ANR-15-CE38-0008

Frédéric Landragin, Juliette Potier & Meryl Bothua

Laboratoire Lattice, Montrouge

Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus

JLC, 4 juillet 2017

Plan

- Identification de stratégies d'annotation
- Expérimentations chronométrées
- Calculs de l'accord inter-annotateurs
- Panorama des pré-annotations envisageables
- Pré-annotation en chunks : aide ou gêne ?
- Conclusion et perspectives : procédure d'annotation

Identification de stratégies d’annotation

Remarques préalables

- Corpus « Democrat » visé
  - Taille d'un million de mots
  - Au moins 100 000 expressions référentielles annotées, si possible de l'ordre de 200 000 (ce qui ferait de l'ordre de 10 000 chaînes), pour permettre à la fois :
    - L'étude comparative de nombreuses chaînes de référence, même si l'on s'intéresse à une période précise et à un genre textuel précis
    - Des études textométriques, avec un nombre suffisant de données pour que les statistiques aient un sens
    - Des applications TAL : identification automatique des expressions référentielles ; identification automatique des chaînes de référence
- Ce qui représente une tâche coûteuse en temps…
  - Trois stages de M1 au Lattice en 2016, pour 6% du corpus annoté, et peut-être seulement 4% si on considère que les annotations restent à compléter par d'autres en temps voulu

Enjeux

- Compte tenu de ces remarques préalables, la tâche d'annotation manuelle doit être la plus efficace possible

- Pour mettre en œuvre les détails de cette procédure, nous avons :
  - Tenu compte des retours d'expérience du projet MC4
  - Tenu compte des retours d'expérience d'ANCOR
  - Mis en œuvre des expérimentations chronométrées pour comparer le temps pris par deux stratégies envisageables
  - Fait appel au calcul de l'accord inter-annotateurs pour pondérer nos décisions

- Au départ, deux stratégies :
  - Systématique
  - Avec filtrage des référents (que les humains, les animés…)

Aspect technique

- Importance décisive de l'ergonomie de l'outil d'annotation
  - On peut affecter un identifiant à chaque expression référentielle (et donc saisir plusieurs fois le même identifiant) ou construire les chaînes en regroupant plusieurs expressions référentielles (qui sont juste délimitées, pas annotées)
  - C'est la rapidité d'action qui compte et qui permet de choisir entre les deux schémas d'annotation :
  - Construire une chaîne est délicat et prend du temps
  - Saisir plusieurs fois le même identifiant est redondant, mais s'avère très rapide si l'outil permet la complétion automatique en temps réel

- Pour cette présentation
  - Toutes les expérimentations ont été faites avec ANALEC
  - Désormais : plusieurs outils en parallèle, et notamment TXM

1. Stratégie systématique

- Base d'expérimentation
  - Textes littéraires, narratifs, écrits en français contemporain (débuts de romans libres et gratuits, sur wikisource) + 1 bloc de *L'Est républicain*
  - Blocs d'environ 10 000 mots (entre 8 500 et 13 000, la variation nécessitant une pondération pour les études textométriques mais *a priori* sans conséquences pour l'étude comparative des chaînes)
  - Méthode d'échantillonnage : 10 000 premiers mots du texte (début) si on en prend un seul extrait ; début et fin si on garde deux extraits ; début, milieu et fin si on garde trois extraits

- Choix d'annoter toutes les expressions référentielles
  - Avantage 1 : le travail de délimitation des expressions référentielles demande peu de réflexion (aspect « robotique »), et l'annotateur se concentre sur la résolution de la référence
  - Avantage 2 : le corpus est complet (pour le problème de la référence) et permet des applications textométriques et TAL qu'un filtrage des référents ne permet pas
  - Inconvénient : tâche d'annotation supposée beaucoup plus longue

