Seen2Unseen at PARSEME Shared Task 2020: All Roads do not Lead to Unseen Verb-Noun VMWEs

Caroline Pasquer
University of Tours, LIFAT
France
first.last@etu.univ-tours.fr

Agata Savary
University of Tours, LIFAT
France
first.last@univ-tours.fr

Carlos Ramisch
Aix Marseille Univ, Université de Toulon,
CNRS, LIS, Marseille, France
first.last@lis-lab.fr

Jean-Yves Antoine
University of Tours, LIFAT
France
first.last@univ-tours.fr

Abstract

We describe the Seen2Unseen system that participated in edition 1.2 of the PARSEME shared task on automatic identification of verbal multiword expressions (VMWEs). The identification of VMWEs that do not appear in the provided training corpora (called *unseen* VMWEs) – with a focus here on verb-noun VMWEs – is based on mutual information and lexical substitution or translation of *seen* VMWEs. We present the architecture of the system, report results for 14 languages, and propose an error analysis.

1 Introduction

The identification of multiword expressions (MWEs) such as ***spill the beans*** is a challenging problem (Baldwin and Kim, 2010; Constant et al., 2017), all the more so for verbal MWEs (VMWEs) subject to morphological (***spill the bean***) and syntactic variability (***the beans*** *were* ***spilled***). The PARSEME shared task (PST) provided training, development and test corpora (hereafter Train, Dev, and Test) manually annotated for VMWEs.[1] Our system aimed at identifying every VMWE in Test which also appears in Train or Dev, including possible morphological or syntactic variants (henceforth *seen* VMWEs) or not present in Train/Dev (*unseen* VMWEs). Unseen VMWE identification, the main focus of this PST edition, is harder than seen VMWE identification, as shown by previous results (Ramisch et al., 2018).

We submitted two systems: Seen2Seen (closed track) and Seen2Unseen (open track). Seen2Unseen relies on Seen2Seen for the identification of seen VMWEs and has an additional module for unseen ones. Its best global unseen F-score (i.e. not only for verb-noun constructions) was obtained for Hindi (42.66) and it reached 25.36 in French, which was our main focus. Despite the lower global MWE-based F1-score of Seen2Unseen (63.02) compared to Seen2Seen (66.23), we describe the former (Sec. 2), analyse its interesting negative results (Sec. 3), and conclude with ideas for future work (Sec. 4).

2 System Description

While describing the architecture of our system, we use the notions of a VMWE *token* (its occurrence in running text) and a VMWE *type* (abstraction over all occurrences of a given VMWE), as introduced by Savary et al. (2019b). We represent VMWE types as multisets of lemmas and POS.[2] Our system uses a mixture of discovery and identification methods, as defined by Constant et al. (2017). Namely, VMWE *discovery* consists in generating lists of MWE types out of context, while VMWE *identification* marks VMWE tokens in running text. The system is freely available online (https://gitlab.com/cpasquer/st_2020).

[1] http://hdl.handle.net/11234/1-3367

[2] VMWEs are represented as multisets (i.e. bags of elements with repetition allowed), since the same lemma and/or POS can occur twice, as in ***appeler*** **<u>*un*</u>** **<u>*chat*</u>** **<u>*un*</u>** **<u>*chat*</u>** '*to call a cat a cat*'⇒'to call a spade a spade'.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 124–129
Barcelona, Spain (Online), December 13, 2020.

Seen2Seen in a nutshell Seen2Seen is a VMWE identification system dedicated to only those VMWEs which have been previously seen in the training data. Its detailed description is provided in Pasquer et al. (2020), but a brief overview is included here to make the current paper self-contained. Seen2Seen extracts lemma combinations of VMWEs seen in Train, looking for the same combinations (within one sentence) in Test, with an expected high recall. To improve precision, up to eight independent criteria can be used: (1) component lemmas should be disambiguated by their POS, (2) components should appear in specific orders (e.g. the determiner before the noun), (3) the order of "gap" words possibly occurring between components is also considered, (4) components should not be too far from each other in a sentence, (5) closer components are preferred over distant ones, (6) components should be syntactically connected, (7) nominal components should appear with a previously seen inflection, and (8) nested VMWEs should be annotated as in Train. We select the combination of criteria with maximal performance on Dev among all $2^8 = 256$ possibilities. The candidates remaining after applying the criteria are annotated as VMWEs. This relatively simple system relying on morphosyntactic filters and tuned for 8 parameters was evaluated on 11 languages of the PARSEME shared task 1.1 (Ramisch et al., 2018). Seen2Seen outperformed the best systems not only on seen (F=0.8276), but even on all seen and unseen VMWEs (F=0.6653).[3] In edition 1.2 of the PARSEME shared task, Seen2Seen scored best (out of 2) in the global ranking of the closed track and second (out of 9) across both tracks. It outperformed 6 other open track systems, notably those using complex neural architectures and contextual word embeddings. We believe that these competitive results are due to carefully taking the nature of VMWEs into account (Savary et al., 2019a). Since Seen2Seen, by design, does not account for unseen VMWEs, its score in this category is very low (F=1.12).[4] Therefore, it was later extended with a VMWE discovery module. Seen2Unseen is precisely this extended system. It relies on Seen2Seen for seen VMWEs and on discovery methods described below for unseen VMWEs.

