Adaptation au domaine et combinaison de modèles pour l'annotation de textes multi-sources et multi-domaines

Tian TIAN

sous direction de: Thierry Poibeau, Marco Dinarelli, Isabelle Tellier†

14-11-2019

Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Introduction

Thèse CIFRE en collaboration avec Synthesio (E-Réputation)
Traitement automatique des textes par apprentissage automatique :

- Reconnaissance d'entités nommées avec les Champs Conditionnel Aléatoire (CRFs)
- Etiquetage morpho-syntaxique
- Normalisation lexicale de textes de réseaux sociaux avec les réseaux de neurones

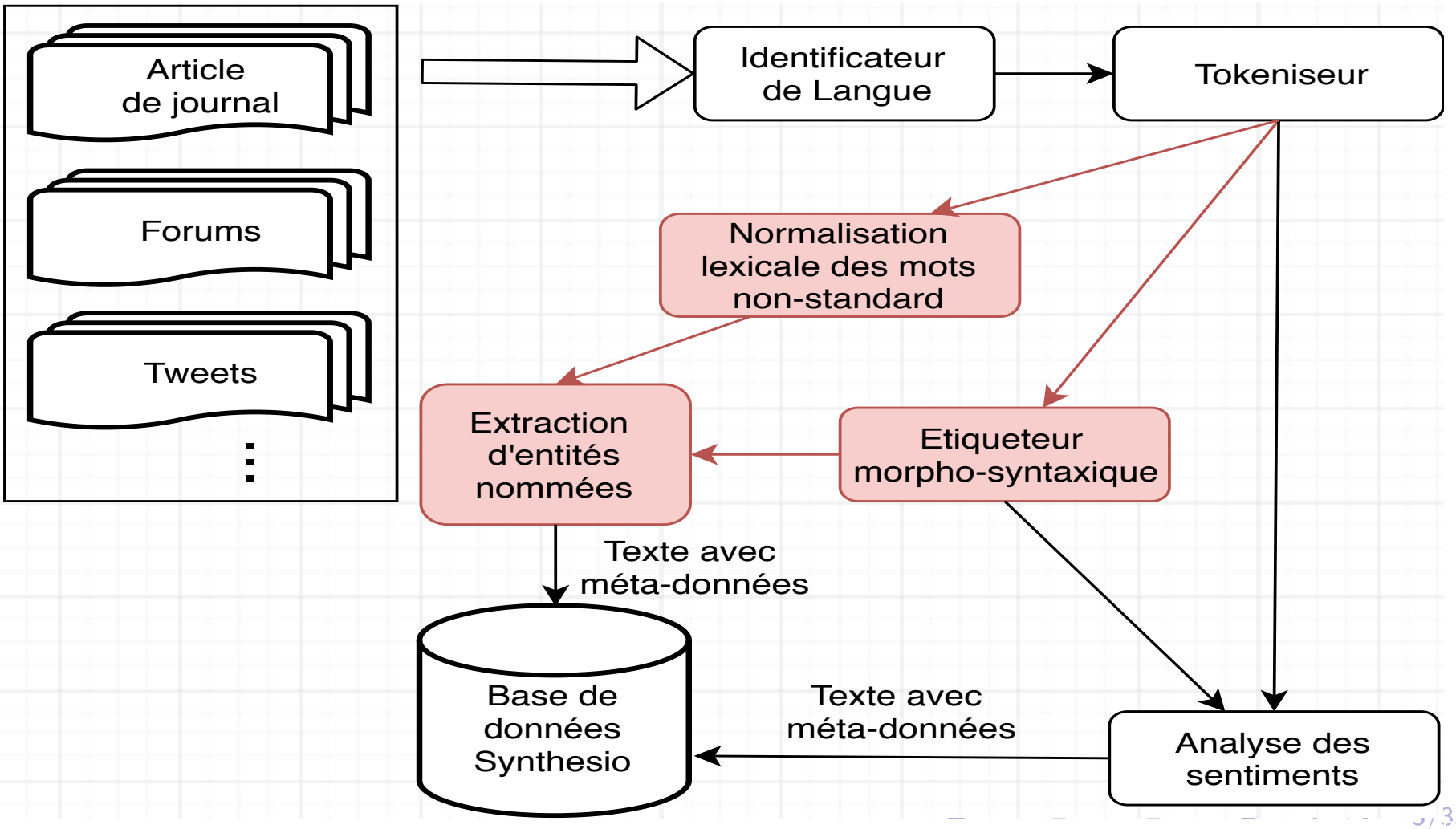
Utilisation d'étiquetage morpho-syntaxique et de normalisation lexicale pour aider à la reconnaissance d'entités nommées

Etat de l'art

Reconnaissance d'entités nommées (REN)

- Extraction d'information [Chinchor and Robinson,1998]
- Annotation séquentielle (Champs Conditionnels Aléatoires) [Tellier, Isabelle and Marc Tommasi (2011)]
- Modèles appris sur les corpus bien formés fonctionnent mal face aux données bruitées (de réseaux sociaux) [Ritter2011]
- Normalisation lexicale : détecter les mots non-standard et proposer des formes correctes par réseaux de neurones [Baldwin, Timothy et al. (2015)]

Chaîne de traitements automatiques des textes



Adaptation au domaine

- Textes bien écrits :
  articles journal, poème, reportage
- Domaine spécifique :
  texte bio-médical, bulletin de météo
- Réseaux sociaux :
  forum de discussion, tweets

Adaptation au domaine en apprentissage automatique :
apprendre un modèle avec des textes du domaine source puis adapter ce modèle pour le domaine cible par apprentissage supervisé ou semi-supervisé

Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Corpus d'évaluation Synthesio

