Joining Forces for Multiword Expression Identification

Leonardo Zilio[1], Rodrigo Wilkens[1(✉)], Luís Möllmann[1], Eric Wehrli[2], Silvio Cordeiro[1], and Aline Villavicencio[1]

[1] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
{zilio,rodrigo.wilkens,luis.mollmann,srcordeiro, avillavicencio}@inf.ufrgs.br
[2] Department of Linguistics, University of Geneva, Geneva, Switzerland
eric.wehrli@unige.ch

Abstract. Multiword Expressions (MWEs) display some kind of linguistic and statistical markedness that may influence the effectiveness of techniques that automatically identify them in texts. While parsing-based techniques for MWE identification are considered to be better at handling long-distance dependencies, passivization and internal modification, statistics-based techniques use association measures to detect statistical markedness regardless of syntactic form. In this paper we compare these two approaches focusing on nominal compounds in Portuguese. We compare the accuracy of each method and propose that combining the strengths of both for increased accuracy.

1 Introduction

Multiword Expressions (MWEs) are sequences of words that must be treated as a unit at some level of linguistic processing [2]. Since tasks like parsing and automatic translation depend on resources such as dictionaries and grammars, the automatic MWE identification can enhance the quality of such resources, especially in uncommon languages and specialized domains, by pointing out expressions that need to be treated as a single unit.

Various techniques for automatic identification of Multiword Expressions (MWEs) from corpora have been investigated, varying in the amount of linguistic or statistical knowledge they employ as basis for the identification [8,10]. Their effectiveness seems to depend on variables like the particular language, the MWE type and the size of the corpus in which the identification is performed [3,7,8]. As a consequence no consensus about the best technique that works for all languages and MWE types has been reached. On the one hand parsing-based techniques, where identification is based on syntactic structure, seem to produce more precise results and capture long-distance dependencies for MWEs and other grammatically well-formed expressions [10]. However, they depend on parsing coverage for a given language and there are other types of MWEs for which they cannot be used. For instance, they cannot easily retrieve

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 233–238, 2016.
DOI: 10.1007/978-3-319-41552-9_23

compounds which are not grammatically well-formed, such as *by and large*, and do not work well for named entities. On the other hand, statistics-based techniques, while generally less precise and more recall-oriented, rely on association measures to capture the statistical markedness and can help to identify candidates whose components co-occur more often than chance, regardless of their syntactic structure. They may also be combined with linguistic patterns, for greater precision. One important advantage is that they are language and domain independent [8], as for many languages and domains a parser may not be available.

In this paper we compare a precision-oriented parsing-based technique with a language-independent recall-oriented statistics-based technique. The main goal here is to combine results from both techniques, observing how each technique complement the other. The focus lies on MWEs in Portuguese, more specifically on compound nouns, like *carro de polícia* (*police car*).

By examining the combination of these complementary techniques we aim to obtain the basis for a more precise MWE resource. This paper is structured as follows: we discuss the materials and methods used in this work, and the evaluation and results obtained respectively in Sects. 2 and 3. We wrap it all up by presenting final remarks and future work Sect. 4.

2 Corpus and Methodology

For the MWE identification process we used the Europarl corpus [4], which was pre-processed with the PALAVRAS parser [1]. The corpus contains 67 million tokens from which MWE candidates were extracted.[1] The selection of the Europarl corpus was driven by the possibility of future comparison with other languages, since it presents alignments between the EU member languages.

The corpus was then processed using both a parsing-based and a statistic-based approach for identificating MWE candidates. For the parsing-based identification we used the Fips "deep" linguistic parser [11,12] is a symbolic parser system which recognizes structures corresponding to grammatically well-formed phrases and uses statistical information for detecting recurrent sequences based on corpus frequency. For the statistics-based identification we used the mwetoolkit [9], which allows the definition of linguistic (POS tag) patterns to extract an initial list of candidates and the use of statistical association measures to identify candidates whose components co-occur more often than chance. Measures implemented in the toolkit include Pointwise Mutual Information, Student's t-test score, Dice's coefficient and Maximum Likelihood Estimation.

