Création d'un étiqueteur automatique pour la détection d'expressions polylexicales

Manon Scholivet
Mémoire encadré par Carlos Ramisch

15 janvier 2018

Plan

Détection des expressions polylexicales (MWEs)

Approches existantes

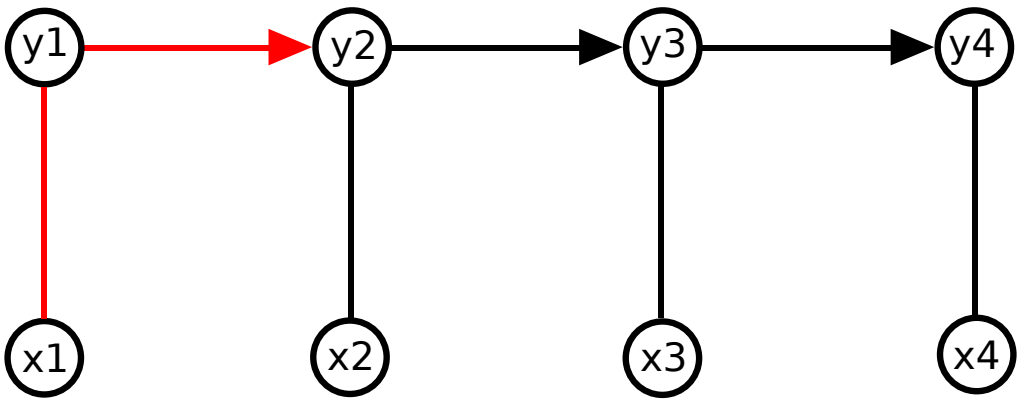
- Dictionnaires
- Analyseur syntaxique
- Apprentissage de modèles de séquence
  - → Requiert peu d'annotations
  - → Peu de contraintes sur la structure
  - → Approche indépendante de la langue

i:	1	2	3	4	5
w_i :	*Le*	*français*	***défend***	*ses*	***couleurs***
MWE :	O	O	B1	O	I1

Plan

Les *Conditional random fields* (CRF)

- Apprennent une fonction f qui prédit une étiquette y_t en fonction des traits $r(x_t)$ de l'entrée x_t et de l'étiquette précédente y_{t-1}
  - $y_t = f(r(x_t), y_{t-1})$



Obstacles

- Dépendances lointaines pas prises en compte :
  Il ***fera*** *cet après-midi une belle* ***présentation***
- Pas de généralisation sur les mots ayant un sens proche :
  Faire *un/une* ***présentation/discours/intervention***

Obstacles/Solutions

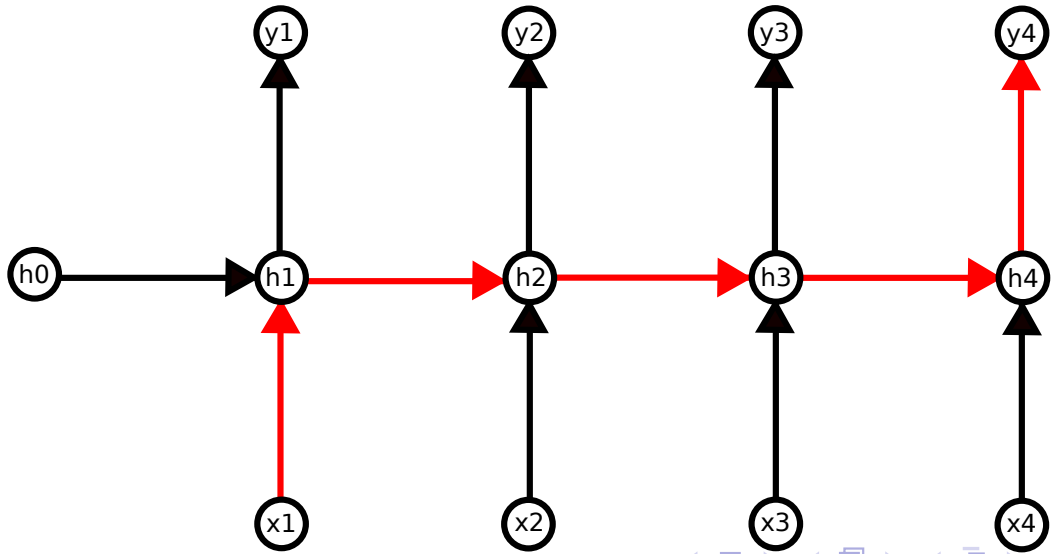
- Dépendances lointaines pas prises en compte :
  *Il **fera** cet après-midi une belle **présentation***
  → Réseaux de neurones récurrents (RNN)
- Pas de généralisation sur les mots ayant un sens proche :
  Faire** un/une **présentation/discours/intervention

