TALN 2004, Session Poster, Fès, 19-21 avril 2004

Résolution automatique d'anaphores infidèles en français : Quelles ressources pour quels apports ?

Susanne Salmon-Alt

ATILF – CNRS (UMR 7118)
44, Avenue de la Libération, B.P. 30687, 54063 Nancy Cedex
Susanne.Salmon-Alt@atilf.fr

Résumé – Abstract

La performance d'une résolution automatique d'anaphores infidèles pour le français pourrait atteindre une F-mesure de 30%. Ce résultat repose toutefois sur une ressource équivalente à un bon dictionnaire de la langue française, une analyse syntaxique de qualité satisfaisante et un traitement performant des entités nommées. En l'absence de telles ressources, les meilleurs résultats plafonnent autour d'une F-mesure de 15%.

A system for solving indirect anaphora in French seems to be able to achieve a F-measure of 30%. However, this result supposes a high quality lexical database (equivalent to a classical dictionary), a good parser and a high precision named entity recognition. In case such resources are not available, the best results are obtained by using simple heuristics and are limited to a F-measure of 15%.

Mots Clés – Keywords

anaphore infidèle, ressource sémantique, résolution d'anaphore, corpus annoté multiniveau

indirect anaphor, lexical database, anaphora resolution, multi-level corpus annotation

1 Problématique

Par anaphore « infidèle », nous entendons, dans la lignée de la linguistique descriptive française, un groupe nominal anaphorique (dont l'interprétation dépend d'une expression nominale du contexte précédent) et coréférentiel (dont le référent est identique à celui de l'antécédent), caractérisé par une tête nominale différente de celle de l'antécédent (*le déficit en pluie – la sécheresse*). Ces reprises partagent la coréférentialité avec d'autres types d'anaphores (pronominales, nominales « fidèles »), mais leur compréhension s'appuie sur des connaissances sémantiques et pragmatiques au-delà des mécanismes mis en jeu lors de l'interprétation des pronoms personnels et reprises directes. S'il a été montré pour différentes langues (anglais, portugais, français) que les définis en anaphore infidèle fournissent en moyenne 25 à 50% des descriptions définies coréférentielles (Poesio et Vieira, 1998 ;

Salmon-Alt et Vieira, 2002), au maximum un tiers parmi eux peuvent être résolus (Poesio et al., 2002), même dans les meilleures configurations s'appuyant sur les ressources et outils à la fois accessibles et correspondant à l'état de l'art anglo-saxon. L'objectif de ce travail est d'évaluer les performances que l'on peut obtenir pour la résolution automatique des anaphores infidèles en comparant trois types des ressources sémantiques différentes.

2 Corpus de test

Les expériences ont été conduites sur un extrait du corpus *Le Monde* (~ 9000 mots, 276 phrases). En préparation l'annotation manuelle des anaphores, ce corpus a été étiqueté et normalisé au niveau structurel, morphologique (Schmid, 1994) et syntaxique (Bick, 2003). Du corpus résultant ont été extraits automatiquement les antécédents potentiels (tous les groupes nominaux : 1924) et les anaphoriques potentiels (groupes nominaux définies : 741).

Les 741 groupes nominaux définis restants ont été soumis, par l'intermédiaire d'un outil à interface graphique (Müller et Strube, 2001) à deux annotateurs pour l'annotation des relations coréférentielles. Sur le total des définis en entrée, respectivement 256 et 247 ont été considérés comme coréférentiels. Ce taux (autour de 30%) est comparable à celui de travaux antérieurs (Poesio et Vieira, 1998 ; Salmon-Alt et Vieira, 2002). En revanche, la répartition entre reprises fidèles vs. infidèles variait de 23 points entre les deux annotateurs. En raison de cette disparité importante, nous avons décidé de retenir uniquement l'intersection des descriptions définies infidèles, soit 78 descriptions définies. Ce nombre, bien que ne représentant plus 10,5% des 741 définis en entrée, constitue donc un noyau sûr de descriptions définies à reprise coréférentielle infidèle : 85,9% se sont d'ailleurs vu attribuer le même antécédent par les deux annotateurs. Le Tableau 1 propose une synthèse des principales caractéristiques des anaphores retenues : une part importante d'antécédents sous forme de noms propres (plus d'un tiers), l'absence de régularité sur l'accord en nombre et genre entre antécédent et reprise et une distance entre antécédent et anaphore variant entre 0 et 27 phrases.

