# Classifying French Verbs Using French and English Lexical Resources

Ingrid Falk†, Claire Gardent‡ and Jean-Charles Lamirel§
†Université de Lorraine/LORIA, ‡CNRS/LORIA, § Université de Strasbourg/LORIA
Nancy, France
{ingrid.falk,claire.gardent,jean-charles.lamirel}@loria.fr

## Goal: build a Verbnet like classification for French

### associating groups of verbs with

- subcategorisation frames
- thematic role sets.

### based on French and English lexical resources

- Manually built
- French: subcategorisation lexicons
- English: Verbnet

### using IGNGF (Incremental Growing Neural Gaz with Feature maximisation) Lamirel et al. (2011)

Growing neural gas clustering method
- based on Hebbian learning
- incremental
- *winning* clusters determined through distance function

#### IGNGF

- uses feature maximisation to determine *winning* cluster
- supports cluster labeling with distinguishing features

## IGNGF: feature maximisation

- Cluster relevant features: cluster maximising features

Feature $f$ maximising cluster $c$

Feature F-measure $FF_c(f)$ higher for $c$ than other cluster

Feature recall $FR_c(f) = \frac{\text{verbs in } c \text{ having } f}{\text{all verbs having } f}$

$$FR_c(f) = \frac{\sum_{v \in c} W_v^f}{\sum_{c' \in C} \sum_{v \in c'} W_v^f}$$

Feature precision $FP_c(f) = \frac{(f,\ \text{verb}) \text{ combinations in } c}{\text{all (feature, verb) combinations in } c}$

$$FP_c(f) = \frac{\sum_{v \in c} W_v^f}{\sum_{f' \in F_c, v \in c} W_v^{f'}}$$

## Lexical Resources: French and English, manually developed → extract features for IGNGF clustering

### French: Dicovalence, Ladl tables and TreeLex → features:

1. subcategorisation frames (**scf**)
2. additional syntactic (**synt**) and semantic (**sem**) features

| | Feature | Description | VN class | Feature | Description | VN class |
|---|---|---|---|---|---|---|
| Mostly syntactic (**synt**): | **ArgNbr** | 4 or more arguments | *get, send, ...* | **Event** | clausal argument | *correspond, characterize, say, ...* |
| Mostly semantic (**sem**): | **Loc** | location role | *put, remove, ...* | **Nhum** | concrete object, non human role | *hit, other_cos, ...* |

### English: Verbnet Kipper Schuler (2006) → features:

- translate English Verbnet classes to French using dictionaries
- use SVM to decide: does verb in ⟨verb, VN class⟩ pair have roles of VN class?

3. verb has **grid** feature: verb ∈ translated VN class with thematic role set given by grid.

## Evaluation: Reference and Metrics

### Reference Sun et al. (2010)

- 116 verbs in 16 Levin classes
- Levin classes identified with Verbnet thematic role sets
- Thematic role sets are grouped

→ 116 verbs in 11 classes

Example:

Thematic roles: AgExp, Instrument, Patient

| | |
|---|---|
| hit-18.1: | cogner, heurter, battre, frapper, fouetter, taper, rosser, brutaliser, éreinter, maltraiter, corriger |
| other_cos-45.4: | mélanger, fusionner, consolider, renforcer, fortifier, adoucir, polir, atténuer, tempérer, pétrir, façonner, former |

### Metrics

Supervised: mPUR & ACC

- How well can the clustering be embedded into gold? Cluster C → prev($C$) ∈ gold classification with maximal $|\text{prev}(C) \cap C|$

$$\text{mPUR} = \frac{\sum_{C \in \text{Clustering}, |\text{prev}(C)|>1} |\text{prev}(C) \cap C|}{\sum_{C \in \text{Gold}} \text{Verbs}_{\text{Clustering} \cap C}}$$

- How well can the gold be embedded into the clustering? gold class C → dom($C$) ∈ clustering with maximal $|\text{dom}(C) \cap C|$

$$\text{ACC} = \frac{\sum_{C \in \text{Gold}} |\text{dom}(C) \cap C|}{\sum_{C \in \text{Gold}} \text{Verbs}_C}$$

