Attribution d’auteur par traits linguistiques variés

Ludovic Tanguy et aussi :
Basilio Calderone, Nabil Hathout, Franck Sajous, Assaf Urieli

« Thématiques actuelles de la recherche en TAL » – Octobre 2012

PLAN

- L’attribution d’auteurs en quelques mots
  - Principe général
  - Exemples concrets
  - Méthodes
- Les compétitions PAN 2011 & 2012
  - Principes et tâches
  - 2011 : courriers électroniques d’Enron
  - 2012 : fiction américaine contemporaine
- Méthode
  - Traits linguistiques « riches »
  - Techniques d’apprentissage utilisées
  - Résultats obtenus
- Au-delà de la tâche
  - Examen des modèles
  - Pistes

L’Attribution d’Auteur (AA) : petit résumé

- Définition :
  - Etant donné un texte, quel auteur est le plus susceptible de l'avoir écrit ?
    - *En considérant une liste finie d'auteurs potentiels pour lesquels on dispose d'écrits attestés*
  - Variation : classe fermée ou ouverte
    - *Envisage-t-on que le texte soit écrit par un auteur inconnu (absent de la liste de référence) ?*
  - Tâche connexe 1 : Vérification d'auteur
    - *Etant donné un texte attribué à un auteur, et étant donné un ensemble de textes de cet auteur, confirme-t-on cette paternité?*
  - Tâche connexe 2 : Démêlage d'auteurs
    - *Etant donnés plusieurs textes, sont-ils écrits par un seul et même auteur ?*

- Exemples de cas célèbres
  - *Federalist papers*
    - *85 essais datés de 1787-88, publiés pour promouvoir la ratification de la constitution des Etats-Unis.*
    - *Publication anonyme, mais 3 auteurs au total*
    - *Nombreux débats sur les auteurs d'une douzaine de textes*
  - Molière versus Corneille
    - *P. Louÿs (1919) propose d'attribuer plusieurs pièces de Molière à Corneille, mettant en avant la versification et les choix de vocabulaire*
    - *Selon Dominique Labbé (2001), 16 des comédies les plus connues de Molière sont à attribuer à Corneille (distance intertextuelle basée sur le vocabulaire commun)*

- Exemples génériques
  - Linguistique légale
    - *Utilisation d'expertises linguistiques dans le cadre de procédures judiciaires (lettres de menace, de suicide, faux documents, etc.)*
      - Confirmation/infirmation de l'auteur
      - Profilage d'un auteur
      - Identification d'un locuteur
  - Plagiat
    - *Identification des passages d'un texte directement ou indirectement basés sur ceux d'un autre, sans indication*
    - *Détection de plagiat « interne »*
      - Repérage des changements notables des caractéristiques au fil du texte (sans faire appel à une collection de sources externes)

- Principes généraux :
  - Mesures basées sur des caractéristiques des textes, supposées stables à travers les écrits d'un même auteur (« style »)
  - Classiquement :
    - *Eviter les marques trop liées au contenu thématique*
    - *Privilégier les caractéristiques syntaxiques, ponctuationnelles, rythmiques, etc.*

- Une tâche difficile pour les humains
  - Une exception en TAL : les méthodes automatiques sont plus efficaces que les analyses manuelles
  - Petit jeu en interne à l'ERSS
    - *Deux juges (F et N)*
    - *Cible : des mails en anglais*
  - Règles du jeu :
    - *Une collection de 2000 mails écrits par 2 auteurs différents*
    - *Plusieurs heures d'observation des données (pour identifier des caractéristiques discriminantes)*
    - *100 mails choisis aléatoirement parmi la collection, pas de limite de temps*
  - Résultats :
    - *N = 87%, F=93%, Machine = 97%*

- Principales méthodes employées
  - Mesures de distance entre textes
    - *Pour un texte inconnu, de quel(s) texte(s) connu(s) est-il le plus proche ?*
  - Techniques par apprentissage automatique
    - *Entraînement d'un modèle à partir des données connues*
    - *Application à de nouvelles données pour prendre une décision*
- Utilisation de *descripteurs*
  - Variables calculées pour chaque texte
    - *Pour les données d'entraînement et celles à tester*
  - Représentation vectorielle des textes

- Deux grands types de descripteurs :
  - Descripteurs pauvres, facilement obtenus, généralement très nombreux
    - *N-grammes de caractères*
      - Fréquence des séquences de (n=3) caractères trouvées dans un texte
      - LE type de descripteur le plus utilisé dans les approches de classification en TAL
    - *Vocabulaire (fréquence des mots de chaque texte)*
      - Surtout les mots-outils (prépositions, pronoms, etc.)
    - *Mesures diverses*
      - Signes de ponctuation, longueur moyenne des unités, etc.

