Corpus-Based Acquisition of Support Verb Constructions for Portuguese

Britta D. Zeller and Sebastian Padó

Department of Computational Linguistics,
Heidelberg University, Germany
{zeller,pado}@cl.uni-heidelberg.de
http://www.cl.uni-heidelberg.de

Abstract. We present a resource-poor approach to automatically acquire Support Verb Constructions (SVCs) for European Portuguese with a two-stage procedure. First, we apply a cross-lingual approach with a bilingual parallel corpus: starting with a Portuguese full verb, we use the translations into another language and the corresponding backtranslations to identify Portuguese verb-noun pairs with the same meaning. Since not all of these are SVCs, the candidates are ranked and filtered in a second, monolingual step based on association statistics. We discuss two parametrisations of our procedure for a high-precision and a high-recall setting. In our experiments, these parametrisations achieve a maximum precision of 91% and a maximum recall of 86%, respectively.

Keywords: lexical acquisition, support verbs, multi-word expressions, parallel bilingual data, word alignment, association measures

1 Introduction

Support Verb Constructions (SVCs), like *dar um passeio* 'to take a walk', are verb-noun complexes which occur in many languages. They form a syntactic and semantic unit and act as a multi-word predicate. Their meaning is mainly reflected by the nominal predicate, while the support verb (SV) is often a semantically impoverished verb, e.g., a light verb [3]. The distinction of SVCs from other complex predicates (CPs) or arbitrary verb-noun combinations is not a simple task. On the syntactic level, the difficulty is that SVCs occur in different forms – e.g. with direct object (*dar esperança* 'to give hope') or prepositional object (*estar na dúvida* 'to be in doubt') – and there are exceptions for most syntactic criteria [1]. Semantically, it is challenging to capture the difference between SVCs and a fully compositional construction in a corpus-driven fashion.

SVCs play a role in many natural language processing (NLP) tasks, such as anaphora resolution. Consider the following mini-discourse from Storrer [25], where the nominal of the SVC acts as antecedent of a pronoun: *One should only* ***provide [assistance]***$_1$ *to the children when they need [it]*$_1$*. [It]*$_1$ *can take the form of questions (...)* This construction would not be possible when the full verb *to assist* is used. Similarly, semantic role labelling works differently for full verbs

(where the verb introduces the event and its dependents are arguments) and for SVCs (where the noun introduces the event and arguments are distributed) [22].

In this paper, we present a two-stage approach for the acquisition of SVC lists for Portuguese, a relatively resource poor language. We presuppose only a part-of-speech (POS) tagger and a parallel corpus. We concentrate on SVCs formed with a direct object, a very productive SVC pattern for Portuguese whose SVCs can often be paraphrased with a full verb [8].

2 Related Work

There are many studies about SVCs and other CPs, ranging from manual linguistic and lexicographic work to automatic NLP-oriented studies. On the manual side, Hanks et al. discuss dictionary representations of SVCs [11]. Hendrickx et al. develop a specific annotation layer for Portuguese SVCs on the CINTIL corpus[1], and carry out studies on the manually annotated data regarding syntactic and semantic aspects [12,7][2]. Cinková et al. take a step towards automatisation by developing a component to extract Swedish SVCs semi-automatically [5].

On the automatic side, Duran et al. use POS patterns to identify CPs in Brazilian Portuguese and extract productive patterns for SVCs [8]. Grefenstette and Teufel extract argument structures for SVs [10] by searching for nominalisations of full verbs, e.g. *to appeal* → *appeal*, and then locating the corresponding SV, e.g. *make + appeal*. Krenn and Evert [14,9] and Wermter and Hahn [27] compare association measures regarding their ability to establish rankings for collocations, including SVCs. Generally, the studies find that the choice of the association measure is crucial, but their performance varies across collocations.

