Résolution de la référence dans des dialogues homme-machine : évaluation sur corpus de deux approches symbolique et probabiliste

Alexandre Denis[1] Frédéric Béchet[2] Matthieu Quignard[1]
(1) UMR 7503 LORIA/CNRS – Campus scientifique
56 506 Vandoeuvre-lès-Nancy Cedex
(2) LIA – 339, chemin des Meinajaries, BP 1228,
84 911 Avignon Cedex 9
alexandre.denis@loria.fr, frederic.bechet@univ-avignon.fr,
matthieu.quignard@loria.fr

Résumé Cet article décrit deux approches, l'une numérique, l'autre symbolique, traitant le problème de la résolution de la référence dans un cadre de dialogue homme-machine. L'analyse des résultats obtenus sur le corpus MEDIA montre la complémentarité des deux systèmes développés : robustesse aux erreurs et hypothèses multiples pour l'approche numérique ; modélisation de phénomènes complexes et interprétation complète pour l'approche symbolique.

Abstract This paper presents two approaches, one symbolic, the other one probabilistic, for processing reference resolution in the framework of human-machine spoken dialogues. The results obtained by both systems on the French MEDIA corpus points out the complementarity of the two approaches : robustness and multiple hypotheses generation for the probabilistic one ; global interpretation and modeling of complex phenomenon for the symbolic one.

Mots-clefs : dialogue homme-machine, résolution de la référence, évaluation, compréhension dans le dialogue

Keywords: human-machine dialogue, reference resolution, dialogue understanding, evaluation

1 Introduction

Le projet MEDIA de l'action TECHNOLANGUE propose une méthodologie d'évaluation des systèmes de compréhension de la parole dans un cadre de dialogues homme-machine. Le paradigme est inspiré du projet PEACE (Maynard & Devillers, 2000). Il définit une représentation sémantique commune vers laquelle chaque système devra convertir sa propre représentation.

A partir du corpus collecté (Bonneau-Maynard *et al.*, 2005), le projet MEDIA se divise en deux campagnes d'évaluation : MEDIA-HC (hors contexte) et MEDIA-EC (en contexte). Dans la première campagne les systèmes sont évalués sur leur capacité à produire la forme sémantique désirée sans prendre en compte le contexte du dialogue, chaque énoncé étant considéré comme

indépendant (Bonneau-Maynard *et al.*, 2006). La deuxième campagne évalue la compréhension d'un énoncé dans le contexte du dialogue. En particulier deux aspects sont évalués :

– spécification du sens en contexte des entités détectées lors de la phase MEDIA-HC ;
– résolution des références vers des entités introduites dans des tours précédents du dialogue.

Deux systèmes ont participé à cette évaluation en-contexte, un système réalisé au LIA et un autre au LORIA, chacun implémentant des méthodes différentes : une approche numérique pour le LIA, une approche symbolique pour le LORIA. Cet article décrit ces deux systèmes, l'outil d'évaluation développé pour cette campagne par le LORIA et les premiers résultats obtenus.

2 Problématique et manuel d'annotation

La résolution de la référence est le processus cognitif qui met en relation une expression (linguistique ou non) et la représentation mentale que cette expression désigne (voir pour une définition proche (Reboul & Moeschler, 1994)). Une représentation mentale est définie comme l'aggrégation de données hétérogènes (perceptives, mémorielles, logiques, etc.) sur une entité (Reboul & Gaiffe, 1999). Dans cette campagne, nous nous focalisons sur la représentation des référents acquise par une relation intra-linguistique de *coréférence*, définie entre deux expressions référentielles si ces dernières désignent le même référent (van Deemter & Kibble, 2000).

L'évaluation de la référence au sein du consortium devait répondre à deux objectifs principaux : elle devait être accessible à tous les systèmes participants et elle devait être compatible avec le paradigme d'évaluation hors-contexte et en particulier la représentation sémantique utilisée (Bonneau-Maynard *et al.*, 2005).