Obstacles/Solutions

- Dépendances lointaines pas prises en compte :
  *Il **fera** cet après-midi une belle **présentation***
  → Réseaux de neurones récurrents (RNN)
- Pas de généralisation sur les mots ayant un sens proche :
  Faire** un/une **présentation/discours/intervention
  → Word embeddings

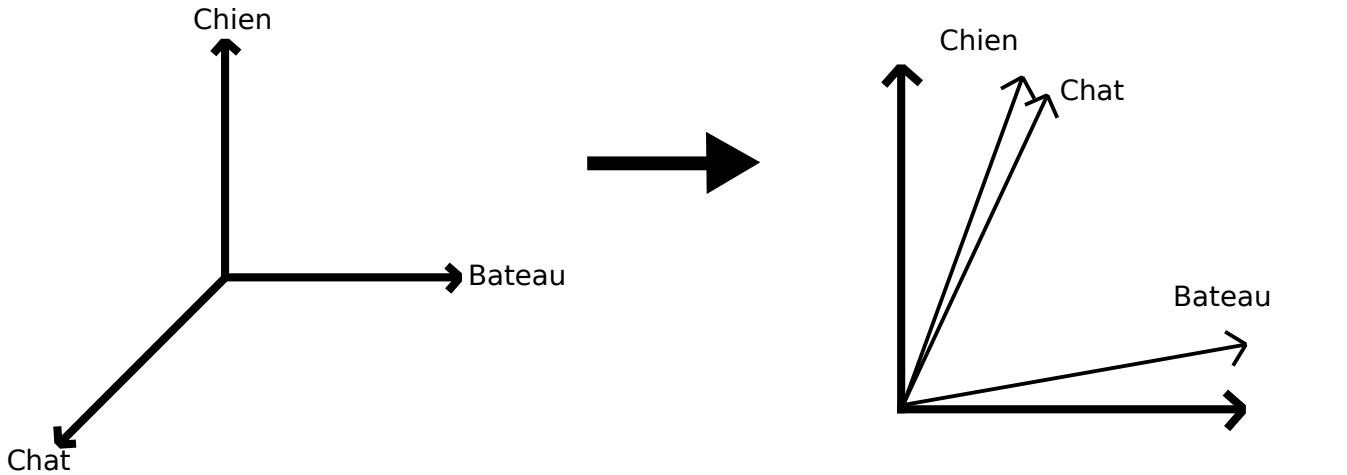
Les réseaux de neurones récurrents (RNN)

- Résolvent le problème des dépendances lointaines
- Apprennent une fonction qui prédit une étiquette y_t en fonction d'un état caché h_t, lui-même dépendant des entrées $r(x_t)$ et de h_{t-1}
  - $h_0 = 0$
  - $h_t = f(r(x_t), h_{t-1})$
  - $y_t = g(h_t)$



Les *Word Embeddings*

→ Résolvent le problème de la généralisation sur les mots de sens proche



One-hot

Word Embeddings

→ Apprentissage par le RNN

→ Pré-apprentissage externe

"You shall know a word by the company it keeps" (Firth, 1957)

Plan

Description du système proposé

- Chaque phrase standardisée à 128 éléments
- Deux entrées : les POS et les lemmes
- Couches d'embeddings, 64 neurones par défaut
- Concaténation des deux couches d'embeddings
- 2 couches de GRU bidirectionnelles
- Couche dense
- Couche d'activation softmax

→ Grande variabilité des modèles produits. Vote majoritaire sur une vingtaine de modèles pour obtenir un unique résultat plus stable.

