Word sense induction for (French) verb valency discovery

Naïma Hassert François Lareau

OLST
Université de Montréal

June 26, 2023

Overview

- We created a method to induce the senses of French verbs without relying on an external resource.
- We clustered contextualized embeddings from three language models with three different clustering algorithms.
- Best results: monolingual language model, CamemBERT, combined with the agglomerative clustering algorithm.
- Score: 58,19 % F_1 on 500 verbs of the *Wiktionnaire*.

Valency dictionaries

Useful in many natural language processing applications.

- Indicate precisely how a predicate expresses its arguments in syntax.
- Include information on selected part-of-speech, preposition or case.

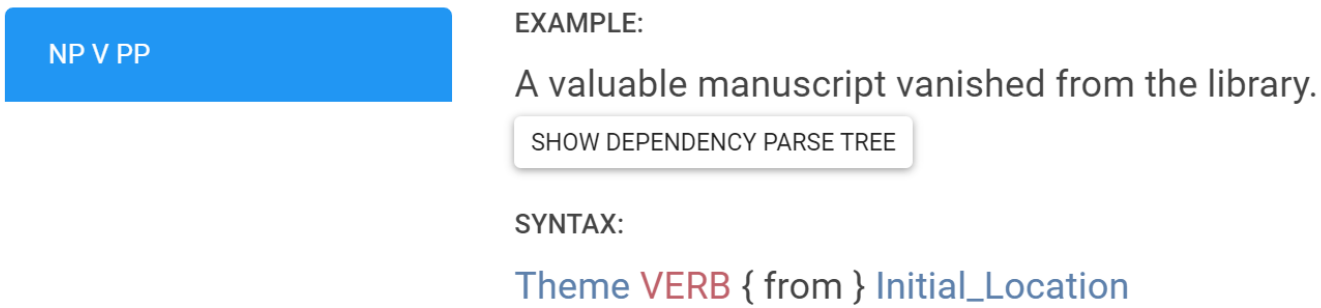


Figure: Valency of the verb *vanish*, from VerbNet

Existing valency dictionaries

Many quality valency dictionaries are already available in a machine-readable format for French :

- Le*fff*
- Dicovalence
- Verbənet (Verbnet's counterpart)
- *Dictionnaire électronique des mots*
- *Les verbes français*

The problem

Those dictionaries are all created at least in part manually.

- Very costly
- Requires highly trained staff
- Harder to update

The Goal: automate the construction of valency dictionaries

One important subtask: automatically identify the polysemy of verbs. Example with the verb CHANGE :

'modify'

The discussion *has* *changed* ***my thinking*** *about the issue.* → S V O

'become different'

She *changed completely as she grew older.* → S V

Word sense disambiguation (WSD)

- Knowledge-based methods: use the content of existing resources to compare with the data on hand and deduce the word sense;
- Supervised methods: rely on sense-annotated data, which comes from an existing bank of sense;
- Popular existing resources are for example: WordNet, BabelNet and FrameNet.

The problem with WSD for dictionary creation

1. The senses listed in major lexical resources are often too fine-grained;
2. Most resources are based on English;
3. Relying on external resources prevents the discovery of new senses.

Word sense induction (WSI)

Context clustering (Schütze, 1998)

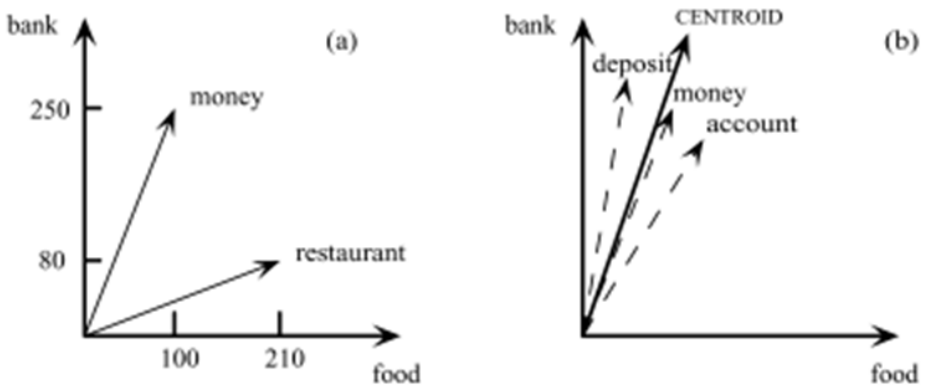


Fig. 15. (a) An example of two word vectors *restaurant* = (210, 80) and *money* = (100, 250). (b) A context vector for *stock*, calculated as the centroid (or the sum) of the vectors of words occurring in the same context.

