This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

A Hybrid Approach for Multiword Expression Identification

Carlos Ramisch[1♣], Helena de Medeiros Caseli[♢], Aline Villavicencio[♣♠], André Machado[♣], Maria José Finatto[♡]

[1]GETALP/LIG, University of Grenoble (France)
[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♢]Department of Computer Science, Federal University of São Carlos (Brazil)
[♠]Department of Computer Sciences, Bath University (UK)
[♡]Institute of Language and Linguistics, Federal University of Rio Grande do Sul (Brazil)
ceramisch@inf.ufrgs.br, helenacaseli@dc.ufscar.br,
{avillavicencio,ammachado}@inf.ufrgs.br, mfinatto@terra.com.br

Abstract. Considerable attention has been given to the problem of Multiword Expression (MWE) identification and treatment, for NLP tasks like parsing and generation, to improve the quality of results. Statistical methods have been often employed for MWE identification, as an inexpensive and language independent way of finding co-occurrence patterns. On the other hand, more linguistically motivated methods for identification, which employ information such as POS filters and lexical alignment between languages, can produce more targeted candidate lists. In this paper we propose a hybrid approach that combines the strenghts of different sources of information using a machine learning algorithm to produce more robust and precise results. Automatic evaluation on gold standards shows that the performance of our hybrid method is superior to the individual results of statistical and alignment-based MWE extraction approaches for Portuguese and for English. This method can be used to aid lexicographic work by providing a more targeted MWE candidate list.

1 Introduction

Recent research on Multiword Expressions (MWEs) has devoted considerable attention to their identification. One of the problems that these works address is that MWEs can be defined as combinations of words that have idyosincrasies in their lexical, syntactic, semantic, pragmatic or statistical properties [1], such as idioms (*make ends meet*), phrasal verbs (*find out*), light verbs (*give a speech*) and compounds (*mother nature*). However, MWEs are very numerous in languages accouting for between 30% and 45% of spoken English and 21% of academic prose [2], and having the same order of magnitude in a speaker's lexicon as the number of single words [3]. Moreover, if we consider that new MWEs are also constantly coined (*credit crunch*), and that for language from a specific domain

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

the specialized vocabulary is going to consist largely of MWEs (*chromosomal mutation*), these estimates are likely to be conservative underestimates.

In this context, especially for NLP tasks that involve some kind of semantic processing, it is important to adequately identify and treat MWEs, as failing to do so may cause serious problems [1]. For example, in order to avoid the generation of unnatural sentences, a Machine Translation system must translate the idiom *to kick the bucket* differently in a sentence like *He was only 39 years old when he kicked the bucket* (meaning *to die*) than in *The janitor kicked the bucket with water*.

Therefore, there is a need for robust (semi-)automatic ways of acquiring lexical information for MWEs that can contribute to improving the quality of NLP systems. In this context, a number of methods for identifying MWEs from corpora have been proposed. For this task they employ information that ranges from purely statistical Association Measures (AMs), to more linguistically-based, such as e.g. Part-of-Speech (POS) patterns, with varying degrees of success [4, 5].

While the former can retrieve a large list of multiword units, more linguistically motivated methods for MWE identification, such as those based on POS filtering or lexical alignment, on the other hand, may result in a more accurate list of candidates.

After evaluating statistical AM and alignment-based approaches separately in previous work [6, 7], in this paper we investigate their weighted combination aiming at a more robust method that could output a more accurate set of MWE candidates than those of the individual methods. The proposed approach can be used to aid lexicographic work by providing a more targeted MWE candidate list to keep lexical resources up to date and also to improving the quality of NLP systems.

The remainder of this paper is structured as follows. In section 2 we briefly discuss MWEs and their identification. Section 3 presents the materials used in our experiments while section 4 describes the hybrid method proposed to extract MWEs. Section 5 presents the results and section 6 finishes this paper with some conclusions and proposals for future work.

