| Introduction | Word sense induction | Метнор | Evaluation | Results | Conclusion |
|--------------|----------------------|--------|------------|---------|------------|
| 0000 | | 000 | 0000 | 0 | 0 |
| | | | | | |

WORD SENSE INDUCTION FOR (FRENCH) VERB VALENCY DISCOVERY

Naïma Hassert François Lareau

OLST Université de Montréal

June 26, 2023

Université de Montréal

<ロト < 団 > < 豆 > < 豆 > < 豆 > < 豆 > < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < □ > < < □ > < □ > < □ > < □ > < < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

| Introduction | Word sense induction | Метнор | Evaluation | Results | Conclusion |
|--------------|----------------------|--------|------------|---------|------------|
| 0000 | | 000 | 0000 | 0 | 0 |
| | | | | | |

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.

▲ロ ▶ ▲ □ ▶ ▲ □ ▶ ▲ □ ▶ ▲ □ ▶ ● ○ ○ ○

- Best results: monolingual language model, CamemBERT, combined with the agglomerative clustering algorithm.
- Score: 58,19% F₁ on 500 verbs of the *Wiktionnaire*.

INTRODUCTION

Word sense induction

OD EVAL 0000

Evaluation 0000 Results 9 Conclusion 0

Université 🎢 de Montréa

▲ロト ▲周ト ▲ヨト ▲ヨト - ヨ - の々ぐ

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.



 $\mathrm{Figure:}\xspace$ Valency of the verb vanish, from VerbNet

Introduction 0●00 Word sense induction

Method 300 Evaluation 0000 Results 0 Conclusion 0

EXISTING VALENCY DICTIONARIES

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

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

Université m de Montréal ・ロト 4 同ト 4 ヨト 4 ヨト ヨークへへ

| Introduction 0000 | Word sense induction 0000 | Method 000 | Evaluation 0000 | Results 0 | $\mathop{\mathrm{Conclusion}}_{O}$ |
|----------------------|---------------------------|---------------|--------------------|--------------|------------------------------------|
| | | | | | |

THE PROBLEM

Those dictionaries are all created at least in part manually.

Univer

・ロト ・ 画 ・ ・ 画 ・ ・ 画 ・ 今々ぐ

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

INTRODUCTION 000 Word sense induction

lethod 30 EVALUATION 0000

Results 0 Conclusion

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. \rightarrow S V O

'BECOME DIFFERENT'

She changed completely as she grew older. \rightarrow S V

> Université m de Montréal

> > Э

nac

| INTRO | |
|-------|--|
| | |

Word sense induction

Method 000 EVALUATION 0000

Results D

▲ロト ▲周ト ▲ヨト ▲ヨト - ヨ - の々ぐ

Conclusion

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.

 Introduction
 Word sense induction
 Method oco
 Evaluation oco
 Results occurse
 Conclusion

 THE PROBLEM WITH WSD FOR DICTIONARY
 CREATION
 CREATION
 CREATION
 CREATION
 CREATION

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

de Montréa

= 900

イロト イポト イヨト

Introduction 0000 Word sense induction

íethod Do Evaluation

Results 0

イロト イポト イヨト イヨト

Conclusion 0

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)

Université de Montréal

Sac

3

Introduction 0000 Word sense induction 0000

Method 200 Evaluation 0000 Results 0 CONCLUSION O

Université m de Montréal

E

990

WORD SENSE INDUCTION (WSI)

Transformers (Vaswani et al., 2017)



| Introduction | Word sense induction | Method | Evaluation | Results | Conclusion |
|--------------|----------------------|--------|------------|---------|------------|
| 0000 | | ●00 | 0000 | 0 | 0 |
| | | | | | |

We tested three language models:

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



< □ > < □ > < □ > < ⊇ > < ⊇ >

Université de Montréal

990

Э

| Introduction | |
|--------------|--|
| 0000 | |

Word sense induction

Method 0●0

Evaluation 0000 Result 0 Conclusion

CLUSTERING

We tested three clustering algorithms :



FIGURE: Affinity Propagation



FIGURE: HDBSCAN



< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

INTRODUCTION WORD SENSE INDUCTION METHOD EVALUATION RESULTS CONCLUSION 0000 0000 000 000 000 0

Agglomerative Clustering



FIGURE: Agglomerative Clustering

Université de Montréal

3

500

(日) (四) (三) (三) (三)

Introduction Word s 0000 0000

Vord sense induction

Метнор 000 Evaluation

Results 0

▲□▶ ▲□▶ ▲三▶ ▲三▶ 三三 - のへ⊙

Conclusion 0

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.

INTRODUCTION WORD SENSE INDUCTION METHOD EVALUATION RESULTS CONCLUSIC 0000 FRENCHSEMEVAL: CLUSTERING THE EVALUATION DATASET

| Clustering algorithm | Т5 | 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

Université m de Montréal

€ 990

イロト イポト イヨト

INTRODUCTION WORD SENSE INDUCTION METHOD EVALUATION RESULTS CONCLUSION 0000 FRENCHSEMEVAL: CLUSTERING EACH VERB INDIVIDUALLY



Université de Montréal

ł

500

イロト イポト イヨト イヨト

Introduction

Word sense induction

Method 000 Evaluation

Results 0 Conclusion 0

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 % |



990

E

< □ > < □ > < □ > < ⊇ > < ⊇ >

| Introduction 0000 | Word sense induction | Метнор 000 | Evaluation 0000 | Results | Conclusion 0 |
|----------------------|----------------------|---------------|--------------------|---------|-----------------|
| D | | | | | |

Results

• We tested our parameters on a subset of the *Wiktionnaire* (500 verbs / 16935 verbs total).

Univer

<ロト < 団 > < 豆 > < 豆 > < 豆 > < 豆 > < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < < □ > < □ > < < □ > < □ > < □ > < □ > < < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

- 58,19 % F₁ B³:
 - 54 % precision
 - 77 % recall
 - 2.23 sense per verb (vs 4.73 in the gold data)

| Introduction 0000 | Word sense induction 0000 | Method 000 | Evaluation 0000 | Results 0 | Conclusion |
|----------------------|---------------------------|---------------|--------------------|--------------|------------|
| | | | | | |

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

▲ロト ▲周ト ▲ヨト ▲ヨト - ヨ - の々ぐ