Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue

Guillaume Wisniewski Nicolas Pécheux Elena Knyazeva
Alexandre Allauzen François Yvon

Université Paris Sud & LIMSI-CNRS

3 juillet 2014

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®

LIMSI

Contexte

- apprentissage statistique :

  corpus étiqueté $\rightarrow$ apprentissage $\rightarrow$ prédictions

- permet de « résoudre » toutes les tâches du TAL...
- ... au moins lorsque l'on dispose d'un grand nombre de données du domaine

Notre problème

Que faire lorsque l'on ne dispose pas de données étiquetées ?

Solution proposée

Que faire lorsque l'on ne dispose pas de données étiquetées ?

⇓

Importer (en les contrôlant) des données supervisées d'une autre langue

Plan

1. Contexte
2. Création d'un corpus d'apprentissage par transfert d'étiquettes
3. Apprentissage faiblement supervisé
4. Résultats expérimentaux
5. Conclusion

Première partie I

Création d'un corpus d'apprentissage par transfert d'étiquettes

Méthode d'étiquetage automatique d'un corpus

Fusion de deux sources d'informations

1. contraintes de type issues d'un dictionnaire ;
2. contraintes de token issues des liens d'alignements ;

Exemple

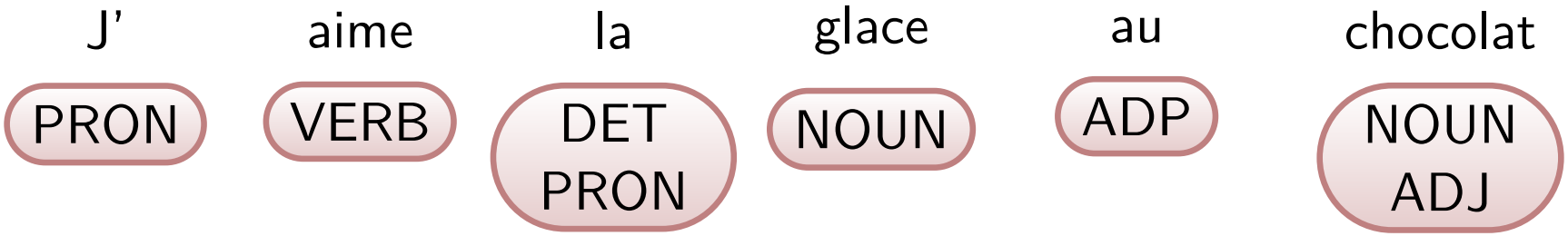
J' aime la glace au chocolat

Méthode d'étiquetage automatique d'un corpus

Fusion de deux sources d'informations

1. contraintes de type issues d'un dictionnaire ;
2. contraintes de token issues des liens d'alignements ;