2 Related Work

MWEs have been the focus of both linguistic and computational work, and they have proved to be a difficult problem to tackle from either field [1]. The different phenomena that are defined as MWEs form a very heterogeneous group, with phrasal verbs, idioms, compounds, among others, each with its particular characteristics. Moreover, even within a single MWE type there is considerable variation in their possible linguistic realizations. Verbal idioms, for example, vary in terms of morphosyntactic and semantic flexibility from more rigid combinations (*kick the bucket*) to more flexible ones (*touch a nerve*). As a consequence MWEs defy attempts to capture them uniformly.

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

Due to the tight connection between the elements of an MWE and their co-occurrence patterns, AMs have been often used to identify them [4, 8], as they are sensitive to such patterns. Since we expect the component words of an MWE to occur frequently together, these measures can give an indication of whether a sequence of words is a MWE. The advantage of using them in the identification of MWEs is that AMs are an inexpensive language and type independent means of detecting recurrent patterns and can be democratically applied to any language and MWE type. The effectiveness of these methods seem to depend on the MWEs themselves (e.g. type, syntactic flexibility) [8], in characteristics of the corpus used (e.g. size and domain) [4], and on the gold standard used for evaluation [7].

A number of these works have also combined these measures with linguistic information such as syntactic and semantic properties of the MWEs [9, 8] or automatic word alignment [10]. Fazly, Cook and Stevenson [8], for instance, use properties like lexical and syntactic flexibility in statistical measures for verb-noun idiom identification. Ramisch et al. [11] combine standard statistical measures with information about syntactic flexibility using a supervised machine learning approach for the identification of Verb-Particle Constructions (VPCs).

Some work has looked for evidence from other language for MWE identification. For instance, Melamed's proposal for the automatic detection of non-compositional compounds (NCC) [12] is based on the idea that their translation to another language does not usually correspond to their word-for-word literal translation. This method can successfully identify many NCCs, but it does not use monolingual information about possible NCCs within a language. The work of Villada Moirón and Tiedemann [10] seems to be the most similar to the approach proposed in this paper. Their method looks at the automatically generated translations of MWE candidates assuming that the translations of idiomatic expressions would be less predictable and less compositional than the non-idiomatic cases. However, while their method uses the alignment information just for ranking the MWE candidates, in this paper, the word alignment is the basis of MWE extraction process.

In this paper we investigate the combination of several sources of information for the identification of MWEs. In particular, we propose that the combination of statistical AMs and alignment-based information has a positive effect on the performance of this task, as they each capture different aspects of MWEs. We also evaluate thoroughly their contributions to the overall performance, looking at factors, such as language and size of the ngram, that influence these results.

3 The Corpus and Reference Lists

For our experiments, we used the Corpus of Pediatrics [13], a parallel corpus composed of 283 pairs of texts in Portuguese (785,448 words) and their translations to English (729,923 words) extracted from the *Jornal de Pediatria*[1]. We

[1] www.jped.com.br

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

use a parallel corpus to evaluate the MWE identification for these two different languages, Portuguese (pt) and English (en), and also to investigate whether the choice of language influences the results obtained.

Our automatic evaluation process uses the Pediatrics Glossary, a domain-specific resource built semi-automatically from the Corpus of Pediatrics for supporting translation studies.[2] The Portuguese Glossary was constructed by first extracting all ngrams (with n ranging from 2 to 4) from the texts which occurred at least 5 times in the corpus, then applying a POS filter to exclude candidates beginning with Article + Noun and beginning or finishing with Verbs and, finally, manually verifying the remaining entries. Subsequently an enrichment process was performed as described in [14] to include all the valid bigrams contained in the trigrams and removed during the construction of the Glossary. The English Glossary was built by a similar process with translations of the ngrams in the Portuguese Glossary. The final versions of the gold standards have 2,150 terms in Portuguese and 883 terms in English [3]. Due to the smaller number of entries in the English Glossary, we also considered as true positives the candidates contained in a general dictionary, such as the Cambridge International Dictionary of Idioms [15], and these two sources are marked in table 1 as specialized and generic respectively.