- Descripteurs riches, résultant de la projection de connaissances linguistiques sur les textes
  - Fréquences d'unités plus complexes
    - *Catégories morphosyntaxiques, structures syntaxiques, suffixes, classes lexicales, etc.*
  - Autres caractéristiques plus synthétiques
    - *Taux d'erreur d'orthographe, complexité syntaxique, cohésion lexicale, organisation du texte, etc.*

- Un débat en cours
  - Les traits pauvres sont majoritaires, avec des approches complexes en apprentissage
  - Les traits linguistiques sont peu ou mal utilisés, et considérés trop coûteux

Exemple de traits calculés

- « *that's what i was afraid of. i'll try again and resend. thanks a bunch.* »
  - Trigrammes de caractères :
    - *tha->2, hat -> 2, at' -> 1, t's -> 1, etc.*
  - Trigrammes de tags :
    - *WDT_VBZ_WP -> 1, VBZ_WP_NNS -> 1, etc.*
  - Longueurs :
    - *3 phrases, longueur moyenne 5,6 mots, etc.*
  - Divers :
    - *Pas de faute d'orthographe, pas de majuscules, formule de politesse = « thanks a bunch »*
    - *Contractions : 2 cas sur 2*

Les compétitions PAN

« *Uncovering Plagiarism, Authorship and Social Software Misuse* »

- Principes d'une compétition (ou tâche partagée) en TAL
  - Une tâche clairement définie
  - Un jeu de données d'*entraînement* distribué en avance aux candidats (avec les réponses)
  - Au jour J, un jeu de *test* distribué (similaire au jeu d'entraînement, mais sans les réponses)
  - Au jour J+n (n étant trop petit), chaque candidat renvoie ses réponses (« *run* »)
  - Calcul des scores (diverses mesures) et classement des runs
- Exemples de compétitions
  - Recherche d'information (TREC, CLEF, INEX, etc.)
  - Extraction d'information (MUC)
  - Etiquetage de textes (GRACE, EASY, PASSAGE, etc.)
  - Traduction, résumé, traitement morphologique, détection d'émotions, etc.
- Avantages
  - Motivation des équipes, factorisation des efforts de collecte et de préparation des données, évaluation commune
- Problèmes
  - Course au résultat plus qu'à l'innovation, données et méthode d'évaluation parfois discutables

- PAN : Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection
  - En place depuis 2007
  - Tâches proposées en 2011 :
    - *Détection de plagiat (données artificielles)*
    - *Détection de vandalisme dans la Wikipedia (données trop complexes, peu linguistiques, et notion trop floue)*
    - *Attribution d'auteur*
  - Tâches proposées en 2012 :
    - *Détection de plagiat externe (« big data »)*
    - *Identification de prédateurs sexuels dans les discussions en ligne (données glauques et langue difficile à traiter)*
    - *Attribution d'auteur « classique »*
- L'axe TAL à PAN
  - Décision prise pour accroître notre visibilité internationale
  - Choix d'une tâche qui permet de multiplier les approches linguistiques à différents niveaux
  - Militer pour une meilleure prise en compte des descripteurs linguistiques

- Vue d'ensemble du travail
  - 1/ Nettoyage du corpus d'entraînement
  - 2/ Analyse automatique (étiquetage, parsing)
  - 3/ Définition et calcul des traits
  - 4/ Apprentissage automatique supervisé pour associer les traits aux auteurs
    - *Choix de la technique, paramétrage*
    - *Evaluation sur le corpus d'entraînement (si possible)*
  - 5/ Réception du corpus de test
  - 6/ Nettoyage, analyse, mesure des traits
  - 7/ Projection du modèle appris
  - 8/ Envoi des résultats

Les données

- PAN 2011
  - Données : emails du corpus « Enron »
    - *Données authentiques*
    - *Courriers rendus publics suite à la banqueroute de la compagnie en 2001 et de l'enquête qui a suivi*
    - *Données disponibles pour la recherche, et grossièrement anonymisées*
  - Sous-tâches
    - *Attribution (qui a écrit ce message ?)*
      - Classe fermée (tous les auteurs sont connus)
        - Petit jeu de données (SmallTest)
        - Grand jeu de données (LargeTest)
      - Classe ouverte (possibilité d'auteur inconnu)
        - Petit jeu de données (SmallTest+)
        - Grand jeu de données (LargeTest+)
    - *Vérification (l'auteur X a-t-il écrit ce message ?)*
      - Verifiy1, Verify2, Verify3

- Description des données
  - A priori en anglais
  - Courrier électronique moyen :
    - *61 mots, 11 lignes, 4.3 phrases*

Tâche	Auteurs	# Messages (Entraînement)	# Messages (Evaluation)
Large	72	9337	1300
Large +	72 + x	9337	1416
Small	26	3001	495
Small +	26 + x	3001	634
Verify1	1 + x	42	92
Verify2	1 + x	55	101
Verify3	1 + x	47	89

Quelques exemples (courts)

- *Call 713-515-0389*
- *<NAME/>,*
  here is the info per our phone conversation earlier today. Call me with any question you may have.
  Regards,
  <NAME/> Hyatt
- *hi mom, yup 16.5 x 33 is my size. thanks! love, coop*
- *Hey bundalicious! How are you doing? We only have one more day to go before the holiday. I can't wait to spend the week with you! I wanted to get you dad's email address.*
  Love ME
- *CONFIDENTIAL -- DO NOT DISTRIBUTE OUTSIDE THE COMPANY.*
  Attached is a spreadsheet of how prices may change with the FERC Refund case. This is still subject to change. Hope this helps with the Negative CTC calculation and negotiations.