La plupart des approches évaluent les relations entre expressions référentielles (Popescu-Belis *et al.*, 2004), et s'appuient sur des formats d'annotation qui se concentrent sur les relations, à l'instar du format des campagnes MUC-6 et MUC-7 basé sur les coréférences (Chinchor & Hirschmann, 1997; van Deemter & Kibble, 2000). Dans cette lignée, le Reference Annotation Framework, RAF (Salmon-Alt & Romary, 2004) définit des catégories de données pour annoter les expressions référentielles (les *markables*) et les relations de différentes natures qu'elles entretiennent (les *referentials links*).

Cependant, comme tous les systèmes ne pouvaient produire ces relations mais étaient tous capables de fournir une représentation sémantique des référents (Bonneau-Maynard *et al.*, 2006), il a été décidé d'évaluer non pas les relations mais les descriptions des référents. L'annotation en contexte est alors vue comme une extension de l'annotation hors-contexte. Nous nous sommes toutefois inspirés des catégories pertinentes de liens référentiels définis dans RAF (identité, codomanialité, partie-tout) pour annoter les formes d'expressions référentielles.

Annotation hors-contexte de la référence. L'annotation hors-contexte consiste à annoter chaque segment signifiant d'un énoncé par un trait sémantique du type $\langle$*mode*, *attribut*, *valeur*$\rangle$ dont les contraintes ont été définies collectivement lors de l'évaluation hors-contexte [1] (Bonneau-Maynard *et al.*, 2005). En ce qui concerne la référence, cette annotation s'est limitée à annoter la présence d'une expression référentielle grâce à un trait d'attribut *lienRef* raffiné

[1] le mode décrit la polarité du trait (positive, négative, interrogative, optionnelle), l'attribut représente la catégorie sémantique mentionnée, et la valeur son instanciation

par la catégorie de l'expression. Les différents raffinements retenus (appelés *spécifieurs*) sont proches des catégories de RAF (voir tableau 1), à l'exception de partie-tout que nous n'avons pas annoté faute de consensus. A la différence des *markables* de RAF, seuls les déterminants des groupes nominaux sont associés au *lienRef* afin de ne pas interférer avec le reste du groupe nominal déjà annoté en sémantique.

TAB. 1 – Types de spécifieurs

Spécifieur	**Signification**	**Expressions référentielles**
coRef	coréférence : l'expression référentielle désigne son référent par référence directe	pronoms, définis, démonstratifs
elsEns	élément-ensemble : l'expression référentielle désigne son référent en vertu de propriétés sémantiques ou indexicales qui l'opposent à d'autres entités dans un ensemble	ordinaux, superlatifs, relatives, certains pronoms démonstratifs
coDom	co-domaine : l'expression référentielle désigne son référent grâce à un marqueur linguistique d'altérité	altérités

Afin de réduire le coût d'annotation, seules les expressions référentielles dont la résolution dépasse le cadre de l'énoncé ont été annotées. Cela exclut dès lors les référents dont l'antécédent a été introduit dans le même énoncé, les entités nommées et les indéfinis[2]. En revanche, les articles définis ont été annotés systématiquement, du moins pour les entités relevant de la tâche.

Annotation en contexte de la référence. Une référence est représentée comme un ensemble de référents, chacun décrit par un ensemble de traits sémantiques. On adjoint un champ *reference* à tous les traits *lienRef*. On notera par exemple {(t1,t2), (t3)} une expression référentielle qui fait référence à deux entités, l'une décrite par deux traits et l'autre décrite par un seul.

Exemple d'annotation hors et en contexte :

je veux / une / chambre double / et / une / chambre simple
+/nombre-chambre : 1
+/chambre-type : double
+/nombre-chambre : 1
+/chambre-type : simple
est-ce que / **ces** / chambres / ont la douche ?
+/lienRef-coRef : pluriel **reference = {**
(+/nombre-chambre :1, +/chambre-type :double),
(+/nombre-chambre :1, +/chambre-type :simple)}
+/objet : chambre
?/chambre-equipement : douche

Afin de raffiner l'expression du formalisme on autorise le *lienRef* a être respécifié dans certains cas. Par exemple, pour représenter l'ambiguïté, en l'absence d'un niveau supplémentaire requis, on approxime l'alternative comme un ensemble de référents en spécifiant le *lienRef* par *ambigu*. Ou encore pour distinguer les référents exclus des référents inclus (en particulier pour l'indéfini avec altérité, voir note 2) on spécifie le *lienRef* par *exclusion* ou *inclusion*.

