Word sense induction for (French) verb valency discovery

Naïma Hassert, François Lareau

OLST, Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal QC, H3C 3J7, Canada
E-mail: first.lastname@umontreal.ca

Abstract

We explore the use of Transformers in word sense induction for the automatic construction of a valency dictionary of French verbs. To account for the way the arguments of a verb change depending on its sense, this type of dictionary must distinguish at least the main senses of a lemma. However, constructing such a resource manually is very costly and requires highly trained staff. That is why one important subtask in the construction of this resource is to automatically identify the polysemy of the verbs. For each of the 2,000 most frequent French verbs, we extract the word embeddings of 20,000 of their occurrences in context found with Sketch Engine, and we cluster those embeddings to find the different senses of each verb. In order to identify the language model and clustering algorithm most suited to our task, we extract the word embeddings of the sentences in the FrenchSemEval evaluation dataset with one language-specific model, CamemBERT, and two multilingual models, XLM-RoBERTa and T5. These vectors are then clustered with three different algorithms that do not require a predetermined number of clusters: Affinity Propagation, Agglomerative Clustering and HDBSCAN. Our experiments confirm the potential of unsupervised methods to identify verb senses, and indicate that monolingual language models are better than multilingual ones for word sense induction tasks involving a single language.

1. Introduction

Valency dictionaries such as DEM (Dubois & Dubois-Charlier, 2010), Dicovalence (van den Eynde et al., 2017), Le*fff* (Sagot, 2010), LVF (Hadouche & Lapalme, 2010), VerbNet (Kipper et al., 2006) and Verbənet (Danlos et al., 2016) are useful in many natural language processing applications, in particular for rule-based natural language generation. This type of dictionary indicates precisely how a predicate expresses its arguments in syntax, including information on selected part-of-speech, preposition or case. However, the way a word expresses its arguments can change significantly depending on its sense. For example, the verb *change* requires a direct object when it means 'modify', as in *The discussion has changed my thinking about the issue*, but with the sense 'become different', as in *She changed completely as she grew older*, then there is no object at all (examples taken from WordNet[1]; Fellbaum 1998). Therefore, a valency dictionary must distinguish at least the main senses of a lemma. Constructing this kind of resource manually, however, is very costly and requires highly trained staff.

Our goal is thus to automate the construction of valency dictionaries. In this paper we focus on how we tackled an important subtask: automatically identifying the polysemy of verbs. Our data is drawn from French, but the method we present here is language-independent.

[1] https://wordnet.princeton.edu/

Since our goal is to produce a resource entirely automatically, we want to use raw data as material and rely on as little external resources as possible. This comes down to a word sense induction (WSI) task. Several WSI techniques have been introduced as early as the 1990s, e.g., context clustering (Schütze, 1998), word clustering (Lin, 1998) or co-occurrence graphs (Véronis, 2004). However, the field has been revolutionized with the arrival of Transformers (Vaswani et al., 2017), which can produce high quality contextualized word embeddings in several languages.

We tackled this WSI task in three main steps: first, we extracted contextualized vectors of the sentences in the FrenchSemEval evaluation dataset (Segonne et al., 2019) with one language-specific model, CamemBERT (Martin et al., 2020), and two multilingual models, XLM-RoBERTa (Conneau et al., 2020) and T5 (Raffel et al., 2020). This dataset is comprised of 66 French verbs in context, each having around 50 sense-annotated examples. Then, we tested three unsupervised clustering algorithms that don't require knowing the number of clusters beforehand: Affinity Propagation (Frey & Dueck, 2007), Agglomerative Clustering (Szekely & Rizzo, 2005) and HDBSCAN (McInnes et al., 2017). The best results were achieved with CamemBERT vectors clustered with Agglomerative Clustering, obtaining a BCubed F_1 score (Bagga & Baldwin, 1998) of 65.20 %. As a comparison, the FlauBERT team (Le et al., 2020), also using CamemBERT vectors, obtained an F_1 score of 50.02 % on the same dataset, although they used a supervised method and measured their results with the traditional F_1 score, which could not be used in our case since we used an unsupervised method. Finally, for each verb present in the evaluation dataset, we add the word embeddings of 20,000 instances of this verb in context extracted via Sketch Engine[2] (Kilgarriff et al., 2014). We then cluster each group of approximately 20,050 verbs separately (the 20,000 verbs in context previously extracted, plus the 50 examples from the evaluation data), and evaluate the performance of the clustering on the evaluation data. Our experiments allow us to pinpoint the best combination of language model, clustering algorithm and parameter to identify the senses of a verb from raw data.