- PAN 2012
  - Données : fiction contemporaine américaine (partie gratuite du site Feedbooks.com)
  - Sous-tâches :
    - *A : 3 auteurs, textes courts (4-6 kmots), classe fermée*
    - *B : 3 auteurs, textes courts, classe ouverte*
    - *C : 8 auteurs, textes moyens (8-15 kmots), classe fermée*
    - *D : 8 auteurs, textes moyens, classe ouverte*
    - *E : démêlage d'auteurs, textes de 30 paragraphes*
    - *F : intrusion d'auteur, textes de 20 paragraphes*
    - *I : 14 auteurs, textes longs (40-170 kmots), classe fermée*
    - *J : 14 auteurs, textes longs, classe ouverte*
  - Données d'entraînement : 2 textes par auteur

- Extrait
  - *Victor Dolor went to the diner because two months ago a man killed five people there. The man was Hugo Herrera. He was forty-one, divorced, recently unemployed from a downsized-factory job, and had finally been diagnosed with post traumatic stress disorder from something that happened when he was a child. Victor scanned several online articles for more specifics about the childhood trauma but found nothing.*
    In response to Hugo's most recent therapy session with some high-priced psychologist, Hugo wrote a letter to The New York Times that said he was "sick of all the fucking shit and finally going to do something about all the worthless shits in the world."
    (A1.txt)

Les descripteurs utilisés

- Notre but : utiliser des traits « Riches »
  - Qui utilisent des connaissances externes
  - Plus complexes qu'une simple liste de mots
- Exemples
  - Morphologie : emploi de suffixes (CELEX)
  - Syntaxe : complexité des phrases (Stanford parser)
  - Sémantique :
    - *Ambiguité et specificité (WordNet)*
    - *Cohésion (liens sémantiques de la Distributional Memory database)*
  - Traits ad hoc
    - *Fautes d'orthographes, formules d'ouverture et de fermeture, etc.*

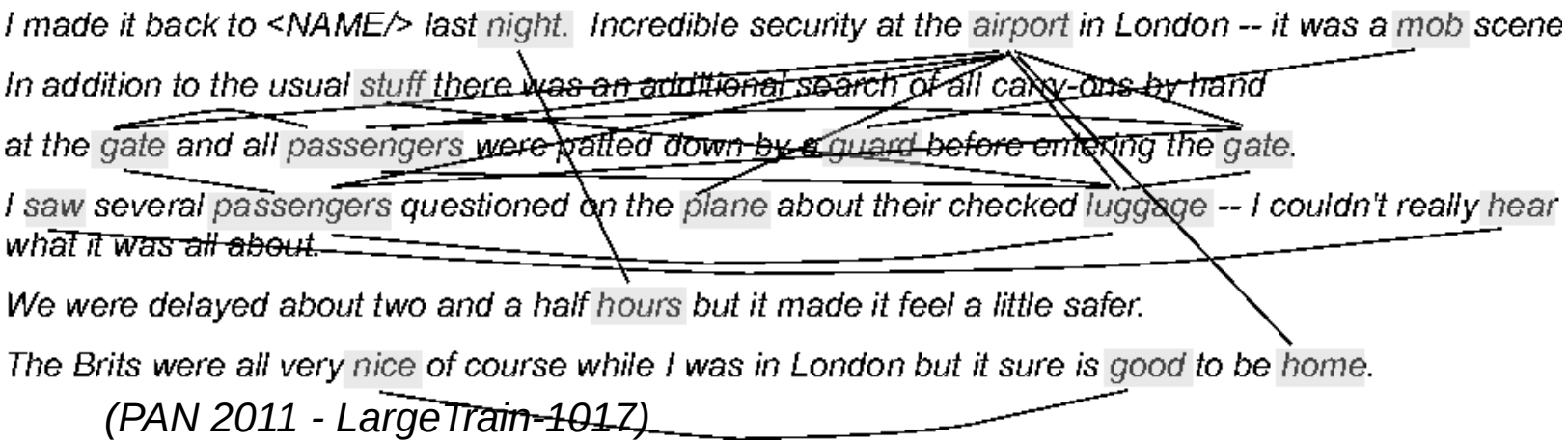
- Complexité morphologique
  - Suffixes extraits de CELEX et projetés
  - Haute fréquence de mots suffixés
    - *<NAME/>, Attached is a clean document for **execution**. If in **agreement**, please sign two **originals** and forward same to my **attention** for final signature. I will return a **fully** executed **agreement** for your records. Do not hesitate to give me a call should you have any **questions** regarding the enclosed.*
      Best regards,
      (PAN 2011 - SmallTrain-2249)
  - Basse fréquence de mots suffixés
    - *Suz, I say lets do it! and so does <NAME/>. I will make Rotel dip and other stuff too. I think it will be fun - and maybe we can carry the party to the hood after! Keep me posted on how your day is going. I kind of hope you get to go today to see your fam. K.*
      (PAN 2011 - SmallTrain-1358)
  - Egalement un trait pour chaque suffixe spécifique (-*ous*, -*ing*, etc.)