From Seen2Seen to Seen2Unseen We assume that seen VMWEs could help identify unseen ones by using (i) lexical variation, tolerated by some VMWEs (e.g. ***take*** *a* ***bath/shower***), and (ii) translation, e.g. (FR) ***prendre décision*** 'take decision' = (PL) ***podejmować decyzję*** = (PT) ***tomar decisão*** = (SV) ***fatta beslut***.[5] We also expect seen and unseen VMWEs to share characteristics, such as the distance between components or their syntactic dependency relations, e.g. nouns often being objects of verbs. The categories that should benefit from our strategy are, mainly, light-verb constructions (LVCs) containing nouns and, in some cases, verbal idioms (VIDs). These categories are universal, so our method can be applied to the 14 languages of the PST. Since LVCs are often verb-noun pairs, Seen2Unseen quasi-exclusively focuses on them.[6] Consequently, we do not aim at exhaustively identifying unseen VMWEs, but at determining to what extent seen verb-noun VMWEs can help us discover new unseen ones.

Resources In addition to the PST Train, Dev and Test corpora, we used the CoNLL 2017 shared task parsed corpora, hereafter CoNLL-ST (Ginter et al., 2017).[7] The CoNLL-ST corpora were preferred over the PST-provided parsed corpora because they are conveniently released with pre-trained 100-dimensional word2vec embeddings for the 14 languages of the PST, which we used to generate lexical variants. Additionally, we used a free library to implement translation towards French and Italian.[8] We automatically translated all VMWEs in the other 13 languages into French (resp. Italian), privileged due to the availability of two Wiktionary-based lexicons in the same format for both languages.[9] These lexicons were used to lemmatize and POS-tag automatic translations, e.g. (PT) ***firmar contrato*** 'sign contract' $\xrightarrow{translation}$ (FR) *a* ***signé*** *un* ***contrat*** $\xrightarrow{lemma,POS}$ ***signer***$_{\text{VERB}}$ ***contrat***$_{\text{NOUN}}$.[10]

[3] In this paragraph we refer to macro-averaged MWE-based F-scores.

[4] The score is not null due to different implementations of unseen VMWEs in the evaluation script and in Seen2Seen.

[5] Languages are referred to with their PST identifier: e.g. FR for French.

[6] We also model inherently reflexive verbs with cranberry words, i.e. verbs which never occur without a reflexive pronoun, e.g. (FR) ***s'évanouir*** vs. **évanouir*. With 1 VMWE discovered in Portuguese and 3 in French, this module is omitted here.

[7] http://hdl.handle.net/11234/1-1989

[8] Googletrans: https://pypi.org/project/googletrans, implementing the Google Translate API.

[9] For French: http://redac.univ-tlse2.fr/lexicons/glaff_en.html, for Italian: http://redac.univ-tlse2.fr/lexiques/glaffit.html

[10] In case of multiple POS or lemmas, the most frequent verb-noun combination in CoNLL-ST was selected.

Unseen VMWE identification To support identification of unseen VMWEs we use a combination of semi-supervised discovery and identification methods: lexical replacement, translation and statistical ranking. For a language L, let $SeenVN^L$ be the set of all seen LVC and VID types having exactly one verb and one noun (and any number of components with other POS tags). Let each type in $SeenVN^L$ be linked with its manually annotated occurrences in Train. This set is used in the following steps:
① *Lexical replacement*: The idea is to observe lexical variability of seen VMWEs and to generate on this basis new potential VMWEs. Let LVC^L_{Vvar} contain LVC types in $SeenVN^L$ that tolerate variation in verbs, e.g. ***accomplir/effectuer***VERB ***mission***NOUN 'fulfil/perform mission'. Similarly, let LVC^L_{Nvar} contain LVCs types with variation in nouns, e.g. ***accomplir***VERB ***mission/tâche***NOUN 'fulfil mission/task'. Then we define two sets of candidates:

- MIX^L combines each verb in LVC^L_{Vvar} with each noun in LVC^L_{Nvar} to predict new combinations. e.g. ***effectuer tâche*** 'perform task'.
- SIM^L contains VMWEs from LVC^L_{Vvar} (resp. LVC^L_{Nvar}) where we replace the verb (resp. noun) by its closest verb (resp. noun) according to cosine similarity in CoNLL-ST word embeddings.[11]

② *Translation*: By translating seen VMWE types in one language we obtain a list of VMWE type candidates in another language:

- $TRANS^L$ is built only for French and Italian, and is empty for other languages. $TRANS^{FR}$ (resp. $TRANS^{IT}$) contains automatic translations of each VMWE in $SeenVN^{L'}$, with $L' \neq$ FR (resp. $L' \neq$ IT), into French (resp. Italian). We eliminate translations which do not contain exactly one verb and one noun (and possible components of other POS), e.g. due to a wrong translation. For the remaining translations, we keep only the verb and the noun lemmas.

③ *Statistical ranking*: This approach is based on statistical characteristics of both seen VMWEs and unseen VMWE candidates. We first calculate 3 sets of features for the whole $SeenVN^L$ list:

- $Dist^L$ is the maximal verb-noun distance for all VMWE tokens occurring at least twice in $SeenVN^L$. This should help eliminate candidates whose components are too distant in a sentence.
- $P^L_{Dep}(Dep_V, Dep_N)$ is the ratio of VMWE tokens in $SeenVN^L$ in which the incoming dependencies of the verb and of the noun are Dep_V and Dep_N. For instance, $P^{FR}_{Dep}(root, obj)$ is higher than $P^{FR}_{Dep}(root, nsubj)$ because, in French, active voice (e.g. ***rendre*** *une* ***visite*** 'pay a visit') is more frequent than passive voice (e.g. ***malediction*** *fut* ***lancée*** 'curse was cast'). We thus favour the most commonly observed VMWE dependencies.
- $P^L_{Dist}(i)$ is the ratio of VMWE tokens in $SeenVN^L$ in which the number of words inserted between the verb and the noun is i. For instance, $P^{FR}_{Dist}(0) = 0.46$, i.e. occurrences in which the verb and the noun are contiguous represent 46% of $SeenVN^{FR}$. This ratio tends to decrease as i increases: $P^{FR}_{Dist}(2) = 0.11$, $P^{FR}_{Dist}(5) = 0.006$, etc. Candidates whose number of intervening words i has higher $P^L_{Dist}(i)$ likely are true VMWEs.

Given these characteristics of seen VMWEs, we proceed to extracting and ranking unseen VMWE candidates. Namely, $Cand^L$ is the list of all occurrences of verb-noun pairs in Test such that: (i) the verb and the noun are directly connected by a syntactic dependency, (ii) the distance between the verb and the noun does not exceed $Dist^L$, and (iii) the verb and the noun never co-occur with a direct dependency link in Train or in Dev. The latter condition excludes both seen VMWEs (already covered by Seen2Seen) and verb-noun constructions not annotated as VMWEs in Train or Dev, i.e. being no VMWEs, e.g. (FR) *avoir an* '*have year*' in *elle a quinze ans* '*she is 15 years old*'. $Cand^L$ is then ranked by considering statistical properties. For each candidate c in $Cand^L$, we calculate three measures:

- $P(c)$ is the estimated joint dependency- and distance-based probability. Suppose that i is the number of words inserted between c's verb and noun, and their incoming dependencies are Dep_V and Dep_N, respectively. Then, $P(c) = P^L_{Dep}(Dep_V, Dep_N) \times P^L_{Dist}(i)$.

[11] In this way, we limit the lexical replacement to only these components whose variability within VMWEs is attested in Train. We previously applied this method to all seen VMWEs but the results were too noisy.

List	DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
MIX^L	0 (0)	0.31 (42)	0.34 (41)	0.57 (21)	0 (0)	0 (0)	0 (0)	0 (0)	0.40 (20)	0.29 (35)	0 (0)	0 (2)	0.48 (21)	0.29 (7)
SIM^L	0 (0)	0 (0)	0.45 (11)	0.17 (6)	0 (0)	0 (0)	0 (0)	0 (0)	0.33 (3)	0.20 (20)	0 (0)	0 (0)	0.33 (6)	0 (0)
$RANK^L$	0.19 (101)	0.05 (228)	0.09 (329)	0.19 (159)	0.21 (137)	0.04 (129)	0.46 (273)	0.11 (163)	0.14 (164)	0.08 (225)	0.03 (422)	0.21 (100)	0.15 (214)	0 (32)

Table 1: Unseen MWE-based precision (and number of predicted VMWEs) in Test for the 14 languages L, when using only MIX^L, SIM^L or $RANK^L$ lists.