This paper begins with a brief summary of previous work in the WSI field in §2. Follows a presentation of the language models (§3.1) and the clustering algorithms §3.2 on which we experimented. Then, we describe in §4 how we evaluated the combinations of those algorithms. Finally, we present in §5 an analysis of our results, and conclude in §6.

2. Automatic identification of the polysemy

2.1 Word sense disambiguation (WSD)

The automatic identification of the sense of an ambiguous word in context has been a research topic for decades and is still an unresolved task. Yet, it is crucial in many applications, such as:

- automatic translation, where a word in a language can have many different translations in another;
- information retrieval, where search queries often contain ambiguous words;
- information extraction, where we want to automatically retrieve specific information related to a specific topic;

[2] https://www.sketchengine.eu

- lexicography, where we often want to obtain lexical information specific to a given word sense.

A common way of tackling this task is by using a knowledge-based method or a supervised one. Knowledge-based methods rely heavily on existing resources like WordNet (Fellbaum, 1998), BabelNet (Navigli & Ponzetto, 2012), FrameNet (Baker, 2014) or other dictionaries, and use the content of those resources to compare with the data on hand and deduce the word sense. Supervised methods rely instead on sense-annotated data, which is then used to annotate raw data. Most state-of-the-art methods are hybrid, i.e., combine features of knowledge-based and supervised methods (Bevilacqua et al., 2021).

In the context of dictionary creation, however, knowledge-based or supervised methods are not necessarily the most appropriate way to identify the sense of an ambiguous word, for the following reasons:

1. **The senses listed in major lexical resources are often too fine-grained.**
   A popular lexical resource in the field of natural language processing (NLP) is WordNet, an electronic dictionary of English based on *synsets*, i.e., sets of synonymous lexemes. If one looks up a word in WordNet, one ends up with all the synsets that contain it. In the case of *change*, for instance, there are 10 synsets related to the noun *change*, and 10 synsets related to the verb *change*. Its multilingual counterpart, BabelNet (Navigli & Ponzetto, 2012), is a result of the merging of WordNet and Wikipedia, where the synsets are provided in part by the human-generated translations provided by Wikipedia and in part by a machine translation system. It has been pointed out, however, that WordNet's senses are very fine-grained, to the point where inter-annotator agreement when using the WordNet inventory is around 70 % (Navigli, 2006), only 5 % more than the most frequent sense (MFS) baseline, which consists of annotating each word with its most frequent sense (Raganato et al., 2017).

2. **Most resources are based on English.**
   WSD systems rely heavily on sense-annotated data. This type of data exists in English, thanks mainly to SemCor (Miller et al., 1993), which is sense-annotated based on WordNet. However, since manual semantic annotation is very costly, this data is scarce or non-existent for languages other than English. As a result, most of the lexical resources in other languages are derived from the English ones, like Europarl (Koehn, 2005), a corpus annotated with the senses of BabelNet (which itself is derived in part from WordNet). Relying on those resources can thus be misleading if we want to do WSD for, say, French.

3. **Relying on external resources prevents the discovery of new senses.**
   As mentioned earlier, lexical resources have the inconvenience that they are costly to create and update. However, new words and senses are created continually. Thus, hand-curated lexical resources can easily become outdated.

