Minimally Supervised Prediction of Coarse Semantic Distinctions

C. Aloui*, L. Barque°, A. Nasr*, C. Ramisch*

* LIS, Université Aix Marseille
° LLF, Université Paris 13

Lundi 8 Juillet 2019
JE Demonext "Sémantique pour les ressources en morphologie dérivationnelle"

Plan

Introduction

Minimally supervised method to predict coarse-semantic distinctions

- Using seed lists and unannotated corpora

Aims

- Cues for (more fine-grained) semantic classes
- Help for semantic processing (WSD, SRL) and NLP tasks involving semantic treatments (MT, IE)

Justification

- French, like many other languages, lacks semantically labelled corpus data

Introduction

- We focus on two coarse distinctions in French:
  - COUNTABILITY : Count Ns (*two maps, several crimes*) **vs.** Mass Ns (*unemployment, some water*)
  - ANIMACY : Animate Ns (*daughter, committee, troll*) **vs.** Inanimate Ns (*tree, weapon, lie*)
- Within both distinctions, nominal forms can pertain to both categories
  - *produce paper*$_{Mass}$ **vs.** *submit two papers*$_{Count}$
  - *a crane*$_{Anim}$ *urgent warning* **vs.** *a crane*$_{Inanim}$ *operator*
- Similar distributions (majority class: ∼78%)
  - Difference : countability is a semantic and a syntactic phenomenon

Introduction

Related work

- Minimally supervised classification
- Supersense tagging
- Animacy and countability detection
  - Lexical acquisition
  - Supervised vs. unsupervised methods
  - Countability detection

	Count	Uncount	Avg
Lapata and Keller 2005	88.62	91.53	90.07
Baldwin and Bond 2003	93.90	95.25	94.57

Introduction

- Representing semantic properties of lexical items as numerical scores denoting coarse distinctions
- Minimally supervised method to predict these scores using seed lists and unannotated corpora
- Evaluation and study of some parameters of our method on (new) datasets annotated for noun animacy and countability in French.

Plan

Method

Our method is composed of the following steps :

1. Build disjoint lists L_0 and L_1 of **seed words** prototypical of each semantic class 0 and 1
2. Locate in a raw corpus C all occurrences of elements of $L_0 \cup L_1$ and annotate them with their class, yielding a **training set** C_A
3. Train a classifier P on C_A that takes as input a context c and returns a **contextual score** $0 \leq s_{cont}(c) \leq 1$
4. Extract from C all contexts $c_1 \ldots c_n$ of a given target word w and predict scores $s_{cont}(c_i)$ with P. These predicted scores are then aggregated in a **lexical score** $0 \leq s_{lex}(w) \leq 1$
5. Devise a **strategy** for annotating the target word's occurrence (w, c), based on $s_{lex}(w)$ and on $s_{cont}(c)$ predicted by P.

Method: illustration from countability data

1. Seed words (0 for count, 1 for mass)

0 : directive, fusil, pic, modèle. . .	**1** : magie, calcium, timidité. . .

2. Training set

de plus amples **directives**$_0$ seront	comme par **magie**$_1$ et m'a
elle prévoit un **pic**$_0$ d'abandon	cette impression de **magie**$_1$ que
viande sur des **pics**$_0$ à brochette	un peu de leur **timidité**$_1$. Les
La **directive**$_0$ européeenne qui	Oui, le **calcium**$_1$ ascorbate peut
blancs, armés de **fusils**$_0$	vitamines, **calciums**$_1$ et sels

3. Learning contextual scores (model 2L0R|f|num)

plus amples **directives**$_{0\,plur}$	comme par **magie**$_{1\,sing}$
prévoit un **pic**$_{0\,sing}$	impression de **magie**$_{1\,sing}$
sur des **pics**$_{0\,plur}$	de leur **timidité**$_{1\,sing}$
La **directive**$_{0\,sing}$	Oui, le **calcium**$_{1\,sing}$
armés de **fusils**$_{0\,plur}$	vitamines, **calciums**$_{1\,plur}$

Method: illustration from countability data

4. Prediction of contextual scores for unseen nouns

Lui, il continue à te causer derrière la **fumée** de sa cigarette [0.67]
mais aussi de sérieux désagréments liés aux **fumées** ! [0.16]
t'avales pas la **fumée**, ça fait fondre la glace ! [0.74]
Des **fumées** s'élevaient près de la gare de triage de Maaskola. [0.15]
On peut citer par exemple le traitement des **fumées** [0.24]
Les premières **fumées** quittent les cheminées et montent dans [0.07]
l'intérêt majeur du système (reposer son pied) part en **fumée**. [0.81]

- $S_{lex}(\text{fumée}) = 0.32$

5. Strategy for annotating a target word's occurrence
   - Priority given to the (discriminant) context
   - t'avales pas la **fumée**$_{sing}$, ça fait fondre la glace !

