Identification of Multiword Expressions in Technical Domains: Investigating Statistical and Alignment-based Approaches

Aline Villavicencio[♣♠], Helena de Medeiros Caseli[♢], André Machado[♣]

[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♠]Department of Computer Sciences, Bath University (UK)
[♢]Department of Computer Science, Federal University of São Carlos (Brazil)
avillavicencio@inf.ufrgs.br, helenacaseli@dc.ufscar.br,
ammachado@inf.ufrgs.br

Abstract. *Multiword Expressions (MWEs) are one of the stumbling blocks for more precise Natural Language Processing (NLP) systems. The lack of coverage of MWEs in resources can impact negatively on the performance of tasks and applications, and can lead to loss of information or communication errors; especially in technical domains where MWE are frequent. This paper investigates some approaches to the identification of MWEs in technical corpora based on: association measures, part-of-speech and lexical alignment information. We examine the influence of some factors on their performance such as sources of information for identification and evaluation. While the association measures emphasize recall, the alignment method focuses on precision.*

1. Introduction

The coverage of lexical resources have a significant impact on the performance of many Natural Language Processing tasks and applications, and much research has therefore been devoted to methods for automating lexical acquisition. In recent years some of these works have also started to target a set of phenomena for which lexical resources are particularly lacking in coverage: Multiword Expressions (MWE) [Baldwin 2005, Villavicencio et al. 2007]. These can be defined as combinations of words that have lexical, syntactic, semantic, pragmatic or statistical idyosincrasies [Sag et al. 2002], and include among them phrasal verbs (*carry up, consist of*), light verbs (*take a walk, make a demo*), compounds (*police car, frying pan*) and idioms (*shoot the breeze, make ends meet*). MWEs are very numerous in languages accouting for between 30% and 45% of spoken English and 21% of academic prose [Biber et al. 1999], and having the same order of magnitude in a speaker's lexicon as the number of single words [Jackendoff 1997]. However, these estimates are likely to be underestimates if we consider that for language from a specific domain the specialized vocabulary is going to consist largely of MWEs (*global warming, protein sequencing*) and new MWEs are also constantly appearing in language (*weapons of mass destruction, axis of evil*).

The impact caused by the lack of coverage in lexical resources can be seen, for instance, in the context of parsing, where in a random sample of 20,000 strings from the British National Corpus (BNC) missing MWEs accounted for 8% of total parsing errors [Baldwin et al. 2004], even with a broad-coverage grammar. Therefore, MWEs should be identified and dealt with adequately, as failing to do so may cause serious problems,

especially for NLP tasks that envolve some kind of semantic processing [Sag et al. 2002]. Robust (semi-)automated ways of acquiring lexical information for MWEs can significantly extend the coverage of resources, and, for example, just extracting VPCs from the BNC doubled the number of verb-particle constructions (VPCs) listed in a dictionary such as the Alvey Natural Language Tools [Baldwin 2005].

In this paper, we investigate some approaches for the identification of MWEs in technical corpora. We evaluate the performance of the proposed approach, examining the impact of the sources of information employed for the task. In particular we compare the results obtained with a domain specific English-Portuguese parallel Corpus of Pediatrics, verifying how a second language can provide relevant cues for this task. We also discuss some aspects that influence a more accurate evaluation of results. Such cost-effective approaches to the (semi-)automatic identification of MWEs can considerably extend the coverage of lexical resources and speed up lexicographic work, providing a more targeted list of MWE candidates.

The remainder of this paper is structured as follows. Section 2 briefly discusses MWEs and some previous works on automatically extracting them. Section 3 presents the resources used in our experiments while section 4 describes the methods proposed to extract MWEs. Section 4.3 presents the evaluation methodology and analyses the results and section 5 finishes this paper with some conclusions and proposals for future work.

2. Related Work

MWEs present a tough challenge for both linguistic and computational work [Sag et al. 2002] due to their heterogeneous features:

- syntactic flexibility: while some MWEs do not present internal variation (e.g. *ad hoc*, others allow different degrees of internal variability and modification (e.g. *touch a nerve* (*touch/find/strike a [raw] nerve*).
- semantic opaqueness: MWEs range from more opaque meanings (e.g. *to kick the bucket* as *to die*) to more transparent cases (e.g. *carry up*, where the particle *up* adds a sense of direction and location to the verb *carry*).