- Complexité syntaxique
  - Basée sur l'analyse du Stanford dependency parser
    - *Profondeur de l'arbre syntaxique*
    - *Distance entre les mots reliés syntaxiquement*
  - Complexité élevée (distance 3.6, profondeur 8.5) :
    - *unfortunate...**but you also don't want to go getting yourself attached to someone whom you ultimately don't have enough in common with to sustain the kind of relationship you're looking for**. off the soapbox.... i'm going to the grocery store (forgot some things), the dry cleaners, running and finishing up laundry detail...so that will take up a bit of time. (PAN 2011 - SmallTrain-2944)*
  - Complexité faible (distance 2.7, profondeur 2.7)
    - *<NAME/>,Seattle was sweet this weekend. I went and saw <NAME/> at the Breakroom...what did you think of Husky Stadium? Woohoo! Man, Thursday...whoa...and think, I went out after that...whoa...but it was my birthday...sorry for calling late. Are you doing anything cool this weekend? Motorcycle dirt track races are on Saturday night at Portland Speedway…I am stoked. Plus the first Husky football game is this weekend in Seattle against Michigan! How are other things going? Hopefully well. Later, <NAME/> (PAN 2011 - SmallTrain-623)*

- Ambiguïté et spécificité sémantique (WordNet)
  - Nombre moyen de synsets par mot, et profondeur moyenne de ces synsets (*specifique* *générique*)
  - Spécificité importante
    - *Hey <NAME/>,*
      I've done some research on the actuals that you make reference to (Vectren). <NAME/>'s sale with Heartland Steel is at the interconnect between Midwestern Gas Transmission and Vectren (formerly know as Indiana Gas). The actual volumes that you are reporting and consider to be your monthly actuals are volumes that I believe are behind Vectren's city gate (which means that you more than likely have an imbalance on Vectren's system). This bears checking with Vectren, regarding an imbalance behind their gate. You should be receiving some type of statement or invoice from Vectren. Per the contract, <NAME/> uses the Midwestern Gas Transmission (pipeline statement) to actualize our monthly invoices to you. I've attempted to draw a diagram for you to make it as clear as I can.
      Let's talk!
      (SmallTrain-929)
  - Vocabulaire générique
    - *I believe that we did have some activity on Blue Dolphin, but it was done by the Wellhead group. You should send the Vol Mgmt people to <NAME/> Smith.*

      (SmallTrain-2579)

- Cohésion
  - Basée sur la mémoire distributionnelle (Baroni & Lenci 2010)
    - *Couple de mots en relation s'ils partagent des contextes syntaxiques dans un corpus de référence*
    - *Nombre de mots ainsi reliés dans le message*
  - Cohésion élevée



(PAN 2011 - LargeTrain-1017)

  - Cohésion faible (aucun lien repéré)
    - *We are OUT of the pool. I want my money back. Prentice, please get your stuff out of my apartment. You can have the cats.*
      Love,
      <NAME/>
      (PAN 2011 - LargeTrain-2285)

- Autres traits utilisés:
  - Pauvres :
    - *Trigrammes de caractères*
    - *Fréquences lexicales*
    - *Signes de ponctuation*
    - *Longueur des mots, des messages, des lignes*
    - *Distribution des lignes vides*
  - Riches :
    - *Etiquettes morpho-syntaxiques (uni- et trigrammes)*
    - *Triplets syntaxiques (mot1-REL-mot2)*
    - *Entités nommées*
    - *Contractions*

- Traits ad hoc (PAN 2011):
  - Vocabulaire britannique vs américain
    - *-ise/–ize, -our/–or, etc.*
  - Fautes d'orthographe
    - *Douublement, omision, inverison, de lettres, etc.*
  - Formules d'ouverture (22):
    - *<NAME/>*
      Dear <NAME/>
      Hello
      <NAME/>,
      Hi,
      Hey <NAME/>,
      Hello <NAME/>
      Hi <NAME/>,
  - Formules de fermeture (44):
    - *thanks,\n*
      Thanx
      <NAME/>\n
      thank you,\n<NAME/>
      Love,\n<NAME/>
      thanks,\n<NAME/>.
      best,\n<NAME/>
      thanks!\n<NAME/>
      Thanks.

- Traits ad hoc (PAN 2012)
  - Discours direct
    - *Ratio de phrases entre guillemets*
  - Narration à la première personne
    - *Ratio de « I » parmi les sujets des verbes conjugués (hors guillemets)*