Règles d'annotation. Nous avons collectivement défini des règles d'annotation pour réduire le risque de désaccord entre annotateurs et s'adapter aux spécificités de l'annotation hors-contexte [3]. Tout d'abord

[2] À l'exception de l'indéfini avec altérité (par ex. "une autre chambre") que nous avons jugé pertinent d'évaluer. Dans ce cas ce n'est pas le référent qui est annoté (il est indéfini) mais l'entité *exclue*.

[3] ces règles sont disponibles sur `http://www.loria.fr/~denis/media.html`, l'accord inter-annotateur est donné dans le tableau 2

la portée de la description d'un référent est constituée uniquement des traits sémantiques présents dans le dialogue antérieur en excluant l'énoncé courant et les énoncés simultanés en cas de chevauchements. Ensuite la couverture de la description est un compromis entre une annotation maximale qui aurait été trop coûteuse à effectuer et une annotation minimale restreinte aux traits discriminants peu intéressante à évaluer lorsqu'il n'y a qu'un seul référent. Nous avons alors distingué entités nommées et non-nommées :
– les entités nommées ou assimilées (hôtel nommé, dates, prix, villes, etc.) ne sont décrites que par un ensemble très restreint de traits comme leur nom ou leur valeur ;
– les entités non-nommées (hôtel non nommé et chambre) sont, elles seules, annotées avec la totalité de la description possible, y compris les traits d'autres référents.

Enfin on contraint la description des référents à être en forme normale, c'est-à-dire principalement non-redondante et non-contradictoire, de manière à homogénéiser les descriptions répétées (comme dans les anaphores fidèles) ou révisées.

3 L'approche symbolique du LORIA

Le système de résolution de la référence développé au LORIA est fondé sur la théorie des domaines de référence (Corblin, 1987; Reboul *et al.*, 1997; Salmon-Alt, 2001). Il postule que la désignation des référents passe par l'identification préalable d'un domaine dans lequel l'expression référentielle isole le référent. Cette vision de la référence se rapproche de la théorie des espaces focaux de Sidner (Grosz & Sidner, 1986). Elle a pour but d'unifier le traitement des différentes modalités ou types de désignation modélisés de manière hétérogène par d'autres systèmes. En particulier les ordinaux et les expressions d'altérité ne nécessitent pas de traitement *ad hoc* et s'intègrent élégamment dans la théorie. L'évaluation MEDIA fut pour nous l'occasion de tester le modèle sur l'anaphore bien qu'il fût à l'origine créé pour la référence multimodale.

Structuration en domaines de référence. Un domaine de référence est constitué d'un support, un ensemble d'objets défini intensionnellement ou extensionnellement, et d'un ensemble de critères de différenciation qui en discriminent les éléments. Chaque désignation active le domaine dans lequel on extrait le référent en le focalisant, le préférant ainsi pour les désignations ultérieures. On représente ici un domaine comme un couple (S, C), où S est l'ensemble support et C un ensemble de critères (formalisés par des relations d'équivalence). Chaque critère sera noté c :F, où c est la relation et F un élément focalisé de la partition opérée par c. Dans l'exemple suivant un seul critère (l'index) sera utilisé pour discriminer les référents.

