HMSid and HMSid2 at PARSEME Shared Task 2020: Computational Corpus Linguistics and *unseen-in-training* MWEs

Jean-Pierre Colson
University of Louvain
Louvain-la-Neuve, Belgium
jean-pierre.colson@uclouvain.be

Abstract

This paper is a system description of HMSid, officially sent to the PARSEME Shared Task 2020 for one language (French), in the open track. It also describes HMSid2, sent to the organizers of the workshop after the deadline and using the same methodology but in the closed track. Both systems do not rely on machine learning, but on computational corpus linguistics. Their score for unseen MWEs is very promising, especially in the case of HMSid2, which would have received the best score for unseen MWEs in the French closed track.

1 Introduction

Although the PARSEME Shared Task 2018 (Savary et al., 2018) produced very interesting results for the extraction of verbal multiword expressions, one important note of caution has to be made: the participating systems produced poor results for unseen MWEs, i.e. expressions that were absent from the training data. As pointed out by the organizers of the new Parseme Shared Task 2020[1], a possible solution to this issue is the recourse to large MWE lexicons.

In this paper, however, we report the results of two systems offering promising results for *unseen* MWEs with no recourse to MWE lexicons: HMSid (*Hybrid Multi-layer System for the extraction of Idioms*) and HMSid2. Both systems are based on computational corpus linguistics: they just used the training data and an additional general linguistic corpus. As the models require a fine-tuned adaptation to each language under study, they were only applied to the French dataset of the PARSEME Shared Task 2020.

HMSid used as an external corpus the French WaCky corpus (Baroni et al., 2009) and was submitted to the PARSEME Shared Task 2020. As there was a recourse to an external corpus, it was logically put in the open track. Thanks to the feedback from the organizers of PARSEME 2020, however, we adapted the system in order to propose it in the closed track: the corpus used was the Wikipedia corpus included in the training data. The new version, HMSid2, was sent to the organizers after the official deadline. In this paper, both the official results of HMSid and the new results from HMSid2 are discussed.

Our theoretical starting point for both systems is that, while Deep Learning will surpass most techniques for reproducing elements that are somehow present in training sets, it will need additional corpus-

[1] Introduction to the PARSEME Shared Task 2020, http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020___lb__COLING__rb__&subpage=CONF_40_Shared_Task

based information for unseen-in-training MWEs. It should also be pointed out that MWE extraction is a daunting practical task, but that the theoretical background is also very complex, as it is related to grammatical and semantic structure. Information retrieval (Baeza-Yates and Ribeiro-Neto, 1999) has shown that semantic relations may be analyzed by very diverse methods, including vector space models and clustering methods. Many of its findings are compatible with the Distributional Hypothesis (Harris 1954): differences in meaning will be reflected by differences in distribution. However, the distribution of words is also affected by existing MWEs, as at least 50 percent of the words from any text will actually be included in MWEs, collocations or phraseological units (Sinclair, 1991). In addition, a wide array of studies in construction grammar (Hoffmann and Trousdale, 2013) strongly suggest that language structure consists of a very complex and probabilistic network of constructions at various levels of abstraction and schematicity.

It is no wonder then that very complex techniques are necessary for extracting MWEs, in much the same way as for the extraction of semantic links. In particular, the complex interplay between 1st-order co-occurrence (words appear together) and 2nd-order co-occurrence (words appear in similar contexts, Lapesa and Evert, 2014) probably requires a hybrid methodology. While deep learning and in particular neural networks are very efficient ways of gaining information from a training set, it may be complemented by a more traditional, corpus-based approach in the case of the extraction of data that are unseen in the training set.

The technical background for HMSid and HMSid2 is a combination of techniques inherited from Information Retrieval, such as *metric clusters* (Baeza-Yates and Berthier Ribeiro-Neto, 1999) and a query likelihood model, with a big data approach, in this case a large (unparsed and untagged) linguistic corpus: the French WaCky for HMSid and the Parseme French training corpus (Wikipedia) for HMSid2. As described in Colson (2017; 2018), a clustering algorithm based on the average distance between the component parts of the MWEs is measured, the *cpr-score (Corpus Proximity Ratio)*:

$$cpr = \frac{n(w_1, w_2, \dots, w_n)}{n\left(x_{t_1} = w_1, x_{t_2} = w_2, \dots, x_{t_n} = w_n \,\middle|\, \max(t_{i+1} - t_i) \leq W;\ i = 1, \dots, n-1\right)}$$

Figure 1. The *cpr-score*

This approach, as opposed to vector models, is a 1st-order model, as it is based on the co-occurrence of words and not on similar contexts. Given a window *W* of x tokens (depending on the language and the corpus, typically set at 20 for MWEs), the score simply measures the ratio between the number of exact occurrences of an n-gram, divided by the number of occurrences with a window between each gram. The main advantages of this metric are that it is not limited to bigrams, and that semantic links may be captured as well by enlarging the window, a point that has also been made by Lapesa and Evert (2014): larger windows may enable 1st-order models to capture semantic associations.