Exploitation des traits : Apprentissage Automatique

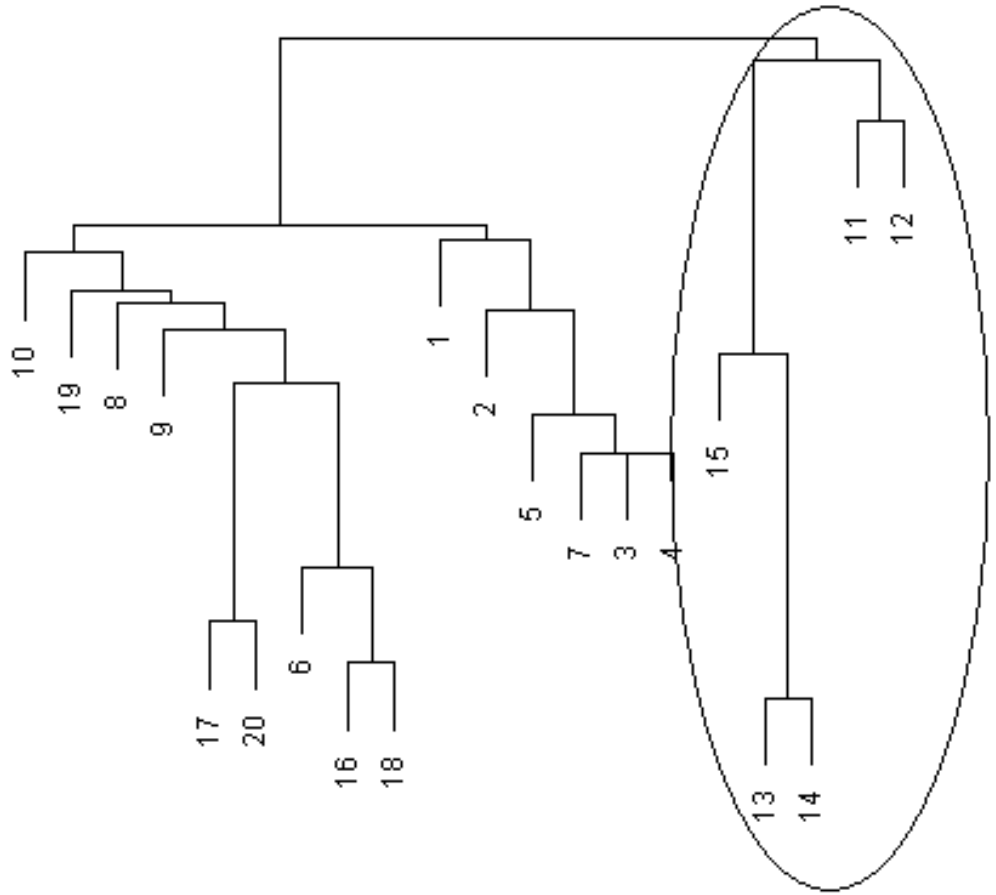
- Principes généraux de l'apprentissage automatique supervisé
  - Etant donné une collection d'éléments décrits par des variables et catégorisés (corpus d'apprentissage), identifier les liens entre les variables et les catégories
  - Pour un nouvel élément à catégoriser (corpus de test) décrit par les mêmes variables, utiliser ces liens pour décider sa catégorie
- Techniques
  - Symboliques (règles) : arbres de décision, règles d'association, règles conjonctives, etc.
    - *Plus anciens, traits peu nombreux, modèle interprétable*
  - Statistiques (poids ou probabilités) : systèmes bayésiens, SVM, réseaux de neurones, entropie maximale, etc.
    - *Plus modernes, traits nombreux, boîte noire*

- Classification par entropie maximale
  - Outil utilisé : *csvLearner*, Assaf Urieli
  - Système probabiliste : calcule une estimation de la répartition des probabilités de chaque trait pour chaque auteur
    - *Matrice de poids trait x auteur*
  - Avantages :
    - *A fait ses preuves pour différentes tâches de TAL (étiquetage, parsing, etc.)*
    - *Robustesse et rapidité*
    - *Capacité à traiter des traits nombreux et redondants*
  - Inconvénients :
    - *Boîte noire (ou au moins très sombre) : pas d'interprétation directe des poids obtenus par apprentissage*
    - *Un peu vieillot au dire des spécialistes (et supposément moins bon que les SVM)*

- Fonctionnement pour chaque tâche :
  - Construire le modèle sur les données d'entraînement
  - Pour chaque item du jeu de test, appliquer le modèle
    - *En sortie : une probabilité pour chaque auteur du jeu d'entraînement*
  - Classe fermée :
    - *Réponse = auteur avec la probabilité la plus élevée*
  - Classe ouverte :
    - *Idem, mais : réponse = « auteur inconnu » si la probabilité maximale est trop faible (seuil statique ou dynamique)*

- Mélange de paragraphes : classification non supervisée (clustering)
  - Principe : regrouper les paragraphes en *clusters*
  - Utilisation du classifieur par entropie maximale pour définir une distance entre les paragraphes (DePauw and Wagacha, 2008) :
  - Entraînement : tous les paragraphes comme items
    - *Classe = numéro du paragraphe*
  - Reclassification : chaque paragraphe analysé par le classifieur
    - *Resultat = Matrice de probabilités (Mp)*
    - *Matrice de distance entre les paragraphes : Md= -log(Mp)*
  - Clustering: regroupement des paragraphes similaires
    - *Classification Hierarchique Ascendante basée sur Md*
  - Resultat : clusters de plus haut niveau

Exemple de clustering de paragraphes (dendogramme)