2.2 Word sense induction (WSI)

When WSD is performed without the help of an external resource, it is called word sense *induction* (or *discrimination*). WSI methods can be a solution to the knowledge acquisition bottleneck, since they only rely on raw, non annotated, data. This method does not assign a sense to a word *per se*: instead, it aims to detect *how many* senses there are, based on the assumption that two occurrences of a word have the same sense if they occur in similar contexts. Common approaches in the field are:

- **Context Clustering**
  This algorithm, developed by Schütze (1998), interprets senses as groups, or clusters, of similar contexts of an ambiguous word. More specifically, each word is represented as a vector whose components are counts of the number of times another word appears in its context (the context can be a sentence, a paragraph, or any other length of text). The original algorithm dealt with vectors built from second-order co-occurrences, i.e., where vectors of the words in the context of the ambiguous word are themselves built from their own context. These context vectors can then be clustered into groups based on their similarity. Each group is represented by the mean of all the vectors of this group, namely the *centroid*. This is the method closest to the one we decided to adopt in this paper.

- **Word Clustering**
  This algorithm has been developed by Lin (1998). It identifies words that are similar to a target word based on their syntactic dependencies. The context is parsed syntactically and represented as triples, each of them consisting of the target word, a syntactic dependent and the syntactic relationship between them. Common information between two words are the triples that appear in the description of both of the words. One can then use this information to calculate the similarity between two words. Finally, a tree is created with the help of a clustering algorithm. The nodes directly under the main node are considered as the different senses of the word.

- **Co-occurrence graph**
  Véronis (2004) presented HyperLex, arguing that the problem with clustering vectors is that it can exclude less frequent word senses, which will tend to be considered as noise by the algorithm even if those senses are not rare ones for an average speaker. In an attempt to solve this problem, a graph is built where the nodes are words and they are connected according to their co-occurrences in a given context size. One ends up with *small worlds*, i.e., highly connected groups that are said to correspond to a sense and that are all linked in some way.

- **Recurrent neural networks**
  Recurrent neural networks (RNNs) are a class of artificial neural networks that recursively define the output at any stage as a function of the previous output. They have been useful for several NLP tasks, but suffer from the vanishing gradient problem, which makes them only possible to use in short sequences. Long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) is a RNN variation that avoids the vanishing gradient problem to a certain extent and allows recurrent networks to learn over many more steps. However, they require a lot of resources and time to train, and still do not have a huge memory.

- **Transformers**
  Transformers were first introduced by Vaswani et al. (2017) and have revolutionized

the field. Their attention mechanism allows them to process the entire input at once, reducing training times drastically and achieving state-of-the-art results in NLP. Transformer models pretrained on huge datasets can be easily downloaded from Hugging Face[3] and further trained.

3. Method

3.1 Word embeddings

We used three different language models for our experiments, all downloaded from Hugging Face. We used one monolingual model for French, CamemBERT, and two multilingual models, XLM-RoBERTa and T5. Monolingual models have been shown to yield better results than multilingual language models such as mBERT (Martin et al., 2020). However, in 2021, XLM-RoBERTa showed impressive results on the SemEval-2021 Task 2: Word in Context Disambiguation (Martelli et al., 2021), including for French, so we included it in our experiments. We also experimented on another multilingual language, T5, released by Google. We used the large version of each model, and got our contextualized vectors by calculating the mean of all hidden layers.

CamemBERT (Martin et al., 2020) is a monolingual model constructed especially for French. Its architecture is based on RoBERTa's, a method that builds on BERT's language masking strategy while modifying key hyperparameters and training with much larger mini-batches and learning rates. RoBERTa reportedly have better downstream task performance than BERT (Liu et al., 2019). CamemBERT is trained on 138 GB of raw data. On its release in 2020, it has improved the state of the art for part-of-speech tagging, dependency parsing, named entity recognition and natural language inference tasks for French.

XLM-RoBERTa (Conneau et al., 2020) is a multilingual language model pretrained on 2.5 TB of data from 100 languages. At its release, it showed a very significant improvement over the multilingual models mBERT and XLM-100, and obtained competitive results over state-of-the-art monolingual models, including RoBERTa, in English. They demonstrated that multilingual models can improve their monolingual BERT counterpart. To the best of our knowledge, however, CamemBERT and XLM-RoBERTa have not been compared for WSI, so we have yet to verify whether XLM-RoBERTa can improve on its monolingual counterpart for French.