Corpus

- Corpus PARSEME, Campagne d'évaluation 2017 (français) :

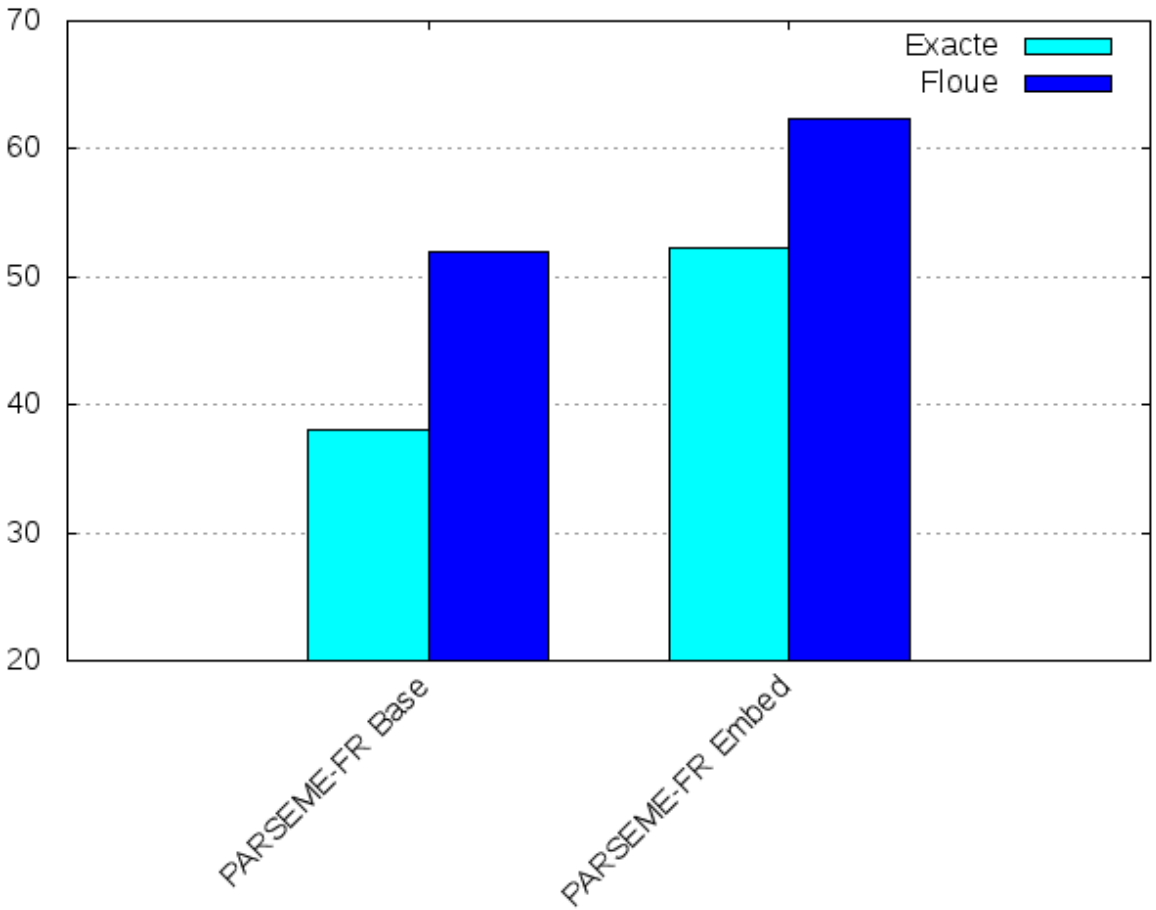
	Train	Test
Nb. phrases	17 880	1 667
Nb. MWEs	4 462	500