A variety of approaches has been proposed for automatically identifying MWEs, differing in terms of the type of MWE and language to which they apply, and the sources of information they use. Some of these works concentrate on particular languages (e.g. English [Baldwin 2005]), but some work has also benefited from information in one language to help deal with MWEs in the other (e.g. [Villada Moirón and Tiedemann 2006, Caseli et al. 2009]). As basis for helping to determine whether a given sequence of words is in fact an MWE (e.g. *ad hoc* vs *the small boy*) some of these works employ linguistic knowledge for the task, while others employ statistical methods (e.g. [Evert and Krenn 2005, Villavicencio et al. 2007]) or combine them with some kinds of linguistic information such as syntactic and semantic properties [Van de Cruys and Villada Moirón 2007] or automatic word alignment [Villada Moirón and Tiedemann 2006]. In this paper we want to determine the influence of different sources of information in the identification task.

Statistical measures of association have been widely employed in the identification of MWEs as they can be democratically applied to any language and MWE type. The idea behind their use is that they are an inexpensive language and type independent

means of detecting recurrent patterns and since we expect the component words of an MWE to occur frequently together, then these measures can give an indication of MWEness. However, some measures seem to provide more accurate predictions of MWEness than others, and there is no consensus about which measure is best suited for identifying MWEs in general. A comparison of some of these measures for the type-independent detection of MWEs indicated that Mutual Information is better at differentiating MWEs from non-MWEs than χ_2 [Villavicencio et al. 2007]. In addition, for MWE identification, the efficacy of a given measure seems to depend on factors like the type of MWEs being targeted for identification, the domain and size of the corpora used, and the amount of low-frequency data excluded by adopting a threshold [Evert and Krenn 2005]. For general MWE identification, the corpus size and nature also seem to have influence over the methods [Villavicencio et al. 2007]. We further investigate some approaches for the identification of MWEs in a technical domain, and look at some aspects for a more accurate evaluation of these methods. For Portuguese, the combination of some frequency-based measures and heuristics to extract terms for building an ontology from a domain-specific text resulted in an F-measure of up to 11.51% for bigrams and 8.41% for trigrams [Vieira et al. 2009].

Among the methods that use additional information to extract MWE, the one proposed in [Villada Moirón and Tiedemann 2006] seems to be the most similar to the alignment-based approach tested in this paper. The main difference between them is the way in which word alignment is used in the MWE extraction process. In this paper, the word alignment is the basis of MWE extraction process while Villada Moirón and Tiedemann's method uses the alignment just for ranking the MWE candidates which were extracted on the basis of association measures (log-likelihood and salience) and head dependence heuristic (in parsed data). Another related work is the automatic detection of non-compositional compounds (NCC) [Melamed 1997] in which NCCs are identified by analyzing statistical translation models trained in a huge corpus by a time-demanding process. In this paper we use an alignment-based approach considering as a MWE candidates the sequences of two or more consecutive source words joined by the aligner regardless of whether they are translated as (aligned with) one or more target words.

3. The Corpus and Reference Lists

In the experiments presented in this paper we used the Corpus of Pediatrics, a Portuguese-English parallel corpus, containing 283 texts (and 785,448 words) in Portuguese and their parallel versions in English extracted from the *Jornal de Pediatria*. To evaluate the Portuguese MWE candidates, we used the Pediatrics Glossary[1], a domain-specific glossary built from the Corpus of Pediatrics for supporting translation studies. The Glossary, contains ngrams from the corpus with frequency higher than 5, filtered using part-of-speech information, and manually checked, in a total of 2,407 terms (from which 1,421 are bigrams and 730 trigrams). The English candidates are evaluated using a general dictionary of English [Cambridge 1994], with 24,160 entries (from which 9,174 are bigrams and 2,946 trigrams).

[1]Produced by TEXTQUIM/TERMISUL `http://www.ufrgs.br/textquim`

4. Experiments and Results

In this section we describe the experiments carried out following two approaches for MWE Identification. The first one, the statistically-driven approach, applies the well-know measures Pointwise Mutual Information (PMI) and Mutual Information (MI) [Press et al. 1992], as implemented in the Ngram Statistics Package [Banerjee and Pedersen 2003]. The second one is based on the automatic lexical alignment of Portuguese and English versions of the Corpus of Pediatrics generated by the statistical word aligner `GIZA++` [Och and Ney 2000]. After extracting a prior list of MWE candidates following each approach these candidates were filtered out based on, for example, their frequencies or some pre-defined part-of-speech (POS) patterns as explained in the next subsections.

