Bilan 1ère année de thèse

...

Brice Olivier
encadré par
Marianne Clausel, Jean-Baptiste Durand, Anne-Guérin Dugué

Sujet de thèse et objectifs initiaux

1. Modéliser conjointement des traces oculométriques et EEGs par modèles de Markov semi cachés et couplés
    a. Modélisation disjointe des traces oculométriques
    b. Modélisation disjointe des EEGs
2. Analyser les co-variables
    a. Analyser la variabilité des textes et individus et leur impact sur le modèle
    b. S'aider des co-variables pour mieux caractériser le modèle
3. Relier les topics models aux comportements des lecteurs

Plan

1. Axes suivis
    a. L'existant : compréhension des données et du modèle
    b. Bibliographie : validation du modèle
    c. Etude d'EM : stabilité des résultats
    d. Enrichissement de l'interprétation du modèle : utilisation des co-variables
    e. Enrichissement du modèle : réduction de la variabilité des textes par clustering
    f. Préparation à l'analyse conjointe des traces oculométriques et EEGs : changement du pas de temps
    g. Création d'un module VPlants : Eye Movement Analysis (EMA)
2. Difficultés rencontrées
3. Bilan administratif
4. Programme de travail pour la 2ème année

a. L’existant : compréhension des données et du modèle

- Déroulement de l’expérience : thème → texte → question
- Construction des textes : à partir de textes issus d’articles LeMonde 1999
- Acquisition des données sous forme de signaux (Eye tracker + 31 canaux EEGs)
- Prétraitements (fenêtre de fixation, suppression d'artefacts, suppression de données aberrantes, création d’un readmode)
- Modèle de semi-Markov caché : Probabilités de transitions, d’observations, loi de temps de séjour

b. Bibliographie : validation du modèle (1/2)

- Hidden Markov Model (Chap. Sequential Data - Bishop 2006, Chap. Markov and hidden Markov models - Murphy 2012, Rabiner 1989, Online courses)
- Hidden Semi Markov Model (Yu 2010)
  - Hypothèses sur le modèle : Marhasev, Residential time HMM, variable transition HMM, **explicit duration HMM**
  - Hypothèses d'initialisation du processus : sur]-inf;+inf[, **sur [1;T]**
- Pairwise Markov Chains : plus efficace pour la segmentation
- Triplet Markov Chains : permet de gérer des processus non stationnaires
- HMM d'ordre M
- Coupled HMMs, Event-Coupled HMMs, Factorial HMMs, I/O HMMs (Zhong 2001)

b. Bibliographie : validation du modèle (2/2)

- Generative vs Discriminative Model (Maximum Entropy Markov Models)
- Hierarchical Dirichlet Process HMM
- Utilisation de HSMM pour des données de traces oculométriques : discrimination de tâches, interprétation/caractérisation de stratégies de lecture, calcul de scanpath moyen (Simola 2008)
- Eye Fixation Related Potential investigation (Frei 2013)
- Faire face à la variabilité matérielle, humaine et au signal overlapping (Frei 2016?)
- Clustering de textes par k-means généralisé pour données temporelles (Soheily Khah 2016) et ontologies (Liu 2014)

c. Etude d’EM : stabilité des résultats

- Problème : EM trouve un minimum local
- Solution : Plusieurs initialisations pour garder celle qui maximise la vraisemblance des données complètes
- Processus génératif : On se base sur les séquences observées pour générer des séquences d’états cachées aléatoires pour ensuite créer un modèle Semi Markovien non caché servant d’initialisation

- Toujours en cours d’implémentation...

d. Enrichissement de l'interprétation du modèle : utilisation des co-variables

- L'utilisation de co-variables, comme le texte, peuvent permettre de **caractériser** et **valider** les stratégies de lecture découvertes
- Importance du ReadMode, de l'entité fixation (saccade entrante + fixation ou saccade sortante + fixation) : influence sur la vitesse de lecture, l'amplitude des saccades
- Les descripteurs pour chaque stratégie de lecture :
  - mots lus (et relus)
  - nombre de fixations
  - durée des fixations
  - amplitude des saccades
  - direction des saccades

e. Enrichissement du modèle : réduction de la variabilité des textes par clustering (1/3)