Plan

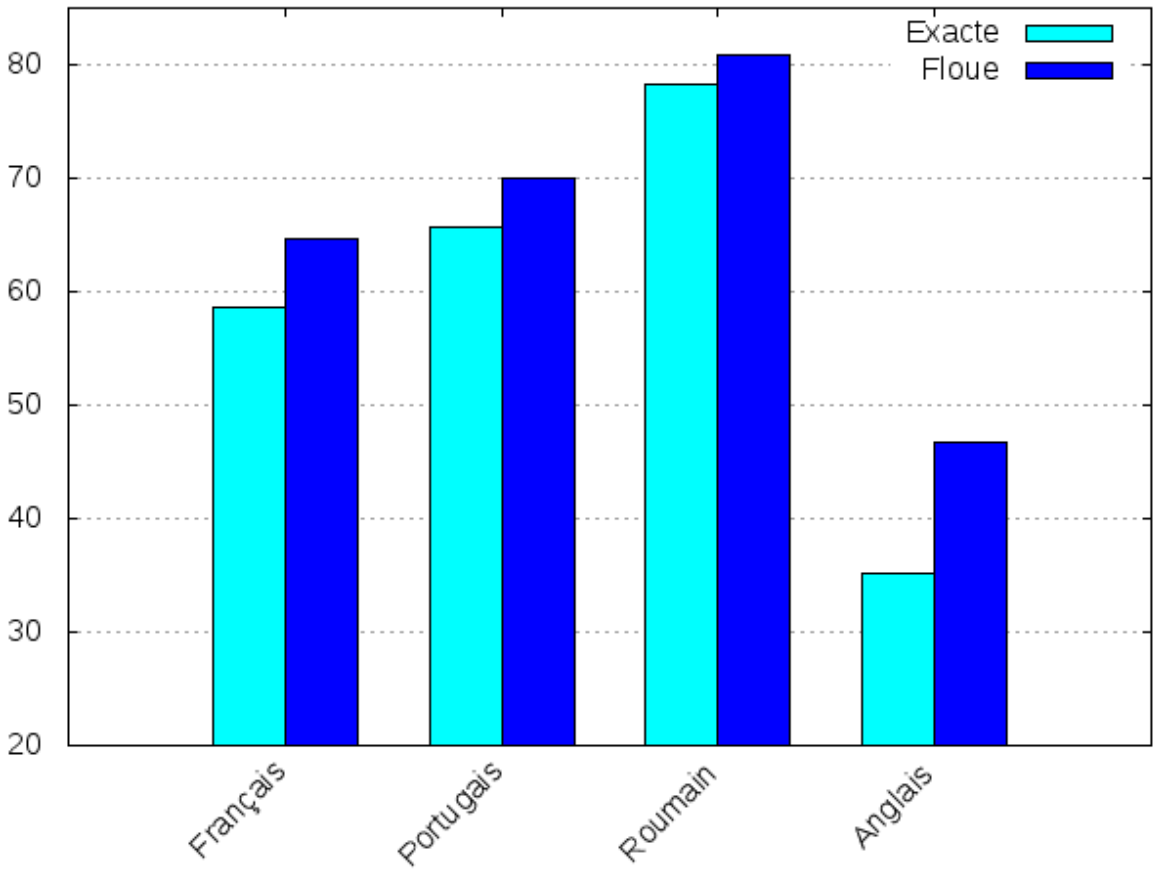
Hypothèses

- Comment la pré-initialisation des embeddings influence les performances ?
- Est-ce que le système est indépendant de la langue ?
- Quelles sont les performances de notre RNN comparées à d'autres systèmes ?
- Le RNN prend t-il mieux en compte :
  - l'historique lointain ? *(gestion des discontinuités)*
  - la variabilité des MWEs ?
  - les nouvelles MWEs ?

