed data. Similar to the system of [48] for French, the system generates SCFs from the dependency relations associated with individual verbs in parsed data. According to the evaluation of Zanette [49], this system performs at around 50.6% F-measure.

We considered all the dependencies of interest and all the SCFs with frequency higher than 5 in our experiments. This yielded 3,779 verb lemmas, 408 basic SCF types and 3,578 preposition-parameterized SCF types. For SP acquisition, we used the method proposed by Sun and Korhonen (2009) (without the automatic definition of best number of clusters in Sun et al. (2010)). The method involves (i) taking the SUBJ, OBJ and IOBJ relations associated with verbs in parsed data, (ii) extracting all the argument heads in these relations, and (iii) clustering the resulting N most frequent argument heads into M classes. We considered frequency higher than 5 and N $\{200, 500\}$ most frequent argument heads and M $\{10, 20, 30, 80\}$ classes. Finally, all feature vectors were normalized by the sum of the feature values before clustering.

4.2 Clustering Algorithms

We used two clustering algorithms in our work: the MNCut spectral clustering algorithm (SPEC) which produced the best results in both English and French [25, 27] and a recent Data-Cluster-Data (DCD) algorithm [50], not previously employed for verb clustering. We wanted to experiment with DCD because it had been shown to work together with SPEC to reduce problems of data or feature sparsity which a less-resourced language, in particular, will suffer from.

In DCD, SPEC is first used to perform dimensionality reduction using measures of distributional similarity. The resulting feature space tends to be dense and the infrequent (and potentially unreliable) features become less important when distributional similarity measures are used. DCD takes the output of SPEC as the initial guess and performs further optimization. In the experiments performed by [50] the method further improved the performance of SPEC on varied datasets (consisting of text, images and other material).

We introduce the two clustering approaches in the below sections, respectively.

Spectral Clustering. Spectral clustering (SPEC) has proved promising in several previous verb clustering experiments, e.g. [27, 51]. Following [27] we used the MNCut spectral clustering [52].

The similarity matrix A is normalized into a stochastic matrix P.

$$P = D^{-1}A \tag{1}$$

The degree matrix D is a diagonal matrix where $D_{ii} = \sum_{j=1}^{N} A_{ij}$.

It was shown by [52] that if P has the K leading eigenvectors that are piecewise constant[2] with respect to a partition I^* and their eigenvalues are not zero, then I^* minimizes the multiway normalized cut which is the sum of transition probabilities across different clusters.

[2] The eigenvector v is piecewise constant with respect to I if $v(i) = v(j) \forall i, j \in I_k$ and $k \in 1, 2...K$.

In practice, the leading eigenvectors of P are not piecewise constant. However, we can extract the partition by finding the approximately equal elements in the eigenvectors using a clustering algorithm like KMeans. KMeans is a simple clustering method that iteratively partitions data in order to minimize the within-cluster sums of point-to-cluster-centroid distance.

Data-Cluster-Data. In DCD[3] given a similarity matrix A of the n verbs, the clustering task is to divide the verbs into r disjoint subsets. The pairwise similarity is measured using Jensen Shannon Divergence as in [27]. The aim of the clustering is to find the probability of assigning the ith verb to the kth cluster $p(k|i)$.

The similarity matrix can be seen as an undirected similarity graph where each node corresponds to a verb. If we augment the similarity graph by r cluster nodes, the connection weight between the verb and the cluster is (assuming uniform prior $p(i) = 1/n$):

$$p(i|k) = \frac{p(k|i)p(i)}{\sum_v p(k|v)p(v)} = \frac{p(k|i)}{\sum_v p(k|v)}$$

The similarity between two verbs can be defined as two-step random walks from ith verb to jth verb via all clusters:

$$p(i|j) = \sum_k p(i|k)p(k|j) = \sum_k \frac{p(k|i)p(k|j)}{\sum_v p(k|v)}$$

The objective of the clustering is to find a good approximation between the input similarity matrix A and the random walk probabilities $\hat{A}$. The difference between the two matrices is measured using the Kullback-Leibler divergence. The learning target can be formulated as the following optimization problem:

