Résolution d'anaphores dans une encyclopédie en langue anglaise : conception, implémentation et évaluation des performances

François-Régis Chaumartin – fchaumartin@linguist.jussieu.fr
Laboratoire LATTICE – Université Paris 7

La résolution d'anaphores est un problème ouvert en TAL. Sa complexité provient du fait qu'elle nécessite des connaissances de plusieurs niveaux, ainsi qu'une « compréhension » du contexte. Dans le cadre d'un projet d'extraction de connaissances encyclopédiques, nous avons mis en œuvre un système complet de résolution d'anaphores et d'identification de chaînes de coréférence. Nous utilisons simultanément des techniques classiques, pauvres en connaissances, et des outils linguistiques évolués (analyse syntaxique en profondeur et lexique sémantique). L'ensemble offre des performances prometteuses dans le cadre d'articles encyclopédiques ; l'ajout prochain d'une heuristique statistique supplémentaire, basée sur la disponibilité récente d'une ressource de large couverture, devrait permettre de les améliorer encore.

1 Complexité de la résolution d'anaphores

Une *anaphore* est un mot ou un syntagme qui, dans un énoncé, assure une reprise sémantique d'un précédent segment appelé *antécédent*. L'utilisation d'anaphore permet d'éviter une répétition lexicale (« L'Amazone est un fleuve très long, seul le Nil **le** dépasse en longueur »)[1]. La résolution d'anaphores (ainsi que son prolongement, l'identification des chaînes de coréférence) est un problème complexe en linguistique (pour la modélisation des phénomènes entrant en jeu) et en TAL (pour l'implémentation de ces modèles et la constitution de ressources électroniques). Elle fait appel, selon le cas de figure, à des connaissances de nature lexicale, syntaxique, sémantique et pragmatique, ainsi qu'à une compréhension du contexte, pour lever les éventuelles ambiguïtés. Par exemple, les trois phrases suivantes partagent la même construction syntaxique ; seul l'adjectif final varie, produisant à chaque fois une interprétation différente du pronom **ils** :

> Les gardiens ont donné les fruits aux singes parce qu'**ils** étaient *pourris*.
> Les gardiens ont donné les fruits aux singes parce qu'**ils** étaient *affamés*.
> Les gardiens ont donné les fruits aux singes parce qu'**ils** étaient *rassasiés*.

Une littérature vaste existe autour de ce sujet, proposant de classifier ces phénomènes (anaphore pronominale, nominale, associative…)[2] ou d'en proposer des modélisations[3].

2 Méthodes classiques de résolution d'anaphores pronominales

Les méthodes de résolution d'anaphores pronominales ne sont pas récentes. Citons notamment :

---

[1] Les exemples sont annotés avec les anaphores en gras et les antécédents identifiés en souligné.

[2] Voir par exemple (Salmon-Alt, 2002) ou (Kleiber, 1994).

[3] Citons les notions de commande (Langacker, 1969) et de c-commande (Reihart, 1983) et la théorie du liage.

- (Hobbs, 1978) propose un algorithme utilisant l'analyse syntaxique d'un texte.[4]
- (Lappin & Leass, 1994) proposent un algorithme en plusieurs étapes[5], nécessitant une analyse syntaxique, et revendiquent une précision de 86% sur un corpus technique en anglais (avec une analyse syntaxique corrigée manuellement).
- (Mitkov, 1998) présente une approche « robuste et pauvre en connaissance », basée sur plusieurs heuristiques[6] qui exploitent une analyse syntaxique superficielle. Evaluée sur un corpus de manuels techniques (en anglais, polonais et arabe), cette approche donne une précision de l'ordre de 90% à 91%.

3 Notre système dans le cadre d'articles encyclopédiques en anglais

Les articles d'encyclopédies ont quelques caractéristiques qui facilitent leur analyse automatique : ils sont (généralement) correctement écrits, avec un style concis, sans humour ; ils relatent des faits, avec des temps de verbe le plus souvent au passé. Les anaphores étant fortement présentes dans de tels articles, leur résolution est indispensable si on souhaite parvenir à une représentation sémantique correcte d'un article. Ces anaphores sont (majoritairement) pronominales, et portent (le plus souvent) sur le titre de l'article, c'est-à-dire son sujet.[7]

Notre projet, en cours de réalisation[8], vise à extraire des connaissances d'une encyclopédie en langue anglaise, puis de les représenter sous forme de graphes conceptuels. L'un des modules utilisés dans la chaîne de traitement[9] concerne la résolution d'anaphores et l'identification de chaînes de coréférence, et tire parti de ces caractéristiques.