S : *je vous propose* l'hôtel ibis *et* l'hôtel lafayette	`H*=( {h1, h2}, {index :{}} )`
U : *est-ce que* le premier hôtel *accepte les animaux*	`H*=( {h1, h2}, {index :{h1}} )`
S : *non* l'hôtel ibis *n'accepte pas les animaux*	`H*=( {h1, h2}, {index :{h1}} )`
U : *ok, je prends* l'autre *alors*	`H*=( {h1, h2}, {index :{h2}} )`

L'expression ordinale "le premier hôtel" réfère à h1 en vertu de son index, ce dernier reçoit donc la focalisation pour le critère "index". L'expression d'altérité recherche un domaine présentant une partition focalisée dans laquelle extraire l'autre élément, ce qui a pour effet de focaliser h2 dans la partition "index" (Denis *et al.*, 2006b).

Projection dans le formalisme. La projection consiste à construire la représentation MEDIA d'un référent. Etant donné que les domaines de référence ne conservent que le point de vue courant sur les référents et non pas les expressions référentielles et le contexte de leur emploi, il était nécessaire de combiner le modèle avec une représentation lexicale et sémantique des référents. Parallèlement à la structuration domaniale de l'espace référentiel, nous avons conservé la structure sémantique des énoncés à laquelle nous avons ajouté des relations référentielles. Cette dernière s'appuie sur le MultiModal

Interface Language, MMIL (Landragin & Romary, 2004) la représentation sémantique utilisée lors de la phase hors-contexte qui permet la représentation d'informations de natures lexicale, syntaxique ou sémantique. Ces liens référentiels nous permettent de parcourir les chaînes de coréférence afin d'annoter les descriptions des référents en fonction de leur représentation sémantique à différents instants du dialogue.

L'algorithme est le suivant :

1. interprétation de l'énoncé et production d'une forme sémantique (Denis *et al.*, 2006a) ;
2. résolution de chaque référent de l'énoncé : identification d'un domaine compatible grâce à la forme sémantique puis extraction et focalisation du référent (Salmon-Alt, 2001) ;
3. mise à jour de l'historique sémantique : création des relations d'anaphores (coréférence, et anaphore associative) entre les instances des référents ;
4. projection par parcours des chaînes de coréférence : aggrégation des informations sémantiques hors-contexte relatives aux référents dans un voisinage prédéfini correspondant au type d'entités (présence d'une relation sémantique, co-occurrence dans le composant sémantique, co-occurrence dans l'énoncé, présence dans le dialogue antérieur).
   Dans une des conditions du test (*avecHC*, cf. §5), nous intégrons à ce stade les annotations HC fournies par le protocole. Nous n'en tirons pas partie pour identifier les référents.

4 L'approche probabiliste du LIA

Le système d'interprétation du LIA est composé de deux niveaux. Le premier niveau effectue le décodage conceptuel en calculant, à partir d'un message vocal, une liste des n-meilleures séquences de concepts (où chaque concept est représenté par trois champs qui sont la chaîne de mots supports, l'attribut sémantique et la valeur). L'attribution des spécifieurs, du mode et la résolution des références sont pris en compte par le second niveau prenant en entrée la liste des n-meilleures séquences de concepts obtenues à partir du message vocal. Ce système (Servan & Bechet, 2006) est fondé sur une approche probabiliste identique à celle utilisée en Reconnaissance Automatique de la Parole (RAP). Ce système a été développé pour traiter directement des messages vocaux, cependant les résultats présentés dans cette étude sont obtenus sur les transcriptions manuelles du corpus, comme pour toute la campagne MEDIA.

La compréhension en contexte est ici effectuée selon l'approche suivante : la spécification du sens est vue comme une tâche d'étiquetage pouvant être traitée grâce à des techniques d'étiquetage probabilistes. La résolution des références est faite par un certain nombre d'heuristiques décrivant tous les rattachements possibles pouvant être faits entre la liste des n-meilleures interprétations et l'historique du dialogue.