- $AMI(c)$ is the augmented mutual information of c's type in the CoNLL-ST corpus. MWEs are known to have a Zipfian distribution and to often mix very frequent words with very rare ones. AMI is designed specifically to address this phenomenon, so as to leverage the rarely occurring expressions or components (Zhang et al., 2009): $AMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)(1-\frac{P(x,y)}{P(x)})(1-\frac{P(x,y)}{P(y)})}$
- $RR(c)$ is the reciprocal rank combining the two indicators above. Let $rank_P(c)$ and $rank_{AMI}(c)$ be the ranks of c in $Cand^L$ according to the values of $P(c)$ and $AMI(c)$ with $P(c) > 0$ and $AMI(c) > 0$. Then $RR(c) = \frac{1}{rank_P(c)} + \frac{1}{rank_{AMI}(c)}$.

$Cand^L$ is then ranked by $RR(c)$. We keep n top-ranked candidates, where n is estimated by scaling the number (provided the organizers) of VIDs and LVCs in Test – when all the expressions annotated as seen during the Seen2Seen phase have been eliminated – by the recall of our method on Dev on the target constructions (unseen verb-noun LVCs and VIDs).[12] This n-best list is called $RANK^L_n$.

④ *Identification proper*: In step ③ we obtain a list of unseen VMWE candidate tokens $Cand_L$ extracted from Test. The aim of identification is to discriminate among true and false VMWEs on this list. Statistical ranking and retaining top-n candidates is one possible statistically-based criterion. But we hypothesise that some candidates whose rank is worse than n, notably due to data sparseness, can still be correct if they result from lexical replacement or translation of seen VMWEs. Therefore, every c in $Cand^L$ is annotated as an LVC if c belongs to $RANK^L_n$ or if c's type belongs to $MIX^L \cup SIM^L \cup TRANS^L$.

3 Results

Although Seen2Unseen uses 4 lists of candidates, here we analyse their contribution separately, that is, we use one list at a time in step ④ above. We report unseen MWE/token-based precision.[13] Sec. 3.1 analyses the impact of MIX^L, SIM^L and $RANK^L_n$, while Sec. 3.2 discusses $TRANS^L$ for French.

3.1 Impact of MIX^L, SIM^L and $RANK^L_n$

As shown in Table 1, using MIX^L alone leads to precision values above 0.29 for 7 languages out of 14. Conversely, $RANK^L$ alone mostly leads to values below 0.22 (except for Hindi with P = 0.46). The precision using SIM^L alone reaches a maximum of 0.45 for Basque. The error analysis below suggests ways to improve precision.

In French, using MIX^{FR} alone yields 21 candidates in Test. Among the 5 false positives, there is one literal reading (*faire dessin* '*make drawing*'), one omitted VMWE (***recevoir aide*** 'receive help') and three other verb-noun pairs that could have been disregarded (being coincidental occurrences) if we had taken into account not only the existence of the syntactic dependency but also its nature (e.g. *nous avons*$_{\text{VERB}}$ *cinq points à l'ordre*$_{\text{NOUN.xcomp}}$ *du jour* '*we have five items on the agenda*').

This major problem for MIX^L is shared by SIM^L, but a specific drawback with SIM^L is that not all words that occur in similar contexts are actually similar. Indeed, we obtain relevant generated unseen

[12] When the proportion of VIDs and LVCs in Test is unknown, it can be approximated by the analogous proportion in Dev.

[13] Shortly before submitting the final version of this paper the definition of a seen VMWE was updated by the PST organizers. Initially, a VMWE from Test was considered seen if a VMWE with the same (multi-)set of lemmas was annotated at least once in Train. Now, it is considered seen if it is annotated in Train or in Dev. In this paper we report on the evaluation results conforming to the previous definition. The change in definition probably (slightly) impacts the results on seen VMWEs but does not impact the general scores (cf. Sec. 1).

verb-noun pairs, including synonyms, antonyms and hyponyms, but also irrelevant ones. We should therefore either use more reliable resources, such as synonym/antonym dictionaries, and/or disregard frequent verbs (*to have, to do*, etc.). For these frequent verbs, the more reliable equivalences obtained by MIX^L compared to SIM^L should be preferred (*faire* '*do*' $\overset{MIX^{FR}}{=}$ *subir* '*suffer*' vs. *faire* '*do*' $\overset{SIM^{FR}}{=}$ *passer* '*pass*'). Indeed, as shown in Table 1, over 5 languages with MIX^L and SIM^L candidates, 4 exhibit a better precision and higher number of candidates for MIX^L.