- Multi-domaines : 5 industries de 4 domaines
- Multi-sources :
  - Textes longs de forums, Facebook, presque toujours dans un écrit correct
  - Textes courts de Twitter, avec beaucoup de mots non-standard

Textes longs	Deezer	Dunkin Donuts	Land Rover	Mattel	Nissan
#phrases	146	208	183	174	123
#tokens	2314	3116	3836	2958	2166

Textes courts	Deezer	Dunkin Donuts	Land Rover	Mattel	Nissan
#phrases	52	50	59	57	74
#tokens	854	827	1048	1123	1127

- Nombre total de phrases : 1126, nombre total de tokens : 19k

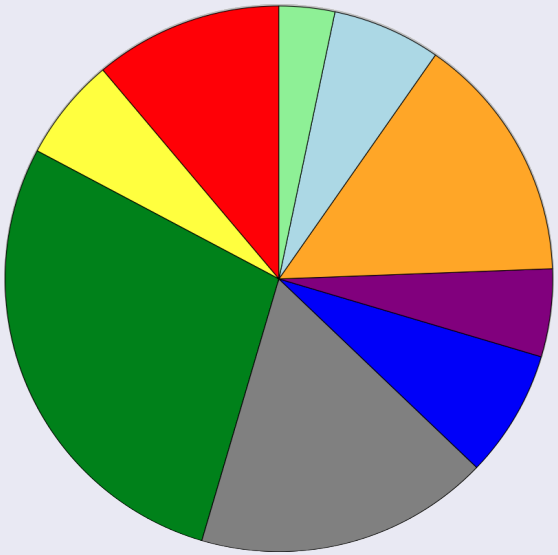
Statistiques sur les corpus

	Ritter	Synthesio	
#Phrases	2194	1126	
#Tokens	48k	19k	
#Entitiés	1672	1218	
Company	186	496	Noms d'entreprise
Product	102	484	Noms de Produits
Person	472	83	Noms des personnes
Geo-loc	291	66	Nom de lieu, pays ou ville
Media	126	41	Noms des journaux, musiciens, artistes
Job-title	87	18	Noms d'emploi comme directeur, PDG
Other	246	13	Noms de fêtes, événement
Facility	107	11	Noms des organisations
Sportsteam	55	6	Noms d'équipes de sports

Répartition des types d'entités

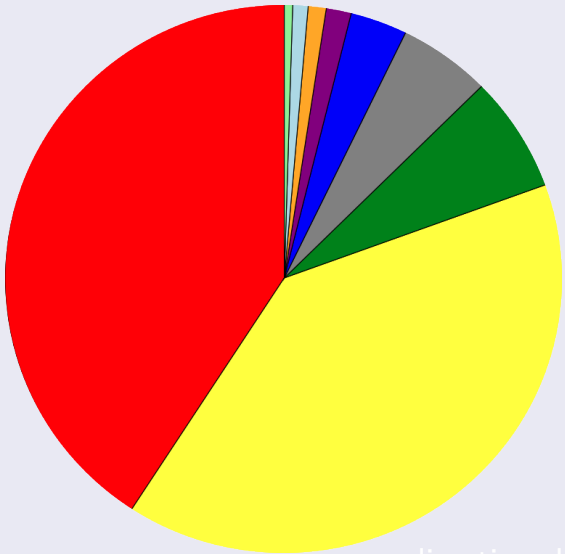
Corpus Ritter

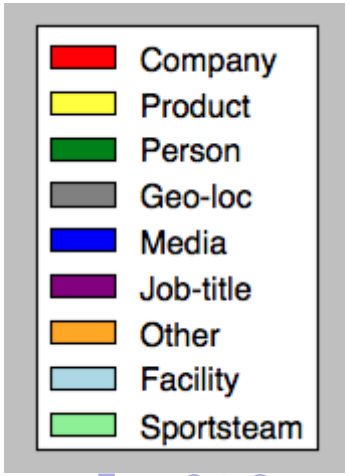
- Les types Person et Geo-loc types sont les plus fréquents
- D'autres types d'entités assez homogènes



Corpus Référence Synthesio

- Les types Company et Product types sont les plus fréquents
- D'autres types sont beaucoup moins fréquents