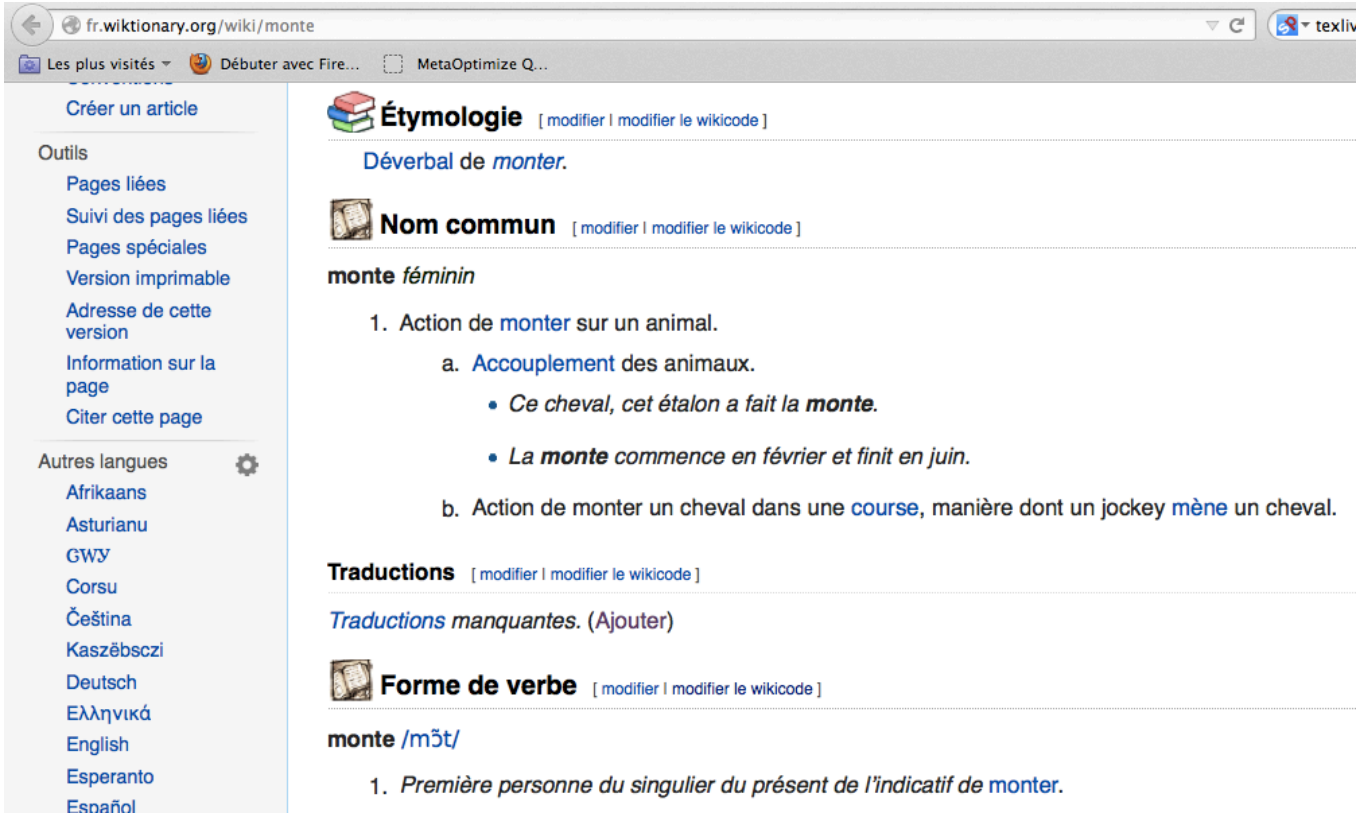
Exemple



1re étape : on utilise un dictionnaire pour étiqueter les mots

$\Rightarrow$ étiquetage ambigu

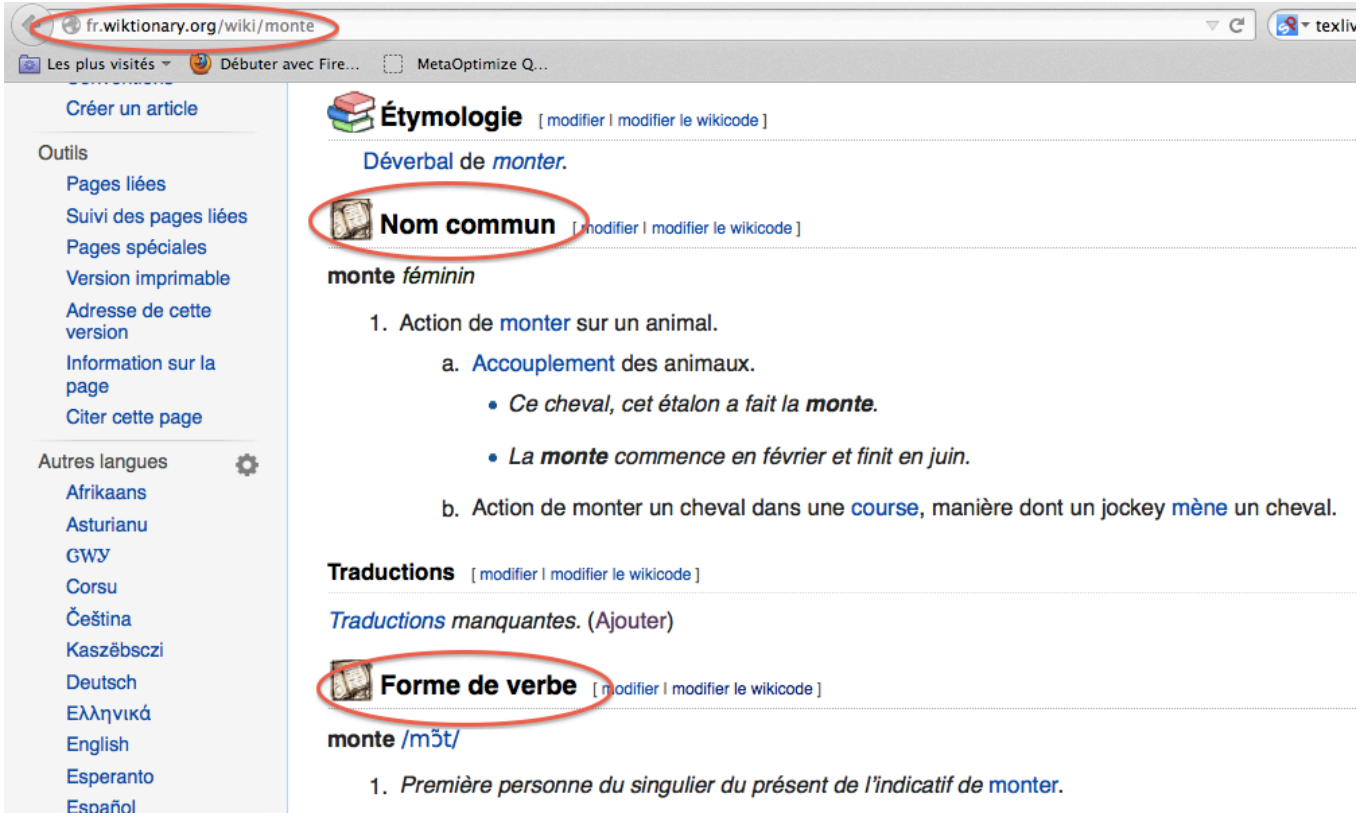
Source de contraintes de type : Wiktionary