Other studies have used bilingual parallel corpora. Villada Moirón and Tiedemann distinguish literal from idiomatic multi-word expressions (MWEs) [26]. Mukerjee et al. detect Hindi CPs as multi-word units aligned to English verbs in an English-Hindi parallel corpus [17]. Sinha first determines Hindi light verbs employing parallel data and subsequently uses them to retrieve CPs [24]. Bannard and Callison-Burch acquire within-language paraphrases from parallel corpora by observing which expressions share the same translation, which they call *pivot* [2]. Zarrieß and Kuhn apply this idea to the acquisition of MWEs but require dependency parses in both languages [28]. In sum, parallel data can provide strong clues to the identification of MWEs, but comes with problems inherited from the reliance on word alignments (e.g., bad performance for infrequent words).

3 A Two-Stage Strategy for the Acquisition of SVCs

Our goal in this paper is to generate lists of *non-prepositional* SVCs which semantically correspond to a given full verb. Our assumption is that there are full verbs which approximately correspond to the meaning of one or several SVCs, as

[1] `http://catalog.elra.info/product_info.php?products_id=1102`

[2] These annotations could be used in the future to evaluate SVC extraction methods.

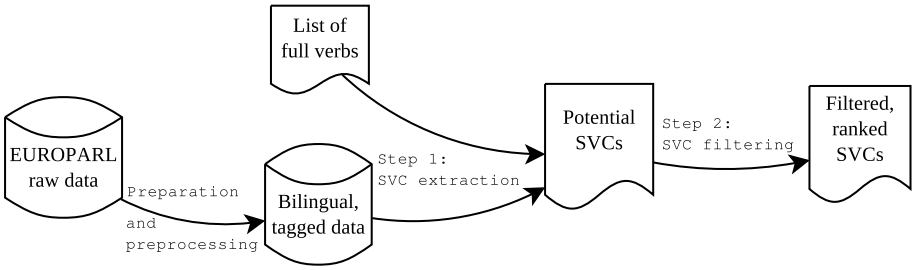

Fig. 1. Overall structure of the SVC acquisition procedure

in *Responda-me!* 'Answer me!' and *Dá-me uma resposta!* 'Give me an answer!' [1]. The resulting lists can be used, for example, to combine statistics collected for different surface forms of the same underlying predicate, or conversely, to generate alternative surface forms for a predicate. To do so, we combine the two main approaches introduced in Section 2: the monolingual and the cross-lingual one:

- From cross-lingual data, we obtain information about *semantic equivalence*, i.e. whether two expressions have (approximately) the same meaning;
- From monolingual data with part-of-speech tags, we obtain information about the *strength of correlation* and *syntactic status* of a given expression.

Combined, these complementary types of information allow us to identify SVCs reliably even in the absence of deeper linguistic analysis, which makes it suitable for languages with few resources like Portuguese.

Figure 1 shows the overall structure of our extraction procedure. The first step is a cross-lingual one, inspired by Bannard and Callison-Burch's proposal to use translations in parallel corpora as pivots for paraphrase extraction [2], adopting their setup specifically to SVCs (cf. Section 3.2 for details). The resulting list contains many SVCs, but also other types of paraphrases that are not SVCs (i.e., that are false positives). The second, monolingual step applies association measures that encode our assumptions about the nature of SVCs to filter out the 'true' SVCs. Our results show that a combination of bi- and monolingual approaches leads to sizable improvements over just the cross-lingual method.

3.1 Data Preparation and Alignment Analysis

For the bilingual step, we use the Portuguese and German (PT–DE) portion of Europarl[3] [13]. We expect that this language pair shows sufficient typological differences so that direct 1-to-1 translation (which would lead to low variation) is unlikely but still close enough so that word alignment is still reliable.

We first align the PT–DE Europarl data using sentence alignment scripts provided on Europarl's web site[4] and the word alignment toolkit GIZA++ [18].