Pré-initialisation des embeddings



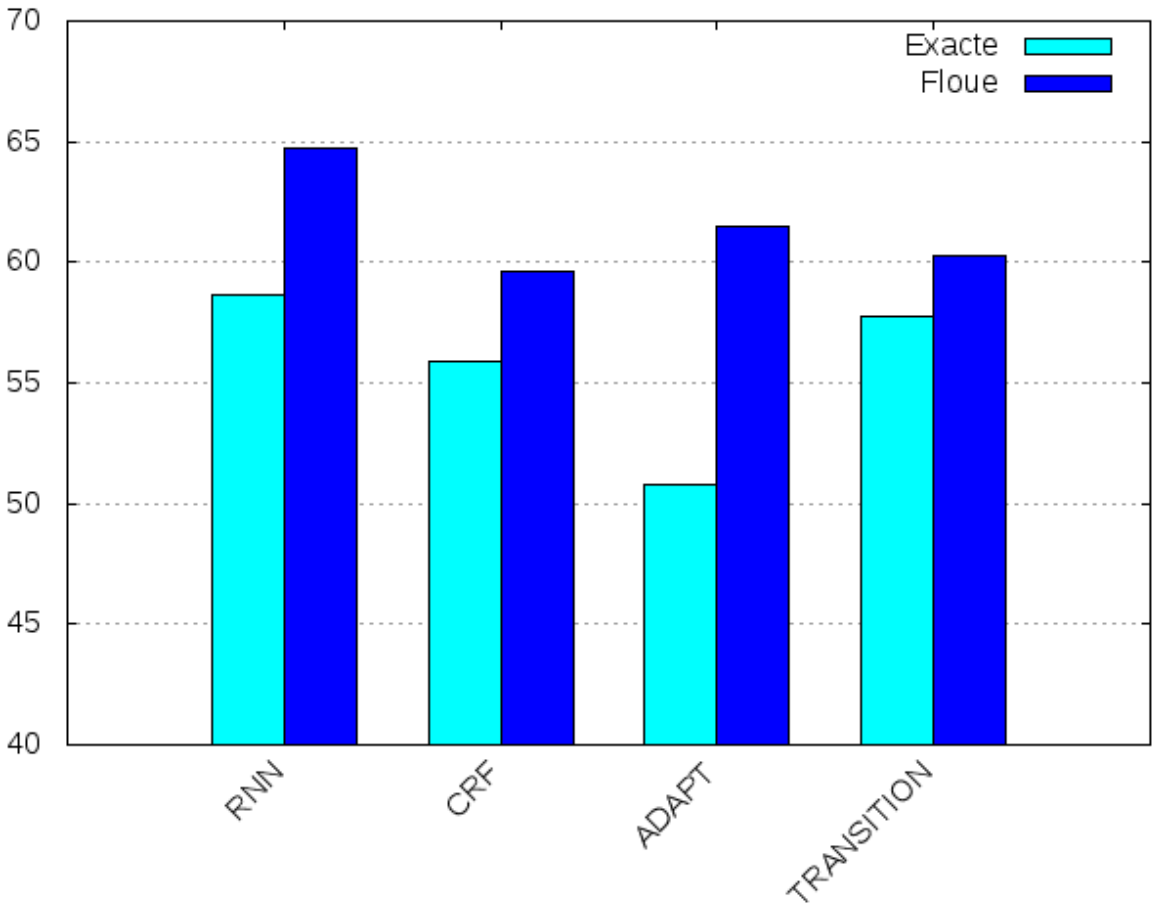
Modèle indépendant de la langue



→ Meilleurs résultats de la campagne d'évaluation :

- Portugais : 67,33% Exacte, 70,94% Floue
- Roumain : 77,21% Exacte, 83,58% Floue
- Anglais : 57,24% Floue