L'architecture macroscopique de fonctionnement est classique :

1. Analyse du texte (au choix : simple étiquetage morphosyntaxique ; analyse syntaxique superficielle ; analyse syntaxique en profondeur[10]),
2. Parcours du texte
    a. Détection des pronoms personnels et possessifs,
    b. Détermination du caractère « anaphorique » du pronom, par élimination des *it* pléonastiques (« *It is possible that…* »)[11] et impersonnels (« *It rains…* »),

---

[4] Etant donné un pronom dans une phrase, l'algorithme effectue un parcours de l'arbre syntaxique de la phrase (et éventuellement de la phrase précédente) à la recherche de son antécédent.

[5] Identification des pronoms pléonastiques ; algorithme de liage identifiant l'antécédent d'un pronom réfléchi ou réciproque dans la même phrase ; assignation d'une valeur de saillance pour chaque syntagme nominal.

[6] Une heuristique est une règle qu'on a intérêt à utiliser en général, parce qu'on sait qu'elle conduit souvent à la solution, bien qu'on n'ait aucune certitude sur sa validité dans tous les cas.

[7] Nous pourrions nous risquer à proposer comme algorithme non-subtil de résolution d'anaphores dans le texte de l'article, un brutal *recherche & remplace* du pronom qui y apparaît le plus fréquemment, par son titre.

[8] Nous visons à disposer fin 2008 d'une indexation sémantique de 15 000 articles de la Wikipedia en anglais.

[9] La chaîne de traitement se base sur la boîte à outil pour l'anglais : aNteLoPe (http://www.proxem.com).

[10] En utilisant le *Link Grammar Parser* ou le *Stanford Parser* qui produisent un graphe de dépendances.

[11] Approximativement 3% des « *it* » (mesure sur 20 articles choisis au hasard dans la Wikipedia en anglais).

3. Pour les pronoms retenus comme « anaphoriques »
    a. Marquage des différents antécédents candidats,
    b. Vérification des contraintes syntaxiques (notamment de c-commande),
    c. Vérification de l'accord en genre et en nombre,
    d. Application de différentes heuristiques qui augmentent ou diminuent la saillance de chaque candidat ; celui présentant la saillance la plus élevée est retenu.
4. Extraction des chaînes de coréférences (composantes connexes du graphe des anaphores).

La conception de notre système est basée sur des bonnes pratiques de génie logiciel. Les interfaces de programmation sont modélisées indépendamment de leur implémentation.

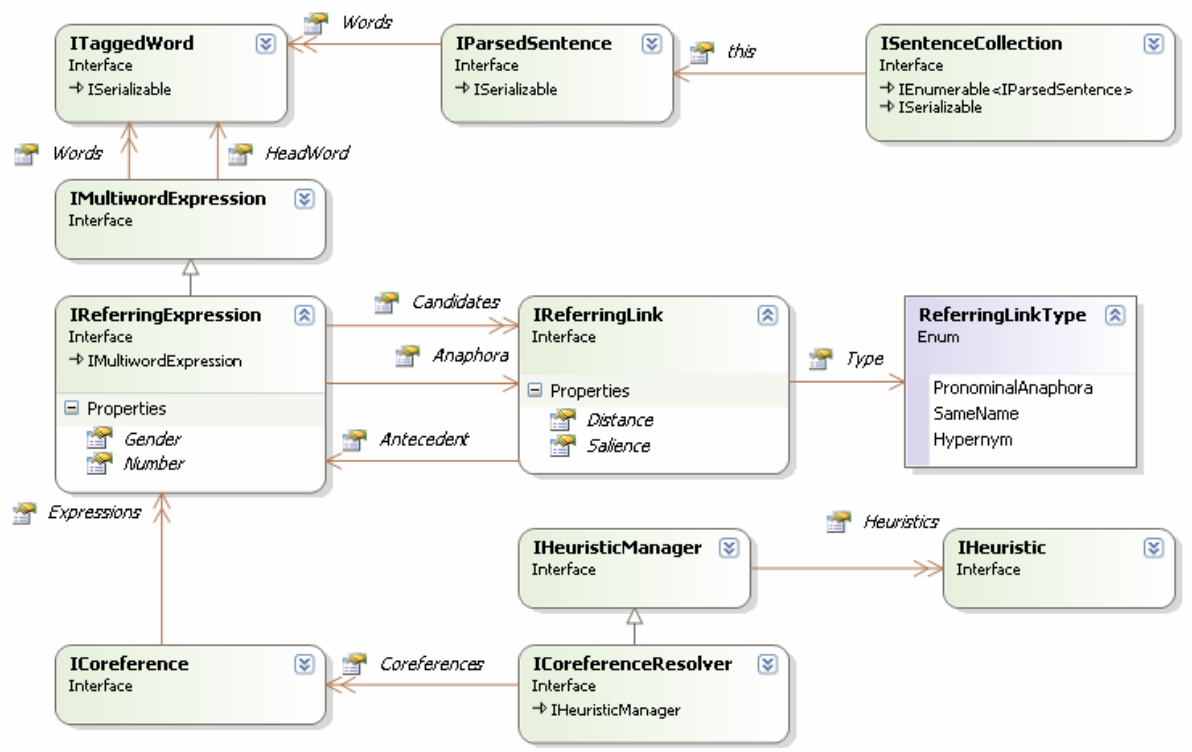

Figure 1 : le modèle de programmation par interfaces.

