# Attribution d'auteur : multiplier les traits linguistiques pour identifier les auteurs de courriers électroniques

Ludovic Tanguy, Assaf Urieli, Basilio Calderone,
Nabil Hathout, Franck Sajous

et tous les autres membres de l'axe qui ont participé à ce travail
(C. Adam, C. Fabre, B. Gaume, L.M. Ho-Dac, F. Morlane-Hondère,
M.P. Péry-Woodley, N. Tulechki)

Séminaire du M2 ECIL – Octobre 2011

## PLAN

- L'attribution d'auteurs en quelques mots
  - Principe général
  - Exemples concrets
  - Méthodes
- La tâche de la compétition PAN 2011
  - Corpus
  - Sous-tâches
- Méthode
  - Traits linguistiques
  - 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*
  - Tâche connexe : 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?*
- Méthodes classiques :
  - Mesures statistiques basées sur les choix de vocabulaire et autres caractéristiques des textes, supposées stables à travers les écrits d'un même auteur (« style »)

- 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)*

5

à

- 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)

6

- Principales méthodes employées
  - Mesures de distance entre deux textes
  - Techniques par apprentissage automatique
    - *Catégorisation de textes*
- Principaux descripteurs utilisés
  - Vocabulaire (fréquence des mots de chaque texte)
  - Caractéristiques synthétiques
    - *Longueur des mots, des phrases, signes de ponctuation, erreurs, distribution des mots grammaticaux, etc.*
  - « Modèles de langue »
    - *N-grammes de caractères, mots, étiquettes morpho-syntaxiques, etc.*

7

## La compétition PAN 2011

8

- 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

9

- PAN : Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection
  - En place depuis 2007
  - 3 tâches proposées en 2011 :
    - *Détection de plagiat (données artificielles…)*
    - *Détection de vandalisme dans la Wikipedia (données trop complexes, peu linguistiques, et notion floue)*
    - *Attribution d'auteur*
- L'axe TAL à PAN
  - Décision prise l'année précédente pour accroître notre visibilité internationale
  - Choix d'une tâche qui permet de multiplier les approches linguistiques à différents niveaux

10

- La compétition vue de plus près :
  - 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*
      - 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*
      - Verifiy1, Verify2, Verify3

11

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

| Tache | Auteurs | Messages (Entrainement) | Messages (Evaluation) |
|---|---|---|---|
| Large | 72 | 9337 | 1300 |
| Large + | 72+ | 9337 | 1416 |
| Small | 26 | 3001 | 495 |
| Small + | 26+ | 3001 | 634 |
| Verify1 | 1 | 42 | 92 |
| Verify2 | 1 | 55 | 101 |
| Verify3 | 1 | 47 | 89 |

12

## 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. <NAME/> this helps with the Negative CTC calculation and negotiations.*

13

- Vue d'ensemble du travail
  - 1/ Nettoyage des courriers 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*
  - 5/ Réception du corpus de test
  - 6/ Nettoyage, analyse, mesure des traits
  - 7/ Projection du modèle appris
  - 8/ Envoi des résultats

14

# Traits linguistiques

15

- Principe général
  - Définir un ensemble de caractéristiques
  - Calculer automatiquement leur valeur pour chaque message
- Quels traits utiliser ?
  - Grande variété à travers l'histoire de la discipline
  - Actuellement, prédominance notable des traits pauvres (trigrammes de caractères)
    - *Efficacité « prouvée », pas de traitement en amont, indépendant de la langue, …*

16

- 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.*

17

- 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, (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. (SmallTrain-1358)*
  - Egalement un trait pour chaque suffixe spécifique (*-ous*, *-ing*, etc.)

18

- 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. (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/> (SmallTrain-623)*

19

- 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)*

20

- 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

*I made it back to <NAME/> last night. Incredible security at the airport in London -- it was a mob scene*
*In addition to the usual stuff there was an additional search of all carry-ons by hand*
*at the gate and all passengers were patted down by a guard before entering the gate.*
*I saw several passengers questioned on the plane about their checked luggage -- I couldn't really hear what it was all about.*
*We were delayed about two and a half hours but it made it feel a little safer.*
*The Brits were all very nice of course while I was in London but it sure is good to be home.*
*(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/>*
      *(LargeTrain-2285)*

21

- Traits ad hoc:
  - Formules d'ouverture (22):
    - *<NAME/>* *<NAME/>,* *Hello <NAME/>*
      *Dear <NAME/>* *Hi,* *Hi <NAME/>,*
      *Hello* *Hey <NAME/>,*
  - Formules de fermeture (44):
    - *thanks,\n* *thanks,\n<NAME/>.* *Thanks.*
      *Thanx* *best,\n<NAME/>*
      *<NAME/>\n* *thanks!\n<NAME/>*
      *thank you,\n<NAME/>*
      *Love,\n<NAME/>*