- Modèle de clustering
  - a. Matrice Termes Documents (TF-IDF)
  - b. LSA avec D=300 permet de garder assez de thèmes pour ne pas en agréger
  - c. Calcul de l'évolution similarité sémantique entre thème et texte (cos. inst. / cum.)
  - d. Clustering de courbes par HAC + DTW
- a. TF-IDF vs Entropy
- a. Utilisation du corpus LeMonde 1999 pour garder le même vocabulaire que celui des textes
- b. LSA vs modèles plus compliqués
  - textes mono-thématiques
  - texte avec des mots incongrus = pluri-thématique ?

e. Enrichissement du modèle : réduction de la variabilité des textes par clustering (2/3)

- c. Traduction d'un sigle ou non
  - e.g. HCR dans le thème des réfugiés
  - permet d'augmenter significativement le cos inst. / cos. cum.
  - Biaisé si l'utilisateur ne connaît pas le sigle
- c. Clustering a priori - vérifier le profil théorique vs. a posteriori - vérifier les affectations a priori
- c. Définition de la similarité sémantique entre thème et texte
- c. Définition d'un mot lu ⅓ du début ou ⅔ de la fin en français : zone fovéale
- c. Une fixation peut s'effectuer sur un mot ou un groupe de mots
- c. Le cos. cum. est beaucoup plus influencé par un cos. inst. faible que élevé

e. Enrichissement du modèle : réduction de la variabilité des textes par clustering (2/3)

- d. Enjeux du clustering : Capturer la dynamique globale ou locale ?
- d. Méthodes de clustering envisagées
  - K-means généralisé et pondéré pour des séries temporelles
  - K-médoides
  - Ontologies
- d. Encadrement groupe Ensimag : Meilleurs résultats obtenus avec la méthode initiale

f. Préparation à l'analyse conjointe des traces oculométriques et EEGs : changement du pas de temps

- Pour permettre la modélisation conjointe des données, il faudra trouver un pas de temps commun
- Redéfinition des données de traces oculométriques au pas de temps
  - on considère qu'un read mode reste le même au cours d'une même fixation
  - cependant la stratégie de lecture peut changer (ou pas)
- Résultats obtenus :
  - Lois des temps de séjour : Binomiales négatives. Espérance 200-400 ms
  - Probabilités de transition : Alternance entre 3 états. Les 2 autres sont similaires mais pas d'auto-transition
  - Lois d'observation : 5 processus catégoriels : 1 état caché engendre tout le temps le même état observé
- Implication d'une série par une autre : causalité de Granger ?
  Changement de similarité sémantique → Onde cérébrale → changement de stratégie de lecture

g. Création d'un module VPlants : Eye Movement Analysis (EMA)

- Compréhension du code existant (fait par Jean-Baptiste)
- Exploration des packages HSMM existants : hsmm (R), mhsmm (R), sequence_analysis (Python)
- Implémentation d'un module pour l'analyse des traces oculométriques
  - Représentation des données sous une structure permettant de faciliter le calcul de descripteurs
  - Calcul de descripteurs pour les stratégies de lecture
  - Initialisation aléatoire d'EM (en cours)
  - Clustering de textes (en cours)

Difficultés rencontrées

OPENALEA

VPLANTS

SEQUENCE ANALYSIS

STAT TOOLS

LE CODE C++

DEBUGGING DE CODE PYTHON/C++

Bilan administratif

- Summer School à Madrid
  - Validation des crédits de formation scientifique
- Label RES (à venir)
  - Validation des crédits d'insertion professionnelle
  - Validation des crédits de formations transversales par l'intermédiaire de formations pédagogiques
- Activités annexes
  - Séminaires proba-stats du LJK
  - Journal oculo au GIPSA-lab
  - Présentation de la thèse au séminaire LJK en Décembre (Merci Jean-Baptiste)
  - Présentation de l'article de Simola au journal oculo en Mars
  - Co-encadrement de groupes d'étudiants Ensimag sur le clustering de textes en Mai

A voir ensemble : papiers pour la réinscription en deuxième année !

Programme de travail pour la 2ème année

- 3 enseignements (dont 2 identiques)
- Finir le debugging du code pour obtenir des résultats stables
- A terminer : Enrichissement de l'interprétation du modèle : utilisation des co-variables
- A terminer : Enrichissement du modèle : réduction de la variabilité des textes par clustering
- Rédiger un rapport de manière formelle sur les axes suivis
- Analyse conjointe des traces oculométriques et EEGs

- Suggestions ?

Merci pour votre encadrement !