L'implémentation des heuristiques évoquées en 3.d. est le cœur de traitement du système. Nous avons basé notre première implémentation sur les heuristiques de Mitkov[12]. Toutefois, disposant d'une boîte à outil autorisant une analyse syntaxique en profondeur, nous avons pu y ajouter certaines des caractéristiques de l'algorithme de Lappin & Leass.

Nous avons également implémenté une résolution des anaphores nominales, qui permet par exemple d'identifier correctement l'anaphore dans : « *As Abraham Lincoln sat in the balcony, Booth crept up behind the* ***President****'s box…* ». Ce sous-module utilise le lexique sémantique WordNet, sur lequel une mesure de similarité[13] a été définie sur la hiérarchie de noms.

---

[12] Obliqueness, definiteness, lexical reiterations, section heading, referential distance, boost pronouns, collocation match, parenthesis…

[13] Proche de celle décrite dans (Lin, 1998), et utilisant la distance entre deux synsets dans la hiérarchie de noms.

4 Evaluation & perspectives

The longest **river** in the world, **it** is about 4132 miles (6650 km) long from **its** remotest headstream and 3473 miles (5588 km) from Lake_Victoria to the **Mediterranean_Sea**. **It** flows generally north from eastern Africa through Uganda, The Sudan, and **Egypt**. **It** receives major tributaries, including the Blue_Nile and the **Atbara_River**, before entering **Lake_Nasser** near the Egypt-Sudan border. After the **Aswan_High_Dam** impounds the **lake**, **it** continues northward to **its** delta near Cairo, where **it** empties into the **Mediterranean**. The first use of the **Nile** for irrigation in **Egypt** began when seeds were sown in the mud left after **its** annual floodwaters had subsided. **It** has supported continuous human settlement for at_least 5000 years, with canals and waterworks built in the 19th century. The **Aswan_High_Dam**, built in 1959-- 70, provides flood protection, hydroelectric power, and a dependable water_supply for both crops and humans. The **Nile** is also a vital waterway for the transport of people and goods.

Figure 2 : un exemple d'identification des chaînes de coréférences sur un article portant sur le Nil. (Chacune d'entre elles est affichée avec une couleur distincte.)

Nous avons évalué les performances du système sur 11 articles d'encyclopédie. Sur 47 anaphores relevées lors d'une annotation manuelle, 46 ont été détectées correctement (l'un des pronoms a été incorrectement étiqueté comme pléonastique). L'antécédent a été correctement trouvé dans 43 cas. Nous avons donc, sur ce jeu de tests réduit, une précision de 93% et un rappel de 97%.

Nous prévoyons d'améliorer ce système en utilisant les n-grammes rendus publics en septembre 2006 par Google. Cette ressource (obtenue à partir d'un corpus Web de 1000 milliards de mots) donne les fréquences de toutes les combinaisons de 5 mots anglais (apparaissant plus de 40 fois dans le corpus). Cette ressource devrait nous permettre d'évaluer la probabilité de substitution de l'anaphore par chaque antécédent possible, et de pondérer la saillance en conséquence.

Bibliographie

van Deemter K. & Kibble R. (2000) On Coreferring: Coreference annotation in MUC and related schemes. *Computational Linguistics* 26(4), pp. 615-623.

Hobbs J. (1978) Resolving Pronoun References. *Lingua*, 44 : 311-338.

Kleiber G. (1994) *Anaphores et pronoms*. Duculot, Louvain-la-Neuve.

Langacker R. (1969) On pronominalization and the chain of command. In Reibel and Schane (eds.) *Modern studies in English*, 160-186.

Lappin S. & Leass H.J. (1994) An algorithm for pronominal anaphora resolution, *Computational Linguistics*, 535-561.

Lin D. (1998). An information-theoretic definition of similarity. Actes de *15th International Conf. on Machine Learning*, 296–304.

Miller G. (1995) Wordnet: A lexical database. Actes de *ACM 38*, pp. 39-41.

Mitkov R. (1998) Robust pronoun resolution with limited knowledge, *COLING-ACL*, Montréal.

Salmon-Alt S. (2002) Le projet Ananas : l'annotation anaphorique pour l'analyse de corpus sémantiques. Actes du *Workshop CRAA - TALN 2002*, Nancy.