UNIVERSITÉ SORBONNE NOUVELLE PARIS 3

Membre de l'Université Sorbonne Paris Cité

Bien choisir ses données d'apprentissage pour le TAL en contexte multi-hétérogène : exemple de l'ancien français

Isabelle Tellier

Lattice, Paris

Introduction

- Tâches traditionnelles du **TAL** : annotation morpho-syntaxique (POS), analyse syntaxique
- Il existe des techniques d'**apprentissage automatique supervisé** efficaces (CRF, Mate, Réseaux Neuronaux…) pour les réaliser…
- … si on dispose de (bonnes et nombreuses) **données annotées** !
- Les modèles appris ont des **résultats dégradés** quand on les applique sur des données différentes de celles ayant permis l'entraînement :
  - Variabilité de **modalité** (transcription de l'oral/écrit)
  - Variabilité de **genre** (articles, blogs, forums, Tweets/SMS…)
  - Variabilité de **domaine** (politique, sport, culture, sciences…)
  - Autres **variabilités** (dialectes…)

Introduction

- **Solutions** habituelles
  - **Ré-annoter** de nouvelles données, en corrigeant une annotation proposée ou en les choisissant par **apprentissage actif**
  - Utiliser des techniques de **correction des données non standards** (normalisation des Tweets…)
  - Utiliser des techniques d'**adaptation de domaine**
  - Utiliser des techniques de **transfert**
  - Combiner apprentissage supervisé et **non supervisé**

- Proposition explorée ici : **la sélection des données d'apprentissage**
- Cas d'étude **: corpus multi-hétérogène de l'ancien français**

Plan

- Introduction
- **SRCMF : un corpus multi-hétérogène**
- Premières expériences
- Perspective : apprentissage « sur mesure »
- Conclusion

SRCMF : un corpus multi-hétérogène

- **SRCMF** : Syntactic Reference Corpus of Medieval French
  - projet ANR franco-allemand (A. Stein, S. Prévost) : 2009-12
  - 15 textes, 245 00 mots en tout
  - Annotés manuellement en POS (60 étiquettes) et en dépendances syntaxiques fines
  - Lemmes (non validés) issus de TreeTagger entraîné sur d'autres textes

