Projet TAL et Industrie

Hugues de Mazancourt
Cours M2 TAL
hugues@mazancourt.com

Cas d’utilisation

- Un consortium se forme pour faire une étude linguistique des e-mails
  - par exemple dans le cadre du support client
- On veut étudier comment les dialogues se font, comment les protagonistes s’écrivent l’un l’autre, etc.
- Mais il n’est pas possible de d’ouvrir (« disclose ») le corpus de mails qui sont de la conversation privée
- L’anonymisation brutale (suppression de tous les noms de personnes ou d’entité) n’est pas une solution car elle ne permet pas une analyse linguistique complète
  - qui parle à qui, comment, etc.

hugues@mazancourt.com

Un (autre) exemple d'étude linguistique de mails

- Travail d'Owen Rambow et al. (Stanford) sur le corpus Enron
- Analyse des « expressions manifestes de puissance » (Overt Display of Power)
    - p. ex. « *I need your report by Friday* » vs. « *Could you please try to send your report by Friday?* »
    - Autres ODP « *I need the answer ASAP, as …* », « *can you please keep me in the loop* », « *ok call me on my cell later.* »
    - entrainement de classifieurs (SVM) pour détecter les ODP
    - (présence de ce type de marqueurs dans 5% du corpus)
- Déduction des liens de subordination avec une précision de 73%
    - Plus difficile chez les femmes qui produisent moins de ces expressions

hugues@mazancourt.com

Exemple (mail Enron)

Date: Tue, 4 Jan 2000 04:54:00 -0800 (PST)
From: susan.scott@enron.com
To: kevin.hyatt@enron.com, jeffery.fawcett@enron.com
Subject: Re: The El Paso Controversy Continues
Cc: steven.harris@enron.com

Attached is Shelley Corman's response to Jeff Dasovich, based on conversations with Drew and me. Accordingly, I would advise you not to participate in the conference call Jeff is setting up for tomorrow.

Questions -- give me a call.

---------- Forwarded by Susan Scott/ET&S/Enron on 01/04/2000 12:52 PM ----------

Shelley Corman
01/04/2000 12:50 PM
To: Jeff Dasovich/SFO/EES@EES
cc: Susan Scott/ET&S/Enron@ENRON

Subject: Re: The El Paso Controversy Continues

Jeff,

From the pipeline group's perspective, this transaction is a commercial matter between El Paso and ENA. Accordingly, the gas pipeline group does not believe that it is appropriate to participate in the development of ENA's response.

Exemple (mail Enron)

Date: Tue, 4 Jan 2000 04:54:00 -0800 (PST)
From: EMAIL#1
To: EMAIL#2, EMAIL#3
Subject: Re: The LOC#1 Controversy Continues
Cc: EMAIL#4

Attached is PERS#1's response to PERS#2, based on conversations with PERS#3 and me. Accordingly, I would advise you not to participate in the conference call PERS#4 is setting up for tomorrow.

Questions -- give me a call.

---------- Forwarded by PERS#5/ET&S/Enron on 01/04/2000 12:52 PM ----------

PERS#5
01/04/2000 12:50 PM
To: PERS#2/SFO/EES@EES
cc: PERS#5/ET&S/ORG#1@ENRON

Subject: Re: The LOC#1 Controversy Continues

PERS#4,

From the pipeline group's perspective, this transaction is a commercial matter between LOC#1 and ORG#2. Accordingly, the gas pipeline group does not believe that it is appropriate to participate in the development of ORG#2's response.

Anonymiser « brutalement » fait perdre de l'information

- Perte du lien entre e-mail et nom (*susan.scott@enron.com* vs « *Susan Scott* »)
- Perte du lien entre deux entités nommées (« *Jeff* » vs. « *Jeff Dasovich* »)
- Ajout d'erreurs : qualification de *El Paso* en nom de lieu
- Le corpus anonymisé brutalement devient difficile à utiliser en général pour de l'apprentissage
    - et totalement inexploitable dans le cadre de l'analyse du discours ou du dialogue (qui parle à qui, de quelle façon, comment sont repris les sujets, etc.)

hugues@mazancourt.com

Une solution : la pseudonymisation

- Remplacer des noms par d'autres noms
- De façon cohérente pour permettre une résolution de liens entre mails
- Voire résolution d'anaphores
  - entre Nom Prénom et reprises par le prénom seulement
  - ou nom.prenom@company.com dans le header et reprise par le prénom dans le corps du texte

Avant traitement

Date: Tue, 4 Jan 2000 04:54:00 -0800 (PST)
From: susan.scott@enron.com
To: kevin.hyatt@enron.com, jeffery.fawcett@enron.com
Subject: Re: The El Paso Controversy Continues
Cc: steven.harris@enron.com

Attached is Shelley Corman's response to Jeff Dasovich, based on conversations with Drew and me. Accordingly, I would advise you not to participate in the conference call Jeff is setting up for tomorrow.

Questions -- give me a call.