Dictionnaire multilingue collaboratif

- plus de 150 langues
- plus de 225 000 mots en français

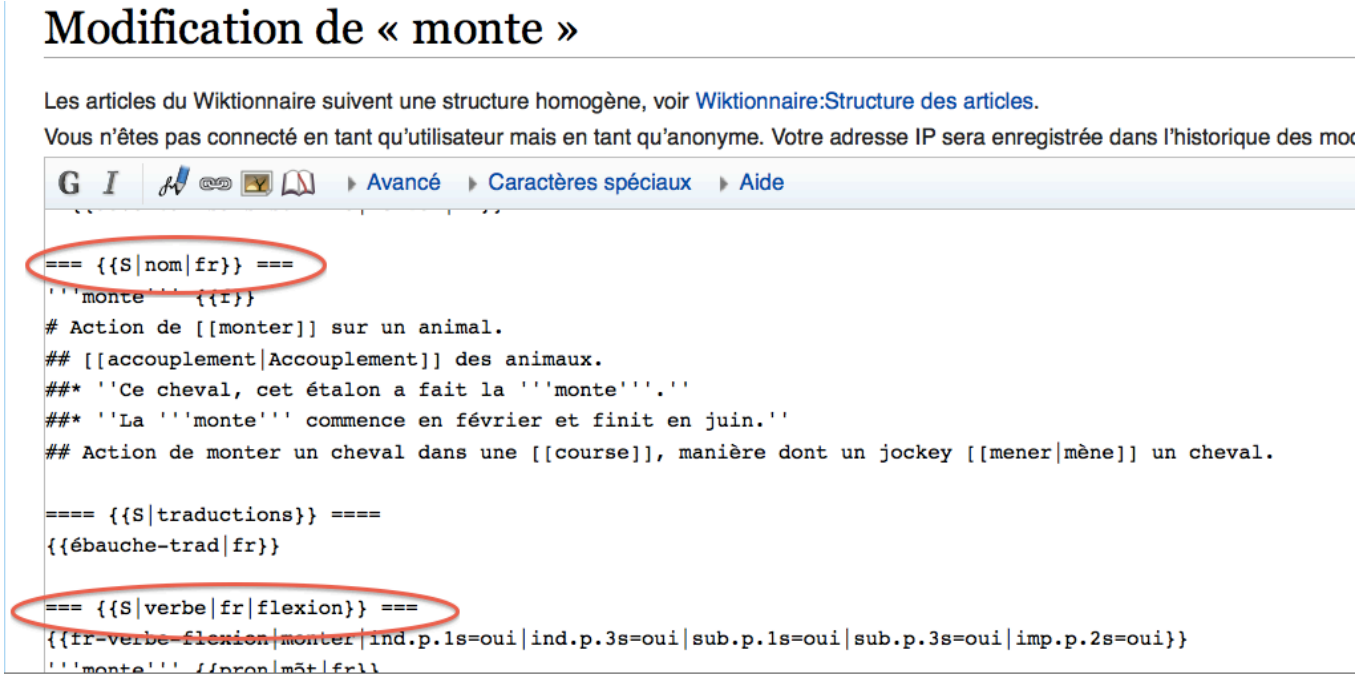
Source de contraintes de type : Wiktionary



Dictionnaire multilingue collaboratif

- plus de 150 langues
- plus de 225 000 mots en français

Extraction des PoS de Wiktionary



Modification de « monte »

Les articles du Wiktionnaire suivent une structure homogène, voir Wiktionnaire:Structure des articles.