Spécification du sens en contexte. Les séquences de concepts produites à partir du graphe de concepts dans la phase de décodage conceptuel ne contiennent aucun spécifieur de sens, hors ou en contexte. Ces spécifieurs sont attribués par un étiqueteur probabiliste basé sur les Champs Conditionnels Aléatoires (ou *Conditional Random Fields*, CRF). Les CRF (Lafferty *et al.*, 2001) ont été utilisés avec succès dans de nombreuses tâches d'étiquetage telles que l'étiquetage morphosyntaxique ou la détection d'entités nommées. L'avantage principal des CRF par rapport à des modèles génératifs tels que les modèles de Markov cachés est la possibilité d'utiliser l'ensemble des observations d'une séquence pour prédire une étiquette. Ce n'est donc pas le seul historique immédiat qui contraint l'attribution d'une étiquette à une observation mais potentiellement toutes les observations précédentes et suivantes. Cela est particulièrement intéressant pour l'étiquetage des spécifieurs dans la mesure où la spécification du sens d'un concept peut se faire avec des éléments situés avant ou après le concept dans l'énoncé courant, ou dans les énoncés précédents. Les liens référentiels sont étiquetés dans cette phase par rapport au type de l'objet référencé et au type de référence.

Le corpus d'apprentissage des CRF est obtenu à partir des corpus MEDIA. Chaque dialogue constitue une séquence où les observations sont les concepts marqués dans la référence et les étiquettes sont soit les spécifieurs attribués aux concepts ; soit le type du ou des objets référencés pour les liens référentiels, ainsi que le type du lien ; soit le symbole NULL si un concept n'a ni spécifieur ni lien référentiel. Lors du traitement d'un message, chaque chaîne de concepts produite par le décodeur mot/concept est traitée par l'étiqueteur CRF et la description des concepts est enrichie avec les étiquettes attribuées. L'étiqueteur développé utilise l'outil `CRF++`[4].

Résolution de la référence. A la suite de la phase précédente les concepts liens référentiels sont étiquetés avec les trois informations suivantes : le type de ou des objets pointés (*chambre*, *hôtel*, *réservation* ou une combinaison de ces valeurs) ; le type du lien référentiel (*ambigu*, *exclusion*, *inclusion*) ; et enfin le nombre (*singulier* ou *pluriel*).

La résolution des références s'effectue alors selon l'algorithme suivant : tous les concepts situés dans l'historique du dialogue (limité aux n énoncés précédents) ayant une étiquette *spécifieur* similaire au type d'objet pointé par le lien référentiel sont associés à ce lien.

Chaque objet est caractérisé par un certain nombre de traits (par exemple la ville, la marque, le nom ou les services associés à un hôtel). L'algorithme d'association fait pointer le lien référentiel vers tous les concepts représentant ces traits. Lorsque tous les traits sont identifiés, l'algorithme s'arrête s'il s'agit d'un lien référence singulier. Sinon un nouvel objet est créé et le processus se poursuit. Aucun contrôle n'est effectué sur le nombre d'objets désignés. Le but ici est de maximiser les mesures de rappel sur les références pour proposer au module de décision (analyseur sémantique, gestionnaire de dialogue) le plus d'associations possibles, chacune avec un score donné par les différents modèles utilisés lors du décodage.

5 Méthodologie

Dans le cadre de la campagne MEDIA un corpus de 1 250 dialogues a été constitué selon un protocole de type *Magicien d'Oz* : 250 locuteurs ont effectués chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain. Pour la campagne MEDIA-EC le corpus d'apprentissage disponible contient 814 dialogues, 11 800 énoncés utilisateurs et 38 800 segments sémantiques dont 2 294 liens référentiels. Le corpus de test MEDIA-EC contient 174 dialogues pour 2 650 énoncés utilisateurs et 7 780 segments dont 910 liens référentiels.

Deux conditions de test sont organisées : la première (*sansHC*) consiste à n'utiliser que les dialogues transcrits, les erreurs de l'étiquetage hors-contexte se cumulant à celles de l'étiquetage en-contexte. La deuxième condition (*avecHC*) consiste à utiliser les dialogues avec leur annotation hors contexte.