To evaluate the efficacy of the investigated approaches in identifying multiword terms in a domain-specific corpus, an automatic comparison was performed using the gold standards for each language cited on section 3. In tables 1 and 2 we show the precision (number of correct candidates among the proposed ones), recall (number of correct candidates among those in the reference lists) and F-measure ($(2 * precision * recall)/(precision + recall)$) figures for both approaches.

4.1. Statistically-Driven Approach

Table 1 shows the number of ngrams extracted from the corpus of Pediatrics using the statistical metrics PMI and MI, after applying the following filters:

- **F1** removing ngrams containing punctuation and numbers
- **F2** using (a) **F1** and (b) a frequency cut-off of 5 occurrences
- **F3** using (a) **F1, F2** and (b) POS tag filters to remove ngrams that begin with determiner, auxiliary verb, pronoun, adverb, conjunction and surface forms such as those from verb to be (*are*, *is*, *was*, *were*), relatives (*that*, *what*, *when*, *which*, *who*, *why*) and prepositions (*from*, *to*, *of*)
- **F3**$_{PMI}$ with (a) **F1, F2** and **F3** and (b) only considering the top n candidates ranked by their PMI score, where n is the number of ngrams in each of the reference lists
- **F3**$_{MI}$ with (a) **F1, F2** and **F3** and (b) only considering the top n candidates ranked by their MI score
- **F3**$_{PMI+MI}$ with (a) **F1, F2** and **F3** and (b) only considering the top n candidates ranked by their average position according to their PMI and MI scores.

For **F3** the candidate ngrams were tagged using the Tree Tagger [Schmid 1994] trained for Portuguese and English, respectively.

4.2. Alignment-Based Approach

Different from the statistical approach, the alignment-based one relies on lexically aligned parallel texts to identify MWE candidates. The lexical aligner searches for correspondences between source and target words and sequences of words in two parallel sentences — a sentence written in one (source) language and its translation to another (target) language. Therefore, taking into account the lexical alignment between a source word sequence S ($S = s_1 \dots s_n$ with $n \geq 2$) and a target word sequence T ($T = t_1 \dots t_m$ with $m \geq 1$), the alignment-based MWE extraction method states that the sequence S will be a MWE candidate. For example, the sequence of two Portuguese words *aleitamento*

Table 1. Ngram MWE Candidates extracted by the statistically-driven approach

MWE candidates		Portuguese		English	
		Bigrams	**Trigrams**	**Bigrams**	**Trigrams**
F1	# proposed ngrams	191,825	356,888	180,046	345,423
F2	# proposed ngrams	20,132	9,593	19,036	10,530
	# correct MWEs	1,400	696	258	58
	precision	6.95%	7.26%	1.36%	0.55%
	recall	98.52%	95.34%	2.81%	1.97%
	F	12.99%	13.48%	1.83%	0.86%
F3	# proposed ngrams	16,102	7,312	10,470	4,511
	# correct MWEs	1,394	696	134	11
	precision	8.66%	9.52%	1.28%	0.24%
	recall	98.17%	95.34%	1.46%	0.37%
	F	15.91%	17.31%	1.36%	0.30%
# ngrams in reference lists		1,421	730	9,174	2,946
$F3_{PMI}$	# correct MWEs	803	195	130	9
	precision, recall and F	56.51%	26.71%	1.41%	0.31%
$F3_{MI}$	# correct MWEs	270	60	126	3
	precision, recall and F	19.00%	8.22%	1.37%	0.10%
$F3_{PMI+MI}$	# correct MWEs	803	196	130	9
	precision, recall and F	56.51%	26.85%	1.42%	0.31%

materno — which occurs 202 times in the corpus used in our experiments — is a MWE candidate because these two words were joined to be aligned 184 times with the word *breastfeeding* (a $2:1$ alignment), 8 times with the word *breastfed* (a $2:1$ alignment), 2 times with *breastfeeding practice* (a $2:2$ alignment) and so on.

Due to its feature of looking for the sequences of source words that are frequently joined together during the alignment despite the number of target words involved, the alignment-based method priorizes precision in spite of recall. In addition to the lexical alignment performed by the statistical word aligner `GIZA++` [Och and Ney 2000], the original corpus was, firstly, sentence aligned by a version of the Translation Corpus Aligner (TCA) [Hofland 1996] and also POS tagged using the morphological analysers and taggers from `Apertium`[2] [Armentano-Oller et al. 2006]. The initial list of MWE candidates was filtered according to several filtering patterns:

F3 is as described above[3]

F4 is the same used during the manual building of the reference lists of MWEs: (a) patterns beginning with Article + Noun and beginning or finishing with verbs and (b) minimum frequency threshold of 5.