Vous n'êtes pas connecté en tant qu'utilisateur mais en tant qu'anonyme. Votre adresse IP sera enregistrée dans l'historique des mod

G I Avancé Caractères spéciaux Aide

```
=== {{S|nom|fr}} ===
'''monte'''  {{f}}
# Action de [[monter]] sur un animal.
## [[accouplement|Accouplement]] des animaux.
##* ''Ce cheval, cet étalon a fait la '''monte'''.''
##* ''La '''monte''' commence en février et finit en juin.''
## Action de monter un cheval dans une [[course]], manière dont un jockey [[mener|mène]] un cheval.

==== {{S|traductions}} ====
{{ébauche-trad|fr}}

=== {{S|verbe|fr|flexion}} ===
{{fr-verbe-flexion|monter|ind.p.1s=oui|ind.p.3s=oui|sub.p.1s=oui|sub.p.3s=oui|imp.p.2s=oui}}
'''monte'''  {{pron|mɔ̃t|fr}}
```

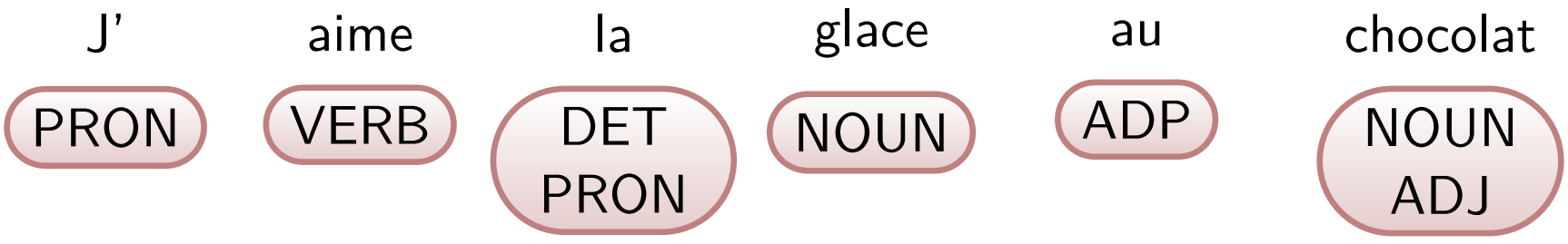
Toute l'information sur les PoS contenue dans des templates

$\Rightarrow$ (relativement) facilement à extraire

Méthode d'étiquetage automatique d'un corpus

Fusion de deux sources d'informations

1. contraintes de type issues d'un dictionnaire ;
2. contraintes de token issues des liens d'alignements ;



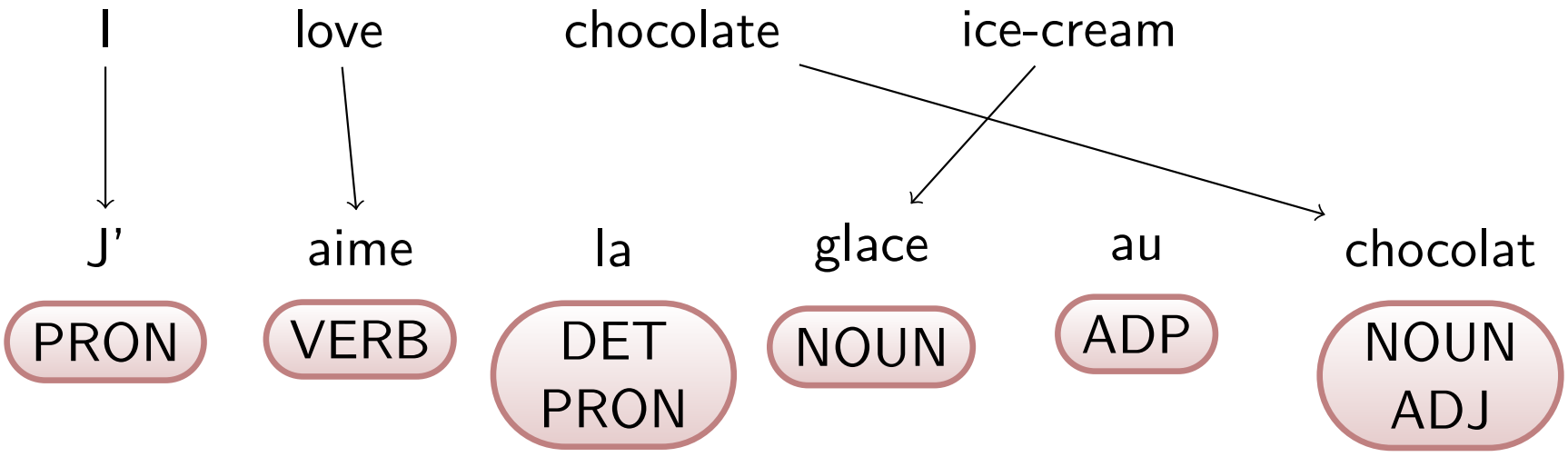
1re étape : on utilise un dictionnaire pour étiqueter les mots

$\Rightarrow$ étiquetage ambigu

Méthode d'étiquetage automatique d'un corpus

Fusion de deux sources d'informations

1. contraintes de type issues d'un dictionnaire ;
2. contraintes de token issues des liens d'alignements ;

