De la coréférence exacte à la coréférence complexe : une typologie et sa mise en œuvre en corpus

Marine Delaborde, Frédéric Landragin

De la coréférence exacte à la coréférence complexe: une typologie et sa mise en œuvre en corpus

Marine Delaborde et Frédéric Landragin

Laboratoire Lattice
CNRS, ENS, Université de Paris 3, PSL Research University, USPC

DRAFT AUTEURS

marine.delaborde@ens.fr, frederic.landragin@ens.fr

L'annotation de la coréférence est une tâche délicate dès lors que l'on rencontre des expressions qui semblent référer à la même entité sans être exactement coréférentes. C'est un problème qui a été soulevé notamment par Recasens et al. (2011) pour le concept de near identity avec une catégorisation de différents types de relations de coréférence dans lesquelles les référents sont proches sans être exactement les mêmes.

Lorsque plusieurs expressions référentielles désignent le même référent, elles sont coréférentes et forment une chaîne de coréférence. Il s'agit principalement de noms communs, de noms propres et de pronoms. Les chaînes de coréférence peuvent s'étendre du début à la fin d'un texte, comme pour un personnage principal dans un roman par exemple, mais elles sont le plus souvent courtes et ponctuelles. Pour qu'il y ait référence, il faut pouvoir identifier un référent qui existe dans le monde ou que l'on peut se représenter (Charolles, 2002). Cependant, il arrive parfois que le référent d'une expression soit difficile à identifier de manière précise. Dans ce cas, la coréférence avec une autre expression référentielle dont on a identifié clairement le référent ne peut pas être stricte. C'est le cas par exemple des anaphores à antécédent flou, à ne pas confondre avec de l'ambiguïté (Fuchs, 1996).

> Elle parle aussi avec une sentimentalité criante. Ma sœur et moi on l'arrête. On l'arrête à temps. Alors elle dit on ne me laisse pas parler ici. Mais ce ne sont pas des paroles qu'on a envie d'entendre, je ne sais pas pourquoi.
> AKERMAN Chantal, Ma mère rit, 2013

Dans cet exemple, les deux premiers on coréfèrent et renvoient à l'antécédent Ma sœur et moi de manière évidente et stricte. En revanche, le troisième on, qui est du discours rapporté, et le dernier on coréfèrent chacun de manière floue aux deux premiers car ils peuvent très bien renvoyer au même antécédent. Cependant le doute persiste en raison du fait qu'ils peuvent tous les deux avoir aussi une valeur générique. Il n'est donc pas

raisonnable d'en faire une seule chaîne, mais les dissocier totalement ferait perdre une information.

L'annotation de la coréférence en corpus implique de faire des choix, notamment au niveau des relations. Le projet ACE (Doddington et al., 2004) distingue 5 types de relations entre les mentions: rôle, partie, localisation, proche et sociale. Le projet OntoNotes (Pradhan et al., 2011) différencie la coréférence identique de la coréférence appositive. C'est aussi le cas dans le corpus WikiCoref (Ghaddar & Langlais, 2016) qui distingue les coréférences identiques, attributives et attributives dans des constructions copulatives. Dans le corpus Phrase Detectives (Chamberlain et al., 2016), les ambiguïtés référentielles sont annotées comme des alternatives avec des scores correspondant aux avis des annotateurs. Pour le polonais, Ogrodniczuk et al. (2015) distinguent les relations de coréférence identique et quasi-identique. Le projet ANCOR (Muzerelle et al., 2013), qui traite du français oral, préconise de caractériser les relations entre les mentions selon 5 types: directe, indirecte, pronominale, associative et associative pronominale. Ces corpus distinguent donc différents types de relations entre les mentions mais toujours de manière stricte.

Annoter les phénomènes de coréférence non stricte et floue permettrait d'obtenir des chaînes de coréférence qui reflètent mieux le texte en gardant l'information que certaines expressions semblent correspondre à une chaîne sans coréférer à ses maillons de manière stricte. C'est pourquoi nous proposons un schéma d'annotation répertoriant trois catégories de coréférence qui prend en compte la coréférence non stricte. La première catégorie correspond à la coréférence exacte: lorsque les mentions réfèrent exactement et sans aucun doute au même référent, comme c'est le cas dans le manuel d'annotation du projet Democrat (Landragin, 2016) par exemple. La seconde catégorie correspond à la coréférence inclusive, elle comprend les cas où un référent en inclut complètement un autre, de manière stricte ou de manière floue. La troisième catégorie correspond à la coréférence intersective. Il s'agit des cas où la coréférence n'a lieu que sur l'intersection de deux référents, de manière stricte ou de manière floue.

Le schéma d'annotation proposé s'inspire du schéma proposé par le projet Democrat, bien que les relations ne soient pas annotées. Il correspond au modèle d'annotation de type Unités-Relations-Schémas (URS), développé à l'origine dans le logiciel Glozz (Widlcher & Mathet, 2009) et implémenté dans le logiciel Analec (Landragin, 2012) puis par extension dans le logiciel TXM (Heiden, 2010). L'annotation du type de coréférence se fait au niveau des relations pour chaque couple de mentions. Chaque catégorie est représentée par un trait: exacte, inclusive ou intersective. Pour les catégories de coréférence inclusive et intersective, il est possible de choisir entre les deux propriétés: stricte et floue. Ce schéma est en cours de test sur une partie du projet Democrat dans le but d'effectuer une comparaison avec un texte déjà annoté en coréférence stricte. Ce schéma d'annotation sera décrit dans un

manuel d'annotation qui pourra répertorier les critères caractérisant chaque catégorie ainsi que des exemples issus du corpus. Pour valider ce schéma, des tests avec de la double annotation sont aussi prévus.

Références bibliographiques

Chamberlain, J., Poesio, M., & Kruschwitz, U. (2016). Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2039?2046. Portoro, Slovenia.

Charolles, M. (2002). *La référence et les expressions référentielles en français.* Ophrys.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 837?840. Lisbon, Portugal.

Ghaddar, A., & Langlais, P. (2016). *WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles.* Présenté à Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portoro, Slovenia.

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM?: Une plateforme logicielle open-source pour la textométrie - conception et développement. *Proc. of 10th International Conference on the Statistical Analysis of Textual Data*, 2, 1021?1032. Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'AFIA*, (92), 11?15.

Landragin, F., Poibeau, T., & Victorri, B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. European Language Resources Association (ELRA). *International Conference on Language Resources and Evaluation*, 357?362. Istanbul, Turkey.

Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2011). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA (éd.), *20e conférence sur le Traitement Automatique des Langues Naturelles* (p. 555?563). Les Sable d'Olonne, France: ATALA.

Ogrodniczuk, M., Glowinska, K., Kopec, M., Savary, A., & Zawislawska, M. (2014). *Coreference: Annotation, Resolution and Evaluation in Polish.* Walter de Gruyter GmbH & Co KG.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning?: Shared Task*, 1-27. Portland, Oregon, USA.

Recasens, M., Hovy, E., & Mart, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6), 1138-1152.

Widlcher, A., & Mathet, Y. (2009). La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence Traitement Automatique des Langues Naturelle, session posters*, 10. Senlis, France.