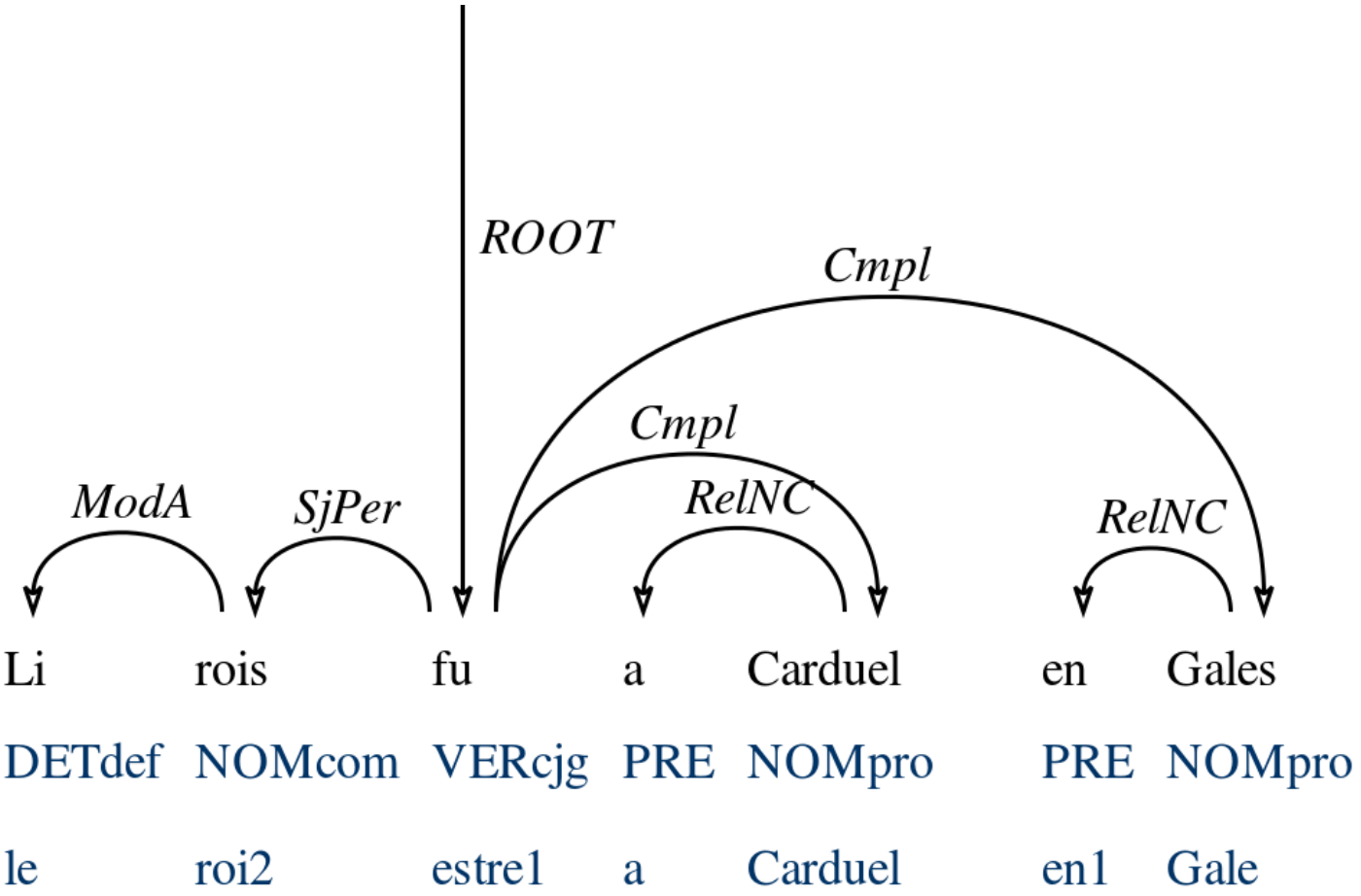
- **Métadonnées (variabilités externes) :**
  - Date d'écriture
  - Forme : vers/prose (les premiers textes sont tous en vers)
  - Dialectes : normand, anglo-normand, champenois, picard…
  - Domaine : religieux, littéraire, historique, didactique…

SRCMF : un corpus multi-hétérogène

- **Quelques propriétés de l'ancien français (variabilités internes) :**
  - Pas de norme orthographique sur l'écriture des mots : 17 formes recensées distinctes (« je », « gié », « jou », « gel »…) pour « je »
  - Même les noms propres sont variables : Yvain(s), Yvein(s)
  - Nombreuses formes contractées (je + le > jel…)
  - Nombreux syntagmes discontinus (« puis…que » pour « depuis que »)
  - Restes de déclinaisons latines (en voie de disparition)
  - Sujets peuvent être nuls
  - Ordre des mots beaucoup plus libre que dans le français moderne

- 9ème-12ème siècle : **émergence du français**, une *autre* période de grande variabilité de la langue !

SRCMF : un corpus multi-hétérogène



SRCMF : un corpus multi-hétérogène (extrait utilisé)

Text	Date	Words	Form	Dialect	Domain
Vie Saint ***Légier***	late 10c.	1388	verse	n/a	religious
Vie de Saint ***Alexis***	1050	4804	verse	normand	religious
Chanson de ***Roland***	1100	28 766	verse	normand	literary
Lapidaire *en prose*	Mid. 12c.	4708	prose	anglo-norm.	didactical
Yvain, Chr. de Troyes	1177-1181	41 305	verse	champenois	literary
La ***Conqueste*** *de Constantinople*, R. de Clari	>1205	33 534	prose	picard	historical
Queste del Saint ***Graal***	1220	40 417	prose	n/a	literary
Aucassin *et Nicolete*	late 12c.-early 13c.	9844	verse & prose	picard	literary
Miracles from Gautier de **Coinci**	1218-1227	17 360	verse	picard	religious
Roman de la ***Rose*** from Jean de Meun	1269-1278	19 339	verse	n/a	didactical