[3] Version 3 (September 2007)
[4] `http://www.statmt.org`

The word alignment is subsequently symmetrised into a single alignment with the 'grow-diag heuristic' [19]. Then we conduct POS tagging and lemmatisation. For Portuguese, we use FreeLing, version 2.2 [4,20] and TreeTagger [23] for German. We take special care of retokenisation issues occurring with FreeLing, i.e. decomposition of contractions and composition of MWEs. These procedures leave us with a parallel corpus of 982,039 sentence pairs.

Qualitative Evaluation of the Alignment. Our bilingual step retrieves Portuguese SVCs through word alignments to German. For pairs of full verbs and SVCs, this involves 1-to-n alignments, which are notoriously unreliable (Zarrieß and Kuhn [28] fall back to syntactic information for this reason). To assess the quality of these alignments, we perform a manual analysis of typical alignments between Portuguese full verbs and their German translations (1-to-1 and 1-to-n). We concentrate on the Portuguese verbs *apoiar* 'to support', *perguntar* 'to ask' and *ler* 'to read' since they are expected to lead to synonymous SVCs. We extract 17,943 sentences, each containing at least one of these full verbs.

We first consider the effect of alignment symmetrisation. It establishes many links which previously do not exist in at least one of the unidirectional alignments, e.g. *apoiar* $\rightarrow \emptyset$ becomes *apoiar* $\rightarrow$ *Beihilfe* 'aid'. Although it also leads to unnecessarily or incorrectly aligned tokens, filling these alignment gaps is strongly desirable. We count 22.9% differences between the symmetrised and the unidirectional alignment for the three full verbs mentioned above. In 10.6%, an alignment is created for an unaligned token. Since after symmetrisation, over 97% of the Portuguese full verbs are aligned with one to four German words and most remaining instances are wrong, we disregard all 1-to-5 (or more) alignments.

We then analyse the translation and word alignment patterns that we find between full verbs and SVCs. Full verbs are often translated as full verbs, mirrored in an 1-to-1 alignment. In the case of a translation as an SVC, the full verb is mostly aligned with the SVC's noun, e.g. *fragen* 'to ask' $\rightarrow$ *pergunta* 'question'. The SVC's verb frequently remains unaligned, which means that one cannot easily effect a large-scale SVC extraction solely from word alignments. The situation is similar for SVC-SVC translations. In most cases, the noun of one SVC is aligned, either to the noun of the corresponding SVC or with the whole SVC. In contrast, the semantically impoverished SV often remains unaligned or is aligned to an SV in the other language. For example, the noun *pergunta* in *fazer uma pergunta* 'to ask a question' is *always* aligned (in 77.5% of cases to a noun), whereas the support verb *fazer* is unaligned at 21.1% and aligned to a verb at 63.9%.

In terms of relative frequencies, about 30% of 1-to-n alignments align a Portuguese verb with a noun-verb combination, i.e., SVC candidates. Most of the remaining 1-to-n alignments are either rejected (since $n > 4$) or due to the fact that Portuguese verbs can incorporate more information than German (as well as English) verbs. E.g. they incorporate person information which must be added in German by a personal pronoun, leading to an 1:2 alignment.

In sum, this analysis suggests that if there is a proper SVC equivalent for a full verb, there are enough and reliable alignments to reveal this equivalence.

Unfortunately, we cannot straightforwardly use them to acquire *complete* SVCs. However, the frequent alignments between full verbs and the SVC nouns can serve as a starting point: a heuristic extension of these alignments can be hoped to improve the retrieval of SVCs. Thus, the acquisition of SVCs, starting from a full verb, is reasonable and promising, even though some effort additionally to the automatic alignment is necessary. We return to this point in Section 3.2.