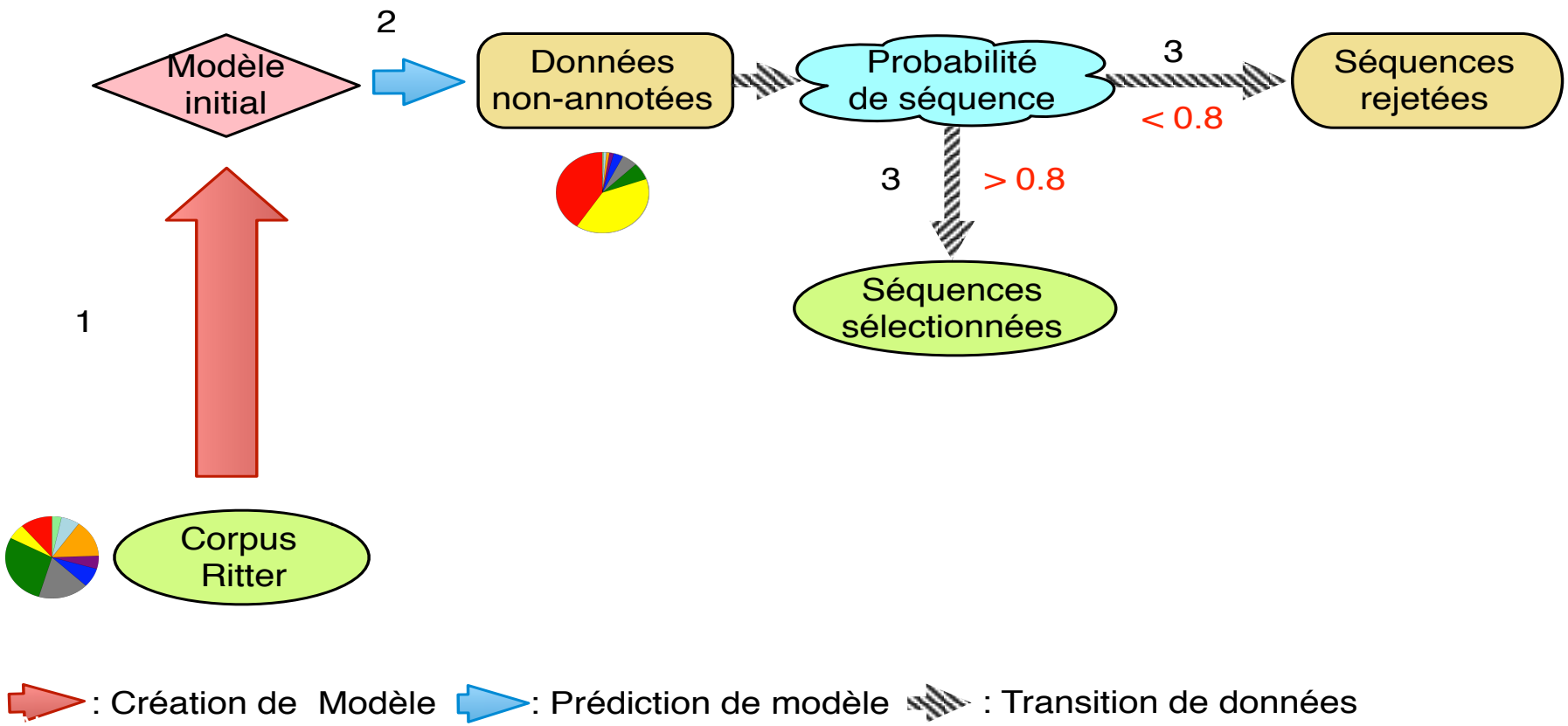




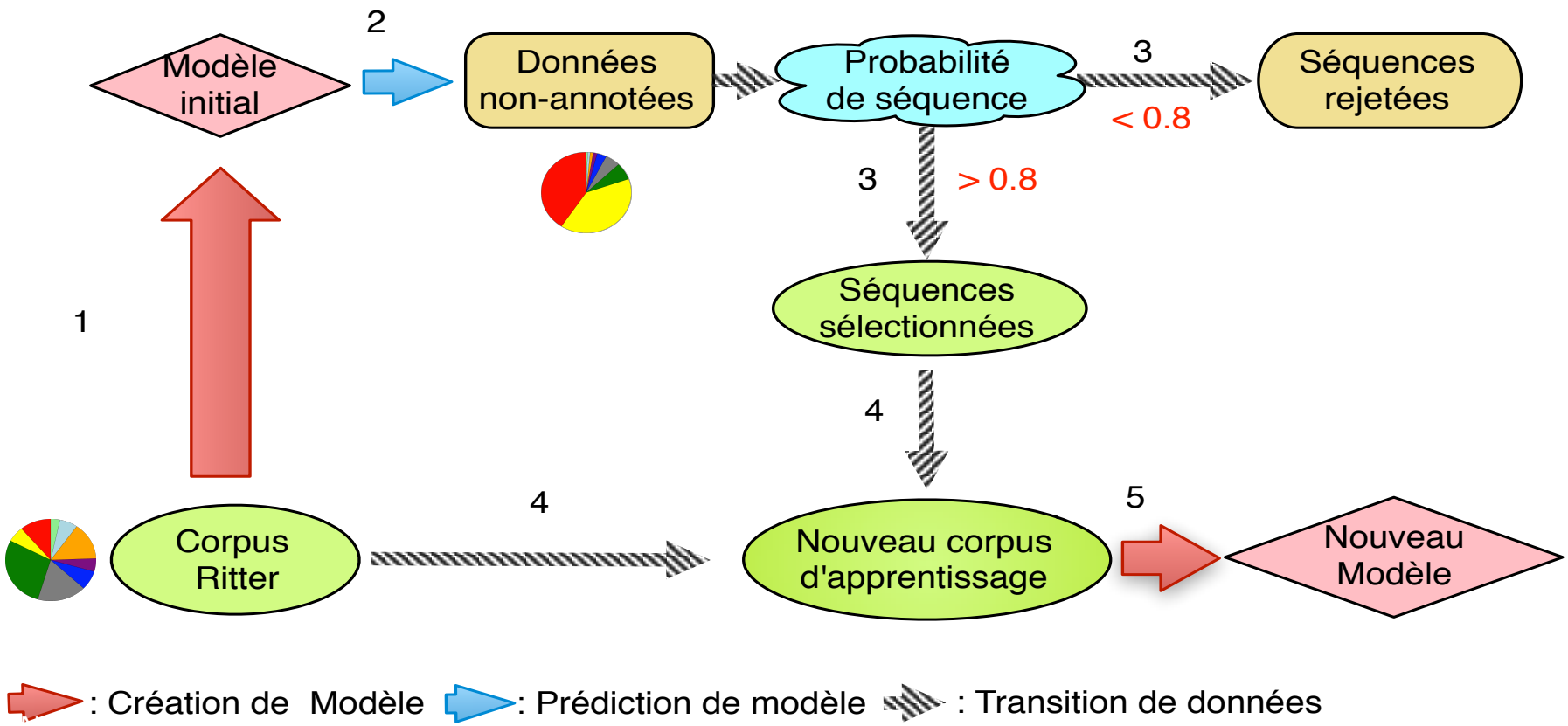
Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

