Impact des représentations des mots sur un tagger d'expressions basé sur les réseaux neuronaux

Nicolas ZAMPIERI (LIS) – Carlos Ramisch (LIS) – Géraldine Damnati (Orange Labs)

PARSEME-FR - 2019

Sommaire

- Introduction
- Description de la tâche
- Description Système
- Expériences et résultats
- Conclusion et perspectives

Description de la tâche

- Tâche:
  - Reconnaître automatiquement les expressions.
    - David fait une présentation.
    - Une présentation est faite par David.
    - Plaidez-vous coupable ou non coupable ?
    - David se fait des idées

Description de la tâche

- Tâche:
  - Reconnaître automatiquement les expressions.
    - David fait une présentation.
    - Une présentation est faite par David.
    - Plaidez-vous coupable ou non coupable ?
    - David se fait des idées.

Description de la tâche

- Pourquoi est-ce une tâche importante ?

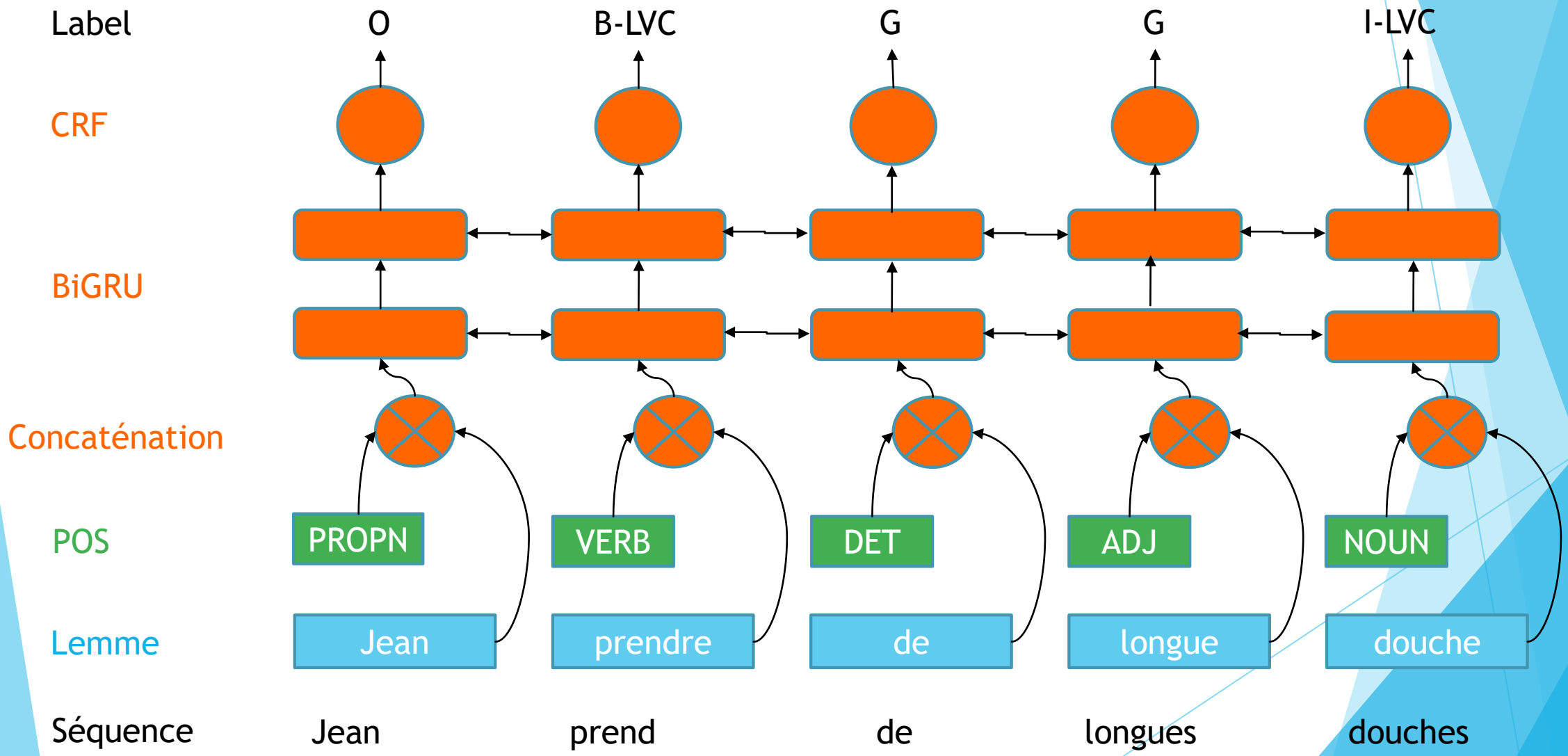
Français	Anglais
Je pense que nous devons prendre le taureau par les cornes	I think we have to take the bull by the horns