2. Stratégie avec filtrage des référents

- 2a. Annoter seulement les référents d'un certain type
  - Avantage 1 : l'annotateur se focalise sur un objet d'étude (l'annotateur peut plus facilement être le chercheur intéressé par les données – mais est-ce vraiment souhaitable ?)
  - Avantage 2 : la tâche est un peu plus rapide
  - Inconvénient 1 : si les types retenus sont nombreux, on a tendance à se mélanger un peu, et, dans tous les cas, la tâche de délimitation demande plus d'attention que pour la stratégie systématique
  - Inconvénient 2 : corpus incomplet pour l'étude de la référence
  - Inconvénient 3 : types variables d'un genre textuel à l'autre, donc l'homogénéité du corpus n'est plus assurée

2. Stratégie avec filtrage des référents

- 2b. Annoter seulement les référents repris
  - Avantage : on a l'impression de ne pas perdre du temps avec les expressions référentielles isolées (« singletons ») et, de fait, la procédure est plus adaptée au repérage des chaînes
  - Inconvénient 1 : il n'est pas possible de savoir à coup sûr si l'expression en cours d'annotation va être reprise ultérieurement ou non. Autrement dit, des oublis sont possibles…
  - Inconvénient 1bis : la tâche peut comporter des retours en arrière dans le texte si l'annotateur s'aperçoit qu'il a oublié une expression (initialement considérée comme singleton et donc ignorée)
  - Inconvénient 2 : se demander à chaque expression si elle a des chances d'être un singleton ou de faire partie d'une chaîne va à l'encontre de l'aspect « robotique » et efficace de la délimitation (se poser trop de questions est parfois contre-productif, et il vaut mieux tout annoter en se posant moins de questions)

Au final

- Stratégie systématique largement préférée
  - Merci de commencer à annoter en suivant la stratégie systématique
  - Pour certaines parties du corpus, on s'autorisera (dans un 2e temps) une stratégie avec filtrage
  - Eventuellement, on pourrait autoriser une stratégie à deux passes :
    - Première passe avec le filtrage, pour aller à l'essentiel,
    - Deuxième passe systématique…
    - …mais deux passes sont parfois plus difficiles à faire qu'une seule…
- Singletons
  - Pour les annoter plus rapidement, on autorise que le champ REF comprenne la valeur « SI » (comme « singleton »)
  - Cela permet de ne pas perdre de temps à trouver un identifiant…
  - …qui de toute façon sera inutile car remplacé automatiquement par SI, de manière à homogénéiser les annotations des singletons dans l'ensemble du corpus

Expérimentations chronométrées

9 blocs annotés (dont 6 pour le corpus Democrat)

- Totaux : 33 000 mentions annotées pour 3 000 chaînes
- Rythme moyen, pour l'ensemble = 550 mentions par jour
  ou 300 chaînes par jour

Textes	Annotatrice	Nombre de mots	Nombre de mentions	Nombre de chaînes
Bouvard et Pécuchet	Juliette Potier	10 086	4280	284
Nemoville	Juliette Potier	12 992	3252	286
De la ville au Moulin	Juliette Potier	10 101	4187	292
The Portrait of Dorian Grey	Michelle Bruni	10 723	3708	304
Le ventre de Paris	Juliette Potier	10 037	3147	329
La capitaine Fracasse	Juliette Potier	8 289	3075	347
La morte amoureuse	Michelle Bruni	12 177	4226	354
L'Est Républicain (2002)	Meryl Bothua	10 360	2626	393
Sarrasine	Michelle Bruni	12 787	4406	399

Avec deux annotateurs : 7 expérimentations de 30 minutes

- Présentées dans l'ordre chronologique…

Textes	Annotateur 1	Annotateur 2	Stratégie suivie
Le collier des jours	95 – 386 mots	93 – 541 mots	Systématique
Boule de suif	100 – 392 mots	78 – 348 mots	Systématique
L'enfant	114 – 338 mots	111 – 345 mots	Systématique
La recherche de l'absolu	85 – 275 mots	80 – 301 mots	Systématique
Manon Lescaut	120 – 397 mots	100 – 390 mots	Objets, humains et animaux
Douce lumière	105 – 543 mots	81 – 311 mots	Systématique avec chunks
Le Capitaine Fracasse	130 – 410 mots	141 – 510 mots	Systématique