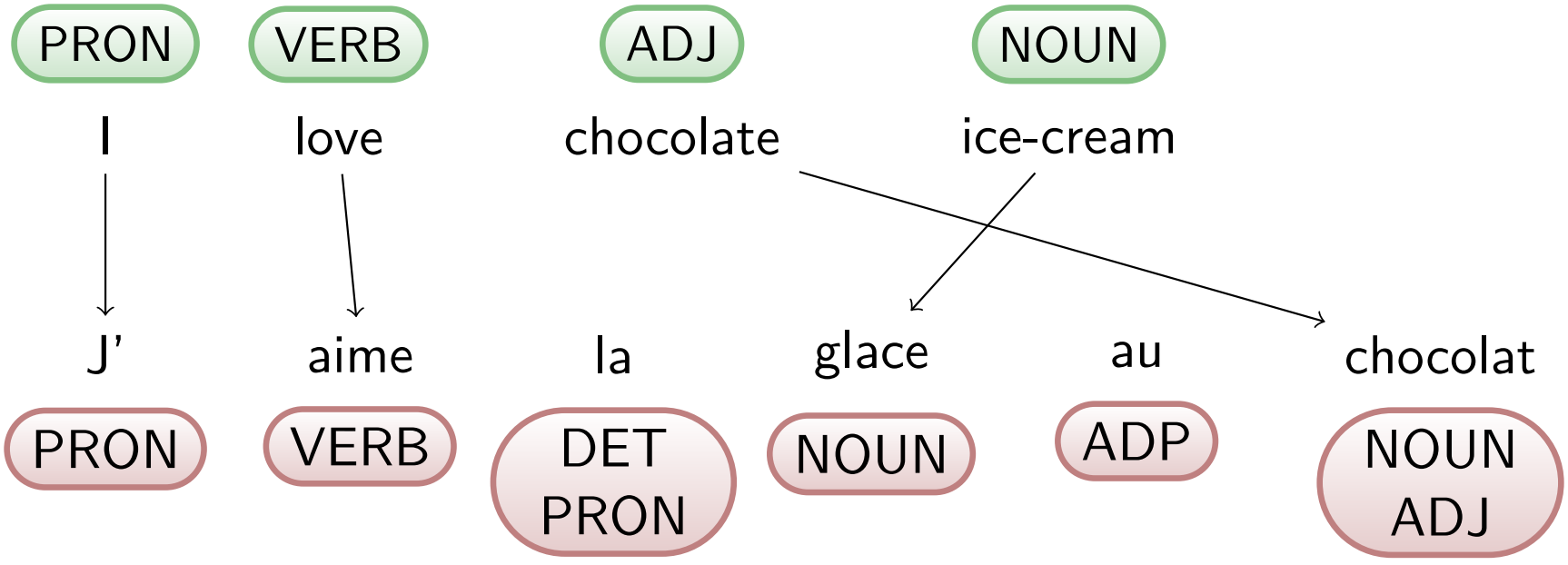


2e étape : on considère les phrases d'un corpus parallèle aligné

Méthode d'étiquetage automatique d'un corpus

Fusion de deux sources d'informations

1. contraintes de type issues d'un dictionnaire ;
2. contraintes de token issues des liens d'alignements ;