3.2 Step one: Bilingual SVC Extraction

Our cross-lingual SVC extraction method is an adaptation of Bannard and Callison-Burch's pivot approach for paraphrase extraction [2]. We start with a quick review of their method, using s to denote source language phrases and t for target language phrases. Their algorithm takes as input an initial phrase s_1 (to be paraphrased). It then locates all target language phrases t aligned with s_1 (*first pivot step*). Next, it gathers all instances of the t phrases and collects their backtranslations into the source language, resulting in a list of source phrases s_2 (*second pivot step*). An example for the language pair English–German: the initial phrase s_1 = *under control* is aligned with t = *unter Kontrolle*, which is backtranslated into s_2 = *in check*. Assuming that a translation is (largely) meaning-preserving, the source language phrases s_2 are considered as candidate paraphrases for s_1 and ranked using probabilities based on relative frequency. An extended version of the model that included word sense disambiguation achieved 70.4% accuracy in an evaluation for correct meaning for English–German.

We apply this model to full verbs as the inputs s_1. For our purpose, we however believe that it makes sense to concentrate on two different parameters of the model than those investigated in detail by Bannard and Callison-Burch.

Occurrence Thresholds. First, instead of using probabilities, we apply some simple occurrence thresholds which indicate how many times an alignment pair must occur to be considered. They are sufficient to counteract the effect of misalignments and overly context-specific translations, both of which are rather infrequent. We use four different thresholds: two each for the first and the second pivot step, respectively. Since there are 1-to-1 as well as 1-to-n translations, both pivot steps contain unigrams (single words) and n-grams (multiple words). We require n-grams to occur at least 6 times in the first and 9 times in the second pivot step. Unigrams are naturally more frequent than n-grams, so that we define a higher threshold for them, i.e. 300 in the first pivot step, and exclude them completely in the second pivot step, for SVCs always consisting of two or more words. Unlike Zarrieß and Kuhn, we do not encounter the problem of losing many n-grams by virtue of these thresholds [28]. Instead, this restriction reliably rejects many arbitrary verb-noun combinations, while not overly lowering recall.

Word Alignment Extension. Our analysis in Section 3.1 has shown that symmetrised alignments provide translations for almost all full verbs and the nominal parts of SVCs but are incomplete with regard to the SVs themselves.

Since it is reasonable that the cross-lingual step should focus on recall – precision can be increased in the subsequent filtering step, if desired – we will focus exclusively on the symmetrised word alignment rather than the unidirectional ones. Furthermore, we strive to further extend the word alignment to support verbs using linguistically motivated rules.

To be able to phrase these rules concisely, we focus on word alignments between parts of speech that are supposed to participate in SVCs, i.e. nouns and verbs (recall that we ignore prepositional SVCs), discarding all others.[5] This leaves us with word alignments of the three following basic structures:

(1) X→ NOUN + VERB (2) X→ VERB (3) X→ NOUN

Alignments of type (1) are already complete. Correct alignments of type (2) occur almost exclusively in the first pivot step and connect Portuguese and German full verbs. Thus, they did not yet lead to extracted SVCs but there is a chance to find an SVC in the second pivot step. Alignments of type (3) are expanded in both directions (i.e., for both pivot steps). The expansion procedure is as follows: if a token X is 1-to-1-aligned with a single noun N, check the tokens in the neighbourhood of N. This neighbourhood is defined as 3 following tokens in German and 6 preceding tokens in Portuguese, respectively, reflecting the different syntactic structures in the two languages: while Portuguese has a rather strict word order and a broader neighbourhood can be considered, the German word order is more flexible; to avoid spurious extensions of N, we consult only a narrow word window. If a verb V occurs within this window, add V to the alignment. We assume that prepositional phrases cannot be inserted into an SVC[6] and that SVCs cannot split across sentences. Hence, the search is stopped after the closest verb is found or as soon as a preposition or a sentence boundary is reached. Finally, we added one lexical restriction: for Portuguese, we exclude occurrences of the verb *ser* ('to be'); according to the literature, *ser* does not form SVCs with direct objects, but it frequently occurs in the corpus.

An exemplary analysis shows that this heuristic increases the recall as intended. We even encounter unexpected SVCs, e.g. *dar assistência* 'to assist' for *apoiar* 'to support'. However, many false positives remain, since the pivoting extracts not only synonymous SVCs but also their antonyms, e.g. *exigir apoio* 'to demand support' for *apoiar*. The second step, filtering, attempts to eliminate these errors.