annotateur	tête de l'antécédent			accord		distance phrastique	
	nom commun	nom propre	autre	nombre	genre	minimale	maximale
A1	55,1 % (43)	37,2 % (29)	7,7 % (6)	57,7 % (45)	34,6 % (27)	0	17
A2	53,8 % (42)	39,7 % (31)	6,4 % (5)	57,7 % (45)	30,8 % (24)	0	27

Tableau 1 : Caractéristique des couples *antécédent – anaphore infidèle*

3 Ressources : Présentation et rendement en situation hors bruit

Ensuite, nous avons testé l'utilité de trois types de ressources lexicales pour la résolution automatique des anaphores infidèles en calculant le nombre de paires *antécédent – anaphore* figurant dans au moins un des lexiques. Menée hors contexte applicatif (donc hors bruit), cette expérience nous a permis d'évaluer le plafond du rappel et de fixer les valeurs limites des paramètres en entrée du résolveur.

La première des ressources testées est une ressource acquise de façon automatique à partir d'un corpus analysé syntaxiquement. Il s'agit d'une liste dite de « similarité sémantique », obtenue par des techniques basées sur des propriétés distributionnelles d'unités linguistiques en corpus (Grefenstette, 1994 ; Gasperin et al., 2001). En appliquant cette méthode sur un extrait de 90.000 mots du corpus *Le Monde* (comprenant les corpus de test), nous avons

obtenu 572 entrées avec 3322 termes proches (moyenne de 5,8 par entrée). Le deuxième lexique a été produit à partir du *EuroWordNet* français dont nous avons extrait 7856 *synsets* comprenant en moyenne 2,3 termes. La troisième ressource est issue d'un dictionnaire « classique ». Il s'agit d'une liste des synonymes extraits manuellement du *Grand Robert* donnant 32.484 entrées avec une moyenne de 4,7 synonymes par entrée.

Le rendement de ces ressources a été testé sur les 78 paires antécédent-reprise. L'entrée était fournie par les têtes non fléchies de l'antécédent (T_A) de la reprise (T_R). Dans 11 cas, T_R était ambigu entre deux expressions, puisque les deux annotateurs avaient choisi des antécédents différents, comme dans *les peines* reprenant, pour un annotateur, *sa condamnation à quatre ans de camp de travail* et pour l'autre *quatre ans de camp de travail*, avec T_{A1} = *condamnation*, T_{A2} = *an* et T_R = *peine*. Pour chaque paire (T_A,T_R) ainsi que les listes de synonymes associés à la tête de l'antécédent S_{TA} et celle de l'anaphore S_{TR}, nous avons considéré que la ressource aide à la résolution de l'anaphore si et seulement si : $T_A \in S_{TR}$ ou $T_R \in S_{TA}$ ou $T_A \in S_X \ni T_R$.

Ressource	SimLists	WordNet	Robert	Cumul ressources
SimLists	2	0	3	5
WordNet	0	0	2	+ 2
Robert	3	2	8	+ 8
Total ressources	5	2	13	15
% (avec NP)	6,4	2,5	16,7	19,2
% (sans NP)	8,5	4,3	27,7	29,8

Tableau 2 : Rendement des ressources

Ressource	Tête de l'antécédent			Accord		Distance phrastique	
	nom commun	nom propre	adjectif	nombre	genre	minimale	maximale
SimLists	80 %	20 %	0 %	100 %	80 %	0	3
WordNet	100 %	0 %	0 %	100 %	50 %	0	3
Robert	92,3%	0 %	7,7 %	92,3 %	46,2 %	0	5

Tableau 3 : Caractéristiques des couples *antécédent – anaphore infidèle* présent dans les ressources