Méthode d'évaluation. L'évaluation de la résolution de la référence s'effectue par comparaison des traits sémantiques proposés pour chaque référent. Nous observons que pour décrire un référent, il faut préalablement l'avoir identifié. De même, pour l'identifier, il faut préalablement avoir repéré l'expression référentielle. Comme ces tâches impliquent potentiellement des capacités différentes, nous avons jugé intéressant d'évaluer la résolution de la référence selon quatre niveaux, chacun donnant lieu à des scores de rappel, précision et f-mesure :

IER Capacité à repérer (ou identifier) les expressions référentielles. Il s'agit d'une capacité hors contexte, qu'on évalue tout de même car l'évaluation de la référence en dépend.

[4] Téléchargeable à `http://chasen.org/~taku/software/CRF++`

DER Capacité à décrire les expressions référentielles identifiées, c'est-à-dire, à fournir les bons spécifieurs (*coRef*, *coDom*, *elsEns*, mais aussi *inclusion*, *exclusion* et *ambigu*). Cette capacité est évaluée sur la base des expressions correctement repérées en IER.

IREF Capacité à identifier les référents, c'est-à-dire à fournir pour chaque référent suffisamment de traits corrects pour qu'il soit couplable avec un référent à trouver. Comme la précédente, cette capacité n'est évaluée que sur les expressions référentielles correctement repérées en IER.

DREF Capacité à décrire *in extenso* les référents. Cette évaluation n'est calculée que sur les référents corrects en IREF.

En procédant ainsi par niveau, nous pouvons mieux apprécier les différentes capacités qu'implique la résolution de la référence. Nous notons qu'il est possible d'avoir un score global de rappel (resp. de précision) en DREF, c'est-à-dire le nombre de traits corrects fournis par un système rapporté au nombre de traits fournis par l'annotation manuelle (resp. fournis par le système), en multipliant les scores de rappel (resp. de précision) obtenus en IER, IREF et DREF.

Pour chaque niveau, nous avons développé les algorithmes suivants :

IER Le but est d'aligner les traits *lienRef* sans s'appuyer ni sur leurs spécifieurs (évalués en DER), ni sur leur contenu référentiel (évalué en IREF et DREF). Or, si l'on retire ces annotations, le trait *lienRef* comporte trop peu d'information pour pouvoir effectuer l'alignement sans risque dans le cas d'une omission ou d'une addition de *lienRef*. Nous effectuons donc un alignement des *lienRef* sur la base des autres traits sémantiques qui sont dans l'intervalle. Nous avons adapté l'algorithme de Levenshtein pour que le gain d'un appariement de *lienRef* soit proportionnel à la valeur d'appariement des traits (non référentiels) qui se trouvent entre un *lienRef* et le suivant[5].

DER Pour chaque couple de *lienRef* appariés selon le procédé ci-dessus, on calcule le nombre d'erreurs qui apparaissent lorsqu'on rajoute les modes et surtout les spécifieurs.

IREF Pour chaque couple de *lienRef* appariés en IER, on effectue un couplage maximal de poids maximal entre référents hypothèse et référents de l'annotation manuelle. La matrice de couplage donne un poids proportionnel aux nombres de traits partagés[6].

DREF Pour chaque couple de référents formé en IREF, on effectue un couplage maximal entre traits. En effet, il n'y a pas d'ordre prescrit pour décrire les référents.

Evaluation de la qualité de l'annotation manuelle. L'annotation EC a fait l'objet de contrôles à trois reprises. A chaque fois, une double annotation a été effectuée sur une dizaine de dialogues puis évaluée à l'aide de l'outil de mesure présenté plus haut. Cette évaluation repose donc sur un échantillon ne représentant que 4% du corpus d'apprentissage (31 dialogues sur 814).

IAG	Dialogues	IER	DER	IREF	DREF
1	10	24 (24)	24 (24)	23 (25)	23 (23)
2	10	29 (29)	27 (29)	32 (33)	72 (108)
3	11	27 (27)	25 (27)	34 (36)	135 (150)
Total	31	80 (80)	76 (80)	89 (94)	230 (281)
		100%	95%	95%	82%

TAB. 2 – Accord inter-annotateurs

Les résultats sont présentés dans le tableau 2. L'accord inter-annotateurs est globalement très bon, surtout dans les niveaux supérieurs (IER, DER et IREF). Le score en DREF, plus faible que les autres, traduit la difficulté de fournir unanimement une description complète des référents.