Plan

- Introduction
- SRCMF : un corpus multi-hétérogène
- **Premières expériences**
- Perspective : apprentissage « sur mesure »
- Conclusion

Premières expériences (réalisées par G. Guibon, TALN 2015, TLT2015)

- **Buts**
  - Explorer l'impact des métadonnées sources de variabilités externes (date, forme, dialecte, domaine) sur les modèles appris

- **Protocole pour chaque métadonnée (ex : forme)**
  - Pour chaque valeur de la métadonnée (ex : vers/prose), répartir les données équitablement en entraînement/test
  - Entraînement d'un CRF (pour les POS) et de Mate (pour le parser) sur les données d'entraînement pour chaque valeur de la métadonnée (ex : pour vers et pour prose)
  - Tests pour toutes les valeurs de la métadonnée (ex : vers/prose)
  - « Matrice d'évaluation » sur les valeurs des métadonnées

Premières expériences

Train \ Test		Prose [test]	Verse [test]
Prose	UAS	**85.47**	76.33
	LAS	**74.96**	62.96
	ACC	**91.36**	83.61
	Unknown | known words	16.49 | **83.51**	**21.26** | 78.74
	Different | shared lexicon	57.02 | **42.98**	**77.05** | 22.95
	Unknown | known words UAS	**73.76** | **87.78**	65.87 | 79.15
	Unknown | known words LAS	**55.48** | **78.81**	46.37 | 67.44
	Unknown | known words ACC	**77.33** | **94.14**	76.78 | 85.46
Verse	UAS	**83.12**	82.79
	LAS	**71.52**	71.40
	ACC	90.06	**90.78**
	Unknown | known words	**18.81** | 81.19	14.03 | **85.97**
	Different | shared lexicon	**66,47** | 33,53	42.52 | **57.48**
	Unknown | known words UAS	**73.43** | **85.37**	72.39 | 84.49
	Unknown | known words LAS	55.45 | **75.24**	**55.62** | 73.98
	Unknown | known words ACC	81.02 | **92.15**	**84.13** | 91.86

Premières expériences

Train \ Test		12th century [test]	13th century [test]
12th century	UAS	**88.81**	83.14
	LAS	**79.91**	71.93
	ACC	**94.69**	89.62
	Unknown | known words	**91.39** | 08.61	78.72 | **21.28**
	Different | shared lexicon	**61.20** | 38.80	28.59 | **71.41**
	Unknown | known words UAS	**81.05** | **90.00**	71.60 | 85.13
	Unknown | known words LAS	**66.42** | **81.47**	54.18 | 75.72
	Unknown | known words ACC	**87.29** | **95.39**	78.14 | 92.73
13th century	UAS	82.24	**89.07**
	LAS	69.24	**80.75**
	ACC	88.67	**94.62**
	Unkonwn | known words	73.83 | **26.17**	**92.25** | 07.75
	Different | shared lexicon	33.96 | **66.04**	**50.12** | 49.88
	Unkonwn | known words UAS	**76.94** | 86.61	74.35 | **88.77**
	Unkonwn | known words LAS	56.75 | 75.84	**57.96** | **80.46**
	Unkonwn | known words ACC	80.13 | 91.69	**85.31** | **95.41**