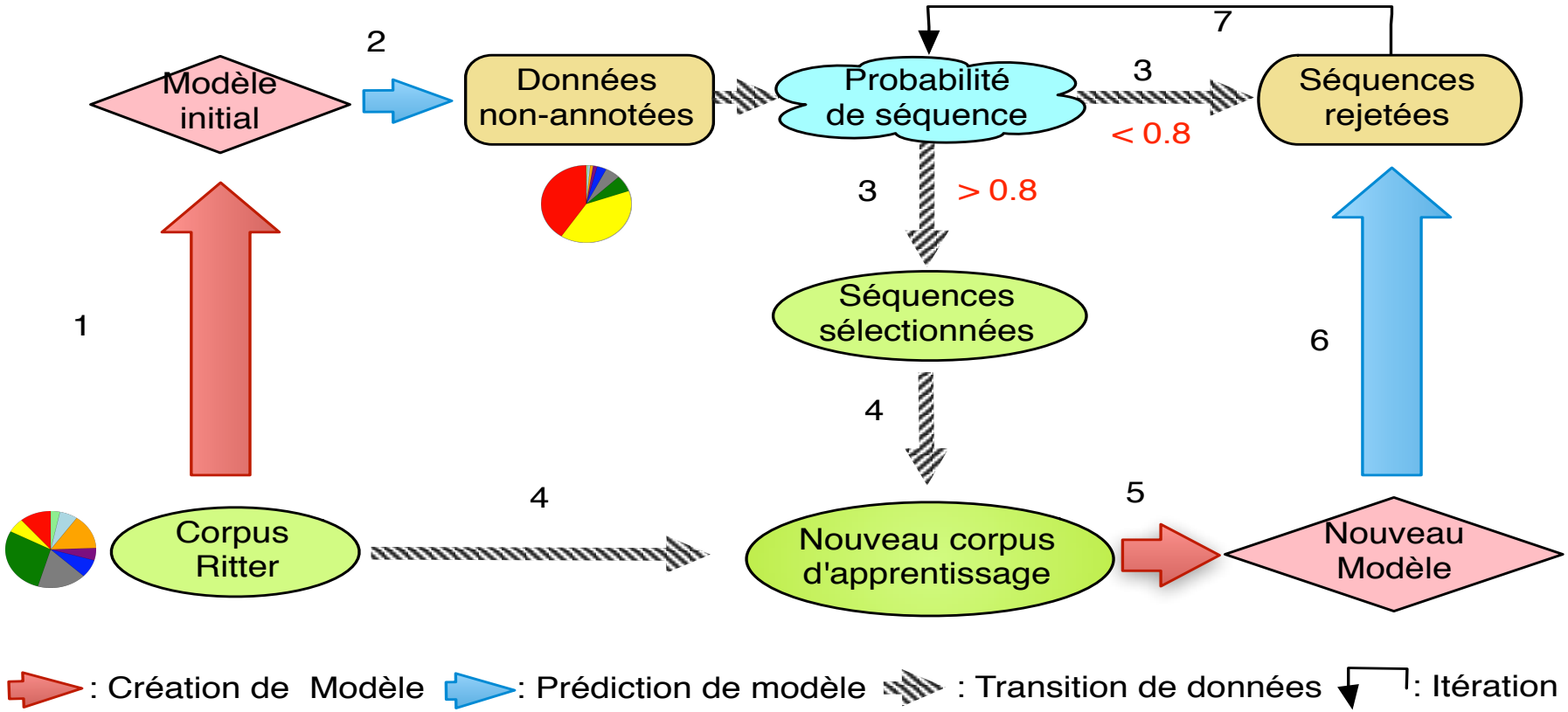
Adaptation au domaine semi-supervisé (boostrapping) avec reconnaissance d'entités nommées



Adaptation au domaine semi-supervisé (boostrapping) avec reconnaissance d'entités nommées



Adaptation au domaine semi-supervisé (boostrapping) avec reconnaissance d'entités nommées



Exemples de séquences correctes et d'erreurs de reconnaissance

- Prédiction correcte malgré l'absence de l'entité nommée dans les données d'apprentissage

  Argyle fan in **North Yorkshire**.

- Erreurs
  - Entités manquantes : Company (Nissan) et Product (Note Nismo)

    Nissan Note Nismo Coming This Fall in **Japan**.

  - Frontières : witch doctor -> **doctor**

    but people are saying that she's a witch **doctor**.

Amélioration des performances par bootstrapping

L'ajout de séquences du domaine Deezer prédites améliore la performance de reconnaissances d'entités nommées des autres domaines

Textes longs	Deezer	Dunkin Donuts	Land Rover	Mattel	Nissan
Précision	0.56/0.68	0.09/0.5	0.31/0.23	0.18/0.55	0.02/0.58
Rappel	0.14/0.13	0.06/0.24	0.05/0.08	0.07/0.28	0.01/0.09
F1-measure	0.224/0.218	0.07/0.32	0.08/0.12	0.08/0.37	0.01/0.16
Textes courts	Deezer	Dunkin Donuts	Land Rover	Mattel	Nissan
Précision	0.74/0.5	0.39/0.6	0.56/0.05	0.19/0.49	0.06/0.3
Rappel	0.09/0.08	0.08/0.23	0.05/0.06	0.02/0.19	0.02/0.09
F1-measure	0.16/0.14	0.13/0.33	0.09/0.05	0.03/0.27	0.03/0.14

Boostrapping sur d'autres domaines possible

Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Apprentissage avec propriétés réduites

- Domaine cible : séquences prédites de Synthesio
- Toutes propriétés utilisées

Token	Features					Label
	Begin with uppercase	Prefix	Suffix	Other features	Pos tag	
By	1	By	By		Prep	O
noon	0	noo	oon		Noun	O
in	0	in	in		Prep	O
California	1	Cal	nia		PropN	B-Geo-loc
,	0	,	,		Ponct	O
it	0	it	it		Pron	O
was	0	was	was		Verb	O
all	0	all	all		Adv	O
over	0	ove	ver		Adj	O
:	0	:	:		Ponct	O

