MWEs discovery, using semantic slusters, association measures, compositionality, and lexicons

Manon Scholivet

June 19, 2024

Plan

Motivation

- Too few annotated multiword expressions (MWEs)
- Difficult detection of unannotated MWEs
- How to maximize the diversity of MWE predictions nonetheless ?
- Question : Can we integrate unsupervised methods for MWE detection ?

Plan

Candidate Extraction

- The methods are intended to be generic and can be applied to both verbal and non-verbal expressions.
- Form of a candidate :

  $\longrightarrow$ Lemma1_Lemma2_... Sorted alphabetically. The same lemma can appear multiple times

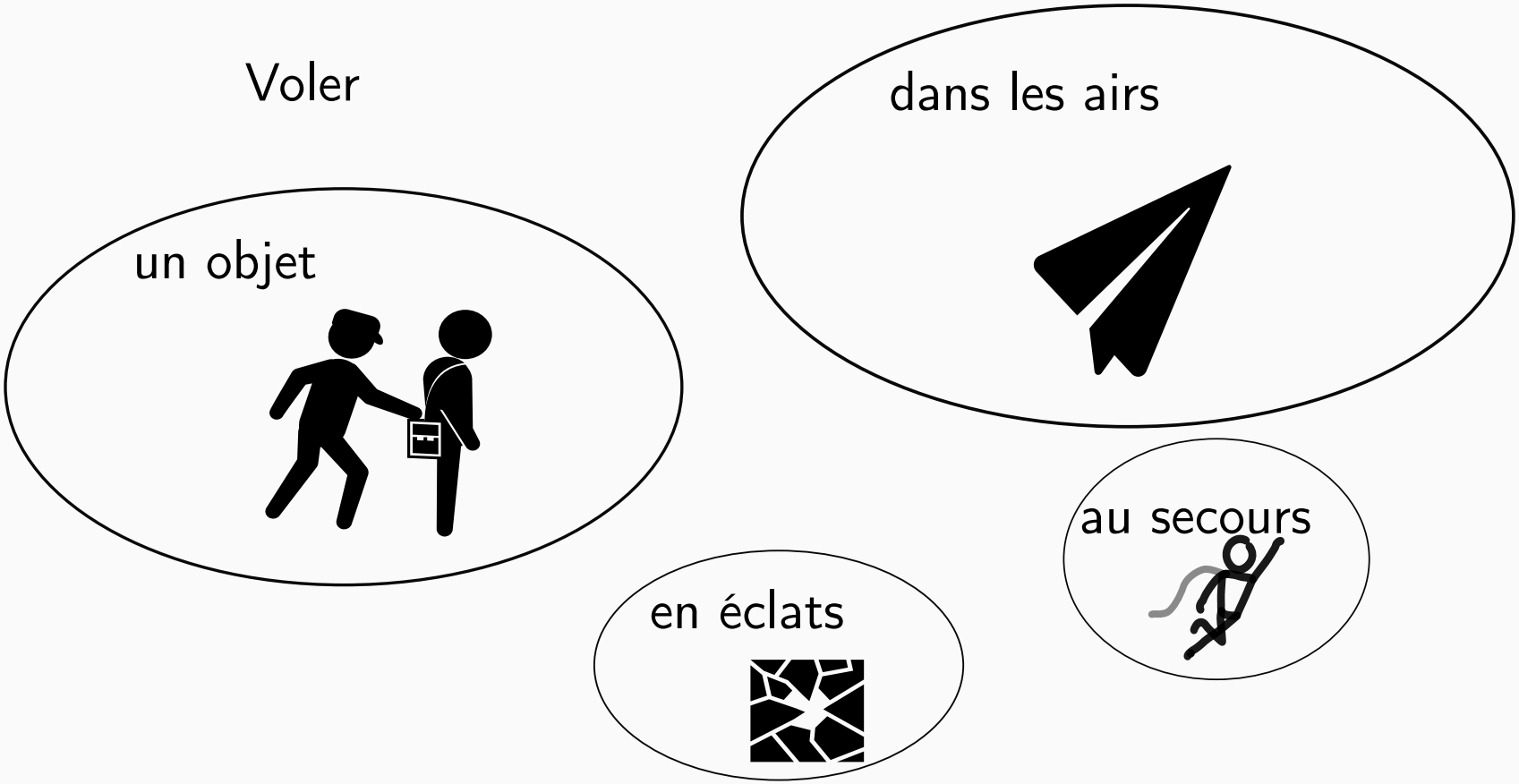
  - à_le_secours_voler 
  - appeler_chat_chat_un_un 

Candidate Extraction

- Methods :
  - Clusters
  - Association Measures
  - Compositionality
  - Lexicons
- Majority Voting ?