     → occurrence of a mass noun

Method

The classifier P

- Multilayer perceptron (MLP)
- Context's word embeddings and simple grammar features

The lexical score $s_{lex}(w)$

- An occurrence is labeled 1 if its contextual score is > 0.5 and labeled 0 if ≤ 0.5
- We define w's lexical score as the ratio $\frac{n_1}{n_0+n_1}$
- Non informative contexts can be ignored by introducing a **lexical threshold** $0 \leq T_{lex} \leq 0.5$
  - Ex. if $T_{lex} = 0.35$
    - n1 : occurrences whose contextual score is ≥ 0.85
    - n0 : occurrences whose contextual score is ≤ 0.15
    - Contexts whose predicted scores fall within the range 0.16 and 0.84 are discarded

Method

Attributing a class to an occurrence of word w in context c:

- **Back-off strategy**: given an occurrence (w, c), the context c is examined first. If its score $s_{cont}(c)$ is sufficiently informative, then the occurrence is annotated with the class predicted for its context. Otherwise the lexical score $s_{lex}(w)$ is used
- A **contextual threshold** $0 \leq T_{cont} \leq 0.5$ is introduced in order to decide whether a context is informative or not
- If $s_{lex}(w)$ cannot be calculated for w, then the majority class is predicted as a fallback

Plan

Data: seed lists

Seeds are selected manually for their univocity (non ambiguous) from a list containing the most frequent nouns in the FrWaC corpus, according to linguistic tests

COUNTABILITY seed lists:

- 196 count Ns, 200 mass Ns
- Linguistic tests : 1) for count N, 2) for mass N, but not both
  1. *un/des/trois N* $\emptyset$
  2. *un peu de* N_{sing}, V_{trans} *du/de la N*

ANIMACY seeds lists:

- 201 animate Ns, 267 inanimate Ns
- Linguistic tests : 1) for anim N, 2) for inanim N, but not both
  1. *det N a décidé de, det N a volontairement V*
  2. *#det N a décidé de, #det N a volontairement V*

Data: training corpus

Corpus:

- FrWaC (Baroni et al. 2009)
- Segmented, tokenized, POS-tagged and lemmatized with TreeTagger (Schmid, 1994)

Lemmatized N from seed lists frequence:

- Average number of occurrences: 90,116
- 12 out of the 845 nouns occur less than 1000 times

Skewed distribution of the target phenomena

- Balanced sample of each class in the training set
- 7,876,629 sentences to learn countability and 21,219,489 sentences to learn animacy

Data: evaluation sets

COUNTABILITY evaluation set

- Manual annotation of 5000 occurrences (50 × 100 N) from the frWaC according to the following strategy:
  i) if the morphosyntactic context is discriminant for countability → contextual annotation
  ii) if the morphosyntactic context is neutral *wrt* the mass/count distinction → lexical annotation
    - Discarded: 226 undetermined occurrences (e.g. *épilepsie, cécité*) + 33 ill-formed sentences

- Occurrences

Count	Mass	Total
3,813	928	4,741

- Lemmas

Count	Mass	Both	Total
71	2	26	99

Data: evaluation sets

ANIMACY evaluation set

- Available evaluation set for animacy in French
  - Manual annotation of occurrences of nouns and pronouns from the Sequoia Corpus (L. Barque, M. Candito, V. Segonne)
  - 1,093 different noun lemmas in the set (493 occur only once)

- Occurrences

Inanimate	Animate	Total
2,613	767	3,380

- Lemmas

Inanimate	Animate	Both	Total
865	183	45	1,093

Plan

Experiments: Set up

Classifiers: simple MLP with two hidden layers containing respectively 300 and 150 neurons

Word Embeddings: 200-dimensional randomly initialized real vectors which are updated through backpropagation

- ReLU activation function
- No dropout
- Keras'categorical cross entropy loss function

Experiments: Results

Accuracy for countability and animacy on the test sets, with $T_C = T_L = 0.4$

	Countability	Animacy
Majority class baseline	80.43	77.31
Best	90.06	92.63
Model	4L0R-LF-num	4L4R-LF-num

Experiments: model features

Influence of the model parameters on the accuracy for COUNTABILITY with $T_C = T_L = 0.4$

	context	word repr.	morpho	accuracy
1	4L0R	LF	num	**90.06**
2	2L0R	LF	num	89.58
3	3L0R	LF	num	88.58
4	3L0R	LF	none	86.62
5	3L0R	F	num	86.50
6	3L0R	L	num	80.37
7	3L3R	LF	num	79.79

Experiments: model features

Influence of the model parameters on the accuracy for ANIMACY with $T_C = T_L = 0.4$

	context	word repr.	morpho	accuracy
1	4L4R	LF	num	**92.63**
2	3L3R	LF	num	92.18
3	4L4R	LF	none	92.07
4	4L4R	L	num	90.59
5	4L4R	F	num	90.32
6	2L2R	LF	num	89.14
7	3L0R	LF	num	88.66

Experiments: Seeds lists size and composition

Influence of the seed list size and composition on accuracy for COUNTABILITY with model `3L0R-LF-num`

	50	100	150	200
1	85.42	87.65	87.54	
2	83.23	86.20	87.12	
3	82.91	85.42	86.00	
Average	83.85	86.10	86.68	88.58

Plan

Conclusion

- Relatively inexpensive method for predicting coarse semantic categories
- Results of the intrinsic evaluation on French data are similar to the state of the art of minimally-supervised methods applied to other languages
  - 90.06% for countability and 92.63% for animacy
- Encouraging results on extrinsic evaluations (parsing and MWE detection)

Future Work

- Studying context's influence for ambiguous words
- Supersense tagging
  - Animacy: $\{$*Person*, *Animal*, *Institution*$\}$ vs others
  - Countability: $\{$*Substance*, *Food*, *Felling*$\}$ vs others
- Lexical semantics representation
  - Supersense embeddings (Flekova&Gurevych 2016)
  - Supersenses scores

	Person	Artifact	Cognition	Event	State	...
cuisinière	0.65	0.47	0.03	0.12	0.09	