For generating the MWE lexicon we selected predefined patterns for Portuguese, such as noun-preposition-noun (e.g. *política de desenvolvimento*, *condição de trabalho*), excluding proper nouns. An initial set of candidates was obtained, and it was automatically validated according to the agreement between

[1] Parsing increased the number of tokens from 49 to 67 million words as contractions like those involving preposition and determiner are expanded to *de + o* by the parser (e.g. *do* (of the)).

the syntactic and statistics-based methods, and, for the cases where they disagree, a weighted voting scheme was used based on how confident they are about each candidate. From the resulting lists of MWE candidates we selected a set of the top 2,000 MWE candidates from each approach. We then created a third set of MWE candidates based on the agreement from the two processes. Finally, we evaluated the quality of the three sets automatically using BabelNet [5] and Onto.PT [6]. A smaller portion of the results (100 top candidates from each process) was also selected for manual validation[2].

3 Results

Table 1 presents the results from the automatic validation of the top 2,000 MWE candidates identified by each of the process. This automatic evaluation process was done by checking if at least one of the resources contained the MWE candidate. As we can see, the intersection of both processes present a more accurate result, rising the accuracy of MWE identification to 17.9 %.

Table 1. Automatic validation of the top 2000 candidates identified by each method

Noun-Prep-Noun	Candidates	In resources	% MWEs
Parsing-based	2000	198	9.9 %
Statistics-based	2000	319	16.0 %
Intersection	797	143	17.9 %

As described in the methodology, the top 100 MWE candidates identified with each approach were manually evaluated by a linguist that is a native-speaker of Portuguese. As some of these candidates are already automatically validated against Onto.PT and BabelNet, only the MWE candidates that were not in these resources were manually evaluated. Results of this evaluation are listed in Table 2.

Since the idea is to combine both MWE identification techniques to look for improvements on the quality of the resource, we identified in Table 2 the number of MWE candidates that are among the top 100 for both approaches (their intersection), those that occur in only one of the lists (their differences) and those that occur in at least one list (their union), in relation to those that were judged to be valid. The results obtained are that they complement each other, since most candidates (61 %) are unique to each technique, but 39 % are common to both, and from these 79.5 % are valid MWEs.

This means that, again, by intersecting parsing-based with statistics-based information, it is possible to obtain a more accurate list of MWE candidates, that

[2] Among the 100 top candidates from each process, some were already validated in the automatic validation step. The manual validation was applied only to those MWE candidates that were not prevalidated.

Table 2. Validation of the top 100 candidates in each system

Noun-Prep-Noun	Candidates	Valid candidates	% MWEs
Intersection	39	30	76.9 %
Parsing-based only	61	33	54.1 %
Statistics-based only	61	36	59.0 %
Union	161	99	61.5 %

presents the strength of both methods. The precision for the 100 candidates can be seen in Fig. 1. By uniting both results, while we have an increase in coverage (presenting a larger list of MWEs), there is also a decrease on precision when compared to the individual systems. Since we are looking for a more accurate resource, the intersection of both is the one we are focusing on right now.

A collateral result from the manual evaluation was the observation that the lexical resources we used for evaluating the MWE candidates lack coverage in terms of MWEs. Table 3 shows the number of automatically and manually validated MWE candidates, demonstrating the increase of correct MWEs from the automatic compared to the manual evaluation. It is important to stress that all MWEs that were manually validated were not present in the lexical resources, since the manual validation was done only on those MWE candidates among the top 100 that were not automatically validated.