$$\begin{aligned} &\min_{w \geq 0} D_{KL}(A||\hat{A}) = \sum_{ij} (A_{ij} \log \frac{A_{ij}}{\hat{A}_{ij}} - A_{ij} + \hat{A}_{ij}) \\ &\text{s.t.} \sum_k W_{ik} = 1, i-1, ..., n, \end{aligned} \tag{2}$$

where we define $W_{ik} = p(k|i)$, and thus

$$\hat{A}_{ij} = \sum_k \frac{W_{ik} W_{ij}}{\sum_v W_{vk}} \tag{3}$$

By dropping the constant terms, the optimization problem with dirichlet prior on W is equivalent to minimising:

$$J(W) = -\sum_{ij} A_{ij} \log \hat{A}_{ij} - (\alpha - 1) \sum_{ik} \log W_{ik} \tag{4}$$

[3] For more detailed information about DCD that that we are able to provide in this section, please see [50].

where α is the parameter of the dirichlet prior.

Taking the constraint in equation 2 into account, the optimization object becomes:

$$L(W, \lambda) = J(W) + \sum_i \lambda_i (\sum_k W_{ik} - 1) \tag{5}$$

[50] proved that the L is non-increasing under the update rule of W and λ detailed in algorithm 1.

Algorithm 1. Optimization Algorithm for DCD [50]

Require: similarity matrix A, number of clusters r, nonnegative initial guess of W
  repeat
    $Z_{ij} = (\sum_k \frac{W_{ik}W_{jk}}{\sum_v W_{vk}})^{-1} A_{ij}$
    $s_k = \sum_v W_{vk}$
    $\nabla^-_{ik} = 2(ZW)_{ik} s_k^{-1} + \alpha W_{ik}^{-1}$
    $\nabla^+_{ik} = (W^T ZW)_{kk} s_k^{-2} + W_{ik}^{-1}$
    $a_i = \sum_l \frac{W_{il}}{\nabla^+_{il}}$, $b_i = \sum_l W_{il} \frac{\nabla^-_{il}}{\nabla^+_{il}}$
    $W_{ik} \leftarrow W_{ik} \frac{\nabla^-_{ik} a_i + 1}{\nabla^+_{ik} + b_i}$
  until W is unchanged
  return cluster assigning probabilities W

The initial guess of W can be produced from the result of another clustering algorithm. As suggested by [50], the result of the spectral clustering can be used for the initialisation of W. Thus, by using the result of spectral clustering as a starting point, we can expect DCD to further improve the clustering result. We convert the result of SPEC to an $n \times r$ binary indicator matrix, and add a small positive random number to all entries. This matrix is used as the input W matrix for the algorithm 1.

4.3 Evaluation Metrics

We evaluate the results of the clustering against the gold standard using F-Measure as in [27] and [25] to facilitate meaningful comparison against previous works. F-measure provides the harmonic mean of precision (P) and recall (R). P is calculated using modified purity – a global measure which evaluates the mean precision of clusters. Each cluster ($k_i \in K$) is associated with the gold-standard class to which the majority of its members belong. The number of verbs in a cluster (k_i) that take this class is denoted by $n_{prevalent}(k_i)$.

$$P = \frac{\sum_{k_i \in K : n_{prevalent(k_i)} > 2} n_{prevalent(k_i)}}{|\text{verbs}|}$$

R is calculated using weighted class accuracy: the proportion of members of the dominant cluster DOM-CLUST$_i$ within each of the gold-standard classes $c_i \in C$.

$$R = \frac{\sum_{i=1}^{|C|} |\text{verbs in DOM-CLUST}_i|}{|\text{verbs}|}$$

We calculate the random baseline as follows: 1/number of classes. We also calculate the statistical significance of the results by using the one-tailed McNemar's test [53], with the extension of [54]. We considered p-value lower than 0.05.

5 Results

Table 2 shows the F-measure results for both clustering algorithms with different featuresets (see Section 4.1 is for legends of the different featureset codes). We can see that SPEC performs better than DCD with nearly all the featuresets. This difference is statistically significant for the majority of features and becomes more pronounced as we move towards more sophisticated featuresets. The only featureset on which DCD seems outperform SPEC is F1 which consists plain SCFs, but this difference in performance is not statistically significant. While the poor performance of DCD may seem surprising in the light of the good results of [50], our dataset, focusing solely on natural language, is different in nature and also multiple times smaller than the data employed by Yang and Oja.