[5]Ce procédé évite aux systèmes de produire une double annotation (HC et EC), l'IER devant être *a priori* calculée sur les *lienRef* corrects en HC et non sur l'annotation EC.

[6]Il suffit donc qu'un référent ait un trait correct pour être candidat. Pour être plus précis, il faudrait ne conserver que les traits permettant de discriminer un référent parmi les autres référents proposés.

6 Résultats

TAB. 3 – Résultats en précision et rappel de la tâche de résolution des références pour les systèmes du LIA et du LORIA. L'accord est une f-mesure des scores de rappel d'un système par rapport à l'autre.

	LIA		LORIA		Accord
sansHC	*prec*	*rappel*	*prec*	*rappel*	*f-mes*
DER	71.4	71.4	50.9	50.9	51.6
IREF	74.1	61.9	65.2	44.3	49.8
DREF	67.3	55.2	68.9	48.3	53.3

	LIA		LORIA		Accord
avecHC	*prec*	*rappel*	*prec*	*rappel*	*f-mes*
DER	86.5	86.5	86.5	86.5	86.5
IREF	77.1	73.8	62.4	40.8	44.0
DREF	74.1	64.0	76.5	43.3	51.6

Le tableau 3 présente les résultats des systèmes du LIA et du LORIA pour les deux conditions du test : *sansHC* et *avecHC*. En DER, le LORIA a un score médiocre dans la phase *sansHC*, qui s'améliore considérablement par la connaissance des traits HC. Pour l'identification des référents (IREF), le système symbolique pèche considérablement en rappel, avec un score comparable dans les deux phases *sansHC* et *avecHC*. Cela s'explique par le fait que l'annotation HC n'est pas utilisée pour la résolution de la référence proprement dite et n'intervient que pour améliorer la description des référents lorsque ceux-ci ont été identifiés. À l'inverse le système statistique tire bien meilleure partie des informations hors-contexte et parvient à augmenter ses scores de rappel d'une dizaine de points sur les trois catégories.

Pour investiguer plus profondément la complémentarité des systèmes, nous les avons évalués l'un par rapport à l'autre. Les résultats sont donnés dans la dernière colonne de chaque tableau. Il s'agit d'une moyenne (f-mesure) des scores de rappel d'un système par rapport à l'autre. Les scores ne sont ni vraiment supérieurs aux valeurs de chaque système, ni vraiment inférieurs. Il semble donc que les systèmes sont aussi distants entre eux que de l'annotation de référence. Ils ne produisent donc pas les mêmes erreurs (l'accord n'augmente pas). De même, leurs sorties ne sont pas fondamentalement complémentaires (l'accord reste du niveau du score de rappel le plus bas).

Système LIA Les résultats de la condition *avecHC* nous permettent de distinguer les erreurs dues uniquement à la spécification du sens et à la résolution des références en contexte. L'étiquetage des liens référentiels (DER) est correcte à 86.5%. En ventilant ce résultat par rapport au type de liens référentiels (les spécifieurs), on remarque que les résultats sont très disparates : le taux d'étiquettes correctes atteint 95.2% pour les étiquettes *lienRef* sans spécifieur qui représentent 57% des liens référentiels du corpus ; il n'est que de 25% pour le spécifieur *ambigu*, associé à seulement 7.2% des liens du corpus. Cette faible représentativité de certains phénomènes pose problème aux méthodes probabilistes qui ont besoin d'un grand nombre d'exemples pour apprendre les modèles. L'identification des référents obtient des scores de précision/rappel convenables étant donné la simplicité des heuristiques mises en œuvre. Une analyse plus fine nous montre également que le système fait peu d'erreurs sur les références directes, qui sont aussi les plus nombreuses. Beaucoup d'erreurs se concentrent sur les liens *ambigus* et les liens contenant le spécifieur *inclusion* sont mieux traités que l'*exclusion*. Enfin notons que pour la condition de test *sansHC* la dégradation des résultats est limitée (-3% de precision, -12% de rappel) sachant que le taux d'erreur de l'annotation hors-contexte est de l'ordre de 20%. Ce dernier point souligne la robustesse de l'approche.