- Domaine source : Ritter
- Nombre de propriétés réduites

Token	Features					Label
	Begin with uppercase	Prefix	Suffix	Other features	Pos tag	
Cant					Verb	O
wait					Verb	O
for					Prep	O
the					Det	O
ravens					PropN	B-Sportsteam
game					Noun	O
tomorrow					Noun	O
....					Ponct	O
go					Verb	O
ray					PropN	B-Person
rice					PropN	I-Person
!!!!!!!					Ponct	O

None

Résultats d'apprentissage avec propriétés réduites

Données d'apprentissage

- Domaine source : Ritter
- Domaine cible : 1608 phrases sélectionnées de textes longs de domaine Deezer

Données d'évaluation :

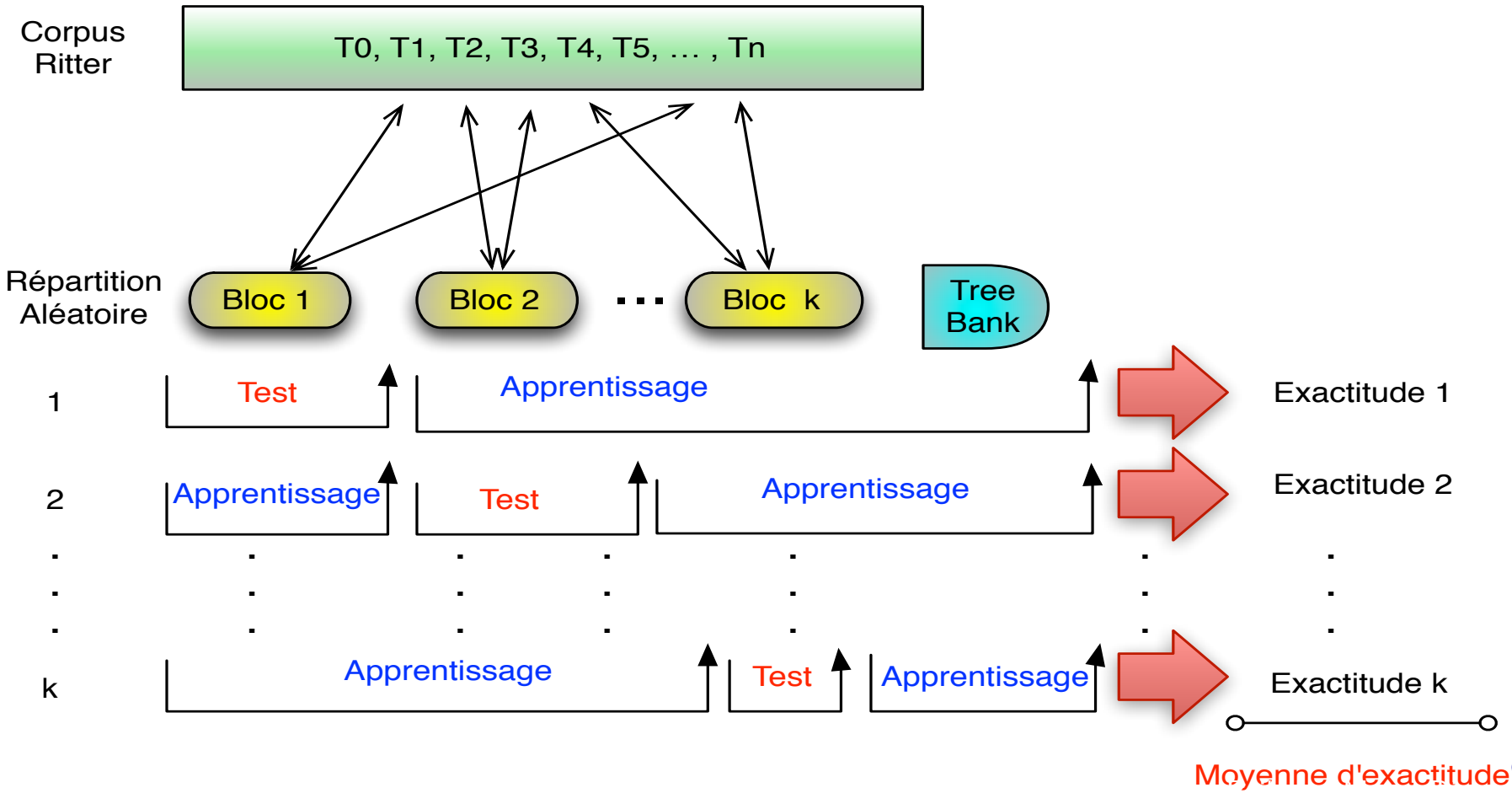
Corpus d'évaluation Synthesio : textes longs/courts de Deezer

Modèle	Deezer de Synthesio	Précision	Rappel	F1
toutes propriétés	texte long	0.83	0.08	0.15
	texte court	0.42	0.03	0.06
token1+pos5	texte long	0.63	0.06	0.11
	texte court	0.52	0.03	0.06
token3+pos3	texte long	0.67	0.06	0.11
	texte court	0.41	0.04	0.07

Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Adaptation au domaine supervisé pour l'étiquetage morpho-syntaxique



Résultats de modèles mixtes

L'ajout de données d'apprentissage permet d'améliorer la performance de l'étiquetage morpho-syntaxique

Jeu d'étiquettes	**#Séquences Penn : Ritter**	**Moyenne d'exactitude**
Ritter (45)	1 :1	85.40%
	4 :1	86.72%
	9 :1	87.18%
Universal (13)	1 :1	89.11%
	4 :1	89.27%
	9 :1	88.95%

Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Utilisation de l'étiqueteur morpho-syntaxique pour aider à la reconnaissance d'entités nommées

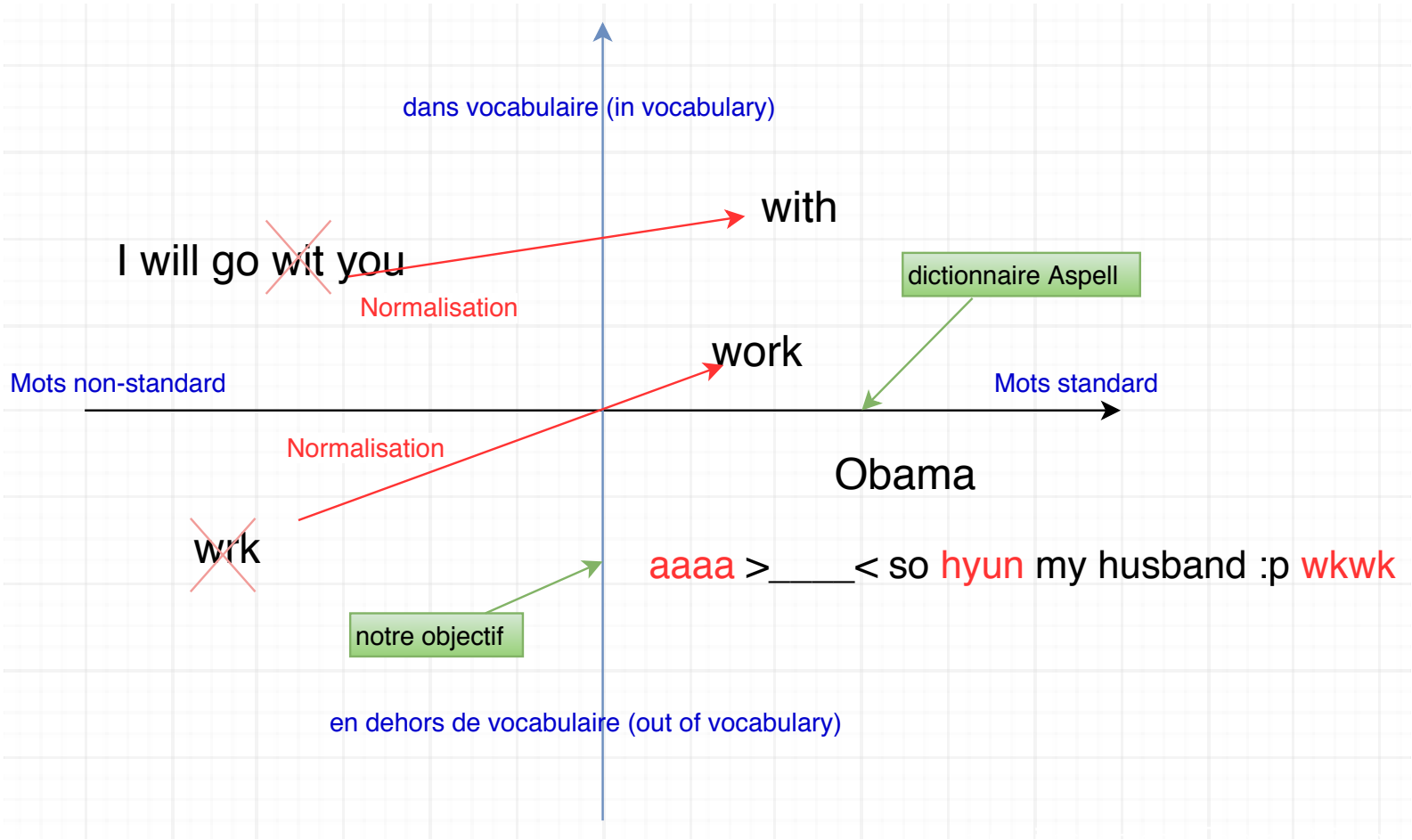
Propriétés utilisées dans modèle CRF :
Etiquettes possibles dans Penn Treebank/ étiquette prédite par l'étiqueteur morpho-syntaxique

	Précision	Rapell	F1-Mesure
Company	0.125/1.0	0.03/0.03	0.05/0.05
Person	0.36/0.35	0.29/0.41	0.32/0.38
Other	0.5/0.5	1.0/1.0	0.67/0.67
Product	0.5/0.5	0.05/0.03	0.10/0.05
Media	0/0.5	0/0.07	0/0.12
Geo-Location	0/0	0/0	0/0
Job title	0/0	0/0	0/0
Micro-Average	0.28/0.34	0.09/0.1	0.11/0.13