T5 (Raffel et al., 2020) is an alternative to BERT. Instead of having class label or a span of the input as outputs, as with BERT-style models, T5 has text string only as input as well as output. T5-large, the checkpoint that we used, has 770 million parameters. T5 was trained on a dataset containing 4 languages: English, French, German and Romanian.

3.2 Clustering

Clustering methods aim at finding structure in a set of unlabeled data. Several clustering algorithms need a number of cluster beforehand; however, since our goal is to eventually be able to find senses that have not been listed in a lexical resource, we tested three clustering algorithms that can choose the optimal number of clusters without being explicitly told.

[3] https://huggingface.co

3.2.1 Affinity propagation

Affinity propagation (Frey & Dueck, 2007) is a clustering algorithm that exchanges messages between data points until members of the input that are representative of clusters, “exemplars”, are obtained. There are two parameters that can be tuned: damping and preference. Damping affects the convergence of the algorithm. Preference adds noise to the similarity matrix, and thus affects the number of clusters.

3.2.2 Agglomerative clustering

Agglomerative clustering (Szekely & Rizzo, 2005) is a type of hierarchical cluster analysis that uses a bottom-up approach. It begins by considering every element in the data as its own cluster and successively agglomerates similar clusters until all clusters have been merged into a single one that contains all the data. In scikit-learn (Pedregosa et al., 2011), instead of specifying the number of clusters, one can simply specify the distance threshold, i.e., the linkage distance threshold above which clusters will not be merged. Basically, it indicates the limit at which to cut the dendrogram tree. We used the default linkage parameter, namely “ward”, and tested distance thresholds ranging from 10 to 300,000.

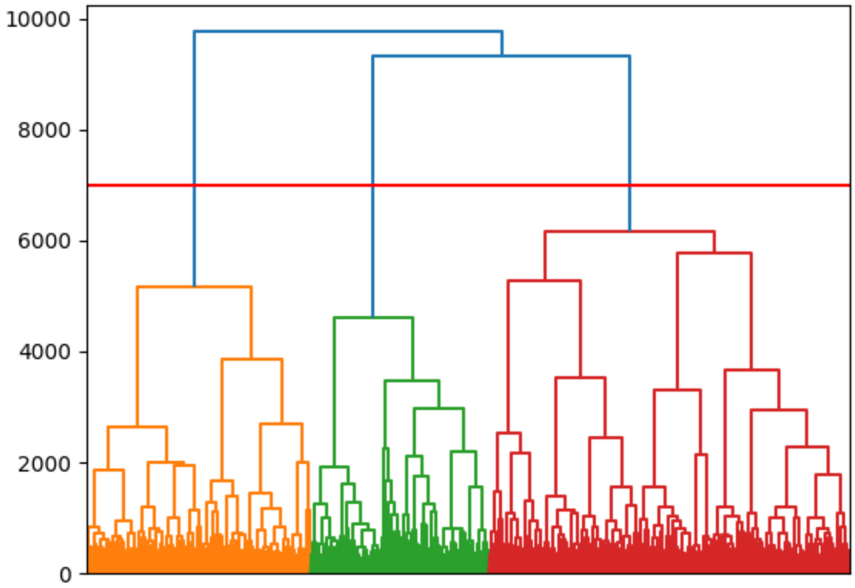

Figure 1: Dendrogram representing the clustering of 20,000 contextual embeddings of the verb *adopter* (‘adopt’) with agglomerative clustering. The horizontal red line represents the final number of clusters (3) obtained with a distance threshold of 7000.

3.2.3 HDBSCAN

HDBSCAN (Campello et al., 2013) is a clustering method that extends DBSCAN by converting it into a hierarchical clustering algorithm. It assumes there is a significant number of noise points, among which one can find islands of higher density. Density methods have the advantage of being efficient even when the data is not clean and the clusters are weirdly shaped. It begins by identifying the densest parts of the data space, and deciding if those densest parts should be merged or kept separate. The algorithm produces a probability score for each data point of belonging to their cluster, and a cluster quality score.