Anglais	Français
I think that we must grasp the nettle	Je pense que nous devons saisir l'ortie

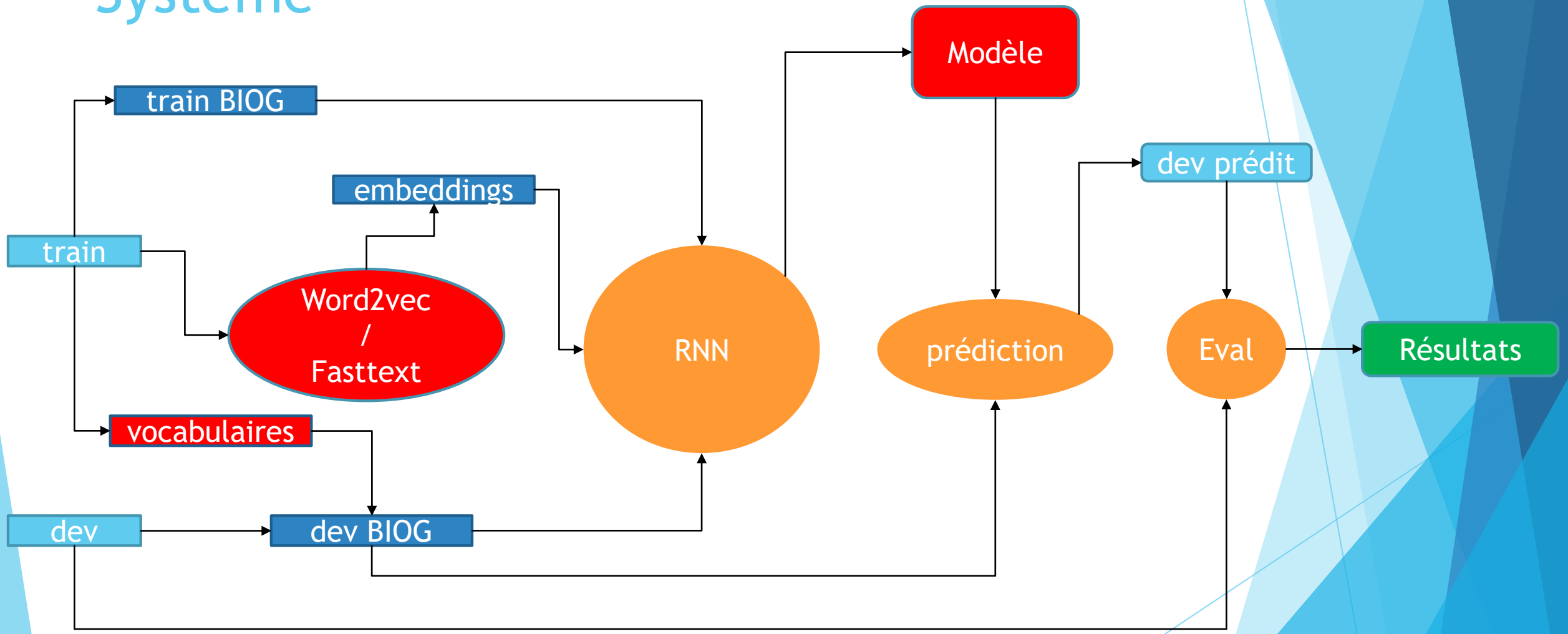
Système

- Système de base
  - https://github.com/zamp13/Veyn
- Représentation des mots (embeddings)
  - *Word2vec*
  - *FastText*