22

- Autres traits pauvres utilisés:
  - Trigrammes de caractères
  - Fréquences lexicales
  - Signes de ponctuation
  - Etiquettes morpho-syntaxiques (uni- et trigrammes)
  - Entités nommées
  - Longueur des mots, des messages, des lignes
  - Distribution des lignes vides
  - Contractions
  - Vocabulaire britannique vs américain
- Total : plus de 26000 traits (dont ~ 300 riches)

23

# Apprentissage Automatique

24

- Principes généraux de l'apprentissage automatique supervisé
  - Etant donné une collection d'éléments décrits par des variables et catégorisés (corpus d'apprentissage), identifier les relations 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.
  - Statistiques (poids ou probabilités) : systèmes bayésiens, SVM, réseaux de neurones, entropie maximale, etc.

25

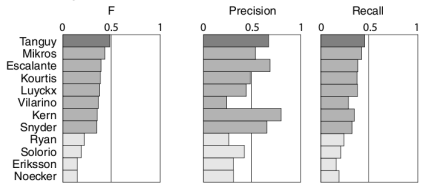
- Préparation
  - Utilisation d'une partie du corpus d'entraînement pour configurer le système
  - Sélection de traits, normalisation, réglage des paramètres
  - Evaluation des performances par mesures de précision, rappel et F-score
    - *Moyenne sur l'ensemble des auteurs*
- Concrètement :
  - Pour les tâches d'*attribution* : classifieur par entropie maximale (*MaxEnt*)
  - Pour les tâches de *vérification* : systèmes symboliques (arbre de décision et règles conjonctives)

26

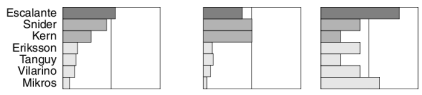
**Résultats**

**The PAN Competition**

*Authorship attribution:*




*Authorship verification:*




27

- Classification par entropie maximale
  - Système probabiliste : calcule une estimation de la répartition des probabilités de chaque trait pour chaque auteur
    - *Matrice de poids trait x auteur*
  - Pour un message à catégoriser, calcule une probabilité pour chaque 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*

28

- Gérer les auteurs inconnus
  - Pour les auteurs connus, choix de celui qui a la probabilité maximale
  - Corrélation remarquable entre les basses probabilités maximales et les cas d'erreur ou d'auteur inconnu
  - Pour les tâches à classe ouverte, si la probabilité maximale est inférieure à un seuil, répondre « inconnu »
    - *Deux « runs » avec des seuils différents*

| Tâche | Seuil | Précision | Rappel | F-score |
|---|---|---|---|---|
| SmallTest+ | 66% | 82.4 | **45.7** | **58.8** |
| SmallTest+ | 95% | **96.6** | 18.0 | 30.3 |
| LargeTest+ | 40% | 77.9 | **47.1** | **58.7** |
| LargeTest+ | 75% | **92.4** | 29.9 | 45.1 |

29

- Vérification d'auteur
  - Entraînement : messages de l'auteur cible + 1500 messages choisis aléatoirement
  - Comparaison de différentes méthodes en fonction des auteurs :
    - *Auteur 1 : arbre de décision (C4.5)*
    - *Auteurs 2 et 3 : règles conjonctives (RIPPER)*
  - Scores corrects sur le corpus d'entraînement
  - Scores très faibles sur le corpus de test

30

- Utiliser l'entropie maximale au lieu des règles
  - Apprentissage avec 100 messages aléatoires
  - Résultats si on l'avait fait

| Tâche | Méthode | Précision | Rappel | F-score |
|---|---|---|---|---|
| Verify1 | Arbre de décision (soumis) | 0.09 | 0.33 | 0.143 |
| | ***Entropie maximale*** | ***0.33*** | ***0.66*** | ***0.444*** |
| Verify2 | RIPPER (soumis) | 0.1 | 0.2 | 0.133 |
| | ***Entropie maximale*** | ***1*** | ***0.40*** | ***0.571*** |
| Verify3 | RIPPER (soumis) | 0.08 | 0.25 | 0.125 |
| | ***Entropie maximale*** | ***0.25*** | ***0.25*** | ***0.25*** |

- Au moins doublé les performances (et fini 1er ou 2ème…)