Table 2. Results for Br. Portuguese verb clustering, considering 16 clusters (number of classes in the gold standard) – * means no statistically significant difference between the algorithms

Feature	Spectral Cluster	DCD
F1*	33.62	**35.08**
F2*	**39.16**	36.79
F3	**42.27**	40.94
F7	**35.79**	32.23
F13	**39.77**	37.13
F16	**41.23**	37.55
F17 (N=200, M=10)	**38.66**	35.99
F17 (N=200, M=20)	**41.15**	35.62
F17 (N=200, M=30)	**39.70**	34.37
F17 (N=200, M=80)	**39.34**	39.26
F17 (N=500, M=10)	**38.54**	35.66
F17 (N=500, M=20)	**42.06**	34.33
F17 (N=500, M=30)	**42.77**	35.92
F17 (N=500, M=80)	**38.51**	35.33

Regarding features, the best individual featureset is F17 with N=500 and M=30 which yields F-measure of 42.8 with SPEC. This is the most sophisticated featureset which incorporates preposition-prameterized SCFs with SPs. Also F3 (SCFs parameterized for prepositions) and F16 (F3 parameterized for LPs) yield F-measure which is

Table 3. The results for Br. Portuguese (BP), English and French (* inducates the best results for each language) using SPEC

Feature	BP	French	English
F1	33.62	42.4	57.8
F2	39.16	45.9	46.7
F3	42.27	50.6	63.3
F7	35.79	55.1	-
F13	39.77	52.7	74.6
F16	41.23	53.4	73
F17	42.77*	54.6*	80.4*

clearly above 40, but there is a statistically significant difference between the performance of these featuresets and that of F17.

To provide an idea of how this performance compares with results obtained for resource-richer languages, Table 3 shows the previously reported results for English [27] and French [25]. Although the same feature sets and the same SPEC algorithm were used here for all the languages, it is important to note that the results are not directly comparable due to differences in data, NLP tools and gold standards (which are not identical, even though they were derived from the same gold standard). However, this table serves to give an idea of the general performance level and the best performing features for the three languages.

We can observe that clustering performs clearly the best for English (with top performance at 80.5F) which has the largest corpus data and the highest quality NLP tools among the three languages. French does not perform equally well (with top performance at 54.6F), with errors reported due to the poor quality of parsing and data sparsity [25] [4]. Br. Portuguese falls behind English and French in performance. Yet, the best results for this language, obtained without any language specific feature engineering, are (in contrast to the earlier verb clustering experiment for Spanish [24]) well beyond the random baseline. This together with the fact that similar feature sets tend to obtain the highest and lowest results among the three languages is encouraging and also demonstrates the cross-linguistic potential of Levin's classification.

The lower quality NLP tools are likely to be one explanation for the lower performance for Br. Portuguese. For example, the SCF system of Zanette, used for many of the featuresets, was reported to perform only at around 50.6% F-measure. The featuresets which were created using multiple NLP tools are likely to suffer from this problem the most due to error propagation.

Another explanation is the small corpus size (our corpus is e.g. four times smaller than that used in the French experiment) since verb clustering is known to be sensitive to data sparsity. We therefore conducted another experiment on the full set of verbs where we investigated the effect of instance filtering on the performance of the best features sets: F3, F13, F16 and F17.

The results shown in Table 4 demonstrate the strong effect of data size of the performance. The results for high frequency verbs are considerably better than those for

[4] Note that higher performance at 65.4 F was reported for high frequency French verbs.

lower frequency verbs. SPEC, in particular, is able to take advantage of big data size. While DCD and SPEC perform quite similarly for for the full set of verbs, the difference between the two methods gets more pronounced when the scope is restricted to high frequency verbs (in particular those that have 2000 or more occurrences). For the highest frequency group (verbs with 4000 or more occurrences), SPEC performs considerably better than DCD for all featuresets.

The most sophisticated featureset F17 performs clearly the best with SPEC, obtaining its top performance for verbs which have 4000 or more occurrences at 75.2 F. The second best featureset is F16 (at 68.3 F) – another linguistically sophisticated featureset which refines preposition-parameterized SCFs with lexical preferences. All the four featuresets obtain results of over 63 F at the highest frequency group.