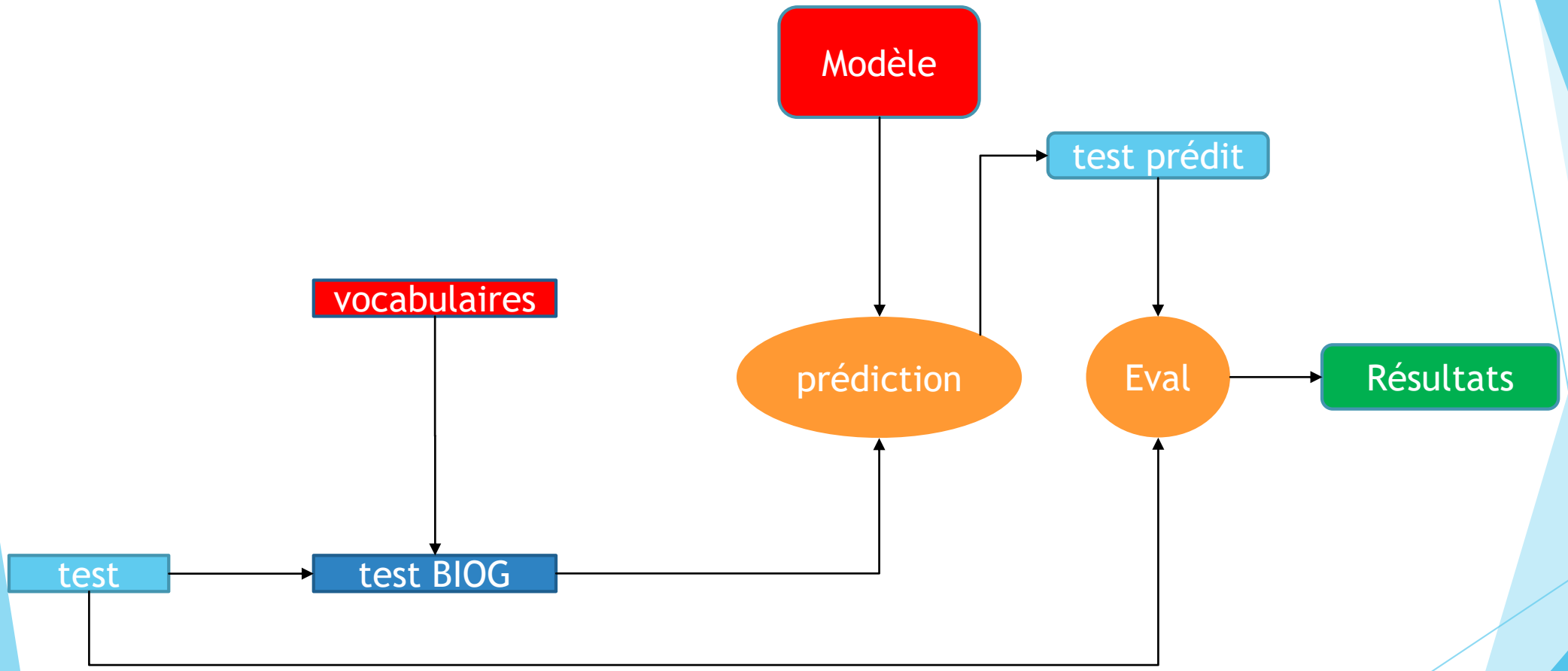
Système


Label
O
B-LVC
G
G
I-LVC
CRF
BiGRU
Concaténation
POS
PROPN
VERB
DET
ADJ
NOUN
Lemme
Jean
prendre
de
longue
douche
Séquence
Jean
prend
de
longues
douches