Les résultats, pour chacune des ressources isolées, les ressources croisées et les ressources cumulées, sont donnés dans le Tableau 2. Il apparaît que l'ensemble des ressources ne contient que 15 des paires recherchées, soit un rappel plafonné à 19,2% avec et à 29,8% sans les antécédents sous forme de noms propres, l'apport de *EuroWordNet* étant entièrement neutralisé par le *Robert*. En vue de la détermination des paramètres du calcul automatique, nous avons analysé certaines propriétés linguistiques (accord en nombre et genre, distance phrastique, nature de la tête de l'antécédent) des cas couverts (Tableau 3).

4 Rendement des ressources en résolution automatique

4.1 Stratégie et paramètres

Afin d'évaluer l'apport des différentes ressources lexicales pour la résolution d'anaphores nominales infidèles en situation réelle (contrairement à une situation hors bruit telle que supposée dans la section 3), nous avons implémenté un système de résolution automatique. Il prend en entrée les descriptions définies classifiées comme anaphores infidèles par les deux annotateurs ainsi que des connaissances associées (identifiants de la phrase et de l'article

« hôte », fonction syntaxique, tête du groupe syntaxique, déterminant, nombre, genre). La stratégie de résolution consiste à chercher, pour chacune d'entre elles, un antécédent correspondant à la première expression en amont remplissant un certain nombre de contraintes, définies en fonction des caractéristiques mises en lumière dans le Tableau 3 :

L'imposition d'un accord en genre entre antécédent et anaphore semble peu adéquate (mois que 50% pour les paires trouvées dans le *Robert*). En contrepartie, l'effet d'un accord en nombre A_N pourrait s'avérer intéressant pour augmenter la précision et fera l'objet de variation systématique. Dans tous les cas, la fenêtre de recherche pour l'antécédent peut être limitée à cinq phrases en amont. Sachant par ailleurs que 93,3% des antécédents se trouvent à une distance phrastique inférieure à quatre phrases et que seulement 6,6% se trouvent dans la même phrase, on fera varier les contours de cette fenêtre afin de trouver le meilleur rapport entre rappel et précision : W_{max}= {S_{-5}, S_{-3}} ; W_{min} = {S_{-0}, S_{-1}}. Enfin, la nature de la tête de l'antécédent peut être restreinte à un nom commun, un nom propre ou un adjectif. Toutefois, les reprises de noms propres pourraient être traitées plus avantageusement par des outils de reconnaissance d'entités nommées, atteignant un taux de reconnaissance supérieur à 90%. (Fourour, 2002). Au vu de la quasi-absence de nom propres dans les ressources, ces cas risquent fort de pénaliser la précision. Pour cette raison, nous ferons également varier les contraintes sur la nature de la tête de l'antécédent (T_A = +/-nom propre).

Pour toutes les combinaisons de ces paramètres, les sorties du système ont été évaluées dans 4 configurations de ressources lexicales (R_L) – absence de ressource (*0*), liste de similarité (*S*), *EuroWordNet* (*WN*), *Robert* (*R*) – par le calcul du rappel, de la précision et de la F-mesure.

4.2 Résultats

Le point de comparaison pour évaluer l'utilité des ressources lexicales est l'application des heuristiques sans prise en compte des lexiques. Nous avons donc retenu comme réponse le premier groupe nominal remplissant toutes les contraintes spécifiées en 4.1. Le tableau 4 montre que les meilleurs résultats – 15% à 20% – s'obtiennent en imposant l'accord en nombre. Par ailleurs, l'espace de recherche peut être réduit aux trois phrases précédentes pour un gain en temps de calcul.