Three main parameters can be fined-tuned in HDBSCAN: the minimum cluster size (the smallest size grouping that is to be considered a cluster), the minimum number of samples (the higher the value, the more conservative the algorithm will be and the more data will be considered as noise) and the clustering selection method (by default "eom"—for *excess of mass*—and can be changed to "leaf", which tends to produce more fine-grained clustering).

4. Evaluation

4.1 FrenchSemEval

FrenchSemEval (Segonne et al., 2019) is an evaluation dataset constructed specifically for the WSD of French verbs. It was built after the authors of this dataset inspected Eurosense (Delli Bovi et al., 2017), a multilingual corpus extracted from Europarl (Koehn, 2005) and automatically sense-annotated using the BabelNet multilingual sense inventory. This resource presented good results in terms of inter-annotator agreement, and for English, the high-precision Eurosense annotations cover 75 % of the content words and have a precision score of 81.5 %. As can be expected, though, the French results are lower: coverage is 71.8 % and precision is 63.5 %. Furthermore, the situation gets worse with verbs, which can be expected since the disambiguation of verbs is known to be more difficult (Raganato et al., 2017). When the authors examined the verbs that had been automatically annotated in Eurosense, they realized that the proportion they judged correct was only 44 %. They also confirmed that BabelNet had a very high number of senses per verb; indeed, on a sample of 150 sentences, they found that the average number of BabelNet senses per verb type occurring in these sentences was 15.5, and that the difference between the senses was sometimes difficult to perceive. In short, like most of the available resources, Eurosense is a resource based on English and thus of a lesser quality for French, and in which senses are too fine-grained.

In contrast, Segonne et al. (2019) observed that in Wiktionary,[4] the granularity level of the senses was usually quite natural and that the sense distinctions were easy to grasp. They thus decided to use the Wiktionary senses as a basis for manual annotation. FrenchSemEval is the result of this effort. It consists of 3,121 sense-annotated sentences, with 66 different verb forms, each having an average of 3.83 senses. All of those verbs were present in the 2,000 most frequent verbs we had identified via Sketch Engine. As indicated in the paper, the MFS baseline for this is 30 % in accuracy.

4.2 MCL-WiC

The Multilingual and Cross-Lingual Word-in-Context Disambiguation task (Martelli et al., 2021) is the first SemEval task to examine the aptitude of systems to discriminate between word senses without any requirement of a fixed sense inventory. The multilingual sub-task is binary: the system must determine if two target words in two different contexts in the

[4] https://www.wiktionary.org

same language has the same meaning or not. The verbs were selected according to their number of senses (it had to have at least three senses in BabelNet) and the sentence pairs were extracted from either the United Nations Parallel Corpus (Ziemski et al., 2016) or Wikipedia. The sentences selected contained sufficient semantic context to determine with certainty the meaning of the target words.

Gupta et al. (2021) got the best result for the Fr-Fr task, attaining 87.5 % accuracy. They obtained fine-tuned contextualized embeddings of the target words from XLM-RoBERTa and passed them to a logistic regression unit. It must be noted that even though we tested XLM-RoBERTa too, our results cannot be directly compared, since we evaluated our results on verbs only (and not on all part-of-speech tags as they did).

4.3 Score measure

In this paper, we use the BCubed F_1 scores. This is because the standard F_1 is designed to compare data that is clustered using the same cluster labels, which is useful if the clusters in question have a specific meaning, but not otherwise. Let us take for example the verb *change* mentioned in the introduction. If we want to find all the tokens that have the sense 'modify' in cluster A and all the tokens that have the sense 'become different' in cluster B, then the cluster labels are important. If all the words put in cluster B should have been instead in cluster A and vice-versa, then the standard F_1 score will be very low.