Système

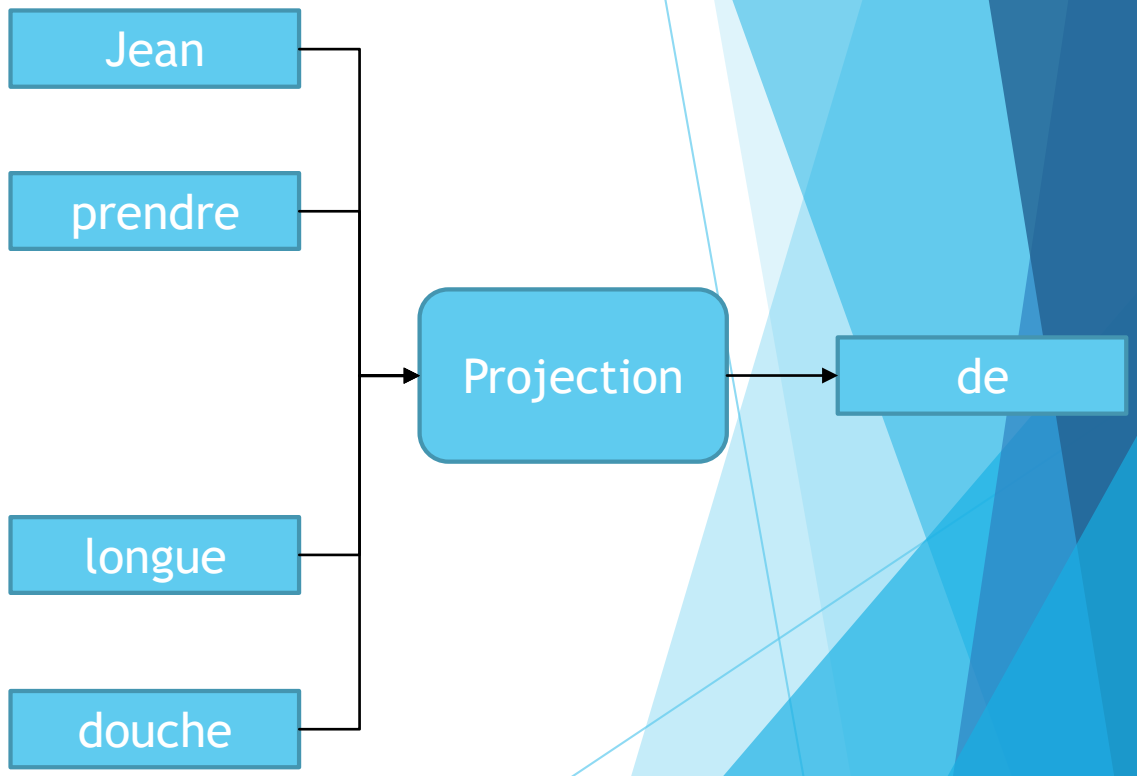

Modèle
train BIOG
dev prédit
embeddings
train
Word2vec / Fasttext
RNN
prédiction
Eval
Résultats
vocabulaires
dev
dev BIOG

Système


Modèle
test prédit
vocabulaires
prédiction
Eval
Résultats
test
test BIOG

Système

- Représentations de mots :
  - Word2vec
    - Représentation des mots classiques
    - Algorithme CBOW
    - Beaucoup de mots inconnus
  - FastText
    - Variante de word2vec
    - Utilisation des n-grams de caractères
    - Algorithme CBOW
    - Possibilité de réduire le nombre de mots inconnus



Expériences et résultats

- Corpus français de la campagne d'évaluation PARSEME 2018
  - Corpus d'entrainement
  - Corpus de développement
  - Corpus de test

- Evaluation :
  - Comparaison de la prédiction avec les expressions de la référence
  - Deux F-mesures, F-MWE et F-TOK, (Savary et al., 2017)

Expériences et résultats

- Détails des corpus utilisés

Corpus	Tokens	VMWE	Vocabulaire		Morpho	OOVs-Vocabulaire		OOVs-Verbes	
			Formes	Lemmes		Formes	Lemmes	Formes	Lemmes
EU-train	117 165	2 832	26 912	11 602	3,32	---	---	---	---
EU-dev	21 604	500	7 766	4 178	1,86	43 %	29 %	32 %	18 %
EU-test	19 038	500	7 226	3 902	1,85	43 %	28 %	32 %	15 %
FR-train	**420 762**	**4 550**	45 166	**33 928**	1,33	---	---	---	---
FR-dev	54 685	629	11 593	8 814	1,32	26 %	27 %	23 %	12 %
FR-test	38 402	498	8 160	6 052	1,35	20 %	19 %	23 %	16 %
PL-train	220 352	4 122	**48 211**	21 795	2,21	---	---	---	---
PL-dev	26 014	515	10 007	5 955	1,68	34 %	19 %	42 %	14 %
PL-test	27 661	515	10 285	6 408	1,61	40 %	25 %	44 %	16 %

Expériences et résultats

- Détails des corpus utilisés

Corpus	Tokens	VMWE	Vocabulaire		Morpho	OOVs-Vocabulaire		OOVs-Verbes	
			Formes	Lemmes		Formes	Lemmes	Formes	Lemmes
EU-train	117 165	2 832	26 912	11 602	3,32	---	---	---	---
EU-dev	21 604	500	7 766	4 178	1,86	43 %	29 %	32 %	18 %
EU-test	19 038	500	7 226	3 902	1,85	43 %	28 %	32 %	15 %
FR-train	420 762	4 550	45 166	33 928	1,33	---	---	---	---
FR-dev	54 685	629	11 593	8 814	1,32	26 %	27 %	23 %	12 %
FR-test	38 402	498	8 160	6 052	1,35	20 %	19 %	23 %	16 %
PL-train	220 352	4 122	48 211	21 795	2,21	---	---	---	---
PL-dev	26 014	515	10 007	5 955	1,68	34 %	19 %	42 %	14 %
PL-test	27 661	515	10 285	6 408	1,61	40 %	25 %	44 %	16 %