The results for high frequency verbs approach those obtained for all verbs in English (with the majority of verbs having frequency more than 1000 in English), showing that big data size can compensate for the lower quality NLP tools.

Table 4. The effect of verb frequency on clustering performance

Freq.	Verbs	F3		F13		F16		F17	
		DCD	SPEC	DCD	SPEC	DCD	SPEC	DCD	SPEC
50	454	40.82	41.79	36.50	35.44	37.23	39.44	32.06	35.20
100	371	37.74	43.72	37.59	41.43	37.84	41.11	32.41	37.66
150	321	39.41	42.27	35.62	41.62	38.80	42.97	32.38	37.83
200	290	39.53	41.42	36.64	39.17	36.29	39.37	30.84	37.03
400	222	43.67	43.98	36.86	44.77	41.42	39.89	31.80	40.51
1000	131	41.11	44.52	42.13	44.37	45.43	46.85	40.99	47.62
2000	82	42.63	55.54	46.20	50.26	49.02	59.12	40.20	52.08
3000	63	44.32	53.21	55.60	57.48	50.32	62.03	50.50	57.50
4000	46	45.32	64.66	53.77	63.29	48.24	68.31	43.90	75.21

6 Conclusion

In this paper we have presented the first work of clustering verbs into Levin-style classes in Br. Portuguese. We have explored, in particular, the cross-linguistic potential of Levin style classification and how this could be exploited in the creation of a gold standard as well as in the development of a verb clustering approach for this language.

We first created a gold standard for evaluation of clustering by translating it from English to Br. Portuguese, showing that the two gold standards share nearly all of their classes and the majority of member verbs. The gold standard was extended further mainly because many member verbs were low in frequency. We then used existing Br. Portuguese NLP tools to extract similar feature sets as previously used for English and French, and clustered them using similar clustering algorithms. The results for different feature sets were in line with those obtained earlier for English and French. In particular, the most sophisticated features – the ones which are in the closest agreement with Levin's original features – produced the best results across the three languages.

The level of clustering performance for Br. Portuguese was considerably lower than that for English and French when the full set of verbs was considered. However, when the scope was restricted to high frequency verbs the results were substantially better, demonstrating that big data size is important for this task. The top performance was, again, obtained using linguistically sophisticated features.

In the future, to improve the results for Br. Portuguese, we plan to use larger corpus data (e.g. supplement existing corpora with text from the web), to improve the accuracy of feature extraction (e.g. SCF acquisition), and to investigate whether it is possible to refine feature sets with language specific constraints.

We have shown that it is possible to adopt a verb clustering method developed for resource-rich language and apply it to a less-resourced language without language specific feature engineering, and obtain a useful result. Future work should investigate the applicability of this approach to other, more distant languages and language families. Such investigations can be highly valuable for the majority of world's languages that suffer from the lack of NLP resources, and could greatly benefit from techniques for (semi-)automatic lexicon development.

Acknowledgements. This work was supported by FAPESP/Brazil (No. 2010/03785-0 and No. 2011/22882-0), EXPERT (EU Marie Curie ITN No. 317471) project and the Royal Society University Research Fellowship (UK).

References

1. Fellbaum, C.: WordNet: An electronic lexical database. MIT Press, Cambridge (1998)
2. Baker, C.F., Fillmore, C.J., Lowe, J.F.: The Berkeley Framenet Project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, University of Montréal, Canadá, pp. 86–90 (1998)
3. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics 31(1), 71–106 (2005)
4. Kipper-Schuler, K.: Verbnet: A broad coverage, comprehensive verb lexicon. Doctor of philosophy, University of Pennsylvania (2005)
5. Levin, B.: English Verb Classes and Alternation, A Preliminary Investigation. The University of Chicago Press, Chicago (1993)
6. Crouch, D., King, T.H.: Unifying Lexical Resources. In: Interdisciplinary Workshop on the Identication and Representation of Verb Features and Verb, Saarbruecken, Germany, pp. 32–37 (2005)
7. Swier, R., Stevenson, S.: Unsupervised Semantic Role Labelling. In: EMNLP 2004, Barcelona, Spain, pp. 95–102 (2004)
8. Yi, S., Lopper, E., Palmer, M.: Can Semantic Roles Generalize Across Genres? In: NAACL HLT 2007, Rochester, NY, USA, pp. 548–555 (2007)
9. Shi, L., Mihalcea, R.: Putting Pieces Together: Combining Framenet, Verbnet and Wordnet for Robust Semantic Parsing. In: 6th International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, pp. 99–110 (2005)
10. Girju, R., Roth, D., Sammons, M.: Token-level Disambiguation of Verbnet Classes. In: Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbruecken, Germany (2005)
11. Abend, O., Reichart, R., Rappoport, A.: A Supervised Algorithm for Verb Disambiguation into Verbnet Classes. In: LREC 2008, Manchester, UK, pp. 9–16 (2008)