Bilan

- Trop de variabilité…
  - Il est difficile de conclure quoi que ce soit de fiable
  - Surtout que la concentration de l'annotateur à l'instant t peut varier. C'est pourquoi 30 minutes semble une durée à ne pas dépasser : elle encourage une concentration en continu – mais cela requiert aussi de ne pas faire plus de deux expérimentations par jour…
  - Néanmoins, les résultats nuancent la plus grande rapidité supposée de la stratégie avec filtrage

Calcul de l’accord inter-annotateurs

Un problème complexe

- Remis en question récemment par les concepteurs de l'outil Glozz
  - The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment (revue *Computational Linguistics*)
  - Évaluation des annotations : ses principes et ses pièges (revue *TAL*)
  - Problème de l'accord « faible » pour des tâches complexes comme celles liées à l'anaphore et à la coréférence : Inter-Coder agreement for Computational Linguistics (R. Artstein & M. Poesio, article de 2008 dans la revue *Computational Linguistics*).

- Deux préoccupations pour l'évaluation
  - *Unitizing*
  - *Categorizing*

  - Pour notre champ REF : encore un autre cas de figure !

Quelques chiffres pour Democrat

- Au tout début, avant que les deux annotatrices ne se mettent d'accord sur certains détails de l'annotation (corrigés depuis dans le manuel d'annotation)

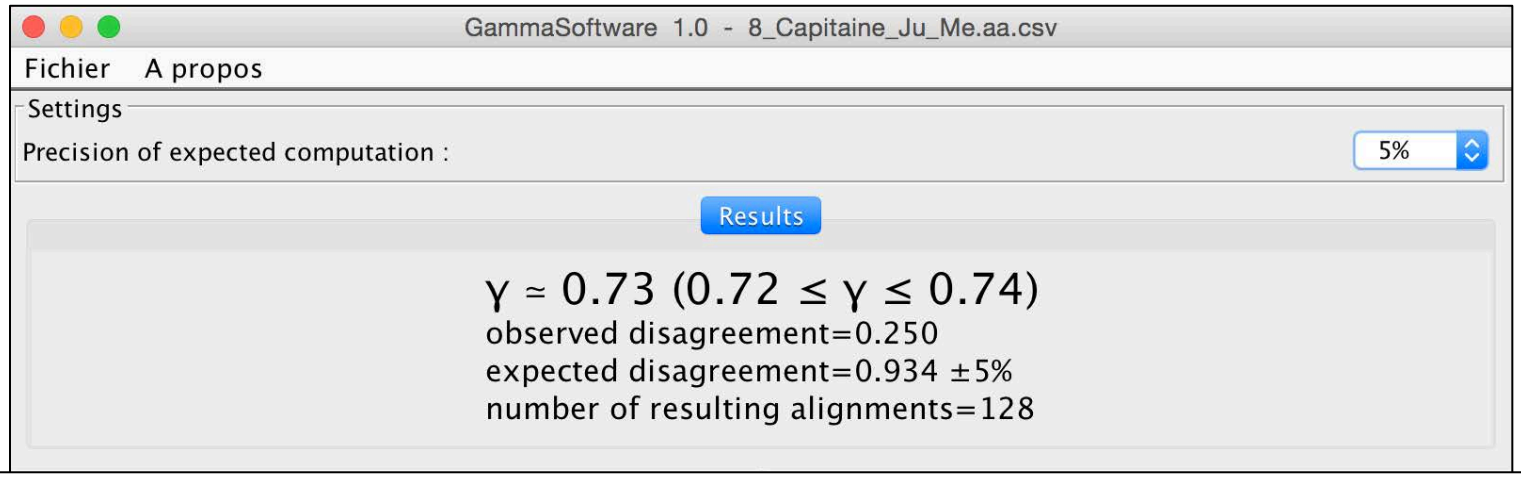
ReCal 0.1 Alpha for 2 Coders
results for file "fichier-input-recal2.csv"