3.3 Step two: SVC Filtering with Association Measures

As stated above, the purpose of the monolingual filtering is to increase the precision of the SVC candidate list created by the cross-lingual extraction step.

There are at least two possible approaches to this task: either with linguistic heuristics or statistically. In line with our strategy in Step 1, we first adopted a linguistically informed strategy that checked whether extracted candidates were likely paraphrases for the initial full verbs. The goal of our strategy was

[5] If more than one verb or noun are co-aligned, only the first hit is kept.

[6] This is an oversimplification, but serves successfully to identify clear true positives.

to first detect the candidates' arguments through POS patterns which typically surround SVCs, and then to compare the candidates' argument heads with the argument heads for the full verbs. Very similar arguments indicate similar meaning [15], which we would expect for SVCs but not for compositional noun-verb combinations. Unfortunately, we found that the actual corpus occurrences of the SVC candidates showed too much variance, and we were unable to make reliable decisions based on the shallow linguistic information available to us.

We therefore adopted a statistical approach, more specifically one based on association measures (AMs). AMs model the common information of two words, that is, how predictable one word is given the other. We expect that SVCs will be recognisable by a predictability between verb and noun that is higher than for compositional verb-noun combinations.

The rest of this section discusses the two main design decisions for this step. The first one is the choice of association measures. A number of AMs have been investigated by Krenn and Evert [14], among which *(relative) frequency*, *pointwise mutual information (PMI)* and *student's t-test*. We decided to experiment with these three measures, the latter two of which are defined as:

$$\text{PMI} = \frac{p(v,n)}{p(v)p(n)} \qquad \text{t-test} = \frac{p(v,n) - p(v)p(n)}{\sqrt{s^2/N}}$$

where $p(v,n)$ is the *observed* co-occurrence probability of verb and noun and $p(v)p(n)$ can be interpreted as the *expected* co-occurrence probability. s^2 is the sample variance and N is the sample (corpus) size. Not surprisingly, all AMs involve co-occurrence frequencies. We only count directly adjacent noun-verb co-occurrences, since we found that intervening words degrade results.

The second design decision is the optimisation of the filtering step for either precision or recall. We indicated in Section 3.2 that the filtering step can be used to improve precision, which corresponds to aggressive filtering. However, for some settings (e.g. for manual post-processing), it might be better to filtering only leniently in order to keep recall high. We define two settings: a high-recall setting (*hiRec*) and a high-precision setting (*hiPrec*). The parameters of the filtering procedure that are varied between the two settings are as follows:

- Since many AMs are known to be oversensitive to low-probability (i.e., unreliable) events, we introduce a minimum verb-noun co-occurrence threshold and discard unfrequent pairs. Specifically, we set it to 2.5 co-occurrences per million words for *hiPrec* and to 1 for *hiRec*.
- Other studies show that there are two categories of SVCs [8,25]: The first one consists of SVCs where the SVs are light verbs, which have a very high context diversity, e.g. *dar apoio* 'to give support', *dar resposta* 'to give an answer', *dar um passo* 'to take a step'. The second category contains SVCs of nearly idiomatic meaning where the SV has a very low context diversity. An example of this type is *correr um risco* 'to run a risk', whose SV *correr* 'to run' occurs in no other verb-noun pair with a co-occurrence frequency >2.5 per million words. In contrast, verbs which cooccur with an average number of nouns are not likely to be part of an SVC. To capture this fact, we

compute the 'diversity' for each verb as the number of different noun lemmas it occurs with in the complete corpus. For the *hiPrec* setting, we retain only those SVC candidates whose diversity is either 1, or higher than the median diversity. For *hiRec*, no filtering takes place.

4 Evaluation

4.1 Creating the Gold Standard

To evaluate our approach, we need a gold standard of SVCs. Since it is impossible to determine how many 'gold SVCs' exist for a given full verb, we took the output of the cross-lingual step to be the basis for manual annotation. Against this gold standard, we can compute precision and (relative) recall, i.e., recall relative to the extraction procedure as defined in Pantel et al. [21].