In our case, though, the cluster labels have no significance: all that matters is to group all the tokens that have similar senses. That is when the BCubed F_1 comes in handy. Instead of calculating the precision and recall based on the number of true and false positives and negatives in all the examples, these scores are calculated for each element individually. The numbers computed for each example in the document are then averaged to produce the recall and precision scores for the entire dataset. The formulas to compute the final BCubed recall and precision are the following:

$$\text{Precision} = \sum_{i=1}^{N} w_i \times Precision_i$$

$$\text{Recall} = \sum_{i=1}^{N} w_i \times Recall_i$$

The formula for the BCubed F_1 score does not differ from the standard one:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In other words, standard F_1 is perfect to evaluate the performance in a WSD task, where the classes are already determined. However, for WSI, we want to evaluate the performance of an algorithm that creates clusters from scratch against an evaluation dataset that will necessarily have its own cluster labels. In this case, BCubed F_1 must be used.

5. Results

5.1 FrenchSemEval

5.1.1 Clustering the evaluation dataset

We clustered the 3,121 sentences of the test set of FrenchSemEval, and calculated BCubed F_1. We report in table 1 the best results for each language model-clustering algorithm combination.

Clustering algorithm	**T5**	**CamemBERT**	**XLM-RoBERTa**
Affinity Propagation	14.86	**14.87**	14.86
Agglomerative Clustering	46.02	**65.39**	56.06
HDBSCAN	30.41	33.76	**35.30**

Table 1: Best BCubed F_1 scores on the FrenchSemEval dataset

As we can see, the combination of Agglomerative Clustering and CamemBERT is by far the best one for our task, yielding impressive results for an unsupervised method. Indeed, the FlauBERT team (Le et al., 2020), using a combination of CamemBERT and a supervised method, attained an F_1 score of 50.02 %. For this clustering method, the distance threshold parameter that allowed each language model to attain the best score varies: in the case of CamemBERT, it was of 650; for T5, it was 100,000; and for XLM-RoBERTa, it was 725.

The worst results we obtained were with Affinity Propagation. Even by doing a grid search with various values of damping and preference, we were not able to achieve more than 14.87 % BCubed F_1. The algorithm achieved a good precision, but a really poor recall in every parameter combination, which indicates that the algorithm was not able to generalize, assigning instead approximately one sense per sentence.

HDBSCAN had the opposite effect: recall was generally much higher than precision, which indicates that it tends to assign only one sense to the entirety of the dataset. The best result with this algorithm was obtained by XLM-RoBERTa, with a minimum cluster size of 10, a minimum number of samples of 2 and the "leaf" cluster selection method. We can also note that an enormous amount of data is considered as noise by the algorithm, and that in almost every parameter configuration, the "leaf" cluster selection method yields much better results than "eom".

5.1.2 Clustering each verb individually

For each verb in the FrenchSemEval dataset, we clustered the 20,000 instances previously collected (cf. §1), to which we had added the 50 or so sentences of the evaluation dataset, and evaluated the performance of our clustering on the evaluation sentences.

It turns out that BCubed F_1 is maybe not the best indicator of the quality of the clustering for our purpose. Indeed, when we increase the distance threshold, recall approaches 100 %, which boosts the score. But it only means that the clustering is more and more severe, so that there is only one cluster or two remaining. For example, if we set the distance threshold to 19,000, it gets to 67.68 %, which is better than our results on the entire dataset. But if we look at the mean number of clusters, we realize that it is not a good clustering: on average, each verb has only one cluster (which is likely not better than the MFS baseline).

The goal could then be to get a number of clusters that is approximately the same as the mean number of clusters for the FrenchSemEval dataset, which is 3.83. We achieve a mean number of clusters of 3.89 with a distance threshold of 6000, with a BCubed F_1 score of 59.93 %, which is still satisfactory. The mean number of clusters goes down as the distance threshold goes up, while, on the contrary, BCubed F_1 goes up (as shown in figure 2); at 7000, the mean number of clusters is 3.13, with a score of 62.75 %.