F5 is F3 plus: (a) patterns beginning or finishing with determiner, adverb, conjunction, preposition, verb, pronoun and numeral and (b) a minimum frequency threshold of 2.

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

[3] Using (a) **F1, F2** and (b) POS tag filters to remove ngrams that begin with determiner, auxiliary verb, pronoun, adverb, conjunction and surface forms such as those from verb to be (*are*, *is*, *was*, *were*), relatives (*that*, *what*, *when*, *which*, *who*, *why*) and prepositions (*from*, *to*, *of*).

Table 2. Ngram MWE Candidates extracted by the alignment-based approach

MWE candidates		Portuguese		English	
		Bigrams	**Trigrams**	**Bigrams**	**Trigrams**
F3	# proposed ngrams	754	110	956	170
	# correct MWEs	95	9	22	3
	precision	12.60%	8.18%	2.30%	1.76%
	recall	6.69%	1.23%	0.24%	0.10%
	F	8.74%	2.14%	0.43%	0.19%
F4	# proposed ngrams	250	19	267	42
	# correct MWEs	48	1	10	1
	precision	19.20%	5.26%	3.75%	2.38%
	recall	3.38%	0.14%	0.11%	0.03%
	F	5.75%	0.27%	0.21%	0.07%
F5	# proposed ngrams	169	20	149	22
	# correct MWEs	65	4	4	0
	precision	38.46%	20.00%	2.68%	0%
	recall	4.57%	0.55%	0.04%	0%
	F	8.18%	1.07%	0.09%	0%

4.3. Discussion about the Results

Tables 1 and 2 show the results of the statistically-driven and alignment-based approaches, and the effects of applying different sources of information to remove the noise from the initial list of ngrams. For instance, the linguistic information in F3 removes noise without excluding many genuine MWEs, while the statistical measures, shown in the bottom half of table 1, significantly increase the precision of the filtered candidates. These measures rank the candidates according to how strong the co-ocurrence of the words are in relation to the frequencies of the individual words. Therefore, in order to indicate how close this ranking is to the gold standard, i.e. how well they measure MWEness, we show the top n candidates of each ranking, where n is the same as the number of ngrams in each reference list (1,421 bigrams and 730 trigrams for Portuguese and 9,174 brigrams and 2,946 trigrams for English). This results in the same values for precision and recall (and consequently for the F-measure) in each of the methods. PMI and MI generate different predictions of MWEness, as can be seen from their combination, which alters the results given by each measure individually, with PMI alone resulting in the most accurate candidate list for both languages.

The low F-scores results for both languages may be explained by the limited coverage of the reference lists that do not contain a significant number of true MWEs among their entries. To verify this, we selected the MWE candidate list with best precision values (from filter F5) to be also analyzed by human experts. From the list of 189 candidates (bi and trigrams), the 122 (63.9%) that were not found in the Pediatrics Glossary (see table 2) were analysed by two native human experts. The judges classified each of the 122 candidates as true, if it is a multiword expression, or false, otherwise independently of being a Pediatrics term. For the judges, a sequence of words was considered a MWE mainly if it was: (1) a proper name or (2) a sequence of words for which the meaning cannot be obtained by compounding the meanings of its words.

The judgments of both judges were compared and a disagreement of approximately 12% on multiwords was verified. This disagreement was also measured by the

kappa (K) measure [Carletta 1996], with k = 0.73, which does not prevent conclusions to be drawn, since a value of k between 0.67 and 0.8 seems to indicate a good agreement [Carletta 1996].

In order to calculate the percentage of true candidates among the 122, two approaches can be followed, depending on what criteria one wants to emphasize: precision or coverage (not recall because we are not calculating regarding a reference list). To emphasize precision, one should consider as genuine MWEs only those candidates classified as true by both judges, on the other hand, to emphasize coverage, one should consider also those candidates classified as true by just one of them. So, from 191 MWE candidates, considering the extended gold standard (the reference lists and the human judgements), 126 (65.97%) were classified as true by both judges and 145 (75.92%) by at least one of them, with a significant improvement in the figures reported, compared to the F5 filter. For Portuguese this may be explained as a result of the focus of the Pediatrics Glossary on domain-specific terms. In comparison, the statistical and alignment based approaches identify both domain-specific and general MWEs, which for NLP tasks are also of importance and must be treated accordingly. Therefore, a more accurate evaluation of MWE identification in technical domains requires both a domain specific and a general lexicon to be part of the gold standard.