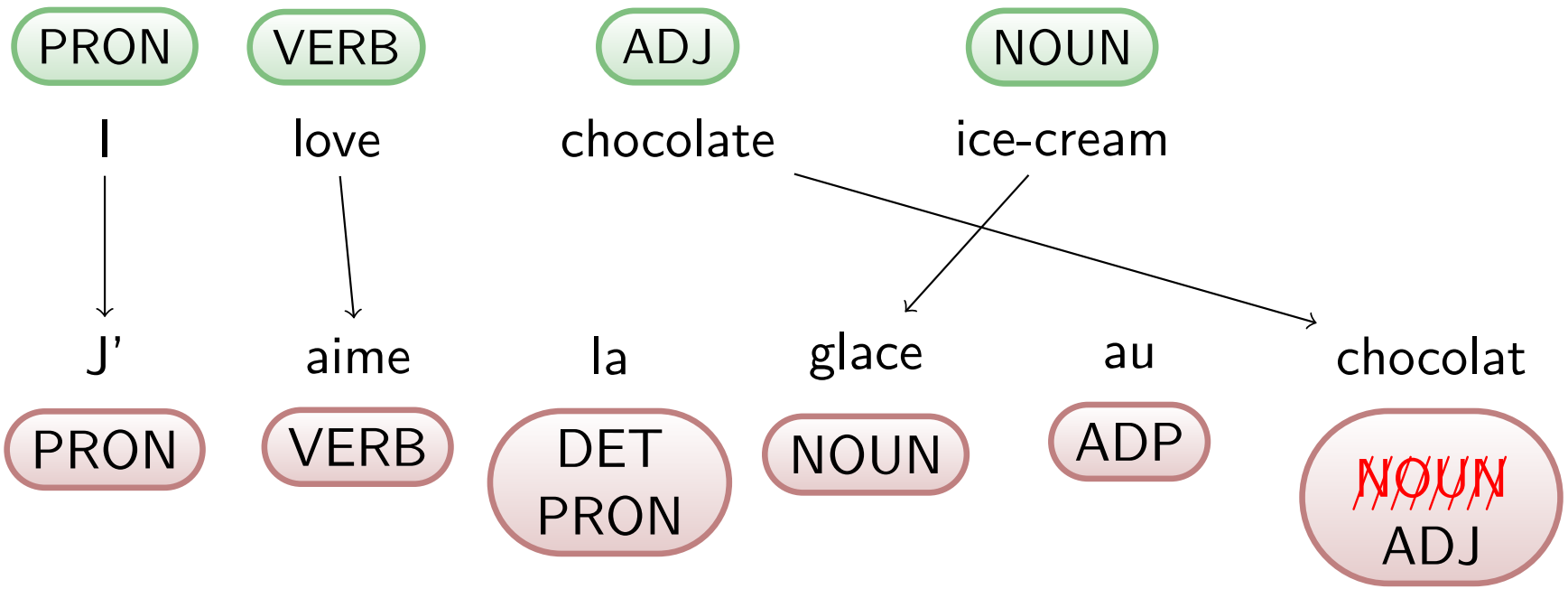


3e étape : on étiquette les phrases source

Méthode d'étiquetage automatique d'un corpus

Fusion de deux sources d'informations

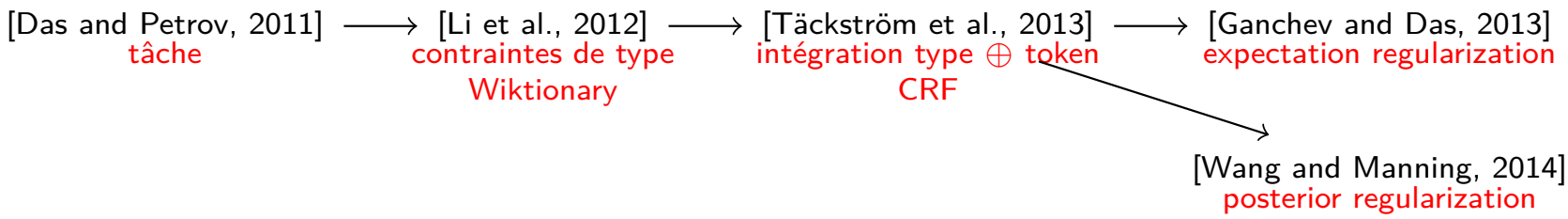
1. contraintes de type issues d'un dictionnaire ;
2. contraintes de token issues des liens d'alignements ;



4e étape : on projette les étiquettes en suivant les liens d'alignements

$\Rightarrow$ réduction de l'ambiguïté

État de l'art



Principaux résultats de l'état de l'art :

1. il est possible d'apprendre un étiqueteur morpho-syntaxique de manière faiblement supervisée
2. fusionner contraintes de type et de token améliore fortement les résultats

Notre contribution : nouvelle méthode d'apprentissage, plus simple avec de meilleures garanties théoriques