12. Chen, L., Eugenio, B.D.: A Maximum Entropy Approach to Disambiguating Verbnet Classes. In: Proceedings of the 2nd Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features, Pisa, Italy (2010)
13. Brown, S.W., Dligach, D., Palmer, M.: Verbnet Class Assignment as a WSD Task. In: IWCS 2011, Oxford, UK, pp. 85–94 (2011)
14. Jackendoff, R.: Semantic Structures. MIT Press, Cambridge (1990)
15. Taulé, M., Martí, M.A., Borrega, O.: Ancora-net: Mapping the spanish ancora-verb lexicon to verb-net. In: The Workshop on Verbs. The Identification and Representation of Verb Features, Pisa, Italy (2010)
16. Liu, M.C., Chiang, T.Y.: The construction of mandarim verbnet: A frame-based study of statement verbs. Language and Linguistics 9(2), 239–270 (2010)
17. Mousser, J.: Classifying arabic verbs using sibling classes. In: International Workshop on Computational Semantics, Oxford, UK (2011)
18. Kingsbury, P., Kipper-Schuler, K.: Deriving Verb-Meaning Clusters from Syntactic Strucutres. In: The Workshop on Text Meaning, in Conjunction with NAACL HLT 2003, Edmonton, Canad (2003)
19. Sun, L., Korhonen, A.: Hierarchical Verb Clustering Using Graph Factorization. In: EMNLP 2011, Edinburgh, UK, pp. 1023–1033 (2011)
20. Reichart, R., Korhonen, A.: Improved lexical acquisition through dpp-based verb clustering. In: ACL 2013, Sofia, Bulgaria (2013)
21. Korhonen, A., Krymolowski, Y., Collier, N.: The choice of features for classification of verbs in biomedical texts. In: COLING 2008, Manchester, UK (2008)
22. Guo, Y., Korhonen, A., Poibeau, T.: A weakly-supervised approach to argumentative zoning of scientific documents. In: EMNLP 2011, Edinburgh, UK (2011)
23. Shutova, E., Sun, L.: Unsupervised metaphor identification using hierarchical graph factorization clustering. In: NAACL 2013, Atlanta, USA (2013)
24. Ferrer, E.E.: Towards a semantic classification of spanish verbs based on subcategorisation information. In: The Workshop on Student Research, in Conjunction with ACL 2004, Barcelona, Spain, pp. 163–170 (2004)
25. Sun, L., Korhonen, A., Poibeau, T., Messiant, C.: Investigating the cross-linguistic potential of Verbnet-style classification. In: The 23rd International Conference on Computational Linguistics, Beijing, China, pp. 1056–1064 (2010)
26. Falk, I., Gardent, C., Lamirel, J.C.: Classifying french verbs using french and english lexical resources. In: ACL 2012, Jeju, Republic of Korea, pp. 854–863 (2012)
27. Sun, L., Korhonen, A., Krymolowski, Y.: Improving verb clustering with automatically acquired selectional preferences. In: EMNLP 2009, Singapore, pp. 638–647 (2009)
28. Merlo, P., Stevenson, S.: Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics 27(3), 373–408 (2001)
29. Li, J., Brew, C.: Which Are the Best Features for Automatic Verb Classication? In: ACL 2008 (2008)
30. Joanis, E., Stevenson, S., James, D.: A General Feature Space for Automatic Verb Classication. Natural Language Engineering (2008)
31. Sun, L., McCarthy, D., Korhonen, A.: Diathesis alternation approximation for verb clustering. In: ACL 2013, Sofia, Bulgaria, pp. 736–741 (2013)
32. Sun, L., Korhonen, A., Krymolowski, Y.: Verb class discovery from rich syntactic data. In: The 9th International Conference on Computational Linguistics and Intelligent Text Processing, Haifa, Israel, pp. 16–27 (2008)
33. Schulte im Walde, S.: Experiments on the Automatic Induction of German Semantic Verb Classes. Computational Linguistics 32(2), 159–194 (2006)
34. Vázquez, G., Fernández, A., Castellón, I., Martí, M.A.: Clasificasión verbal: Alternancias de diátesis. Quaderns de Sintagma, Universitat de Lleida (2000)