L'apport des listes de similarité a été testé dans les mêmes circonstances, en réduisant les réponses aux groupes pour lesquels les têtes (T_A,T_R) remplissant les contraintes introduites dans la section 3. Comme le montre le Tableau 4, on constate, même pour les meilleurs résultats, une baisse de performance de plus de 10 points par rapport à la configuration de base. Cette baisse s'explique à la fois par un rappel inférieur (le système trouve moins de bonnes réponses) et une précision inférieure (la part des fausses réponses augmente). L'inclusion des anaphores ayant des noms propres comme antécédents creuse encore cet écart. Si les résultats doivent être interprétés prudemment en raison des faibles effectifs, ils remettent tout de même en question l'acquisition automatique de listes de similarité sémantique pour la résolution anaphorique : en l'absence de ressources syntaxiques disponibles et d'une taille largement supérieure à la taille des corpus utilisés ici, cette solution présente un déséquilibre important entre efforts (préparation des corpus, annotation syntaxique, transfert de formats, implémentation des calculs de similarité) et effets (résultats largement moins bons que sans ressources).

Les résultats des mêmes expériences conduites sur la base de *EuroWordNet* montrent qu'il n'y a pas de changement en ce qui concerne le taux de rappel très bas. En revanche, la F-mesure est légèrement supérieure, et ceci en raison d'une précision élevée.

Notons toutefois que cette précision est essentiellement due à la couverture insuffisante de la ressource. Le principal constat reste donc une perte de performance de 10 points par rapport aux heuristiques sans ressources.

R_L	A_N	W_{min}	W_{max}	+ noms propres			- noms propres		
				R	P	F	R	P	F
0	non	0	5/3	0,077			0,106		
	non	1	5/3	0,090			0,106		
	oui	0	5/3	0,141			**0,191**		
	oui	1	5/3	**0,141**			0,149		
S	non	0/1	5/3	0,013	0,033	0,019	0,021	0,063	0,032
	oui	0	5	0,013	0,038	0,019	0,021	0,083	0,034
	oui	1	5	0,026	0,077	0,038	0,043	0,167	0,068
	oui	0	3	0,013	0,048	0,020	0,021	0,091	0,034
	oui	1	3	0,026	0,010	**0,040**	0,043	0,182	**0,069**
WN	non	0	5/3	0,026	0,050	0,049	0,043	0,500	0,078
	non	1	5/3	0,013	0,050	0,025	0,013	0,500	0,041
	oui	0	5/3	0,026	0,666	**0,049**	0,043	0 ,666	**0,080**
	oui	1	5/3	0,013	1,000	0,025	0,021	1,000	0,042
R	non	0	5	0,128	0,263	0,172	0,213	0,385	0,274
	non	1	5	0,115	0,243	0,157	0,191	0,360	0,250
	non	0	3	0,115	0,281	0,167	0,191	0,409	0,261
	non	1	3	0,103	0,258	0,147	0,170	0,381	0,235
	oui	0	5	0,128	0,333	**0,185**	0,213	0,500	**0,299**
	oui	1	5	0,115	0,310	0,168	0,191	0,474	0,273
	oui	0	3	0,115	0,391	0,178	0,191	0,563	0,286
	oui	1	3	0,103	0,364	0,160	0,170	0,533	0,258

Tableau 4 : F-mesure pour la résolution d'anaphores infidèles selon différentes ressources lexicales

De façon peu surprenante, l'utilisation des synonymes du *Robert* permet d'obtenir des résultats plus intéressants : dans la meilleure configuration (accord en nombre, mais fenêtre de recherche large), l'amélioration par rapport à la stratégie de base varie de 5 à 10 points. Elle est essentiellement due à une meilleure précision (forte diminution des fausses réponses), allant jusqu'à une bonne réponse sur deux. Par ailleurs, l'amélioration est plus importante pour la version excluant les noms propres ce qui est cohérent puisque ce type d'information n'y figure pas. Le taux de succès reste néanmoins intéressant pour la version avec noms propres (+ 4 points). C'est un point important car le filtrage *a priori* des anaphores ayant des noms propres comme antécédents demande des outils supplémentaires pour la reconnaissance des entités nommées et n'est pas entièrement automatisable.