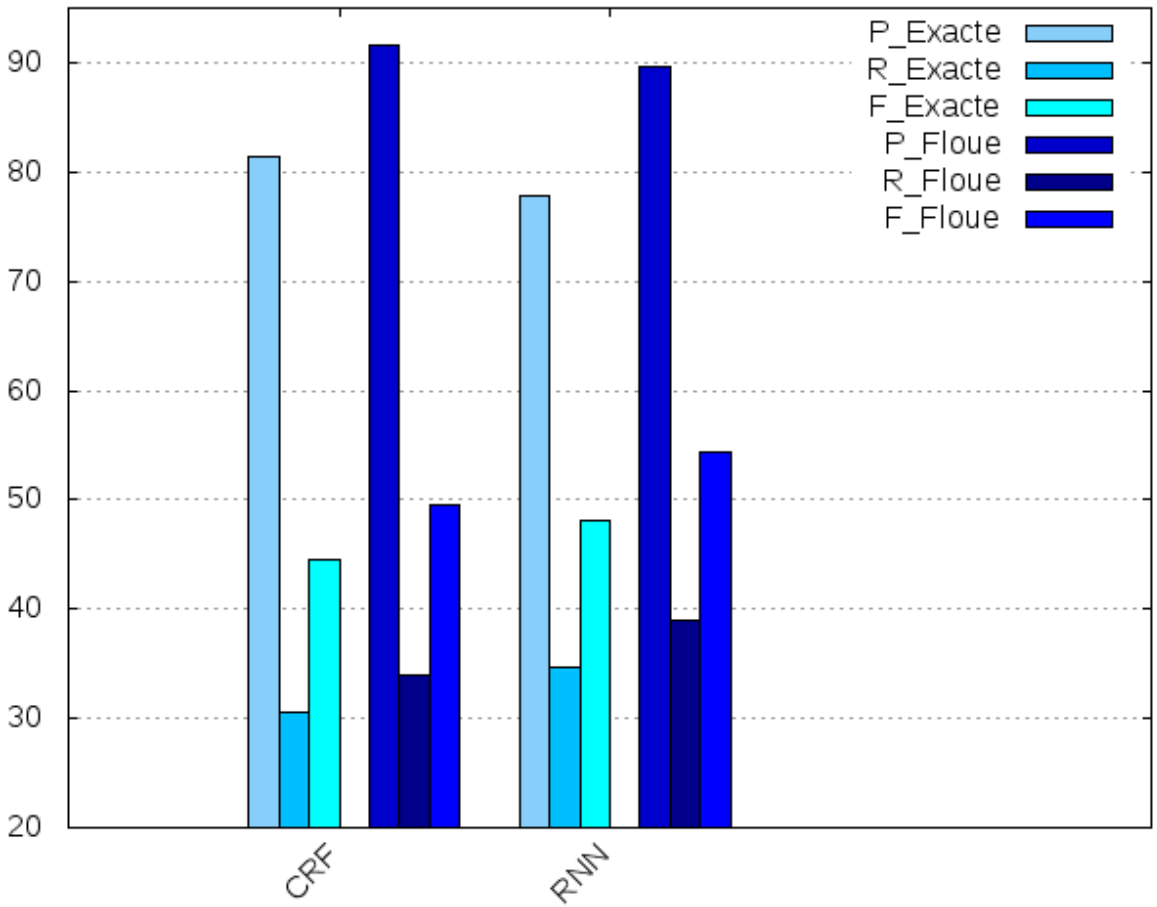
Comparaison avec d'autres systèmes sur PARSEME-FR



Études de phénomènes spécifiques

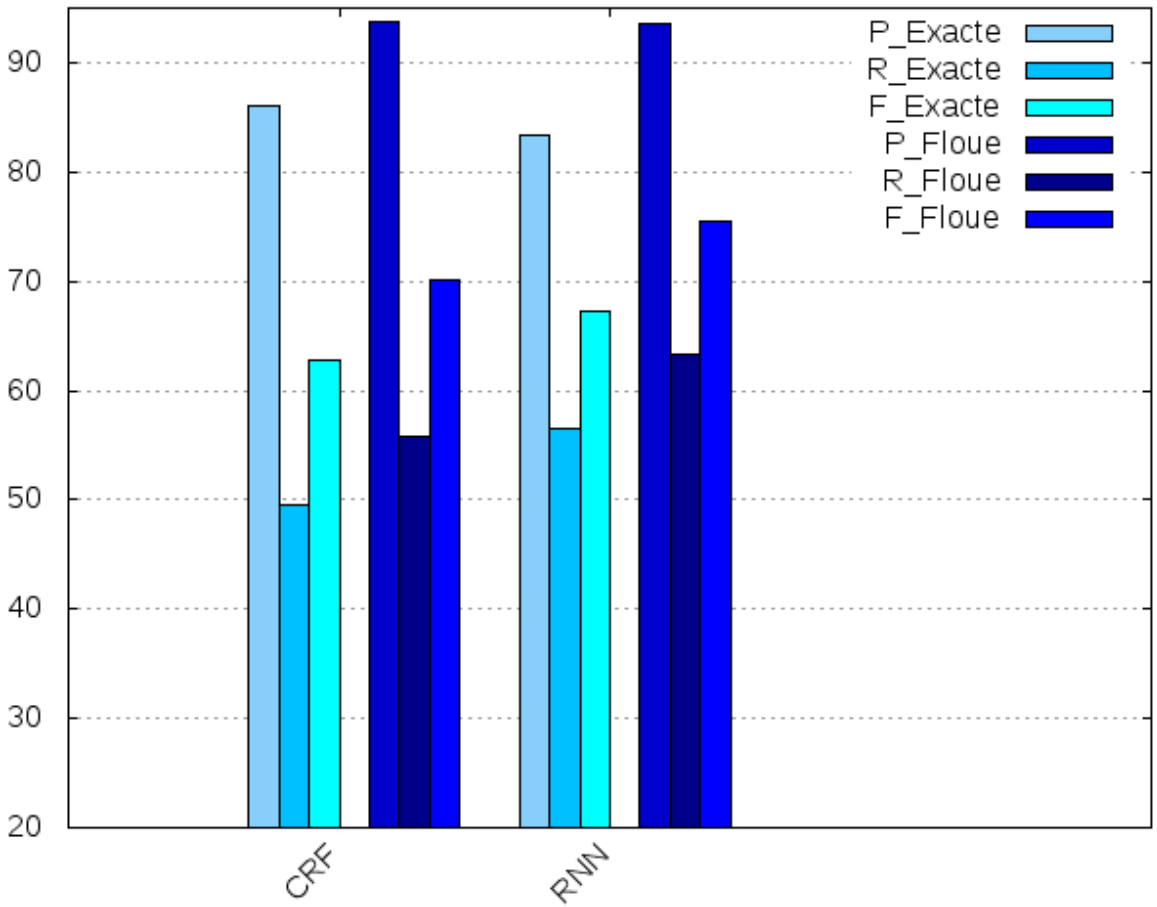
- Création de sous-corpus avec forte densité de :
  - → MWEs discontinues (89.81%)
  - → MWEs variables (82.80%)
  - → MWEs jamais vues (88.00%)