	Specialized	Generic	Total
pt	2150	—	2150
en	883	1382	2190

Table 1. Number of reference entries in each gold standard.

4 MWE extraction methodology

In this paper we propose a hybrid method that combines two independent approaches for MWE identification using a Bayesian network classifier. The first approach applies well-know AMs to all the bigrams and trigrams generated from each corpus: Pointwise Mutual Information (PMI), Mutual Information (MI), t-score, χ^2, Dice coefficient, Fisher's exact test, Poisson-Stirling measure (PS) and Odds ratio, as implemented in the Ngram Statistics Package [16].

The second one, the alignment-based approach, is based on the automatic lexical alignment of Portuguese and English versions of the Corpus of Pediatrics generated by the statistical word aligner `GIZA++` [17]. The hypothesis is that when the lexical aligner encounters a sequence in the source language that cannot be resolved by aligning the target words individually, this sequence is taken to

[2] `www6.ufrgs.br/textquim/Dicionarios/DicPed`
[3] `www.inf.pucrs.br/~ontolp/downloads-ontolplista.php`

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

statistical						
	no filter			filters		
	$n=2$	$n=3$	Total	$n=2$	$n=3$	Total
pt	244420	513494	757914	11290	4553	15843
en	230130	492154	722284	10311	4526	14837
alignment-based						
	no filter			filters		
	$n=2$	$n=3$	Total	$n=2$	$n=3$	Total
pt	15333	7373	22706	12154	5518	17672
en	16345	7469	23814	12222	5154	17376
statistical ∩ alignment-based (filters)						
	$n=2$		$n=3$			Total
pt	1376		134			1510
en	1921		109			2030

Table 2. MWE candidates per method, language and ngram size

be a MWE candidate. Thus, the alignment-based approach considers as MWE candidates the sequences of two or more consecutive source words joined by the aligner regardless of whether they are aligned with one or more target words.

The original corpora were POS tagged using the `Apertium`[4] tools [18] with augmented lexicon [19]. Morphological information was used to filter the candidate lists of both approaches. The filters were applied uniformly, removing:

- punctuation, numbers and special characters (dashes, slashes, brackets,...);
- candidates below a certain threshold (5 occurrences in corpus for AMs, and 5 occurrences as alignment for alignment-based method);
- candidates starting with function words (determiners, auxiliary verbs, pronouns, adverbs, conjunctions, forms of the verb *to be* and prepositions *from, to* and *of*. In this we follow Caseli et al. [20] who found that these patterns are effective for filtering out noise, without removing many false positives.

The corpora were then independently given as input to each of the approaches, and as a result, two lists of MWE candidates for each language were generated. Table 2 shows the number of original candidates extracted for each language before and after filtering. Both languages have about the same number of candidates for each approach, and filtering considerably reduces the candidate lists, especially for the AMs. The last section of table 2 shows the intersection between the two methods, which indicates that their candidates are essentially different: less than 15% of the candidates extracted by the alignment-based method are also captured by the statistical method and vice versa. One difference between the approaches is that the alignment-based approach is able to extract non-contiguous sequences. Therefore when it detects an ngram from the source

[4] `Apertium` is an open-source machine translation engine and toolbox available at `http://www.apertium.org`.

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

ngram	align	statistical								Class
		Dice	Odds	PMI	PS	t-score	MI	χ^2	Fisher	
abnormal findings	Yes	.03	114.1	6.74	25.70	2.62	0	734.73	0	No
adrenal insufficiency	No	.46	10376	11.6	371.9	7.28	.0008	160784	0	Yes
óxido nítrico	Yes	.95	8553397	14.5	289.3	5.66	.0006	733177	0	Yes
academia americana	No	.52	74302	13.3	197.4	4.9	.0004	244244	0	No

Table 3. Sample of the English training set.

aligned with an ngram in the target language (an n:m alignment), if there are intervening words between the two ngrams, the candidate will not include them. For example, from *a mild pain* and *a characteristic pain* aligned with the Portuguese *uma discreta dor* and *uma dor característica* the aligner proposes *a pain* as a candidate even though these two words never occur adjacently in the English corpus, and consequently the statistical method does not propose them as a candidate ngram.