La meilleure performance globale du système (F-mesure de 0.319) s'obtient par une configuration en cascade : utilisation du *Robert* (fenêtre phrastique 0 à 5), puis des listes de similarité (fenêtre phrastique 1 à 3), puis des heuristiques de récence (fenêtre phrastique 0 à 3), le tout en excluant les noms propres et en imposant un accord en genre. Ce résultat se rapproche de ceux obtenus pour l'anglais par Poesio et al. (2002) en combinant un thesaurus acquis sur grand corpus avec *WordNet 1.6*.

5 Discussion

Au final, seuls les synonymes extraits du *Grand Robert*, c'est-à-dire d'une ressource lexicographique générale, conçue par et pour des humains, apportent un gain de performance significatif pour la résolution automatique des anaphores infidèles en français. En l'absence de ressources de qualité et couverture comparables, la meilleure stratégie reste encore la plus simple : l'association du principe de récence à des contraintes morpho-syntaxiques élémentaires permet d'obtenir une F-mesure de l'ordre de 0,15%. L'absence de publications sur la même tâche[1] ne nous permet pas de confronter ce taux à d'autres expériences sur le français. Néanmoins, l'écart de seulement 8 points entre notre meilleur rappel (ressources en cascade) et celui obtenu par Poesio et al. (2002) pour l'anglais peut être considéré comme encourageant. Toujours est-il que ce taux est pour l'instant largement insuffisant pour des applications en grandeur réelle et souligne la distance à parcourir pour rendre vraiment opérationnels des travaux sur l'interprétation et la génération automatique d'expressions référentielles, supposant en général l'existence de ressources sémantiques de bonne qualité sur lesquelles viennent se greffer des mécanismes d'inférence élaborés (Danlos, 1999 ; Salmon-Alt, 2001). Ce constat devra donc encourager à explorer différentes pistes : reconnaissance des entités nommées, prise en compte d'autres relations sémantiques, prise en compte non seulement de la tête, mais des modifieurs des groupes nominaux et intégration d'une analyse morphologique dérivationnelle.

Références

BICK E. (2003). A CG & PSG Hybrid Approach to Automatic Corpus Annotation. In: Kiril Simow & Petya Osenova: *Proc. of SProLaC2003*, pp. 1-12. Corpus Linguistics 2003, Lancaster.

DANLOS L. (1999). Event Coreference between two sentences. *Proc. of International Workshop on Computational Semantics.* Tildburg.

FOUROUR N. (2002). Némésis, un système de reconnaissance incrémentielle d'entités nommées pour le français. *Actes TALN 2002*, Nancy.

GASPERIN C., GAMALLO P., AGUSTINI A., LOPES G., LIMA V. (2001). Using syntactic contexts for measuring word similarity. *Proc. of the Workshop on Semantic Knowledge Acquisition and Categorisation*. ESSLLI 2001, Helsinki, Finland.

GREFENSTETTE G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

MÜLLER C., STRUBE M. (2001). MMAX: A Tool for the Annotation of Multi-modal Corpora. *Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Seattle, Wash., 45-50.

POESIO M., ISHIKAWA T., SCHULTE IM WALDE S., VIEIRA R. (2002). Aquiring Lexical Knowledge for Anaphora Resolution. *Proc. of LREC 2002*. Las Palmas, Spain.

POESIO M., VIEIRA R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24, n° 2.

POPESCU-BELIS A. (1999*). Modélisation multi-agent des échanges langagiers : application au problème de la référence et son évaluation*. Thèse d'université, Université de Paris XI (Paris-Sud).

SALMON-ALT S. (2001). Reference Resolution within the Framework of Cognitive Grammar. *Proc. of International Colloquium on Cognitive Science*. San Sebastian, Spain, May 2001.

SALMON-ALT S., VIEIRA R. (2002). Nominal expressions in multilingual corpora : definites and demonstratives. *Proc. of LREC 2002*. Las Palmas, Spain.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of International Conference on New Methods in Language Processing*. September 1994. Manchester, UK.

[1] Le seul système de résolution d'anaphores nominales du français évalué finement est celui de Popescu-Belis (1999). Toutefois, il n'y a pas d'évaluation consacrée exclusivement aux reprises infidèles.