Experiments with large datasets of idiomatic MWEs have shown (Colson, 2018) that most formulaic and idiomatic constructions can be captured by co-occurrence clusters, provided that the corpus used is sufficiently large (at least 1 billion tokens). In order to reach a good compromise between results that could be extracted from co-occurrence in large corpora and recurrent patterns with specific categories of MWEs, a hybrid methodology was used, as detailed in the following section.

2 Methodology used for HMSid and HMSid2

In the PARSEME Shared Task 2020 for French, the following categories of verbal MWEs had to be extracted from the test set: IRV (inherently reflexive verbs, as in the English example *to help oneself*), LVC.cause (light-verb constructions in which the verb adds a causative meaning to the noun, as in the English *to grant rights*), LVC.full (light-verb constructions in which the verb only adds meaning expressed as morphological features, as in *to give a lecture*), MVC (multi-verb constructions, as in *to make do*) and VID (verbal idioms, e.g. *to spill the beans*).

After a number of preliminary tests, we decided to extract French MWEs from the test set in a two-step process. The first step concerned all categories of verbal MWEs, as described above, except the last one (VID, verbal idioms). The second step was just devoted to verbal idioms.

This two-step approach was motivated by the unpredictable character of verbal idioms: contrary to the other categories of MWEs used for the PARSEME Shared Task, idioms display a very irregular number of elements, of which the syntactic structure is also diverse.

During the first step, we used a Perl script and the Data::Table module[2] for storing each sentence at a time in RAM memory. For the categories IRV, LVC.cause, LVC.full and MVC, the specific syntactic features of these categories were taken into account by the algorithm: in the case of IRV, for instance, the parsed sentences provided by the PARSEME dataset made it easy to extract all pronouns preceding or following the verbs, and an additional check was performed in order to determine whether those pronouns were indeed French reflexive pronouns, including elision (e.g. the pronominal form *s'* instead of *se*). For LVC.cause, a list of French causative verbs was extracted from the training data (for instance *apporter, causer, créer, entraîner*). In the extraction phase, all objects depending on such causative verbs were measured by our co-occurrence score, the *cpr-score* (Colson, 2017; 2018) and the highest values were considered as cases of LVC.cause constructions. For LVC.full, a similar methodology was used, taking into account all subjects (for passive constructions) and objects (for direct object constructions) depending on verbs, excluding causative verbs, with a medium-range association between the subject/object and the verb (computed by the *cpr-score*). In the same way, the MVC category was extracted on the basis of the degree of association between two successive verbs, as in *faire remarquer* (to point out).

In the second step of our extraction methodology, verbal idioms were extracted and added to the results. This made it possible to add the category of verbal idioms in the labels of the final results if and only if the results had not yet received another category label, for instance LVC.full. Preliminary tests on the basis of the training data indeed revealed that our algorithm tended to assign the VID category quite often, whereas the annotators of the gold set had been rather strict as to the idiomatic character of verbal MWEs. Using two separate scripts was a simple way of avoiding interference in the results.

In the Perl script devoted to the extraction of VIDs, we also used the Data::Table module and selected in the parsed data all verbs, all their complements, and all complements of each complement. Extensive testing with the training data showed that this approach yielded higher scores than an n-gram based approach, in which the successive grams of each verb were analyzed left and right.

3 Results and discussion

Table 1 below displays the results obtained for HMSid, our system that was officially sent to the PARSEME Shared Task 2020. As explained in section 2, HMSid relied on an external corpus and was therefore placed in the open track.

Table 2 shows the results obtained with HMSid2, using the same methodology but relying solely on the training data and the training corpus, and therefore belonging to the closed track. The results with HMSid2 were sent to the organizers of the Shared Task after the deadline.

System	Track	Unseen MWE-based				Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank	P	R	F1	Rank
HMSid	open	27.73	53.33	**36.49**	4	63.85	67.84	**65.79**	5	66.4	67.81	**67.1**	5

Table 1: Global results obtained with HMSid at the PARSEME 2020 Shared Task (French).

[2] https://metacpan.org/pod/Data::Table

System	Track	Unseen MWE-based				Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank	P	R	F1	Rank
HMSid2	closed	32.53	49.33	**39.21**	1	68.90	72.04	**70.43**	2	71.10	72.63	**71.86**	2

Table 2: Global results obtained with HMSid2 at the PARSEME 2020 Shared Task (French).