In French, by dividing n by 4 in $RANK_n^{FR}$, the precision would have increased from 0.19 to 0.45 (18 VMWEs over 40 candidates). In other words, using $RANK_n^L$ in step ④ can slightly increase recall but causes a drop in precision, unless n is low. Hindi appears as an exception: no negative impact is observed with $RANK_n^{HI}$ due to a bias in the corpora (*compound* mentioned in the dependency label).

3.2 Impact of $TRANS^L$: (IT) *Traduttore, traditore* 'translator, traitor'?

With translational equivalences, we hypothesized that $TRANS^L$ would lead to situations such as:

- exact matches: (PT) ***cometer crime*** '*commit a crime*' → (FR) ***commettre crime*** ,
- partial matches leading to VMWEs nonetheless: (PT) ***causar problema*** '*cause problem*' → (FR) ***causer ennui***, instead of ***causer problème***,
- no match, but another VMWE: (PT) ***ter destaque*** '*highlight*' → (FR) ***mettre en évidence***.
- literal, non-fluent or ambiguous translations (Constant et al., 2017): (PT) ***jogar o toalha*** '*throw the towel*'⇒'give up' → (FR) *jeter la serviette* instead of ***jeter l'éponge*** '*throw the sponge*',
- non-existing VMWEs in the target language: (TR) ***el atma*** → (FR) *lancer main* '*throw hand*'

We focus on French due to the high number of candidates in $TRANS^{FR}$. In Test-FR, among the 44 annotated verb-noun candidates using $TRANS^{FR}$ alone, 18 are actually VMWEs and 3 partially correspond to VMWEs due to omitted determiners, yielding an unseen MWE-based precision of 0.41 and an unseen token-based precision value of 0.48. These 21 candidates are mainly provided by Greek (10 vs. 6 from PT and 0 from IT or RO). Thus, the size of the training corpora may have more influence on the probability to obtain good translations than the source language family.

The 23 false positives include (i) 13 candidates that can be VMWEs or not depending on the context, including coincidental co-occurrences, literal readings and errors in the manually annotated reference Test corpus, and (ii) 10 candidates that are not VMWEs, whatever the context, e.g. the inchoative *commencer recherche* '*start research*' (from Hebrew) or *payer taxe* '*pay tax*'(from (PL) ***uiszczać opłatę***).

Consequently, translation may be a clue to discover unseen VMWEs, since 78% of $Cand^{FR} \cap TRANS^{FR}$ are VMWEs out of context, but barely half of them were manually annotated in context. As highlighted above, a restriction to the most frequent VMWE syntactic relations could help filter out coincidental occurrences corresponding to 39% of false positives (e.g. *lancer la balle à la main*$_{\text{OBL:MOD}}$ '*throw the ball with the hand*').

4 Conclusions and Future Work

We proposed an error analysis for our system Seen2Unseen dedicated to unseen verb-noun VMWE identification. It reveals that lexical variation and translation can produce valid unseen VMWEs but their ambiguity in context must be solved: we should take into account both the dependency labels (to avoid coincidental occurrences) and the probability of the verb to be light in Train (to avoid frequent co-ocurrences like *fumer cigarette* '*smoke cigarette*'). Using contextual rather than non-contextual word embeddings might also be helpful, even if computationally more intensive. We could also combine $TRANS^L$ and $MIX^L \cup SIM^L$ by applying lexical substitution to the translated VMWEs.

Acknowledgements

This work was funded by the French PARSEME-FR grant (ANR-14-CERA-0001). We are grateful to Guillaume Vidal for his prototype, and to the anonymous reviewers for their useful comments.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)*, online, December. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. ACL. `https://aclweb.org/anthology/W18-4925`.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019a. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy, August. Association for Computational Linguistics.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa I nurrieta, and Voula Giouli. 2019b. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54, April.

Wen Zhang, Taketoshi Yoshida, Tu Bao Ho, and Xijin Tang. 2009. Augmented mutual information for multi-word extraction. *International Journal of Innovative Computing, Information and Control*, 5(2):543–554.