For combining the different methods, a classifier was constructed using the Weka package [21]. The input for each language was the set of filtered ngrams (15,843 for Portuguese and 14,837 for English) annotated with the values of the statistical measures and the judgement of the lexical aligner as to whether the ngram is a possible MWE candidate. Table 3 shows some examples of English and Portuguese entries from the training set. As discussed in the next section, the data sets are unbalanced, with a much larger proportion of non-MWEs than MWEs. Therefore, a Bayesian Network classifier is used to combine the different approaches, since it has been found to be robust and less sensitive to highly unbalanced classes.[5]

5 Experiments and Results

We evaluate the efficacy of the combined approach for MWE identification (from §4) in a domain-specific corpus using the gold standards for each language (§ 3). The results are reported in terms of precision (#correct vs #proposed candidates), recall (#correct vs #candidates in gold standard) and F-measure $((2 * precision * recall)/(precision + recall))$.

The baseline for comparison is obtained by evaluating the individual approaches independently. Table 4 shows the number of True Positives (TPs) in each candidate list considering both the MWEs in the specialized Portuguese and English Pediatrics Glossaries (pt_spec and en_spec) and those in the general English Dictionary (en_spec+gen), while table 5 shows the Mean Average Precision of each AM taken independently, which range from 11.23% for Fisher's exact test to 55.83% for Odds ratio.

[5] Using, e.g. decision trees on the English data generated a single class model guessing "No" for all candidates.

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

In the results for Portuguese the statistical approach captures 86.14% of the MWEs in the text with a precision of 11.69%, while for English only 68.06% of the true instances are captured with a precision of 4%. The differences in the results for these languages can be explained to a large extent due to the differences in coverage of the gold standards, with the Portuguese Glossary containing a much larger number of entries. Indeed, using the extended English gold standard with both specialized and generic MWEs improves the F-measure for both approaches, and this extended resource is adopted in the subsequent evaluations. The alignment-based method has a lower performance partly due to the larger number of candidates to consider (table 2). Although a higher alignment frequency threshold could considerably improve the precision of the aligner [7], more restrictive filters were not applied because we wanted to investigate how much the combination of these methods can filter out the noise in each of the candidates lists.

The results of the Bayesian network classifier using different feature sets and 10-fold cross validation are shown for each language in table 6. For both languages the hybrid model is able to generate much better candidates than the individual methods: e.g. for Portuguese, the Bayesian classifier yields an F-measure of around 50% against 20.59% and 1.78% for the statistical and alignment-based methods, respectively.

To evaluate the contribution of the individual methods to these results we consider four different feature sets: (a) a subset of the Association Measures (subAM), namely PMI, PS and MI, which do not involve the construction of contingency tables and can be straightforwardly applied to ngrams of arbitrary size, (b) the combination of this subset of AMs and the alignment-based approach (subAM + align), (c) all AMs (allAM) which includes subAMs for bigrams and trigrams, and the other AMs only for bigrams (since they rely on contingency tables) and (d) the combination of all AMs and the alignment feature (allAM + align). In terms of individual features, the aligner only improves the performance in subAM for English, where it provides enough extra information for an increase in performance from 0% (subAM) to 4.91% (subAM + align) . This suggests that the alignment information helps to add robustness to the process. To further evaluate the contribution of this feature, we built a decision tree with the same training sets and in the resulting trees the alignment feature is only used after PMI, PS and Dice, which seem to be better predictors of MWEs. The addition of further AMs seems to also have the effect of providing more robustness to

	statistical				alignment-based			
	TP	Precision	Recall	F-measure	TP	Precision	Recall	F-measure
pt_spec	1852	11.69%	86.14%	20.59%	240	0.97%	11.16%	1.78%
en_spec	601	4.05%	68.06%	7.65%	84	0.35%	9.51%	0.68%
en_spec+gen	774	5.22%	35.34%	9.10%	224	0.94%	10.23%	1.72%