Expériences et résultats

- Détails des corpus utilisés

Corpus	Tokens	VMWE	Vocabulaire		Morpho	OOVs-Vocabulaire		OOVs-Verbes	
			Formes	Lemmes		Formes	Lemmes	Formes	Lemmes
EU-train	117 165	2 832	26 912	11 602	3,32	---	---	---	---
EU-dev	21 604	500	7 766	4 178	1,86	**43 %**	**29 %**	32 %	**18 %**
EU-test	19 038	500	7 226	3 902	1,85	**43 %**	**28 %**	32 %	15 %
FR-train	420 762	4 550	45 166	33 928	1,33	---	---	---	---
FR-dev	54 685	629	11 593	8 814	1,32	26 %	27 %	23 %	12 %
FR-test	38 402	498	8 160	6 052	1,35	20 %	19 %	23 %	**16 %**
PL-train	220 352	4 122	48 211	21 795	2,21	---	---	---	---
PL-dev	26 014	515	10 007	5 955	1,68	34 %	19 %	**42 %**	14 %
PL-test	27 661	515	10 285	6 408	1,61	40 %	25 %	**44 %**	**16 %**

Expériences et résultats

- Résultats généraux

Entrées	Embeddings	EU		FR		PL	
		F-MWE	F-TOK	F-MWE	F-TOK	F-MWE	F-TOK
Forme	Word2vec	60,37	70,93	47,41	56,64	42,27	58,23
Forme	FastText	66,52	72,36	52,60	63,47	47,24	56,08
Lemme	Word2vec	53,36	65,37	53,28	63,76	57,82	65,48
Lemme	FastText	62,86	68,79	59,35	**68,60**	**61,49**	63,98
Forme-Lemme	Word2vec	60,56	73,07	56,11	66,31	56,80	**67,16**
Forme-Lemme	FastText	**69,24**	**74,01**	**60,41**	68,39	57,39	64,63

Expériences et résultats

- Expressions Vues (en F-MWE)

Entrées	EU		FR		PL	
	w2v	FT	w2v	FT	w2v	FT
Forme	77,92	80,39	65,82	74,40	69,18	67,47
Lemme	68,83	81,23	79,68	83,23	83,11	78,18
Forme-Lemme	82,80	82,58	76,67	79,80	81,40	78,50

Expériences et résultats

- Expressions Vues (en F-MWE): identique au train

Entrées	EU		FR		PL	
	w2v	FT	w2v	FT	w2v	FT
Forme	94,08	94,50	82,36	**86,63**	88,97	88,49
Lemme	82,81	**91,02**	87,96	**92,20**	89,44	88,73
Forme-Lemme	95,79	95,18	87,11	**90,70**	92,04	**88,65**

Expériences et résultats

- Expressions Vues (en F-MWE): variante au train

Entrées	EU		FR		PL	
	w2v	FT	w2v	FT	w2v	FT
Forme	41,82	50,00	44,14	57,02	50,83	48,37
Lemme	38,83	63,97	68,66	74,56	78,44	69,69
Forme-Lemme	56,33	55,41	62,15	64,26	73,02	70,56

Expériences et résultats

- Expressions non Vues (en F-MWE)

Entrées	EU		FR		PL	
	w2v	FT	w2v	FT	w2v	FT
Forme	3,64	**4,94**	12,88	**18,62**	7,88	10,56
Lemme	3,02	**7,26**	15,87	**18,57**	13,99	16,81
Forme-Lemme	4,10	**5,26**	17,09	**18,57**	11,73	10,96

Conclusion et perspectives

- Annotation d'un corpus de test de la transcription de la parole
- CNN sur les caractères
- Réduction du domaine des mots inconnus :
  - Remplacement des nombres, nom propres, par des balises.
    - ***10*** *tonnes de poissons* → ***<number>*** *tonnes de poissons*

Merci pour votre attention !