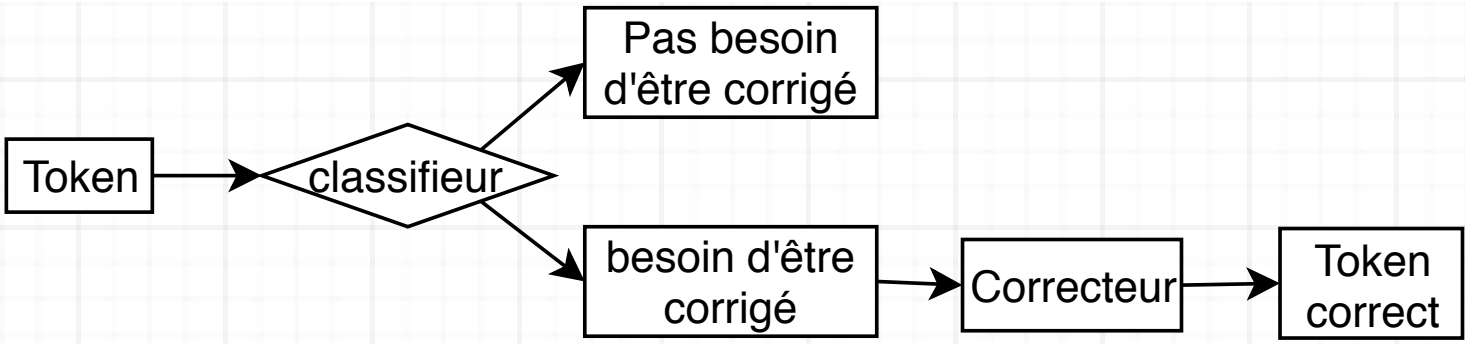
Plan

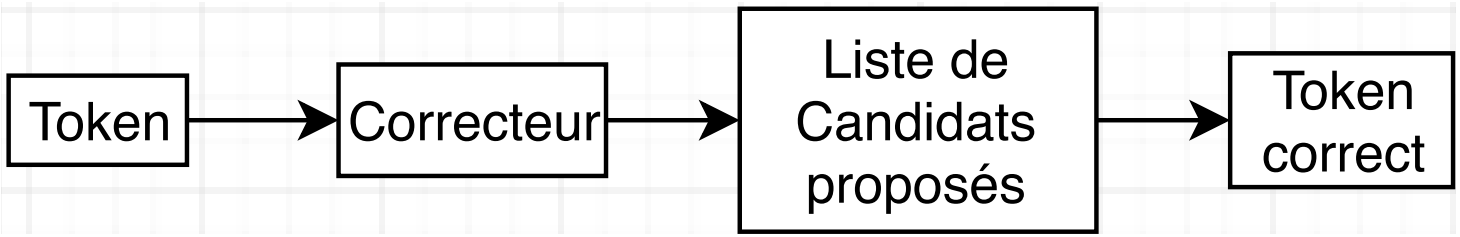
1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Normalisation lexicale - définition : mots non-standard et normalisation lexicale



Procédure générale de la normalisation lexicale





Corpus, expérience et évaluation de la détection des NSWs

nom de corpus	# tweets	# tokens	# NSW	# SW	taille de vocabulaire
WNUT dev	2950	40560	4274	36286	15108
WNUT eval	1698	29421	2776	26645	10765
2577 tweets	2577	44385	3942	40443	8270

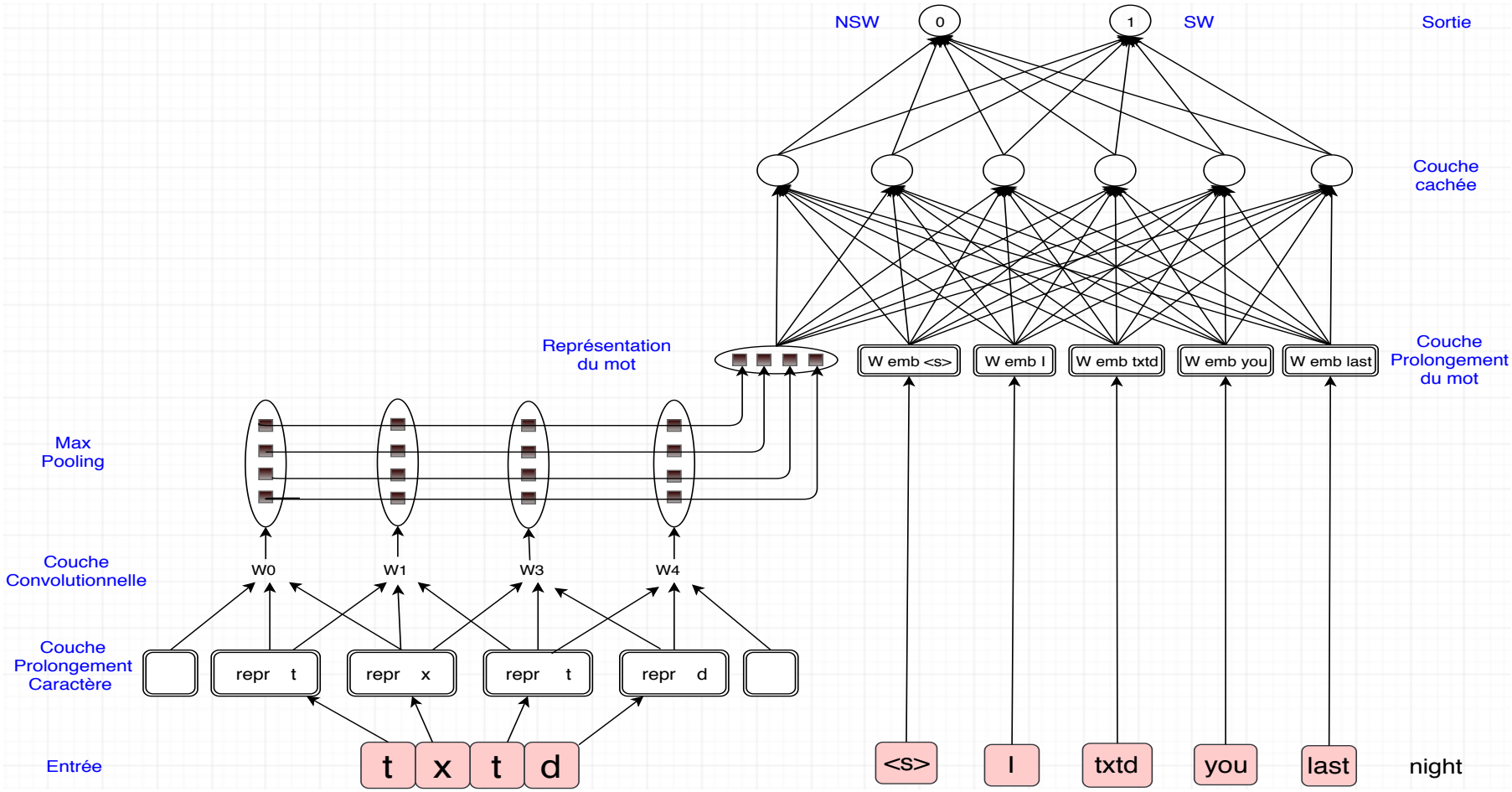
Expériences

- Validation croisée et évaluation avec corpus WNUT eval
- Avec ou sans 2577 tweets (corpus supplémentaire)