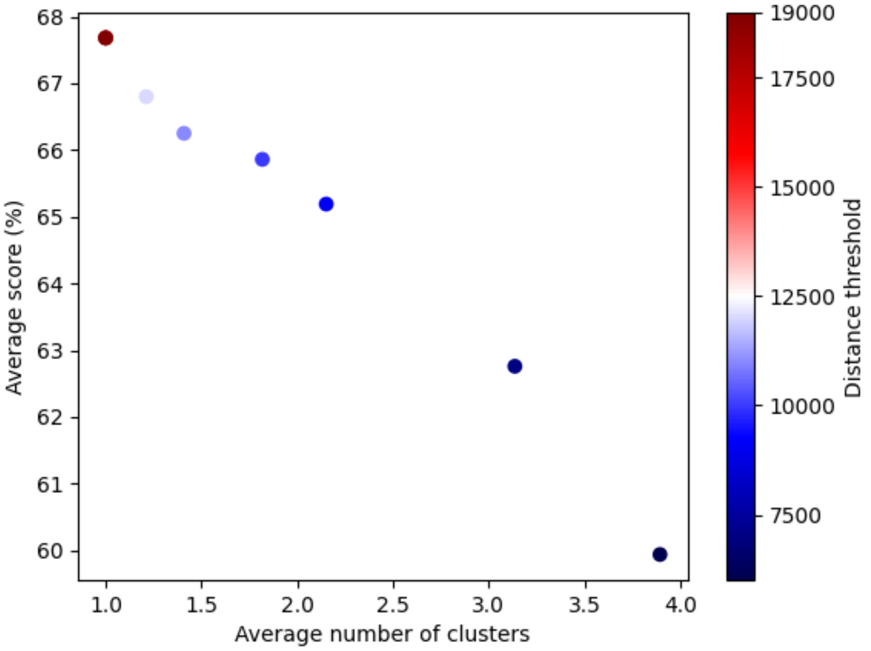

Figure 2: Relation between the score, the number of clusters and the distance threshold when clustering each verb individually with the agglomerative clustering algorithm. The score is expressed in BCubed F_1 and calculated with the FrenchSemEval dataset as a gold standard.

5.2 WiC

After our tests on the FrenchSemEval dataset, we ended up with some uncertainty on the best parameters, knowing we had to strike a balance between the mean number of clusters and BCubed F_1. For this reason, we decided to test our clustering on another evaluation dataset: the Word-in-Context dataset (Martelli et al., 2021). We proceeded the same way as before: for each verb of the dataset, we first extracted the contextualized embeddings of the test sentences with CamemBERT, then merged them with our own CamemBERT embeddings. We then clustered all the embeddings of each verb with the Agglomerative Clustering algorithm, comparing the three possible parameter values: 6000, 6500, and

7000. The results are in table 2. One can observe that the accuracy score is the highest with a distance threshold value of 6000 on this dataset.

Distance threshold	**WiC (accuracy)**	**FrenchSemEval (BCubed F$_1$)**
6000	**61.83 %**	59.93 %
6500	59.92 %	61.30 %
7000	59.54 %	**62.75 %**

Table 2: Performance on the Word-in-Context and FrenchSemEval evaluation datasets according to the distance threshold value of the agglomerative clustering algorithm parameter.

6. Conclusion

In this paper, we have explored how the clustering of contextual embeddings could help discover the senses of French verbs in context. The best results were achieved with a combination of CamemBERT embeddings and the agglomerative clustering algorithm. We noticed that when we augmented the main parameter of the agglomerative clustering algorithm, the distance threshold, the mean number of senses per verb went down while the BCubed F$_1$ score went up when we evaluated ourselves against the FrenchSemEval dataset. Comparing these results with those obtained for the WiC dataset did not really help us to make a wise decision concerning the distance threshold to use on our data, since the tendency was the opposite in this case (in the WiC dataset, the accuracy went down while the distance threshold went up). Since the FrenchSemEval dataset is bigger and has more similarities with the task we want to achieve, we decided to select the distance threshold of 7000, which gives satisfactory results on the FrenchSemEval dataset (62.75 %) while yielding an adequate number of senses per verb. Now that we have identified the best combination of language model, clustering algorithm and parameter for our task, the clustering for the 2000 most frequent verbs can be done.

7. References

Bagga, A. & Baldwin, B. (1998). Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, volume 1*. USA: ACL, pp. 79–85. URL https://doi.org/10.3115/980845.980859.