Candidate extraction : sense clusters

Intuition :



Candidate Extraction : Method 1, Sense Clusters

- ✅ Identification of a list of target verbs

  → 866 verbs from the PARSEME corpus

Candidate Extraction : Method 1, Sense Clusters

- Identification of a list of target verbs
  - → 866 verbs from the PARSEME corpus
- Retrieval of 10,000 diverse examples for each verb
  - → Maximizing entropy based on wordforms in sentences

Candidate Extraction : Method 1, Sense Clusters

- Identification of a list of target verbs
  - → 866 verbs from the PARSEME corpus
- Retrieval of 10,000 diverse examples for each verb
  - → Maximizing entropy based on wordforms in sentences
- Sense clustering for each verb
  - → x-means algorithm, automatically selects the number of clusters (max 15)

Candidate Extraction : Method 1, Sense Clusters

- Identification of a list of target verbs
  - → 866 verbs from the PARSEME corpus
- Retrieval of 10,000 diverse examples for each verb
  - → Maximizing entropy based on wordforms in sentences
- Sense clustering for each verb
  - → x-means algorithm, automatically selects the number of clusters (max 15)
- Extraction of lemma n-grams in each cluster
  - → Punctuation, stop words, and words more than 5 positions away from the target verb are removed
  - → Bigrams : target verb + most frequent word in the cluster
  - → Trigrams and Quadrigrams : target verb + most frequent n-grams

Candidate Extraction : Method 2, Association Measures

For each target verb :

- ✅ Retrieval of all sentences containing the target lemma
- ⋯ PMI computation between the lemma and other words
- ⋯ Candidate : n-grams with a score above a specified threshold

Candidate Extraction : Method 3, Compositionality

Similar method to association measures

Candidate Extraction : Method 4, Lexicons

- Wiktionary (all entries with a space)
- LEFFF (Note : Downloading appears to be currently impossible)

Extraction of Reliable Candidates

A lot of noise in candidate extraction :

- A candidate will be considered "reliable" if it :
  - ✅ Comes from a lexicon
  - ❓ Appears in more than one/two/three methods...
    - → Does appearing in more than one cluster count as multiple methods ?
  - ❓ Achieves a compositional score > threshold
  - ❓ Achieves an association measure score > threshold
  - ❓ Other
- ◔ Majority voting

Plan

Intrinsic Evaluation of Clusters

Using MWE lexicons :

For each MWE in the lexicon including the target lemma :

- Has this MWE been found in at least one of the clusters ?
- Does it appear significantly more often in any of the clusters ?

Allows for evaluating recall and cluster quality.

Extrinsic Evaluation of Clusters (and Candidates)

Using candidates in a task to identify MWEs never seen during training (from PARSEME).

Two methods :

- Data augmentation
- Adding one or more "potential candidate" columns in the data

Extrinsic Evaluation : Data Augmentation

- ✅ Retrieval of all sentences where all the lemmas of a candidate appear
- 💬 Tagging these candidates as gold MWEs
- 💬 Training of a new identification system with more data

Extrinsic Evaluation : "Candidate" Column

- On the gold data, we add a "candidate" column, annotated similarly to the "MWE" column (without the type).
- Reliable candidates are added to this column.
- Addition of a candidate column per extraction method.
- Train the identification system with this additional column.

  $\longrightarrow$ We hope that candidates will capture true annotated MWEs. The system should generally recognize this flag when it encounters an MWE.

Exemple

Train		Le	poulpe	lui	vole	la	vedette
	MWE	-	-	-	1	1	1
	Candidate Cluster	-	-	-	1	1	1
	Candidate Lexicon	-	-	-	1	1	1
	...						
Train		Elle	vole	au	secours	du	poulpe
	MWE	-	1	1	1	-	-
	Candidate Cluster	-	1	-	1	-	-
	Candidate Lexicon	-	1	1	1	-	-
	...						
Train		L'	oiseau	vole	gracieusement		
	MWE	-	-	-	-		
	Candidate Cluster	-	1	1	-		
	Candidate Lexicon	-	-	-	-		
	...						
Test		La	fenêtre	vole	en	éclats	
	MWE	?	?	?	?	?	
	Candidate Cluster	-	-	1	-	1	
	Candidate Lexicon	-	-	1	1	1	
	...						

Plan

Database

Creating a database using sqlite3

- 22 GB
- 54 million sentences
- 1,440,000,000 tokens
- Construction time : 33 hours

Ability to quickly retrieve sentences containing a specific lemma

However...

$\Longrightarrow$ Issues with the initial quality of annotations.

- Wikisource not usable
  - $\rightarrow$ Sentences starting in the middle of a sentence
  - $\rightarrow$ Issues with characters and presence of HTML tags : "*{{nr|ÆNEAS SYLVIUS.|545}}ne me soient retirées.*"
  - $\rightarrow$ Sentences in old french
  - $\rightarrow$ ...
- UDPipe sometimes predicts "PUNCT _" instead of "PONCT PONCT" for rare punctuation marks (%, {, }, etc.)
- Pre-processing issues : segmentation whenever a '.' is encountered
- ...

Evaluation of the annotation quality

30 sentences evaluated by 7 annotators (totaling 210 sentences) for 5,460 tokens

	Accuracy
Correct Lemmas	97.64
Correct POS	95.31
Correct Features	92.05
Problem-free Sentences	84.30

Automatic re-annotation of the corpus

- ✅ Identification of recoverable parts of the Tithir code
  - → UDPipe 1, corpus split
- ✅ Selection of Syntax Annotation
  - → Yes if time permits
- ✅ Choice of tagset
  - → FTB-dep (+ UD if syntax)
- 🕒 Addition of corpus to enhance diversity
- ✅ Addition of features (early stopping, ...)
- 🔴 Full reannotation