File size:	177 bytes
N columns:	8
N variables:	4
N coders per variable:	2

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha (nominal)	N Agreements	N Disagreements	N Cases	N Decisions
Variable 1 (cols 1 & 2)	75%	0.68	0.692	0.72	3	1	4	8
Variable 2 (cols 3 & 4)	50%	0.385	0.429	0.462	2	2	4	8
Variable 3 (cols 5 & 6)	75%	0.68	0.692	0.72	3	1	4	8
Variable 4 (cols 7 & 8)	75%	0.529	0.556	0.588	3	1	4	8

- Calcul du γ : 0.53

Au final



ReCal 0.1 Alpha for 2 Coders
results for file "Accord_Recal_Capitaine_Fracasse.csv"

File size:	2928 bytes
N columns:	6
N variables:	3
N coders per variable:	2

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha (nominal)	N Agreements	N Disagreements	N Cases	N Decisions
Variable 1 (cols 1 & 2)	73.8%	0.732	0.733	0.733	96	34	130	260
Variable 2 (cols 3 & 4)	66.9%	0.66	0.661	0.662	87	43	130	260
Variable 3 (cols 5 & 6)	71.5%	0.318	0.319	0.321	93	37	130	260

Panorama des pré-annotations envisageables

Pourquoi des pré-annotations ?

- Pour plus d'efficacité
  - Aller plus vite grâce à une automatisation (partielle) de la procédure
  - Aider les aspects « robotiques » de la procédure
  - Aider l'annotateur en lui proposant une base de travail plutôt que de partir d'une feuille blanche

- Le mieux : un système de détection automatique des expressions référentielles et des chaînes de référence
  - Pour le français, il n'existe à l'heure actuelle que RefGen, dont le taux d'erreur est important → trop de corrections à faire (problématique car l'annotateur a vite l'impression de passer son temps à corriger plutôt qu'à exploiter directement les pré-annotations)

- A défaut : un système de détection automatique des expressions référentielles
  - Mais ça n'existe pas, en tout cas pas sous la forme qu'on souhaite…

Ce qui s’en rapproche le plus

- Système de reconnaissance des entités nommées
  - Repère et annote les noms propres, les expressions désignant des lieux, des dates, etc.
  - Une annotation : le type d’entité reconnue (« humain »…)
  - Mais ignore tous les pronoms et tous les groupes nominaux sans nom propre ou mot spécifique
- Système de résolution des anaphores
  - Repère et annote les pronoms anaphoriques
  - Annotation : lien avec un antécédent, donc quelque chose que l’on pourrait exploiter pour les chaînes
  - Mais ignore tous les noms propres et tous les groupes nominaux
- Les deux ensemble
  - Il manque toujours la majorité des groupes nominaux

Autres pré-annotations envisageables

- Annotations déjà disponibles dans certains corpus, que l'on pourrait reprendre comme éléments de notre corpus
  - Morphosyntaxe
  - Syntaxe
  - …
  - Base de travail utile mais très éloignée de la détection des expressions référentielles
  - Même si c'était exploitable, le risque serait un manque d'homogénéité des annotations dans le corpus

- Dernière solution envisageable :
  - Système de détection des *chunks* nominaux
  - *Chunk* = plus petite séquence d'unités linguistiques possible formant un groupe avec une tête forte, et qui n'est ni discontinue, ni récursive

Pré-annotation en *chunks* : aide ou gêne ?