Unsupervised: CMP

- Cumulative Micro Precision, Lamirel et al. (2011), unsupervised clustering metric
- Evaluates cluster quality wrt. cluster features

$$\text{CMP} = \frac{\sum_{i=|C_{\text{inf}}|..|C_{\text{sup}}|} \frac{1}{|C_{i+}|^2} \sum_{c \in C_{i+}, f \in F_c} P_c^f}{\sum_{i=|C_{\text{inf}}|..|C_{\text{sup}}|} \frac{1}{|C_{i+}|}}$$

$C_{\text{inf}}$ = smallest cluster size
$C_{\text{sup}}$ = largest cluster size
$C_{i+}$ = clusters with more verbs than $i$
$P_c^f = \frac{|\text{verbs } \in c \text{ having } f|}{|\text{verbs } \in c|}$

## Results

Sample cluster:

```
C6- 14(14) [197(197)]
Prevalent Label — = AgExp-Cause

0.341100  G-AgExp-Cause
0.274864  C-SUJ:Ssub,OBJ:NP
0.061313  C-SUJ:Ssub
0.042544  C-SUJ:NP,DEOBJ:Ssub
**********
**********
0.017787  C-SUJ:NP,DEOBJ:VPinf
0.008108  C-SUJ:VPinf,AOBJ:PP
...
[**déprimer 0.934345 4(0)] [affliger 0.879122 3(0)] [éblouir 0.879122 3(0)]
[choquer 0.879122 3(0)] [décevoir 0.879122 3(0)] ...
```

Impact of feature set

| Feat. set | mPUR | ACC | F (Gold) | #class | CMP at opt (13cl.) |
|---|---|---|---|---|---|
| scf | 0.93 | 0.48 | 0.64 | 17 | 0.28 (0.27) |
| grid, scf | 0.94 | 0.54 | 0.68 | 14 | 0.12 (0.12) |
| grid, scf, sem | 0.86 | 0.59 | **0.70** | **13** | **0.30 (0.30)** |
| grid, scf, synt | 0.87 | 0.50 | 0.63 | 14 | 0.13 (0.14) |
| grid, scf, synt, sem | 0.99 | 0.52 | **0.69** | **16** | **0.50 (0.22)** |
| scf, sem | 0.83 | 0.55 | 0.66 | 23 | 0.40 (0.26) |
| scf, synt | 0.91 | 0.45 | 0.61 | 15 | 0.17 (0.22) |
| scf, synt, sem | 0.89 | 0.47 | 0.61 | 16 | 0.57 (0.11) |
| Sun et al. (2010) | 0.55-0.65, depending on verb frequency | | | | |

- semantic features – **grid, sem** – help clustering
- IGNGF clustering outperforms results in Sun et al. (2010) but not entirely comparable

Comparison with K-means – **grid, scf, sem** feature set

| Method | mPUR | ACC | F (Gold) | #classes | CMP(13cl.) |
|---|---|---|---|---|---|
| IGNGF with IDF and norm. | 0.86 | 0.59 | **0.70** | 13 | **0.30 (0.30)** |
| K-means with IDF and norm. | 0.88 | 0.57 | **0.70** | 13 | **0.10 (0.10)** |
| IGNGF, no IDF | 0.86 | 0.59 | 0.70 | 17 | 0.18 (0.14) |
| IGNGF, no norm. | 0.78 | 0.62 | 0.70 | 18 | 0.15 (0.11) |
| IGNGF, no IDF, no norm. | 0.87 | 0.55 | 0.68 | 14 | 0.21 (0.21) |

- K-means & IGNGF have similar F-measure
- CMP: IGNGF clusters more coherent in terms of features.

## Conclusion

Approach to large scale verb classification for French
- Showed good results on established test set
- Cluster labeling allows association of verb classes with
  - subcategorisation frames
  - thematic role sets

## Future Work

- What is quality of associations of verbs with
  - subcategorisation frames
  - thematic role sets
- Is evaluation on very small test set conclusive?
- Produce overlapping clustering to represent polysemy
- Produce hierarchical clustering – more similar to Verbnet

## References

Kipper Schuler, K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania.

Lamirel, J.-C., Mall, R., Cuxac, P., and Safi, G. (2011). Variations to incremental growing neural gas algorithm based on label maximization. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 956 –965.

Sun, L., Korhonen, A., Poibeau, T., and Messiant, C. (2010). Investigating the cross-linguistic potential of verbnet: style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1056–1064, Stroudsburg, PA, USA. Association for Computational Linguistics.