Baker, C.F. (2014). FrameNet: A Knowledge Base for Natural Language Processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*. Baltimore, MD, USA: ACL, pp. 1–5. URL https://aclanthology.org/W14-3001.

Bevilacqua, M., Pasini, T., Raganato, A. & Navigli, R. (2021). Recent Trends in Word Sense Disambiguation: A Survey. In Z.H. Zhou (ed.) *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. IJCAI Organization, pp. 4330–4338. URL https://doi.org/10.24963/ijcai.2021/593.

Campello, R.J., Moulavi, D. & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining:*

17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17. Springer, pp. 160–172.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL, pp. 8440–8451. URL https://aclanthology.org/2020.acl-main.747.

Danlos, L., Pradet, Q., Barque, L., Nakamura, T. & Constant, M. (2016). Un Verbənet du français. *TAL*, 57(1), pp. 33–58. URL https://hal.inria.fr/hal-01392817.

Delli Bovi, C., Camacho-Collados, J., Raganato, A. & Navigli, R. (2017). EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, volume 2 (short papers)*. Vancouver, Canada: ACL, pp. 594–600. URL https://aclanthology.org/P17-2094.

Dubois, J. & Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*, 179-180(3–4), pp. 31–56.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press. URL https://doi.org/10.7551/mitpress/7287.001.0001.

Frey, B.J. & Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814), pp. 972–976.

Gupta, R., Mundra, J., Mahajan, D. & Modi, A. (2021). MCL@IITK at SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation using Augmented Data, Signals, and Transformers. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. ACL, pp. 511–520. URL https://aclanthology.org/2021.semeval-1.62.

Hadouche, F. & Lapalme, G. (2010). Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages*, 179-180(3–4), pp. 193–220.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), pp. 1735–1780.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: ELRA, pp. 1027–1032. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/468_pdf.pdf.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, pp. 79–86. URL https://aclanthology.org/2005.mtsummit-papers.11.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. & Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC'20)*. Marseille, France: ELRA, pp. 2479–2490. URL https://aclanthology.org/2020.lrec-1.302.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, volume 2*. USA: ACL, pp. 768–774. URL https://doi.org/10.3115/980691.980696.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Martelli, F., Kalach, N., Tola, G. & Navigli, R. (2021). SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. ACL, pp. 24–36. URL https://aclanthology.org/2021.semeval-1.3.

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL, pp. 7203–7219. URL https://aclanthology.org/2020.acl-main.645.

McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *JOSS*, 2(11), p. 205.

Miller, G.A., Leacock, C., Tengi, R. & Bunker, R.T. (1993). A Semantic Concordance. In *Human Language Technology: Proceedings*. pp. 303–308.

Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: ACL, pp. 105–112. URL https://aclanthology.org/P06-1014.

Navigli, R. & Ponzetto, S.P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250. URL https://www.sciencedirect.com/science/article/pii/S0004370212000793.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P.J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. 1910.10683.

Raganato, A., Camacho-Collados, J. & Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 1 (long papers)*. Valencia, Spain: ACL, pp. 99–110. URL https://aclanthology.org/E17-1010.

Sagot, B. (2010). The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: ELRA, pp. 2744–2751. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/701_Paper.pdf.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1), pp. 97–123.

Segonne, V., Candito, M. & Crabbé, B. (2019). Using Wiktionary as a resource for WSD: the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics — Long Papers*. Gothenburg, Sweden: ACL, pp. 259–270. URL https://aclanthology.org/W19-0422.

Szekely, G. & Rizzo, M. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification*, 22, pp. 151–183.

van den Eynde, K., Mertens, P. & Eggermont, C. (2017). Dicovalence. URL https://hdl.handle.net/11403/dicovalence/v1. ORTOLANG.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc., pp. 6000–6010.

Véronis, J. (2004). HyperLex: Lexical cartography for information retrieval. *Computer Speech & Language*, 18, pp. 223–252.

Ziemski, M., Junczys-Dowmunt, M. & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: ELRA, pp. 3530–3534.