Corpus 2577 tweets : Li, Chen and Yang Liu (2014). "Improving Text Normalization via Unsupervised Model and Discriminative Reranking". In : Proceedings of the ACL 2014 Student Research Workshop. Baltimore, Maryland, USA : Association for Computational Linguistics

Corpus WNUT : Baldwin, Timothy et al. (2015). "Shared Tasks of the 2015Workshop on Noisy User-generated Text : Twitter Lexical Normalization and Named Entity Recognition". In : Proceedings of the Workshop on Noisy User-generated Text. Beijing, China : Association for Computational Linguistics

Structure du réseau de neurones



Utilisation de textes normalisés comme propriété pour la reconnaissance d'entités nommées

	Precision	Recall	F-Measure
Company	1.0	0.03/0.05	0.06/ 0.10
Person	0.35/0.38	0.41/0.52	0.38/0.44
Other	0.5	1.0/1.0	0.67/0.67
Product	0.5	0.03/0.04	0.06/0.07
Media	0.5	0.07	0.12
Geo-Location	0/0	0/0	0/0
Job-title	0/0	0/0	0/0
Micro-Average	0.45/0.49	0.07/0.10	0.12/0.17

La normalisation lexicale des corpus d'apprentissage et d'évaluation permet d'améliorer la performance de la reconnaissance d'entités nommées (pour certains types d'entités)

Plan

1. Reconnaissance d'entités nommées (REN)
   - Etat de l'art
   - Chaîne de traitements automatiques des textes
   - Adaptation au domaine (AD)
2. Corpus d'évaluation
3. Adaptation au domaine semi-supervisé pour REN
   - Adaptation au domaine par apprentissage Boostrapping
   - Apprentissage supervisé avec propriétés réduites
4. Adaptation au domaine supervisé pour étiquetage morpho-syntaxique
5. Pour aider à la reconnaissance d'entités nommées
   - Utilisation de l'étiqueteur morpho-syntaxique
   - Utilisation de normalisation lexicale
6. Conclusion et perspectives

Conclusion et perspectives

Conclusion

- L'étiqueteur morpho-syntaxique et la normalisation lexicale permet de corriger les textes bruités et d'améliorer légèrement le résultat de la reconnaissance d'entités nommées
- le Boostrapping est une technique valable pour l'adaptation au domaine

Perspectives

- Vers un corpus d'évaluation mieux adapté à toutes les trois tâches : étiquetage morpho-syntaxique, reconnaissance d'entités nommées et normalisation lexicale
- Vers une modélisation de mots par prononciation pour la normalisation lexicale

Merci de votre attention

Bibliographie

- Chinchor, N. and P. Robinson (1998). "Appendix E : MUC-7 Named Entity Task Definition (version 3.5)". In MUC-7, Fairfax, Virginia,1998
- Baldwin, Timothy et al. (2015). "Shared Tasks of the 2015Workshop on Noisy User-generated Text : Twitter Lexical Normalization and Named Entity Recognition". In : Proceedings of the Workshop on WNUT. Beijing, China
- Li, Chen and Yang Liu(2015). "Improving Named Entity Recognition in Tweets via Detecting Non-StandardWords". In : Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing. Beijing, China
- Ritter, Alan et al. (2011). "Named Entity Recognition in Tweets : An Experimental Study". In EMNLP '11. Edinburgh, United Kingdom
- Putthividhya, Duangmanee (Pew) and Junling Hu (2011). "Bootstrapped named entity recognition for product attribute extraction". In EMNLP '11. Edinburgh, United Kingdom
- Tellier, Isabelle and Marc Tommasi (2011). "Champs Markoviens Conditionnels pour l'extraction d'information". In : Modèles probabilistes pour l'accès à l'information textuelle.

Publications

Communications avec actes dans un congrès national

- Tian, Dinarelli, Cardoso, Tellier : Détection des mots non-standards dans les tweets avec des réseaux de neurones, Traitement Automatique des Langues Naturelles (TALN 2017, papier court), Orléans.
- Tian, Dinarelli, Tellier, Cardoso 2015 : Etiquetage morpho-syntaxique de tweets avec des CRF, Traitement Automatique des Langues Naturelles (TALN 2015, papier court), Caen.
- Marty, Tian, Tellier 2014 : Extraction de propriétés de produits, Conférence en Recherche d'Information et Applications (CoRIA 2014), Nancy.

Communications dans un congrès international sans actes

- Tian, Tellier, Dinarelli 2016 : Understanding Social Media Texts with Minimum Human Effort on #Twitter, PLIN Linguistic Day, Louvain-la-Neuve (Belgium).

Communications (orale ou par poster) avec actes dans un congrès international

- Tian, Dinarelli, Tellier, Cardoso 2016 : Domain Adaptation for Named Entity Recognition Using CRFs, LREC, Portoroz (Slovenia).
- Tian, Dinarelli, Tellier 2015 : Lattice : Data Adaptation for Named Entity Recognition on Tweets with Features-Rich CRF, Shared task on the 2015 Workshop on Noisy User-generated Text : Twitter Lexical Normalization and Named Entity Recognition, ACL Workshop, Beijing (China), 2015.