Premières expériences

Train \ Test	**Champenois[test]**	**Normand[test]**	**Picard[test]**
Champenois [train]			
UAS	**86.07**	78.61	76.66
LAS	**76.30**	61.93	63.63
ACC	**93.41**	81.17	84.02
Unknown/known words	10.23 | **89.77**	**51.05** | 48.95	31.20 | 68.80
Different/shared lexicon	51.09 | **48.91**	**82.09** | 17.91	79.56 | 20.44
Unknown/known words UAS	**73.83** | **87.46**	72.83 | 84.63	66.38 | 81.32
Unknown/known words LAS	**59.14** | **78.25**	51.34 | 72.98	46.29 | 71.49
Unknown/known words ACC	**84.57** | **94.41**	72.59 | 90.12	67.99 | 91.30
Normand [train]			
UAS	74.54	**88**	73.77
LAS	59.31	**77.96**	60.48
ACC	81.12	**93.31**	82.55
Unknown/known words	34.14 | 65.86	11.25 | **88.75**	**38.77** | 61.23
Different/shared lexicon	82.24 | 17.76	43.90 | **56.10**	**87.05** | 12.95
Unknown/known words UAS	64.37 | 79.81	**78.53** | **89.20**	64.19 | 79.84
Unknown/known words LAS	45.30 | 66.58	**60.21** | **80.21**	46.86 | 69.11
Unknown/known words ACC	72.54 | 85.57	**82.01** | **94.74**	72.50 | 88.92
Picard [train]			
UAS	77.35	79.41	**85.14**
LAS	63.46	63.20	**75.90**
ACC	84.40	82.11	**93.25**
Unknown/known words	24.58 | 75.42	**45.57** | 54.23	11.16 | **88.84**
Different/shared lexicon	74.51 | 25.49	**82.42** | 17.58	60.03 | **39.97**
Unknown/known words UAS	66.15 | 81.00	**72.60** | 85.24	71.49 | **86.86**
Unknown/known words LAS	47.03 | 68.81	51.47 | 72.98	**55.29** | **78.49**
Unknown/known words ACC	75.34 | 87.34	72.93 | 89.78	**80.56** | **94.85**

Premières expériences

Train \ Test	Didactical[test]	Historical[test]	Literary[test]	Religious[test]
Didactical [train]				
UAS	**81.78**	78.88	80.11	70.05
LAS	**71.23**	67.28	66.67	55.04
ACC	**90.75**	87.58	87.08	80.80
Unknown | known words	16.53 | **83.47**	**31.15** | 68.85	26.08 | 73.92	30.58 | 69.42
Different | shared lexicon	50.19 | **49.81**	78.05 | 21.95	**83.85** | 16.15	69.67 | 30.33
Unknown | known w. UAS	**71.68** | **83.78**	69.12 | 83.29	70.13 | 83.63	59.80 | 74.57
Unknown | known w. LAS	**53.93** | **74.66**	52.89 | 73.79	50.69 | 72.29	38.46 | 62.34
Unknown | known w. ACC	80.89 | **92.70**	**81.30** | 90.43	77.53 | 90.44	66.96 | 86.89
Historical [train]				
UAS	67.49	**90.07**	73.03	32.29
LAS	51.12	**82.20**	57.30	45.08
ACC	72.74	**95.66**	76.67	69.93
Unknown | known w.	41.09 | 58.91	08.08 | **91.92**	38.66 | 61.34	**42.57** | 57.43
Different | shared lexicon	81.94 | 18.06	46.67 | **53.33**	**90.46** | 09.54	79.84 | 20.16
Unknown | known w. UAS	58.08 | 74.05	**80.16** | **90.94**	65.06 | 78.05	52.80 | 69.33
Unknown | known w. LAS	38.24 | 60.11	**63.70** | **83.92**	45.20 | 64.93	31.56 | 55.10
Unknown | known w. ACC	62.67 | 79.77	**87.50** | **96.38**	66.95 | 82.80	57.20 | 79.38
Literary [train]				
UAS	77.22	82.02	**84.79**	73.09
LAS	64.07	70.79	**73.63**	59.01
ACC	85.10	88.95	**91.93**	83.25
Unknown | known w.	27.01 | 72.99	**27.35** | 72.65	14.42 | **85.58**	27.16 | 72.84
Different | shared lexicon	68.17 | 31.83	73.58 | 26.42	**75.36** | 24.64	65.96 | **34.04**
Unknown | known w. UAS	66.17 | 81.31	74.07 | 85.02	**74.28** | **86.56**	61.18 | 77.53
Unknown | known w. LAS	46.03 | 70.74	**57.21** | 75.90	56.25 | **76.55**	40.67 | 65.84
Unknown | known w. ACC	73.61 | 89.35	80.72 | 92.04	**82.50** | **93.51**	69.04 | 88.55
Religious [train]				
UAS	74.99	79.76	79.52	**80.72**
LAS	61.61	67.94	65.94	**69.35**
ACC	83.31	87.62	85.91	**90.16**
Unknown | known w.	29.01 | 70.99	**29.50** | 70.50	26.58 | 73.42	14.07 | **85.93**
Different | shared lexicon	71.66 | 28.34	76.56 | 23.44	**85.05** | 14.95	43.47 | **56.53**
Unknown | known w. UAS	63.98 | 79.48	**70.17** | **83.77**	69.16 | 83.28	68.61 | 82.70
Unknown | known w. LAS	44.10 | 68.76	**53.31** | **74.06**	49.00 | 72.07	49.58 | 72.59
Unknown | known w. ACC	71.73 | 88.05	**81.80** | 90.06	75.87 | 89.55	75.87 | **92.50**