Chunks nominaux

- Avantages
  - Utiliser un *chunker* nominal permet de délimiter un grand nombre de groupes nominaux et de pronoms. C'est une façon de contourner les inconvénients (pour notre tâche) des systèmes de reconnaissance des entités nommées
  - Annotation ajoutée : le type de *chunk* – en gros GN ou pronom
  - Il en existe beaucoup, et parmi eux SEM a de très bonnes performances (et s'applique à l'oral transcrit)

- Inconvénients
  - Pas d'enchâssement : « le N de N » conduit à l'identification soit d'un seul *chunk*, soit de deux *chunks* non enchâssés. Dans les deux cas, une rectification des bornes est nécessaire
  - Tous les pronoms sont repérés, même les impersonnels
  - Parmi les GN repérés, certains sont bien sûr non référentiels
  - Comme tout système de TAL, il y a des erreurs

Expérimentations effectuées

- 1. Pas de pré-annotation en *chunks* nominaux
- 2. Pré-annotation laissée telle quelle
  - L'annotateur ne corrige pas les erreurs du *chunker*
  - Il ajoute (délimite) les *chunks* non repérés qui sont référentiels
- 3. Pré-annotation aux délimitations corrigées
  - L'annotateur corrige les erreurs de bornes (modifieur ignoré, etc.)
  - Il ajoute (délimite) les *chunks* non repérés qui sont référentiels
- 4. Pré-annotation totalement corrigée
  - L'annotateur corrige les erreurs de bornes
  - Il corrige les erreurs de catégorisation (type de *chunk*)
  - Il ajoute (délimite) les *chunks* non repérés, référentiels ou non

Au final

- L’annotateur corrige les erreurs de bornes
- L’annotateur ajoute les expressions référentielles non repérées
- Et c’est déjà pas mal…
- Dans tous les cas, l’annotateur garde le choix d’utiliser ou non un *chunker*

Bilan et perspectives

- Tout ceci n'a qu'un but : rendre l'annotation la plus rapide et la plus efficace possible, tout en gardant des possibilités intéressantes d'exploitation des données annotées
  - Exploitation linguistique
  - Exploitation textométrique
  - Exploitation TAL

- Quelques mots sur la procédure d'annotation
  - Tout est fait pour minimiser l'annotation manuelle et favoriser une annotation automatique dès qu'elle peut s'envisager
  - D'où : alternance rationnelle de phases manuelles et de lancements de scripts – procédure optimale, mais pas forcément facile à comprendre au premier abord…

Références bibliographiques

- Artstein, R., Poesio, M. (2008). Inter-Coder agreement for Computational Linguistics, *Computational Linguistics, 34*, 555-596.
- Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 22*, 249-254.
- Charolles, M. (2002). *La référence et les expressions référentielles en français.* Paris : Ophrys.
- Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues, 55(2),* 97-121.
- Heiden, S., Magué, J.-P., Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In: *Proceedings of Tenth International Conference on the Statistical Analysis of Textual Data*, Vol. 2, 1021-1032.
- Krippendorff, K. (2012). *Content analysis: an introduction to its methodology (third edition).* Thousand Oaks : Sage Publishing.
- Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus, 10*, 61-80.
- Landragin, F., Poibeau, T., Victorri, B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turquey, 357-362.
- Mathet, Y., Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. *Traitement Automatique des Langues, 57(2),* 73-98.
- Mathet, Y., Widlöcher, A., Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics, 41(3),* 437-479.
- Müller, C., Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (Eds.). *Corpus technology and language pedagogy: New resources, new tools, new methods*. Frankfurt : Peter Lang.
- Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J. (2014). ANCOR CENTRE, a large free spoken french coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.
- Nouvel, D., Ehrmann, M., Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues*. Londres : Éditions ISTE.
- Tellier, I., Duchier, D., Eshkol, I., Courmet, A., Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble.
- Todiraşcu, A., Longo, L. (2011). RefGen, outil d'identification automatique des chaînes de référence en français. 18e Conférence sur le Traitement Automatique des Langues Naturelles, session des démonstrations industrielles, Montpellier.
- Widlöcher, A., Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In : *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles*, Senlis.