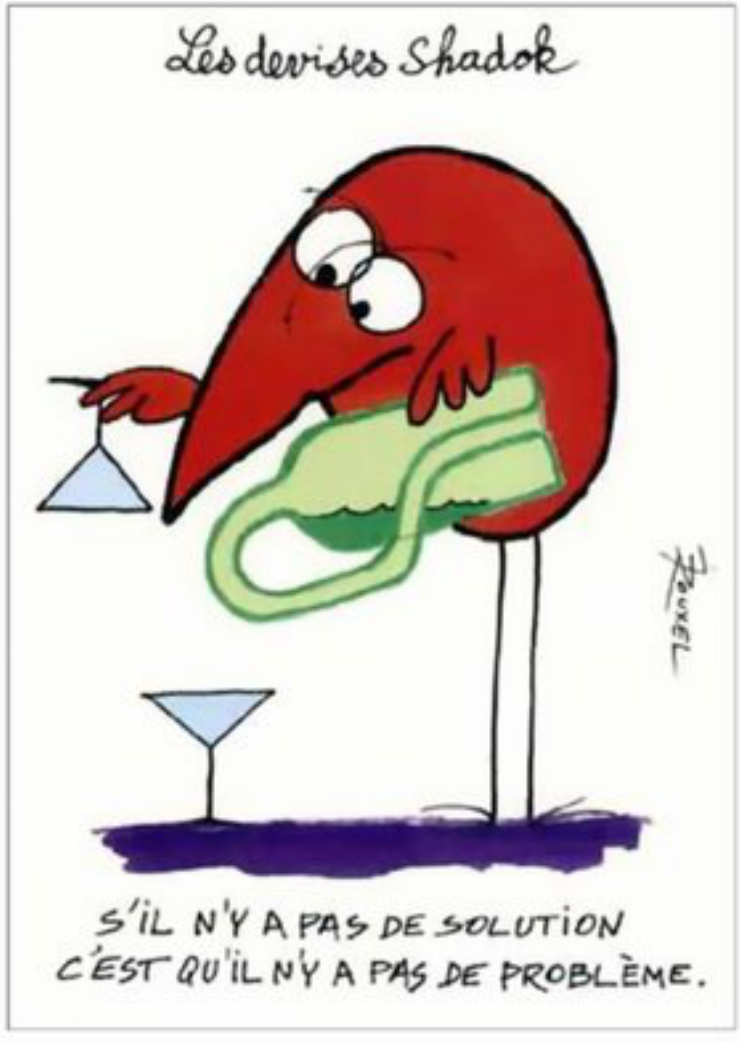
Jeu d'étiquettes

- Notre approche suppose un jeu commun d'étiquettes morpho-syntaxiques en source et en cible
- Plusieurs jeux d'étiquettes universelles ont été proposées
- Nous utiliserons les 12 étiquettes universelles définies par [Petrov et al., 2012]

Deuxième partie II

Apprentissage partiellement supervisé

Le problème



- apprendre à partir d'une séquence d'observations partiellement supervisée
- chaque observation est associée à une ou plusieurs étiquettes
- une seule de ces étiquettes est la bonne

$\Rightarrow$ problème de l'apprentissage ambigu [Bordes et al., 2010], [Cour et al., 2011]

Modèle à base d'historique : inférence

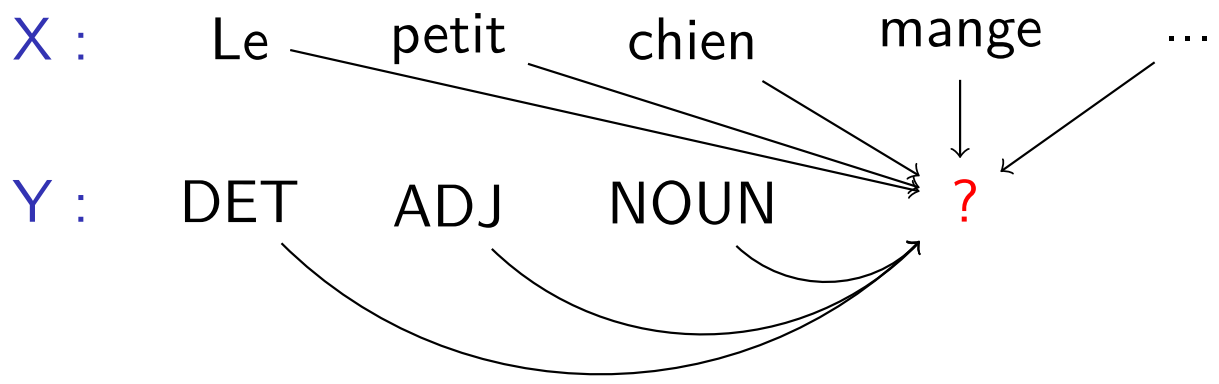
X :	Le	petit	chien	mange	...
Y :	DET	ADJ	NOUN	?	

$$y_i = ?$$

Principe

- réduction de la prédiction d'une séquence à une séquence de prédictions multi-classe
- étant données : une séquence d'observations et une séquence d'étiquettes partielle, choisir l'étiquette suivante

Modèle à base d'historique : inférence



$$y_i = \underset{y \in \{\text{NOUN, VERB, ...}\}}{\arg\max} F(\mathbf{x}, y, y_{i-1}, y_{i-2}, ...)$$

Principe

- réduction de la prédiction d'une séquence à une séquence de prédictions multi-classe
- étant données : une séquence d'observations et une séquence d'étiquettes partielle, choisir l'étiquette suivante

Modèle à base d'historique : apprentissage

- classifieur linéaire $y_i^* = \arg\max_{y \in \mathcal{Y}} \langle \mathbf{w} | \phi(\mathbf{x}, i, y, h_i) \rangle$
- apprentissage en ligne (perceptron)
  1. on prend une décision
  2. si la décision n'est pas bonne $\rightarrow$ correction de la fonction de prédiction = mise-à-jour des poids
- deux types de mise à jour :
  1. dans le cas supervisé :

$$\mathbf{w} \leftarrow \mathbf{w} + \phi\left(\mathbf{x}, i, \hat{y}_i, h_i\right) - \phi\left(\mathbf{x}, i, y_i^*, h_i\right)$$

     renforce la bonne étiquette au détriment de l'étiquette prédite
  2. dans le cas ambigu :

$$\mathbf{w} \leftarrow \mathbf{w} + \sum_{\hat{y}_i \in \hat{\mathcal{Y}}_i} \left(\phi\left(\mathbf{x}, i, \hat{y}_i, h\right) - \phi\left(\mathbf{x}, i, y_i^*, h_i\right)\right)$$

     renforce toutes les étiquettes possibles au détriment de l'étiquette prédite $\Rightarrow$ mêmes performances que si la référence était connue