Table 4. True Positives (TP), precision, recall and F-measure of individual methods

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

	PMI	MI	PS	Dice	Odds	t-score	χ^2	Fisher
pt	28.18%	17.53%	23.99%	53.47%	55.83%	13.9%	54.38%	11.23%
en	13.17%	12.79%	7.07%	25.09%	25.7%	7.14%	25.82%	5.12%

Table 5. Mean Average Precision of the statistical Association Measures (AMs) taken individually.

the task, as they result in considerably higher F-measures for English unlike for Portuguese. These different performances for the two languages are in line with Evert and Krenn's argument that statistical AMs are highly dependent on type and language [4].

	Portuguese				English			
	TP	Precision	Recall	F-measure	TP	Precision	Recall	F-measure
subAM	1102	48.29%	51.26%	49.73%	0	—	—	—
subAM + align	1103	47.98%	51.30%	49.58%	62	16.49%	2.88%	4.91%
allAM	1100	43.51%	51.16%	47.03%	464	19.74%	21.58%	20.62%
allAM + align	1084	43.41%	50.42%	46.65%	465	19.68%	21.63%	20.61%

Table 6. Bayesian network classifier for different feature sets, in Portuguese and in English.

Some of the statistical measures used as features are based upon contingency tables and are therefore not straightforwardly applicable to trigrams.[6] Therefore, in the training set, these measures are represented with a missing value ("?") for trigrams. In order to verify in more details whether these extra values affect the results obtained we performed a second evaluation where we only analyzed candidates which have non-null values for these features (i.e. bigrams).

The performance of the classifiers built on the bigram data set only are summarized in table 7[7]. The results further confirm that in some cases these extra features are adding enough information for the performance of the Bayesian Networks to improve. This is particularly clear in the case of English subAM, for which the F-measure improves by as much as 16% (for the case without alignment information). The difference in the results obtained by only considering the bigrams suggests that the methods propose a larger and more accurate set of bigram candidates, but they do not seem as effective for trigrams. Further investigation would have to be conducted to properly assess which factors play a role in the lower performance for trigrams.

[6] NSP, for example, does not implement these measures for trigrams.
[7] Recall considering only bigrams in the gold standards.

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

	Portuguese				English			
	TP	Precision	Recall	F-measure	TP	Precision	Recall	F-measure
subAM	1021	49.11%	71.90%	58.36%	228	32.66%	16.06%	21.53%
subAM + align	1026	45.72%	72.25%	56.00%	267	26.67%	18.80%	22.06%
allAM	1113	43.04%	78.38%	55.57%	459	19.94%	32.32%	24.66%
allAM + align	1113	42.68%	78.38%	55.26%	459	19.93%	32.32%	24.66%

Table 7. Classifier performance for different feature sets and languages, only bigrams.

6 Conclusions and Future Work

In this paper we presented an inexpensive, language independent hybrid approach for the identification of MWEs, that combines the strenghts of statistical measures with alignment-based information. The results obtained with a Bayesian network classifier confirm the improved performance of the hybrid approach over the individual methods. The use of the alignment information, as well as a larger set of AMs, seem to add robustness to the task, providing enough additional confirmation for the classifier. In addition, the methods seem to perform better for bigrams than trigrams. Further investigation needs to be conducted to identify which factors determine the influence of alignment information on the final performance, since this feature can both introduce noise and improve the performance of the classifier according to the language and size of the ngram. In addition, we also plan on investigating the influence of domain in the performance of these methods, verifying whether domain-specific MWEs are easier to extract than general ones.