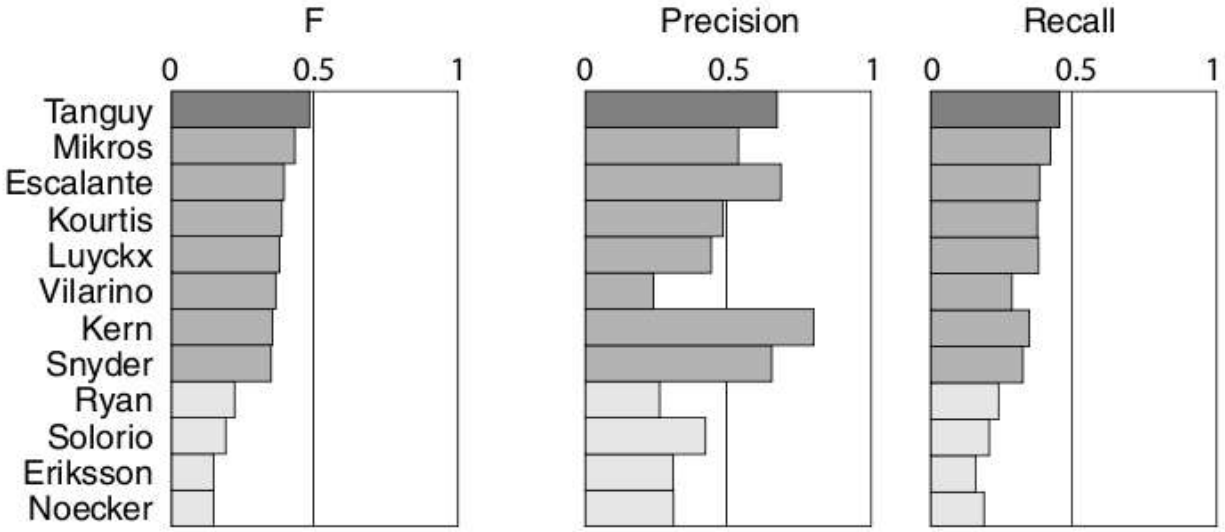


Résultats

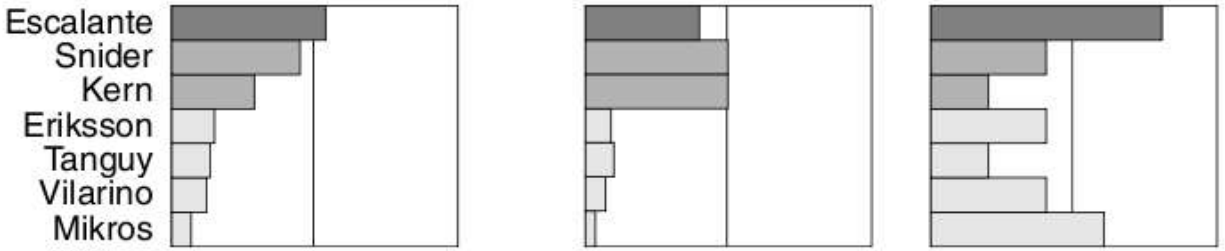
- Mesures d'évaluation
  - Plusieurs scores, en fonction des tâches et des habitudes
  - Choix dans le calcul d'un score global (sur plusieurs documents, sur plusieurs tâches, sur plusieurs auteurs, etc.)
- « Accuracy »
  - Pourcentage de documents pour lesquels l'auteur a été correctement identifié
- Précision/Rappel/F-score
  - Calculés pour chaque auteur
  - Moyenne sur l'ensemble de la collection

The PAN Competition

Authorship attribution:



Authorship verification:



- De très bons résultats !
  - Pour les tâches classiques, classes ouvertes ou fermées
- Un choix malencontreux
  - Choix d'une méthode d'apprentissage symbolique pour la vérification d'auteur
- Mais au final
  - Si nous avions utilisé le classifieur par entropie maximale, nous serions arrivés premiers aussi…

Résultats PAN 2012

TEAM	Overall	Docs corr.
EVL Lab	82,41	88,38
Bar I U	83,40	81,74
Brainsignals	86,37	81,33
CLLE-ERSS 1 (Trigrammes + tous traits riches)	70,81	77,59
Surrey	54,88	75,93
CLEE-ERSS 4 (60 traits synthétiques riches)	67,66	73,03
CLLE-ERSS 2 (Trigrammes)	59,13	68,88
CLLE-ERSS 3 (fréquences des lemmes)	64,71	64,32
Vilarino 2	62,13	63,07
Vilarino 1	50,46	59,75
Zech I-2	17,97	50,62
Zech I-3	17,03	48,13
Zech I-4	16,48	46,47
Brooke	16,63	46,06
Ruseti	57,40	22,82
de Graaff 1	57,55	21,99
Lip6 1	59,77	21,99
Sapkota	58,35	21,58
Lip6 2	54,41	20,33
Lip6 3	52,67	19,50
de Graaff 2	39,48	15,77
Zech terms	43,18	15,77
Zech stats	30,11	11,20
Zech stylo	22,91	8,71
de Graaff 3	2,94	1,66