Figure: From Navigli (2009)

Word Sense Induction (WSI)

Transformers (Vaswani et al., 2017)

We tested three language models:

- One monolingual model:
  - CamemBERT
- Two multilingual models:
  - XLM-RoBERTa
  - T5

Clustering

We tested three clustering algorithms :

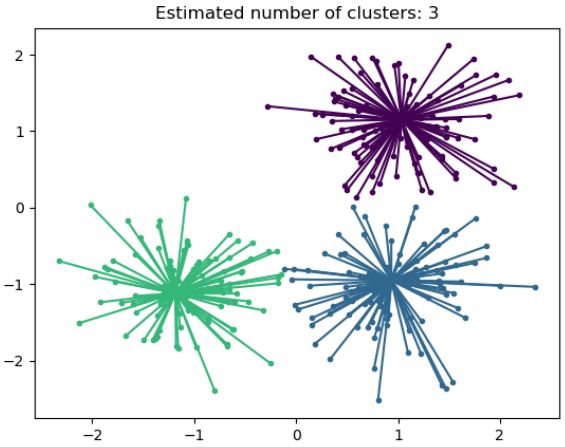


Figure: Affinity Propagation

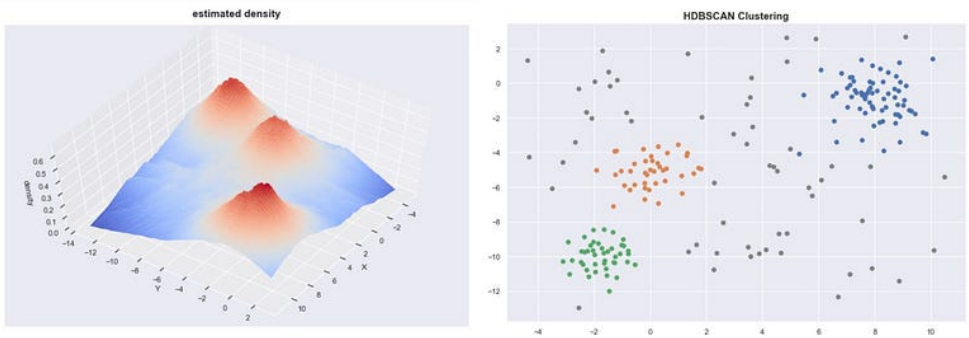


Figure: HDBSCAN

Université de Montréal

Agglomerative Clustering

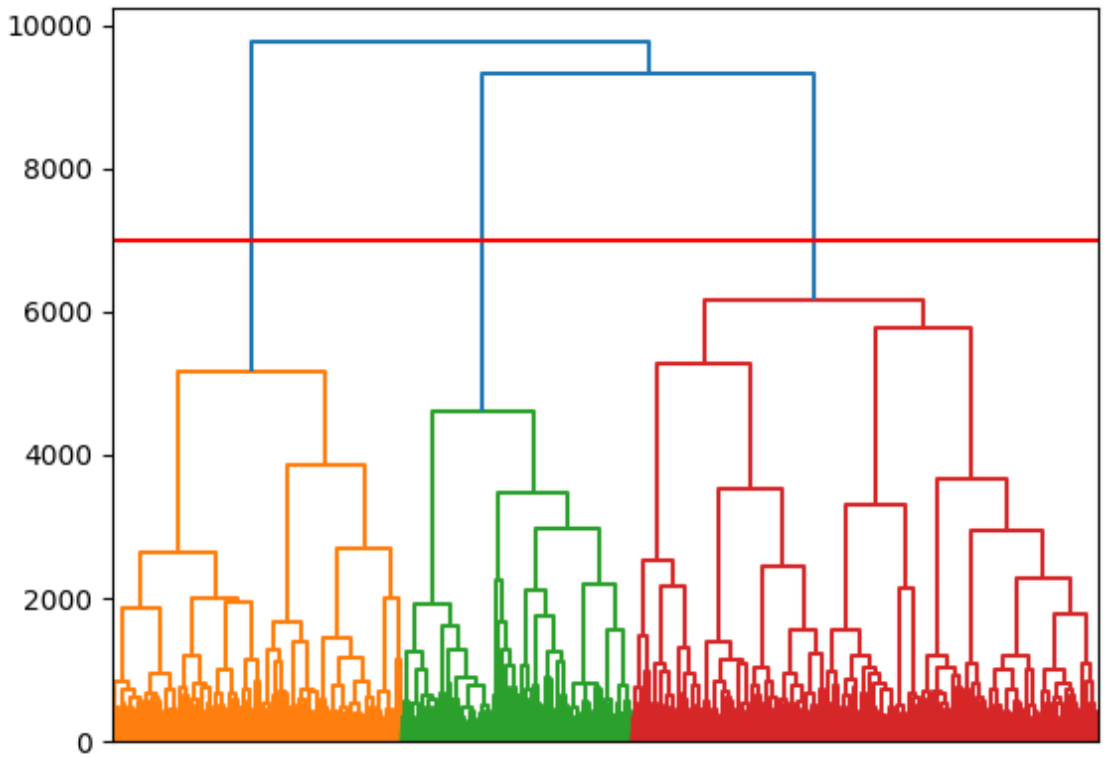


Figure: Agglomerative Clustering

Parameters estimation

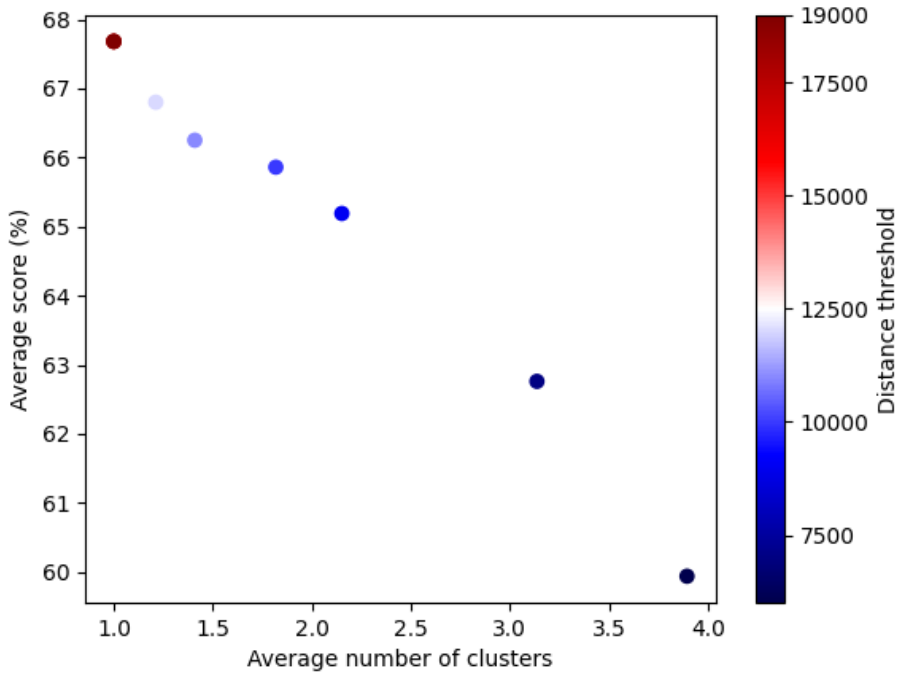
- FrenchSemEval
  - Specifically for French verbs
  - Senses from the Wiktionary
  - 3121 sense-annotated sentences
- Multilingual and Cross-lingual Word-in-Context Disambiguation (WiC)
  - First SemEval task to test the ability of systems to distinguish the different word senses, without an external resource.
  - In the multilingual subtask, the system has to decide if two words in two different contexts in the same language have the same meaning or not.
  - Not only verbs; adjectives, adverbs and nouns too.

FrenchSemEval: clustering the evaluation dataset

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: Best F_1 scores on the FrenchSemEval dataset

FrenchSemEval: clustering each verb individually



WiC vs FrenchSemEval

Distance threshold	WiC (accuracy)	FrenchSemEval (F_1)
6000	**61.83 %**	59.93 %
6500	59.92 %	61.30 %
7000	59.54 %	**62.75 %**

Results

- We tested our parameters on a subset of the *Wiktionnaire* (500 verbs / 16935 verbs total).
- 58,19 % F_1 B^3:
  - 54 % precision
  - 77 % recall
  - 2.23 sense per verb (vs 4.73 in the gold data)

Conclusion

- We propose a method for inducing the senses of verbs that is language-independent
- The parameters chosen can easily be adjusted
- Our results are comparable to the state-of-the-art
- Monolingual language models seem to be more suited than multilingual language models for this task