31

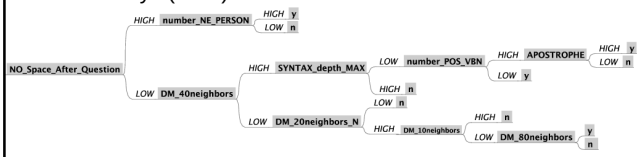
## Observation des modèles

32

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

33

- Modèles obtenus pour la vérification d'auteur
  - Verify1 (C4.5):




  - Verify2 (RIPPER):
    - *1. if DM90neighbors ≥ 0.00493 and DM80neighbors ≤ 9 and APOSTROPHE = 0 then Y*
    - *2. if DM20neighbors ≥ 0.0173 and COLON ≥ 0.0090 and DM10neighbors ≤ 28 then Y*
    - *3. otherwise N*

34

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

35

- Evaluation externe des traits riches

| 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*

36

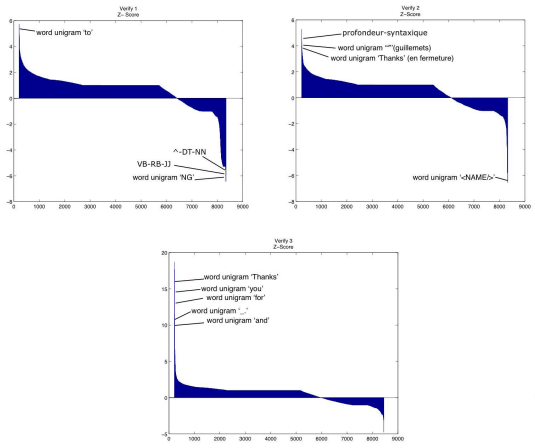
- Distribution des poids sur les différents types de traits
  - *Trigrammes de caractères : 54%*
  - *Fréquences lexicales : 11%*
  - *Trigrammes des POS : 10%*
  - *Fréquences des POS : 4%*
  - *Traits Riches : 22%*
    - Morphologie : 4%
    - Syntaxe : 6%
    - Sémantique : 5.5%
    - Autres : 6%

37

- Zoom sur un auteur particulier
  - 3 auteurs des tâches de vérification
  - Examen manuel des caractéristiques par lecture de l'ensemble des messages
    - *~ 50 messages / auteur*
    - *4 juges sans concertation*
  - Analyse statistique des répartitions de traits

38

**Analyse statistique – traits les plus significatifs**




39

- Analyse manuelle : caractéristiques notables
  - Auteur 1 :
    - *Interrogatives sans « ? » (5/9 interrogatives)*
    - *Messages automatiques (17/42 messages)*
      - *The report named XXX, published as of YYY is now available on the web site….*
  - Auteur 2 :
    - *Phrases courtes dans des messages courts*
    - *Glissement de personne (de « je » vers « nous »)*
      - 10/50 messages
      - ***I** have a few thoughts on the offsite. **I** think **we** could have a theme of restructuring and change. **We** would have to make sure it is forward looking and upbeat in that **we** have learned a lot that will make **us** better in the future.*

40

- Auteur 3 :
  - *Classe de verbes « modalisateurs » à la première personne*
    - 41/105 verb occurrences
    - *Know, hope, doubt, mind, feel, like, think, enjoy, guess*, etc.
  - *Combinaisons de « Let me know » et « if/how/wh.. »*
    - 10/37 messages
    - *If you have any problems, let me know.*
    - *Please let me know if you know where <NAME/> is.*
    - *Let me know if this interferes with any plans.*

41

# Conclusions

42

- Pas mal pour un coup d'essai
- Pas de visibilité claire de la raison de notre victoire :
  - Traits riches
  - Classifieur par entropie maximale
  - Chance du débutant
- Si les traits riches sont une bonne chose
  - Ils ont quand même besoin du support des pauvres
    - *Une des raisons pour lesquelles les systèmes à base de règles ont échoué*

43

- Suite
  - Examen plus détaillé des données
    - *Distribution des traits, sélection de traits pertinents, etc.*
  - Autres données
  - Encore beaucoup d'idées de traits à développer
    - *Structures de phrase, collocations, etc.*
- Lien avec d'autres problématiques
  - La plupart des traits riches sont liés à des « genres » de message :
    - *Courrier formel au client ou au chef,*
    - *Courrier informel à la famille ou à des amis,*
    - *Ordre/requête à des subordonnés,*
    - *Réponse rapide, information directe (URL, téléphone, etc.),*
    - *Lettre d'amour, horoscope, blague du jour*
    - *etc.*
  - Quelle stabilité d'un genre à l'autre ?
  - Traits par genre ?

44