Gestion des MWEs discontinues



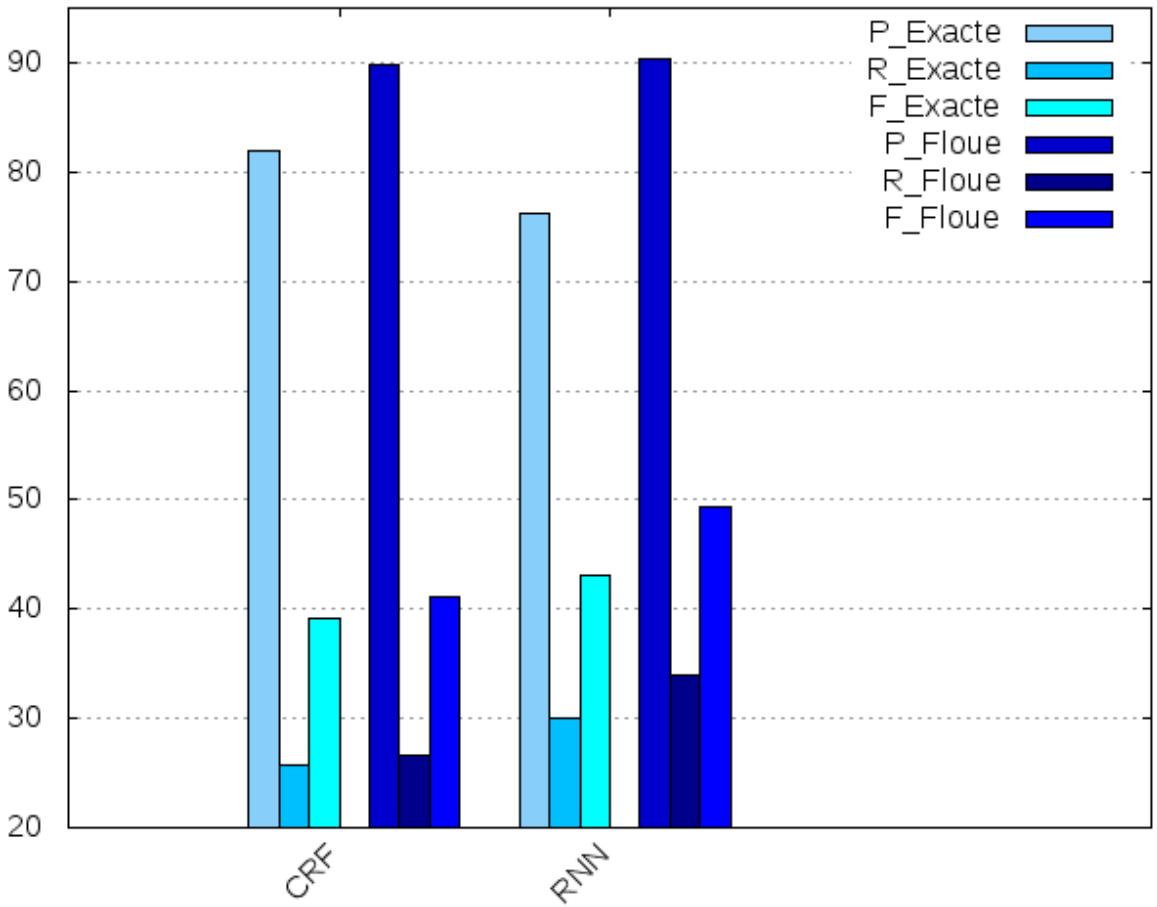
		CRF	RNN
Ex	P	81.36	77.86
	R	30.57	34.71
	F	44.44	48.02
Fl	P	91.67	89.56
	R	33.92	39.06
	F	49.52	54.40