For the annotation, we concentrated on six Portuguese full verbs: *ameaçar* 'to threaten', *apoiar* 'to support', *faltar* 'to lack', *perguntar* 'to ask', *prometer* 'to promise' and *responder* 'to answer'. Each of them has approximately the same meaning as at least one SVC. The retrieved candidate expressions were annotated by two native speakers with professional linguistic knowledge, judging for each expression *i)* whether it was an SVC and *ii)* whether it semantically corresponded to the initial full verb. A total of 84 candidate SVCs have been annotated, ranging from 1 to 64 expressions per verb. The main criterion provided to the annotators was whether the verb can be interpreted as a semantically impoverished SV in the given expression.

We computed inter-annotator agreement (IAA) with Cohen's κ [6] and obtained a value for *i)* of 0.60 and for *ii)* of 0.74. The first κ value is lower than the second one because the decision if an expression is an SVC or not is more general and thus more difficult. These are fairly good IAA rates, regarding the fact that SVC determination is a difficult task because of the fuzziness of SVs [10]. Since other SVC acquisition studies either do not provide IAA rates or have a fairly different setting, we cannot compare our IAA rates. However, Landis and Koch consider these rates as moderate and substantial, respectively [16]. The final gold standard was formed from the intersection of the two annotations, and for all cases in which the evaluators did not agree, we classified the expression by ourselves. This procedure leads to 22 SVCs judged as true positives.

4.2 Results after the Extraction Step

Table 1 shows the results of the cross-lingual extraction step.[7] As noted in section 4.1, the list of candidate SVCs resulting from the pivoting serves as basis for the gold standard. Hence, recall is always 100%. However, precision varies considerably between verbs: the SVC lists for *ameaçar* and *faltar* are already perfect, which speaks to the efficacy of our alignment extension (Section 3.2), but

[7] All results presented in Section 4 refer to our automatically processed corpus. Unfortunately, no numbers are available on the quality of the preprocessing components.

Table 1. Results for the extraction step

	ameaçar	apoiar	faltar	perguntar	prometer	responder	all
Precision	1.00	0.16	1.00	0.71	0.33	0.43	**0.26**
Recall	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**
F_1	1.00	0.27	1.00	0.83	0.50	0.60	**0.42**

Table 2. Overall results of the two-step procedure

	PMI		*Frequency*		*t-test*	
	hiPrec	hiRec	hiPrec	hiRec	hiPrec	hiRec
Precision	**0.91**	0.61	**0.91**	0.61	0.90	0.60
Recall	0.45	**0.86**	0.45	**0.86**	0.41	0.81
F_1	0.61	**0.72**	0.61	**0.72**	0.56	0.69

the results for other verbs are far from perfect. *apoiar* is especially bad: while the other verbs lead to a maximum of 7 candidate SVCs, *apoiar* results in 64 candidates with many false positives. This outlier seems corpus-specific: *apoiar* is strikingly more frequent in EUROPARL than the other full verbs, and two commonly aligned nouns, *apoio* 'support' and *ajuda* 'help', are very frequent as well. The verb-noun pairs in which they occur are often arbitrary, e.g. *encontrar apoio* 'to find support', albeit frequent enough to overcome the thresholds defined in section 3.2. Thus, many false positives slip into the results of Step 1. This also explains the rather low overall precision and f-scores. In sum, the quality of the results of the cross-lingual step depends on the properties of the initial verb.

4.3 Final Results Including the Filtering Step

Recall from Section 3.3 that the filtering step had two parameters: the choice of the AM measure (*PMI*, *frequency* and *t-test*) and the choice between a high-recall and a high-precision setting. Table 2 shows the results for these combinations. We discuss both parameters in turn.