Table 3. Automatic vs. manual validation of the MWE candidates in each system

Noun-Prep-Noun	Automatic validation	Manual validation
Parsing-based	30	31
Statistics-based	22	40

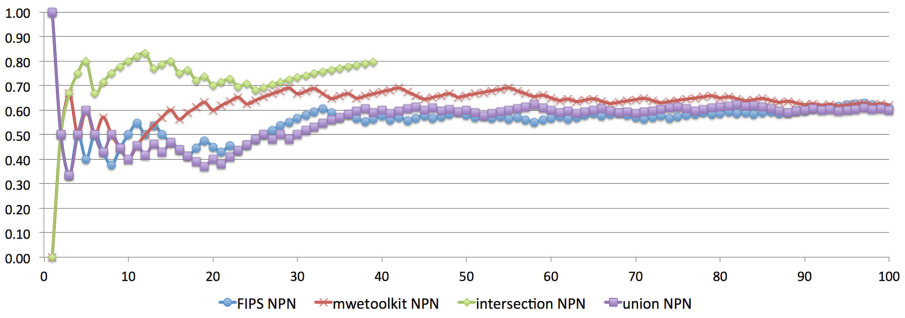

Fig. 1. Precision at 100

4 Conclusions and Future Work

In this paper we compared two approaches for the identification of MWEs, concentrating on nominal expressions in Portuguese. For MWE identification we examined a more precision-oriented with a more recall-oriented technique. We first looked at the first 2000 MWEs candidates and their automatic validation against lexical resources, and then we evaluated again the first 100 candidates, this time using the judgment of a native-speaker expert. The results showed that the combination of both techniques, by intersecting their results, leads to a more accurate result.

Since the corpus used in this work is a parallel corpus with alignment between language pairs, as future work, we plan on investigating how the transfer of information from one language can help identification of MWEs in the other.

Acknowledgments. This research was partially developed in the context of the project *Text Simplification of Complex Expressions*, sponsored by Samsung Eletrônica da Amazônia Ltda., in the terms of the Brazilian law n. 8.248/91. This work was also partly supported by CNPq (482520/2012-4, 312114/2015-0) and FAPERGS.

References

1. Bick, E.: The parsing system Palavras: Automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus Universitetsforlag (2000)
2. Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA), May 2002
3. Evert, S., Krenn, B.: Using small random samples for the manual evaluation of statistical association measures. Comput. Speech Lang. **19**(4), 450–466 (2005)
4. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Conference Proceedings: The Tenth Machine Translation Summit, pp. 79–86. AAMT, Phuket (2005)
5. Navigli, R., Ponzetto, S.P.: BabelNet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 216–225. Association for Computational Linguistics (2010)
6. Oliveira, H.G., Gomes, P.: Onto.pt: automatic construction of a lexical ontology for portuguese. In: Stairs 2010: Proceedings of the Fifth Starting AI Researchers' Symposium, vol. 222, p. 199. IOS Press (2010)
7. Pecina, P.: Lexical association measures and collocation extraction. Lang. Resour. Eval. **44**(1–2), 137–158 (2010)
8. Ramisch, C., Schreiner, P., Idiart, M., Villavicencio, A.: An evaluation of methods for the extraction of multiword expressions. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), pp. 50–53 (2008)
9. Ramisch, C., Villavicencio, A., Boitet, C.: Multiword expressions in the wild?: the mwetoolkit comes in handy. In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, pp. 57–60. Association for Computational Linguistics (2010)

10. Seretan, V.: Syntax-Based Collocation Extraction, Text, Speech and Language Technology, vol. 44, 1st edn, p. 212. Dordrecht, Netherlands (2011)
11. Wehrli, E., Nerima, L.: The fips multilingual parser. In: Gala, E., Rapp, R., Bel-Enguix, G. (eds.) Language Production, Cognition, and the Lexicon. Text, Speech and Language Technology, vol. 48, pp. 473–490. Springer, Switzerland (2015)
12. Wehrli, E., Seretan, V., Nerima, L.: Sentence analysis and collocation identification. In: 23rd International Conference on Computational Linguistics, COLING (2010)