Gestion de la variabilité des MWEs



		CRF	RNN
Ex	P	85.98	83.33
	R	49.46	62.80
	F	56.45	67.31
Fl	P	93.83	93.48
	R	55.88	63.24
	F	70.05	75.44

Gestion des nouvelles MWEs



		CRF	RNN
Ex	P	81.91	76.27
	R	25.67	30.00
	F	39.09	43.06
Fl	P	89.85	90.36
	R	26.70	33.94
	F	41.16	49.34

Les MWEs jamais vues

- *"**La musique** n'**adoucit** pas toujours **les moeurs**."*
  → Aucun de nos systèmes ne détecte cette MWE
- *"... M. Soyer a **fait** l'**historique** de l'école maternelle ..."*
  → Le RNN le détecte, le CRF non
- *"... chacun ne rêvait que de **remettre la main à la pâte** avec Sylvie et Daniel."*
  → Le RNN détecte une partie de la MWE, le CRF aucune

Les MWEs déjà vues

- *"Une **réflexion** commune est **menée** avec les enseignants ..."*
  - → Corpus d'entraînement : *"Nous n'avons **mené** aucune **réflexion** sur le sujet"*
  - → Le RNN détecte la MWE, le CRF non
- *"... elle descendait seule **faire** ses **courses** en centre ville, et **prenait** alors **plaisir** à **faire** un brin de **causette** avec ses copines ..."*
  - → $1^{ère}$ MWE détectée par le CRF, mais mal détectée par le RNN
  - → $2^{ème}$ MWE détectée par les deux systèmes
  - → $3^{ème}$ MWE n'est détectée par aucun système

Plan

Perspectives

- Gestion des chevauchements
  - → *Elle **fera** un **discours** et un **hommage** aux bénévoles*
  - → Utiliser un nouveau système d'étiquettes (ex. *pointer networks*)
- Rareté des données annotées, traitement de nouvelles langues
  - → Projet de thèse sur des modèles profondément multilingues

Merci pour votre écoute !

Table – Détails des résultats des analyses d'erreurs

		Exacte			Floue		
		P	R	F	P	R	F
CRF	Complet	**80.45**	42.80	55.87	**86.45**	45.49	59.61
	Disc	**81.36**	30.57	44.44	**91.67**	33.92	49.52
	Variab	**85.98**	49.46	62.80	**93.83**	55.88	70.05
	US	**81.91**	25.67	39.09	89.85	26.70	41.16
RNN	Complet	75.24	**48.00**	**58.61**	84.67	**52.35**	**64.70**
	Disc	77.86	**34.71**	**48.02**	89.56	**39.06**	**54.40**
	Variab	83.33	**56.45**	**67.31**	93.48	**63.24**	**75.44**
	US	76.27	**30.00**	**43.06**	**90.36**	**33.94**	**49.34**

Table – Détails de la construction des sous corpus

	Disc		Variab		Unseen	
	Nb	% ss-corpus	Nb	% ss-corpus	Nb	% ss-corpus
FR	280	89.81	154	82.80	264	88.00
RO	170	88.08	62	73.81	41	73.21
PT	225	88.93	201	87.39	165	91.67