As shown in Table 1, HMSid obtained a global F1 (Token-based) of 67.1, which puts it in 5th position in the open track. It should be noted, however, that its F1-score on unseen MWEs (36.49) puts it in 4th position (and very close to the 3d one), while its recall for unseen French MWEs is the best of all systems, open or closed track (53.33). This is noteworthy, because HMSid (and HMSid2) do not try to reproduce recurrent patterns from the training set, but rely on statistical extraction from a large linguistic corpus. In other words, both systems do not try to reproduce decisions made by annotators, as reflected in the training set, but are looking for statistical patterns in a large linguistic corpus, regardless of the training set. Of course, the training set was used for fine-tuning the statistical thresholds and deciding whether a combination was a MWE or not, and the different categories (which are in itself debatable, such as the distinction between LVC.full and LVC.cause) were integrated into the statistical extraction. The recall score on unseen MWEs also provides additional evidence of the statistical nature of recurrent MWEs in large linguistic corpora.

This is even more obvious with HMSid2, which used exactly the same methodology, as explained in the above section, but relied on the training corpus provided by the Shared Task (part of the Wikipedia corpus), and would therefore be placed in the closed track. Among the 3 systems submitted to the closed track for French, HMSid2 would receive rank 2 for the global F1-score (MWE-based or Token-Based), and rank 1 for unseen MWEs, with an F1-score (39.21) far better than those obtained by the other systems in the closed track (with respectively 24.4 and 3.67). The best overall system officially submitted to the French closed track (Seen2Seen) has an F1-score of 3.67 for unseen MWEs.

The difference between precision and recall, especially for unseen MWEs, should also be relativized by the choices made in the training and gold set. In spite of the excellent quality of the PARSEME annotated dataset, decisions as to the idiomatic character of a MWE will never be unanimous. In the case of the French dataset, for instance, the notion of verbal idiom (VID) was taken strictly by the annotators, but there are a few notable exceptions. A number of less idiomatic constructions were also labeled as VIDs. For instance, *avoir lieu* (to take place), *il y a* (there is / there are), *mettre en pratique* (to put into practice), *tenir compte de* (take into account), are all considered French verbal idioms in the training data, a choice that may be respected but has consequences on the statistical extraction. The statistical metric indeed had to be more tolerant for weaker associations when assigning the label 'VID', which contributed to a fairly good recall but a slightly lower precision. This appears clearly in all results from Tables 1 and 2, and in particular for unseen MWEs. In this case, one should bear in mind that the algorithm is looking for recurrent patterns in the linguistic system itself, as there are no similar examples in the training set.

Many cases of verbal idioms from the gold set are quite obvious, such as *tourner le dos à* (turn one's back on, lines 5817-19 of the gold set), *il pleuvait des cordes* (it was raining cats and dogs, lines 7415-17) or *sortir le grand jeu* (pull out all the stop, lines 12129-32), all three labelled as VID and also recognized by the algorithm because of the very strong association between the grams: a *cpr-score* of resp. 0.92 / 0.88 / 0.94. In other cases, however, the algorithm and the annotators are at odds. In lines 5868-9, for instance, *rester silencieux* (remain silent, keep quiet) is not considered as MWE by the annotators, but the *cpr-score* contradicts this view: 0.81. The same holds true of many other examples, such as *trouver un compromis* (lines 14387-89), not considered as a MWE in the gold set, but displaying a *cpr-score* of 0.80. In this specific case, it should be reminded that native speakers are not always the best judges of the idiomaticity of their own language. It may be pretty obvious for speakers or French and of English that a compromise may be *found* but a quick look at other European languages reveals that this is far from being the case: in Spanish, for instance, the common construction is *llegar a un compromiso*.

It should also be pointed out that the methodology used for HMSid and HMSid2 is easily applicable to other languages. As a matter of fact, we have already implemented it as an experimental web tool[3], *IdiomSearch* for English, German, Spanish, French, Dutch and Chinese. Measuring associations based on the *cpr-score* is indeed possible for any language, provided that the necessary web corpus is compiled. The only caveat is the goal of the classification. The Parseme Shared Task 2020, as the previous editions, wanted the systems to target very specific categories of verbal expressions, whereas our experimental tool *IdiomSearch* looks for recurrent statistical associations, whatever the precise category may be. Fine-tuning the algorithm to specific categories expected by the gold set, and annotated as such by native speakers of the language requires sophisticated training algorithms such as those used in deep learning.

In conclusion, the most interesting results from HMSid and HMSid2 are those obtained for unseen MWEs. Due to the well-known phenomenon of overfitting, deep learning models often have problems with unseen data, which suggests that a hybrid approach combining deep learning and our model may be useful for future research.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press /Addison Wesley, New York.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43: 209–226.

Colson. 2017. The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In Ruslan Mitkov (ed.), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596*. Springer International Publishing, Cham: 16–28.

Colson. 2018. From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin. In Agatha Savary et al. 2018: 41-50.

Zellig Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Thomas Hoffmann and Graeme Trousdale (eds.). 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Agata Savary, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan and Miriam R. L. Petruck (eds.). 2018. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, *Coling 2018*, Santa Fe NM, USA, Association for Computational Linguistics.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford, Oxford University Press.

[3] https://idiomsearch.lsti.ucl.ac.be