---------- Forwarded by Susan Scott/ET&S/Enron on 01/04/2000 12:52 PM ----------

Shelley Corman
01/04/2000 12:50 PM
To: Jeff Dasovich/SFO/EES@EES
cc: Susan Scott/ET&S/Enron@ENRON

Subject: Re: The El Paso Controversy Continues

Jeff,

From the pipeline group's perspective, this transaction is a commercial matter between El Paso and ENA. Accordingly, the gas pipeline group does not believe that it is appropriate to participate in the development of ENA's response.

Après (idéalement)

Date: Tue, 4 Jan 2000 04:54:00 -0800 (PST)
From: daenerys.targarien@got.com
To: arya.stark@got.com, khal.drogo@got.com
Subject: Re: The Westeros Controversy Continues
Cc: viserys.targarien@got.com

Attached is Jon Snow's response to Tyrion Lannister, based on conversations with Sansa and me. Accordingly, I would advise you not to participate in the conference call Tyrion is setting up for tomorrow.

Questions -- give me a call.

---------------------- Forwarded by Daenerys Targarien/ET&S/GOT on 01/04/2000 12:52 PM ----------------------------

Margery Tyrell
01/04/2000 12:50 PM
To: Tyrion Lannister/SFO/EES@EES
cc: Daenerys Targarien/ET&S/GOT@GOT

Subject: Re: The Westeros Controversy Continues

Tyrion,

From the pipeline group's perspective, this transaction is a commercial matter between Westeros and Lokrum. Accordingly, the gas pipeline group does not believe that it is appropriate to participate in the development of Lokrum's response.

Peudonymisation, la solution idéale ?

- Des éléments non-nommés peuvent permettre d'identifier les individus
- cf. I. Eshkol (corpus oral ESLO 1960/2013) qui définit des éléments d'identification directs
    - *« mon père a fondé le plus grand cabinet d'ophtalmologiste de la ville »*
- ou indirects :
    - *« le locuteur est patron de café au moment de l'enregistrement et il travaillait auparavant dans l'aviation militaire »*
- Ces éléments peuvent permettre la ré-identification avec une connaissance raisonnable du contexte
- Mais, l'un des gros avantages est que ces éléments, ainsi que les éventuelles erreurs d'un traitement automatique, sont noyés dans la masse
    - Donc le processus de ré-identification devient plus coûteux, voire exorbitant

hugues@mazancourt.com

Projet

- Pseudonymiser un extrait du corpus Enron
- De façon à ce que le résultat obtenu soit utilisable pour construire un système qui traiterait des mails
- … sans permettre de ré-identification
- et dans un format proche du format initial

hugues@mazancourt.com

Comment

- Libre choix des moyens (mais le traitement doit rester automatique)
  - Python (nltk, …)
  - Perl (Lingua::EN::NamedEntities)
  - Java
  - …
- Seul ou en groupe
- Ressources : ouvertes (et libres d’usage !)

Données source

- Données source sur http://bit.ly/M2-Diderot
  - ~ 8.000 e-mails au format texte, issus d'une utilisatrice
  - Quelques-un en encodage MIME (voir https://fr.wikipedia.org/wiki/Multipurpose_Internet_Mail_Extensions)
- 8.000 e-mails à traiter, cela oblige à :
  - trouver des solutions efficaces
  - faire des impasses (il faut alors estimer l'impact)
  - (par exemple ceux qui ont un encodage inattendu)

Méthodologie

- Etude de corpus
    - Quelles sont les données nominatives, comment sont-elles exprimées ?
    - Quels types d'information doit préserver une anonymisation ? Expression des anaphores, etc.
- Choix d'outils et de ressources
    - Extraction d'entités nommées ? Autres extracteurs dédiés ?
    - Listes de référence pour les données remplacées (listes de noms, de prénoms, …). Quelles contraintes ?
- Développement du système, test
- Evaluation du système
    - Ce qu'il fait bien, ce qu'il fait mal, ce qu'il ne fait pas.
    - Perd-on des d'informations ? Dans quel cas ?
    - Reste à faire par rapport à un système optimal

hugues@mazancourt.com

Production attendue

- Un rapport décrivant (une dizaine de pages)
  - les moyens mis en œuvre
  - une typologie des limites du système
    - une estimation du coût de développement pour chaque type de limite recensé mis en regard de l'importance du phénomène
- Le corpus anonymisé
- Le logiciel (commenté…) et les moyens de l'exécuter
  - (attention aux pré-requis !)

hugues@mazancourt.com

Quand ?

- 31 Janvier 2018
  - au plus tard…

Parsing des mails

- Il existe des analyseurs de mails (cf. Apache Tika, Email::MIME, …)
- Le format est simple - on peut faire l'hypothèse qu'il n'y a pas de pièce jointe :
  - header+ \n \n contenu
  - header ::= Header-Name ':' valeur \n
- Ne retenir (au plus) que les headers
  - From:
  - To:
  - Subject:
  - Date:
- Le contenu peut être traité comme un seul bloc de texte