- De moins bons résultats
  - Pour les tâches classiques, 10e place
  - Très bons résultats pour le démêlage d'auteurs
  - Résultats très variés pour les tâches classiques

TEAM	A %	B %	C %	D %	E %	F %	I %	J %	Overall
CLLE-ERSS 1 (riches + 3gr)	**100,00**	**60,00**	**37,50**	**41,18**	**73,33**	**93,75**	85,71	75,00	**70,81**
CLLE-ERSS 2 (3gr)	66,67	50,00	25,00	11,76	58,89	**93,75**	85,71	**81,25**	59,13
CLLE-ERSS 3 (lemmes)	**100,00**	**60,00**	**37,50**	11,76	45,56	88,75	**92,86**	**81,25**	64,71
CLLE-ERSS 4 (60 traits riches)	**100,00**	50,00	**37,50**	29,41	67,78	88,75	**92,86**	75,00	67,66
Meilleur score obtenu	100,00	90,00	100,00	76,47	92,22	100,00	92,86	87,50	86,37

Au-delà des scores :
Observation des modèles

- Interprétation des modèles
  - Pour « comprendre » le mécanisme
  - Pour évaluer les traits riches
  - Pour capter le « style » d’un auteur
- Opération très différente en fonction des méthodes
  - Directe avec les méthodes symboliques
  - Complexe voire impossible avec les méthodes numériques

- Observation du classifieur par entropie maximale
  - Méthode « interne » :
    - *A l'intérieur de la boîte noire*
    - *Matrice de poids relatifs trait x auteur*
  - Méthode « externe » :
    - *Etudes de lésion*
    - *Comparer les scores obtenus par des combinaisons de traits différentes*

- Evaluation interne (PAN 2011)
- Distribution des poids sur les différents types de traits
  - *Trigrammes de caractères :* *54%*
  - *Fréquences lexicales :* *11%*
  - *Trigrammes des POS :* *10%*
  - *Fréquences des POS :* *4%*
  - *Traits Riches :* *22%*
    - Morphologie : 4%
    - Syntaxe : 6%
    - Sémantique : 5.5%
    - Autres : 6%
- MAIS nous ne sommes pas certains de la pertinence de ces calculs
  - *Notamment, des traits nombreux et volumineux ont nécessairement un poids plus fort*

- Evaluation externe des traits riches (PAN 2011)

Traits	Efficacité globale	Précision	Rappel	F-score
Riches	61.01	40.13	35.11	36.17
Pauvres	68.08	45.91	37.62	38.03
Tous	70.30	58.28	41.20	43.39
Apport des traits riches	**+2.22**	**+12.37**	**+3.58**	**+5.36**

  - Conclusion:
    - *Traits pauvres légèrement plus efficaces que les traits riches*
    - *Apport notable des traits riches, surtout en précision et surtout pour certains auteurs*

- PAN 2012 : Etude de lésion, tâches A et C
  - Principe : calculer l'efficacité globale du système en supprimant progressivement des traits
    - *Plusieurs centaines de combinaisons testées*
  - Bilan : gain moyen de précision (accuracy) pour chaque groupe de traits :