High Precision vs. High Recall. The figures in Table 2 indicate that the filtering step indeed improves substantially over the results of the cross-lingual extraction step: from an f-score of 0.42, we reach an f-score of 0.72 in the optimal case, corresponding to an error reduction of 50%. The Table also demonstrates that the filtering step can be tuned to the requirements of a particular setting. If high precision is required, the filtering mechanisms we introduced can produce a precision of above 90%, at the cost of a recall of slightly below half. At the same time, the high recall setting can still substantially improve precision (from 26% to 61%) within a rather small loss in recall (from 100% to 86%).

Consider the the verb *perguntar* 'to ask'. The *hiPrec* setting correctly retrieves the SVC *fazer pergunta*. For the *hiRec* setting, the following expressions are found:

fazer pergunta, *levantar questão*, *colocar pergunta*, *colocar questão*, *apresentar pergunta*, and *formular pergunta*. According to our gold standard, only the last expression is a false positive. All SVCs contained in the gold standard are found.

Association Measures. Krenn and Evert [14] did not find any single measure to consistenly outperform the others across all tested collocations. For SVCs, *t-test* and *frequency* worked best, while *PMI* performed poorly, and the authors even suggested to use a modified version of PMI.

In contrast, on our data *PMI* performs very well and does not show the idiosyncrasies observed by Krenn and Evert. It shows essentially identical results to *frequency* in a precision/recall evaluation, while *t-test* performs consistently worse. We also evaluated the lists with average precision, i.e., took the ranking within the lists into account (not shown in the tables). In that case, *PMI* substantially outperforms *frequency* with an AP of 0.33 compared to 0.11 for *frequency*. This indicates that *PMI* does a better job at ranking.

We attribute this difference to the fact that Krenn and Evert re-rank a list of all verb-noun combinations from a corpus, while we only consider the candidates extracted by the cross-lingual step, which are typically located within a fairly narrow range for all AMs. We see this as a further validation of our two-step approach, dividing the work between the cross-lingual alignment-based and the monolingual association-based approach. In sum, the joint application of mono- and cross-lingual methods leads to a very satisfactory overall result.

5 Conclusions and Outlook

This paper has presented a resource-poor two-stage approach to acquire Support Verb Constructions, applied to the Portuguese language. We explored whether cross-lingual techniques are suitable for the extraction of syntactically correct SVCs which semantically correspond to a given full verb, and whether monolingual methods can further improve the cross-linguistically obtained results.

Within the limits of our evaluation, our results indicate that this is indeed the case: word alignment-based extraction is perfectly applicable to the SVC acquisition task without the need for complex preprocessing, while the computation of association measures is capable of ranking and refining the expressions found in the first step. Our approach provides adjustment possibilities for both solid precision and recall values, depending on which focus the user intends.

The main caveat of our approach is that it depends crucially on acquiring reliable translations for the initial full verb. Full verbs which occur in heterogeneous contexts and are translated in many different ways will give rise to noisy candidate lists which cannot be re-ranked successfully. In future work, we plan a corpus-based evaluation (using CINTIL [12]) on a larger number of full verbs, assessing also the distribution of the SVs involved in SVCs.

Another direction for future research is the generalisation of our method to *prepositional* SVCs or a large-scale acquisition of different CPs. This will presumably require better extraction and filtering methods.