Premières expériences

- **Intérêt**
  - Etant donné un nouveau texte connu par les valeurs de ses métadonnées (ex : **13ème siècle**, **en vers**, **picard**, **littéraire**), quelles données choisir pour apprendre un étiqueteur/parser adapté ?
  - Date = **13ème** (meilleure LAS=80.75 entraîné sur 13ème siècle)
  - Forme = **vers** (meilleure LAS=73,63 entraîné sur des vers)
  - Dialecte = **picard** (meilleure LAS=75,9 entraîné sur du picard)
  - Domaine = **littéraire** (meilleure LAS=73,63 entraîné sur textes littéraires)
  - Dans ce cas, la **date** semble le critère le plus **discriminant**

Plan

- Introduction
- SRCMF : un corpus multi-hétérogène
- Premières expériences
- **Perspective : apprentissage « sur mesure »**
- Conclusion

Perspective : apprentissage « sur mesure »

- **Contexte**
  - Soit des données annotées (rares et précieuses) multi-hétérogènes
  - Dont la variabilité est qualifiée par différentes valeurs de métadonnées

- **Problème de l'apprentissage « sur mesure »**
  - Étant donné un nouveau texte caractérisé par ses métadonnées
  - Comment choisir le corpus d'entraînement garantissant **le meilleur modèle possible** pour traiter ce nouveau texte ?

Perspective : apprentissage « sur mesure »

- **Caractéristiques nouvelles du problème**
  - La procédure de décision recherchée s'appliquera sur les données pouvant servir d'entraînement

- **Difficultés**
  - Complexité combinatoire de tous les sous-ensembles possibles
  - Les données disponibles ne sont pas équitablement réparties dans les différentes valeurs des métadonnées
  - Les valeurs de métadonnées ne sont pas indépendantes les unes des autres (ex : les textes les plus anciens sont tous en vers)
  - Autres effets : variabilité interne du texte à traiter, mots connus/ inconnus…

Perspective : apprentissage « sur mesure »

- **Questions ouvertes**
  - Vaut-il mieux utiliser beaucoup de données peu adaptées ou peu très adaptées ?
  - Quelle évaluation prendre en compte sur les valeurs de métadonnées (matrices d'évaluation) ?
  - Quelle forme pour la procédure de décision (doit être applicable à des valeurs nominales de métadonnées) : un arbre de décision ?

Plan

- Introduction
- SRCMF : un corpus multi-hétérogène
- Premières expériences
- Perspective : apprentissage « sur mesure »
- **Conclusion**

Conclusion

- **Synthèse**
  - En apprentissage automatique supervisé, toute l'information repose dans les **données annotées**
  - En contexte de **grande variabilité**, nécessité de modèles **ciblés pour un texte nouveau**
  - toutes les données ne sont pas nécessairement adaptées pour servir d'**entraînement** à un tel modèle ciblé
  - La procédure de décision doit se reporter sur **le choix des données**

Conclusion

- **Perspectives**
  - Projet ANR **PROFITEROLE** (Processing Old French Instrumented Texts for the Representation of Language Evolution) en cours
  - **thèse** à commencer sur ce sujet en septembre (appel à candidature imminent)
  - Evidemment, ce qui vaut pour l'**ancien français** devrait aussi valoir ailleurs…