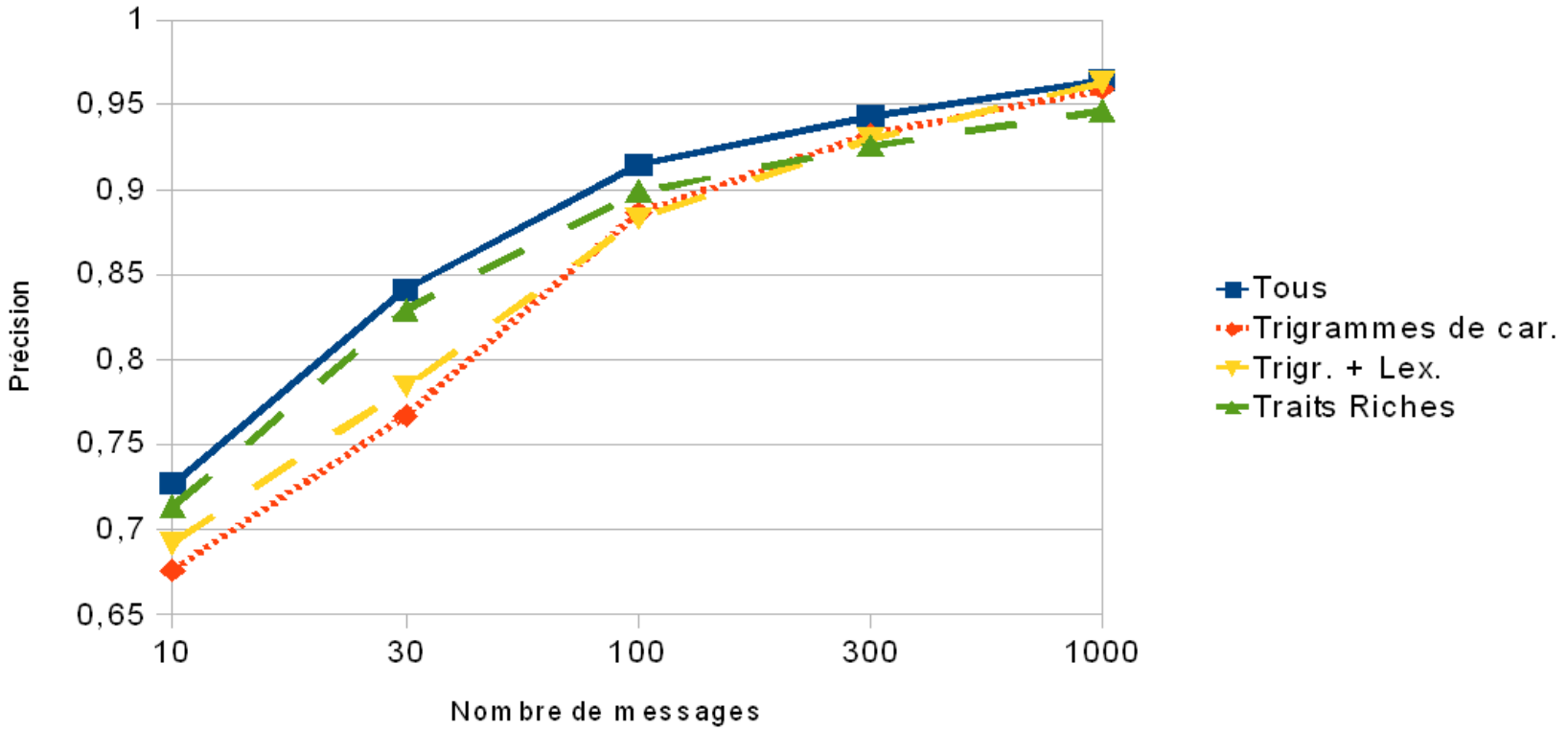
Jeu de traits	%Gain pour la tâche A	%Gain pour la tâche C
Ponctuation et casse	**+0.204**	**-0.040**
Fréquence de suffixes	+0.097	+0.009
Fréquence lexicale de référence	**+0.030**	**-0.003**
Complexité syntaxique	+0.015	+0.006
Ambiguïté / Généricité lexicale	+0.012	+0.008
Cohésion lexicale	+0.002	-0.000
Phrasal verbs (mesure synthétique)	-0.000	+0.022
Complexié morhpologique	-0.005	-0.002
Phrasal verbs (détails)	-0.006	-0.006
Contractions	-0.014	+0.018
Narration à la première/troisième personne	-0.027	-0.026
Trigrammes de POS	**-0.028**	**+0.045**
Trigrammes de caractères	**-0.034**	**+0.206**
Triplets syntaxiques	**-0.059**	**+0.089**

- Et aussi :
  - Des résultats excellents (meilleurs que les officiels) peuvent être atteints avec un très petit nombre de traits (une dizaine, en fonction des tâches)
- Moralité :
  - Les traits « très riches » semblent toujours utiles :
    - *Suffixes, complexité syntaxique, ambiguïté/généricité, cohésion lexicale*
  - D'autres ont des comportement opposés d'une tâche à l'autre
    - *Coefficient de corrélation : r = - 0,48*
- Comment choisir les traits ?
  - Pas assez de données pour effectuer ces tests avant
    - *Normalement : cross-validation sur une partie des données d'entraînement*
  - Seule solution : calcul de la répartition statistique des traits isolés
    - *Mesure de gain d'information, ou autres test statistiques*
    - *Pas de vision globale du comportement des traits en groupes*

Conclusions

- Nos méthodes nous placent clairement parmi les meilleurs systèmes au monde pour cette tâche
- D'où vient notre succès ?
  - Des traits riches (espérons-le)
  - Du classifieur (les spécialistes le trouvent trop rustique)
- Si les traits riches sont une bonne chose
  - Ils ont quand même besoin du support des pauvres
  - De petits ensembles de traits linguistiques bien choisis peuvent être très efficaces

- Les traits riches semblent avoir un avantage pour de petites données d'entraînement
  - Les traits pauvres sont censés capter des phénomènes linguistiques de haut niveau, grâce à la masse de données
  - Exemple : variation de précision pour 2 auteurs, en fonction du nombre de textes utilisés pour l'entraînement (données PAN 2011)



- Des débats intéressants
  - Clivage de la communauté entre deux types d'approches
    - *Traits linguistiques plus ou moins complexes*
    - *Méthode d'apprentissage très sophistiquées avec des n-grammes de caractères*
  - Une tendance forte pour les méthodes du second type
    - *Renforcée par le développement de méthodes d'apprentissage complexes (modèles à noyaux, multi-classification)*
    - *Ces approches rendent totalement impossible toute interprétation des modèles*
- Des encouragements très forts pour nos travaux
  - Pour développer des traits encore plus sophistiqués
  - Pour collaborer avec les spécialistes d'apprentissage et nous « mettre à niveau » dans ce domaine
- Au-delà, une saturation de l'apprentissage ?
  - Choix explicite de données non-apprenables pour un nouveau workshop à CLEF
  - Des approches « manuelles » pour l'identification des prédateurs sexuels