Acknowledgments

We want to thank the financial support of the Brazilian agencies FAPESP, CAPES and CNPq. This research has been partly funded by the FINEP/SEBRAE project COMUNICA.

References

1. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002). Volume 2276 of (Lecture Notes in Computer Science)., London, UK, Springer-Verlag (2002) 1–15
2. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Grammar of Spoken and Written English. Longman, Harlow (1999)
3. Jackendoff, R.: Twistin' the night away. Language **73** (1997) 534–59
4. Evert, S., Krenn, B.: Using small random samples for the manual evaluation of statistical association measures. Computer Speech and Language **19**(4) (2005) 450–466

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

5. Baldwin, T.: The deep lexical acquisition of English verb-particles. Computer Speech and Language, Special Issue on Multiword Expressions **19**(4) (2005) 398–414
6. Caseli, H.M., Villavicencio, A., Machado, A., Finatto, M.J.: Statistically-driven alignment-based multiword expression identification for technical domains. In: Proceedings of the 2009 Workshop on Multiword Expressions (ACL-IJCNLP 2009). (2009) 1–8
7. Villavicencio, A., Caseli, H.M., Machado, A.: Identification of Multiword Expressions in Technical Domains: Investigating Statistical and Alignment-based Approaches. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology, São Carlos, SP (2009)
8. Fazly, A., Cook, P., Stevenson, S.: Unsupervised type and token identification of idiomatic expressions. Computational Linguistics **35**(1) (2009) 61–103
9. Van de Cruys, T., Villada Moirón, B.: Semantics-based Multiword Expression Extraction. In: Proceedings of the Workshop on A Broader Prespective on Multiword Expressions, Prague (June 2007) 25–32
10. Villada Moirón, B., Tiedemann, J.: Identifying idiomatic expressions using automatic word-alignment. In: Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), Trento, Italy (2006) 33–40
11. Ramisch, C., Villavicencio, A., Moura, L., Idiart, M.: Picking them up and Figuring them out: Verb-Particle Constructions, Noise and Idiomaticity. In: Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008). (2008) 49–56
12. Melamed, I.D.: Automatic Discovery of Non-Compositional Compounds in Parallel Data. In: eprint arXiv:cmp-lg/9706027. (June 1997) 6027–+
13. Coulthard, R.J.: The application of corpus methodology to translation: the jped parallel corpus and the pediatrics comparable corpus. Master's thesis, Universidade Federal de Santa Catarina (2005)
14. Lopes, L., Vieira, R., Finatto, M.J., Martins, D., Zanette, A., Jr., L.C.R.: Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. RECIIS - Electronic journal of communication information and innovation in health (English edition. Online) **3** (2009) 76–88
15. Procter, P.: Cambridge International Dictionary of English. Cambridge University Press (1995)
16. Banerjee, S., Pedersen, T.: The Design, Implementation and Use of the Ngram Statistics Package. In: In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. (2003) 370–381
17. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the ACL, Hong Kong, China (October 2000) 440–447
18. Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A.: Open-source Portuguese-Spanish machine translation. In Vieira, R., Quaresma, P., Nunes, M., Mamede, N., Oliveira, C., Dias, M., eds.: Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, (PROPOR 2006). Volume 3960 of Lecture Notes in Computer Science. Springer-Verlag (May 2006) 50–59
19. Caseli, H.M., Nunes, M.G.V., Forcada, M.L.: On the automatic learning of bilingual resources: Some relevant factors for machine translation. In: Proceedings of the 19th Brazilian Symposium on Artificial Intelligence (SBIA). Volume 5249., Springer Berlin / Heidelberg (2008) 258–267

This is a pre-print of an article published in the Proceedings of PROPOR 2016.
The final authenticated version is available online at:
https://doi.org/10.1007/978-3-642-12320-7_9

20. Caseli, H.M., Ramisch, C., Nunes, M.G.V., Villavicencio, A.: Alignment-based extraction of multiword expressions. Language Resources and Evaluation (to appear 2009)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2005)