On the other hand, a comparison of the results for these two languages show much lower F-score values for the English ngrams. The results for English with a general dictionary as a gold standard reflected in much lower F-score values than for the Portuguese with a Pediatrics Glossary, in spite of the large number of entries in the English dictionary and possible differences between these languages. These may be partly explained as a consequence of using a general gold standard that contains MWEs from general language, most of which are not to be found in a domain specific text, since domain-specific terms seem to account for the majority of terms in the domain in both languages. These results confirm the need for developing methods for the semi-automatic identification of MWEs in technical domains to minimize the effect of the lack of coverage of domain-specific terms in general lexical resources.

5. Conclusions and Future Work

In this paper we investigated the impact of different sources of information in the identification and evaluation of MWEs from technical domains, using statistical and alignment-based approaches with parallel corpora. POS filters provided an effective and simple way to remove noise virtually without side-effects. The alignment-based method benefits from the information from the source and target languages to generate a precision-oriented list of MWE candidates, while statistical methods produce recall-oriented results at the expense of precision. Although the majority of the MWEs are domain specific terms, more accurate evaluation was obtained using both domain-specific and general reference lists.

Using the alignment-based extraction method we notice that it is possible to extract MWEs that are Pediatrics terms with a precision of 38% for bigrams and 20% for trigrams, but with very low recall since only the MWEs in the reference lists were considered correct. However, after a manual analysis carried out by two native speakers of Portuguese we found that the percentage of true MWEs considered by both or at least one of them were, respectively, 65.97% and 75.92%. This significant improvement has to be

carefully considered since the human experts classified the MWEs as true independently of being a Pediatrics term. So, as future work, a more detailed analysis of domain-specific terms with experts in Pediatrics will be carried out to evaluate how many of those true MWEs candidates are also Pediatrics terms.

Methods like these presented in this paper can significantly speed up lexicographic work and the results obtained show that, in comparison with the manual extraction of MWEs, the automatic approach can provide also a general set of MWE candidates in addition to the manually selected technical terms.

For future work we plan to investigate a weighted combination of the statistically-driven and the alignment-based methods to produce a set of MWE candidates that is both more precise than the former and has more coverage than the latter. In addition, we intend to apply the results obtained in the semi-automatic construction of ontologies.

Acknowledgments

We would like to thank the TEXTQUIM/UFRGS group for making the Corpus of Pediatrics and Pediatrics Glossary available to us. We also thank the financial support of FAPESP in bulding the parallel corpus. This research has been partly funded by the FINEP project COMUNICA.

References

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., and Scalco, M. A. (2006). Open-source Portuguese-Spanish machine translation. In *Proceedings of the VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR-2006)*, pages 50–59, Itatiaia-RJ, Brazil.

Baldwin, T. (2005). The deep lexical acquisition of english verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Baldwin, T., Bender, E. M., Flickinger, D., Kim, A., and Oepen, S. (2004). Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Banerjee, S. and Pedersen, T. (2003). 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*, pages 370–381.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Grammar of Spoken and Written English*. Longman, Harlow.

Cambridge (1994). *Cambridge International Dictionary of English*. Cambridge University Press.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2):249–254.

Caseli, H. M., Ramisch, C., Nunes, M. G. V., and Villavicencio, A. (2009). Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*.

Evert, S. and Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. In Hockey, S., Ide, N., and Perissinotto, G., editors, *Research in Humanities Computing*, pages 165–178, Oxford. Oxford University Press.

Jackendoff, R. (1997). Twistin' the night away. *Language*, 73:534–59.

Melamed, I. D. (1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *eprint arXiv:cmp-lg/9706027*, pages 6027–+.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, Hong Kong, China.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing. Second edition.* Cambridge University Press.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). 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)*, pages 1–15, London, UK. Springer-Verlag.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

Van de Cruys, T. and Villada Moirón, B. (2007). Semantics-based Multiword Expression Extraction. In *Proceedings of the Workshop on A Broader Prespective on Multiword Expressions*, pages 25–32, Prague.

Vieira, R., Finatto, M. J., Martins, D., Zanette, A., and Jr, L. C. R. (2009). Extração automática de termos compostos para construção de ontologias: Um experimento na área da saúde. *Reciis- Revista Eletrônica de Comunicação Informação e Inovação em Saúde*, 3:76–88.

Villada Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006)*, pages 33–40, Trento, Italy.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007). Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1034–1043, Prague.