Troisième partie III

Résultats expérimentaux

Cadre expérimental

Expérience sur 4 langues : français, espagnol, grec et allemand

Notre méthode nécessite :

- un corpus parallèle — Europarl
- un corpus de test étiqueté — CoNLL'07
- une source de contraintes de type — Wiktionary
- un étiqueteur morpho-syntaxique anglais — Wapiti

Toutes les étiquettes sont mappées vers les POS tags universaux

$\Rightarrow$ nous avons essayé de reproduire (au mieux) le cadre expérimental de [Täckström et al., 2013]

Résultats

Erreur moyenne sur le corpus de test :

	fr	el	es	de
Méthode proposée	**8,9%**	**8,3%**	**7,0%**	10,1%
[Täckström et al., 2013]	11,6%	10,5%	10,9%	**9,5%**
[Li et al., 2012]	—	20,8%	13,6%	14,2%

$\Rightarrow$ Amélioration significative de la précision pour 3 paires de langues sur 4

Interprétation des résultats (pour l'espagnol)

Problème

- amélioration significative de l'état de l'art... $10{,}9\% \rightarrow 7{,}0\%$
- ...mais très loin des performances d'une approche supervisée $1{,}3\%$

Explications

- raisons : évaluation hors-domaine :
  train=Europarl, test=CoNLL
- conséquences :
  1. pas la même *tokenization*
  2. pas le même vocabulaire
  3. conventions d'annotation différentes

Interprétation des résultats (pour l'espagnol)

Problème

- amélioration significative de l'état de l'art... $10{,}9\% \rightarrow 7{,}0\%$
- ...mais très loin des performances d'une approche supervisée $1{,}3\%$

Explications

- raisons : évaluation hors-domaine : train=Europarl, test=CoNLL
- conséquences :
  1. pas la même *tokenization*
  2. pas le même vocabulaire

  3. conventions d'annotation différentes

Les différentes conventions d'annotation

Problème

- approche proposée fusionne trois sources de données
- chaque source de données a été annotée indépendamment ⇒ *mismatches*

Exemples

- en français : les nombres sont étiquetés soit comme ADJ soit comme DET, en anglais, systématiquement en NUM
- en grec : les noms propres étrangers sont étiquetés comme X, en anglais comme NOUN
- en espagnol : *poco* est étiqueté comme DET, alors que *few* est étiqueté comme ADJ et que Wiktionary comme un PRON, un NOUN ou un ADJ

Impact des différences d'annotation

train	étiquettes en train	perf. en test
CoNLL	manuelle	1,3%
CoNLL	restreintes aux étiquettes Wiktionary	4,3%
Europarl	prédites par FreeLing	6,7%
Europarl	transfert cross-lingue	7,0%

$\Rightarrow$ les performances de notre méthode sont sous-estimées

Quatrième partie IV

Conclusions

Conclusions

Contributions

- nouvelle méthode pour apprendre un étiqueteur morpho-syntaxique dans un cadre de transfert cross-lingue
- analyse « fine » des résultats
- résultats proches des meilleures performances atteignables

Perspectives

- évaluation extrinsèque (p.ex. avec une analyse en dépendances)
- transférer d'autres types d'annotations (NER, dépendances, ...)

Bordes, A., Usunier, N., and Weston, J. (2010).
Label ranking under ambiguous supervision for learning semantic correspondences.
In Proceedings of the International Conference on Machine Learning, ICML'10, pages 103–110.

Cour, T., Sapp, B., and Taskar, B. (2011).
Learning from partial labels.
Journal of Machine Learning Research, 12 :1501–1536.

Das, D. and Petrov, S. (2011).
Unsupervised part-of-speech tagging with bilingual graph-based projections.
In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, HLT '11, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ganchev, K. and Das, D. (2013).
Cross-lingual discriminative learning of sequence models with posterior regularization.

In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP'13, pages 1996–2006, Seattle, Washington, USA.

Li, S., Graça, J. a. V., and Taskar, B. (2012).
Wiki-ly supervised part-of-speech tagging.
In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 1389–1398, Stroudsburg, PA, USA.

Petrov, S., Das, D., and McDonald, R. (2012).
A universal part-of-speech tagset.
In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12, Istanbul, Turkey. European Language Resources Association (ELRA).

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013).

Token and type constraints for cross-lingual part-of-speech tagging.
Transactions of the Association for Computational Linguistics, 1 :1–12.

Wang, M. and Manning, C. D. (2014).
Cross-lingual projected expectation regularization for weakly supervised learning.
Transactions of the Association for Computational Linguistics, 2 :55–66.