References

1. Athayde, M. F.: Construções com Verbo-suporte (Funktionsverbgefüge) do Português e do Alemão. In: Cadernos do cieg, vol. 1, pp. 5–68 (2001)
2. Bannard, C., Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 597–604, Ann Arbor, MI (2005)
3. Butt, M.: The Light Verb Jungle. In: Harvard Working Papers in Linguistics, vol. 9, pp. 1–49 (2003)
4. Carreras, X., Chao, I., Padró L., Padró, M.: FreeLing: an Open-Source Suite of Language Analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004)
5. Cinková, S., Pecina, P., Podveský, P., Schlesinger, P.: Semi-automatic Building of Swedish Collocation Lexicon. In: Proceedings of the 5th Conference on International Language Resources and Evaluation, Genoa, Italy (2006)
6. Cohen, J.: A Coefficient of Agreement for Nominal Scales. In: Educational and Psychological Measurement, vol. 20(1), pp. 37–46 (1960)
7. Duarte, I., Gonçalves, A., Miguel, M., Mendes, A., Hendrickx, I., Oliveira, F., Cunha, L. F., Silva, F., Silvano, P.: Light Verbs Features in European Portuguese. In: Proceedings of the 2nd Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Pisa, Italy (2010)
8. Duran Sanches, M., Ramisch, C., Aluísio, S. M., Villavicencio, A.: Identifying and Analyzing Brazilian Portuguese Complex Predicates. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, pp. 74–82, Portland, USA (2011)
9. Evert, S., Krenn, B.: Methods for the Qualitative Evaluation of Lexical Association Measures. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 188–195, Toulouse, France (2001)
10. Grefenstette, G., Teufel, S.: Corpus-Based Method for Automatic Identification of Support Verbs for Nominalizations. In: Proceedings of European Chapter of the Associaton of Computational Linguistics, pp. 98–103, Dublin, Ireland (1995)
11. Hanks, P., Urbschat, A., Gehweiler, E.: German Light Verb Constructions in Corpora and Dictionaries. In: International Journal of Lexicography, vol. 19(4), pp. 439–457 (2006)
12. Hendrickx, I., Mendes, A., Pereira, S., Gonçalves, A., Duarte, I.: Complex Predicates Annotation in a Corpus of Portuguese. In: Proceedings of the 4th ACL Linguistic Annotation Workshop, pp. 100–108, Uppsala, Sweden (2010)
13. Koehn, P.: Europarl: a Parallel Corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit, pp. 79–86, Chiang Mai, Thailand (2005)
14. Krenn, B., Evert, S.: Can We Do Better than Frequency? A Case Study on Extracting PP-Verb Collocations. In: Proceedings of the ACL Workshop on Collocations, Toulouse, France (2001)
15. Lin, D., Pantel, P.: Discovery of Inference Rules for Question Answering. In: Journal of Natural Language Engineering, vol. 7:(4), pp. 343–360 (2001)
16. Landis, J. R., Koch, G. G.: The Measurement of Observer Agreement for Categorical Data. In: Biometrics, vol. 33(1), pp. 159-174 (1977)
17. Mukerjee, A., Soni, A., Raina, A. M.: Detecting Complex Predicates in Hindi Using POS Projection across Parallel Corpora. In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pp. 28-35, Sydney, Australia (2006)

18. Och, F. J., Ney H.: A Systematic Comparison of Various Statistical Alignment Models. In: Computational Linguistics, vol. 29(1), pp. 19–51 (2003)
19. Och, F. J., Tillmann, C., Ney, H.: Improved Alignment Models for Statistical Machine Translation. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 20-28, College Park, MD (1999)
20. Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I.: FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In: Proceedings of the 7th Conference on International Language Resources and Evaluation, Valleta, Malta (2010)
21. Pantel, P., Ravichandran, D., Hovy, E.: Towards Terascale Knowledge Acquisition. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 771–777, Geneva, Switzerland (2004)
22. Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. `https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf` (2010)
23. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994)
24. Sinha, R. M. K.: Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pp. 40–46 , Singapore (2009)
25. Storrer, A.: Corpus-based Investigations on German Support Verb Constructions. In: Fellbaum, Christiane (ed.): Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects, pp. 164-188, London
26. Villada Moirón, B., Tiedemann, J.: Identifying Idiomatic Expressions Using Automatic Word-Alignment. In: Proceedings of the EACL Workshop on Multiword Expressions in a Multilingual Context, Trento, Italy (2006)
27. Wermter, J., Hahn, U.: Collocation Extraction Based on Modifiability Statistics. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland (2004)
28. Zarrieß S., Kuhn, J.: Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pp. 23–30, Singapore (2009)