35. Duran, M.S., Aluisio, S.M.: Propbank-br: A brazilian treebank annotated with semantic role labels. In: LREC 2012, Istanbul, Turkey (2012)
36. Salomao, M.M.: Framenet Brasil: Um trabalho em progresso. Revista Calidoscópio 7(3), 171–182 (2009)
37. Bertoldi, A., Chishman, R.: Frame semantics and legal corpora annotation: Theoretical and applied challenges. Linguistic Issues in Language Technology 7(9) (2012)
38. da Dias Silva, B.C., Felippo, A.D., Nunes, M.G.V.: The Automatic Mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In: Proc. LREC 2008, Marrakech, Morocco, pp. 1535–1541 (2008)
39. Marrafa, P.: Portuguese wordnet: General architecture and internal semantic relations. DELTA 18, 131–146 (2002)
40. Marrafa, P., Amaro, R., Chaves, R.P., Lourosa, S., Martins, C., Mendes, S.: Wordnet.pt new directions. In: The Third Global WordNet Association Conference, Jeju, Republic of Korea, pp. 319–320 (2008)
41. Bentivogli, L., Pianta, E., Girardi, C.: Multiwordnet: Developing an aligned multilingual database. In: The First International Conference on Global WordNet Conference, Mysore, India, pp. 293–302 (2002)
42. Scarton, C., Aluísio, S.M.: Towards a cross-linguistic Verbnet-style lexicon to Brazilian Portuguese. In: The Workshop on Creating Cross-language Resources for Disconnected Languages and Styles, in Conjunction with LREC 2012, Istanbul, Turkey (2012)
43. Aluísio, S.M., Pinheiro, G.M., Manfrim, A.M.P., Genovês Jr., L.H.M., Tagnin, S.E.O.: The Lácio-web: Corpora and Tools to Advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In: LREC 2004, Lisbon, Portugal, pp. 1779–1782 (2004)
44. Muniz, M., Paulovich, F.V., Minghim, R., Infante, K., Muniz, F., Vieira, R., Aluísio, S.: Taming the tiger topic: An xces compliant corpus portal to generate subcorpus based on automatic text topic identification. In: CL 2007, Birmingham, UK (2007)
45. Aziz, W., Specia, L.: Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In: STIL 2011, Cuiabá, MT (October 2011)
46. Bick, E.: The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Doctor of philosophy, University of Aarhus (2005)
47. Zanette, A., Scarton, C., Zilio, L.: Automatic extraction of subcategorization frames from corpora: An approach to Portuguese. In: PROPOR 2012 - Demo Session, Coimbra, Portugal (2012)
48. Messiant, C.: A subcategorization acquisition system for French verbs. In: NAACL HLT 2008, Columbus, OH, pp. 55–60 (2008)
49. Zanette, A.: Aquisiçao de Subcategorization Frames para Verbos da Língua Portuguesa. Projeto de diplomação, Federal University of Rio Grande do Sul (2010)
50. Yang, Z., Oja, E.: Clustering by low-rank doubly stochastic matrix decomposition. In: ICML (2012)
51. Brew, C., Schulte im Walde, S.: Spectral clustering for german verbs. In: EMNLP 2002, pp. 117–124 (2002)
52. Meila, M., Shi, J.: A random walks view of spectral segmentation. In: AISTATS (2001)
53. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2), 153–157 (1947)
54. Dietterich, T.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7), 1895–1923 (1998)