Système LORIA Le dépouillement des résultats du LORIA s'appuie sur une classification des erreurs IREF. Le premier résultat significatif est que 57% de nos erreurs proviennent de la résolution de la référence, alors que les 43% restants sont issus d'erreurs extérieures au module (projection hors-contexte, construction de la forme sémantique, analyse syntaxique ou lexicale). Par exemple si l'analyse

syntaxique oublie un hôtel parmi trois hôtels, le module de référence l'ignorera complètement de telle sorte que l'expression ultérieure "ces trois hôtels" échoue à trouver trois hôtels et ne retournera rien.

Nous avons ensuite raffiné les 57% d'erreurs de référence en vingt catégories classées en deux groupes. Nous distinguons les phénomènes non pris en compte (35%) et les réelles erreurs, problèmes et bugs (65%). Le premier groupe recouvre des cas complexes comme par exemple le générique "la chambre" en anaphore associative qui est mal géré en cas d'antécédent pluriel pour cause d'inconsistence logique. Le second groupe d'erreurs correspond à un fonctionnement anormal du module, par exemple rechercher le référent avec des contraintes qui n'auraient pas lieu d'être (comme dans "je voudrais plus de détails sur les hôtels" où l'on recherche "des hôtels avec plus de détails") ou encore une mauvaise structure domaniale. Nous pouvons conclure de cette analyse que le modèle des domaines de référence est très fin mais peu robuste. A la différence du système statistique du LIA très tolérant face aux erreurs, une erreur en début de dialogue peut entraîner ici une cascade d'erreurs.

Les résultats doivent cependant être compris en considérant la mesure d'évaluation qui a été adoptée. En effet cette dernière définit l'identité de deux référents comme une identité de description, traduisant alors mal le fait que deux référents peuvent être différents tout en se ressemblant. Par exemple, le référent de *"une chambre double à l'hôtel ibis paris"* et celui de *"une chambre double à l'hôtel lafayette paris"* sont similaires à 80% bien qu'ils représentent deux référents distincts. Le système du LIA est insensible au fait qu'il s'agisse de deux référents puisqu'il s'appuie directement sur les traits sémantiques de bon type dans le contexte antérieur, alors que le système du LORIA qui construit une représentation structurée des référents y est au contraire très sensible. Cette mesure ne permet alors d'évaluer que les capacités descriptives des référents, nécessaires dans un système de dialogue mais pas suffisantes. En effet la mesure ne permet pas de comparer les capacités référentielles pour lesquelles il est indispensable d'identifier avec précision le référent (en l'occurrence ne pas réserver une chambre dans le mauvais hôtel). Afin de considérer ces capacités, il est envisageable d'améliorer le couplage IREF pour qu'il ne couple que des référents décrits par les mêmes traits s'il s'agit d'entités nommées ou que des référents décrits par les mêmes traits discriminants s'il s'agit d'entités non-nommées (voir note 6).

7 Conclusion et perspectives

Cette étude présente la comparaison de deux approches très différentes pour résoudre le problème difficile de la résolution des références dans un contexte de dialogue. Les résultats obtenus permettent de mettre en avant les qualités et défauts des deux approches : bonne modélisation de phénomènes complexes mais faible tolérance aux erreurs pour le système symbolique ; bonne intégration avec les modules de décodage de parole et d'étiquetage conceptuel mais mauvaise prise en compte des références complexes pour le système probabiliste. A l'examen détaillé des résultats, la tâche de résolution des références semble être un domaine d'expérimentation prometteur pour étudier l'intégration des approches numériques et symboliques : en générant des listes d'hypothèses valuées, un système probabiliste peut être utilisé en entrée d'un système symbolique chargé de vérifier la cohérence des hypothèses générées, en supprimer certaines, afin de fournir les analyses les plus probables au gestionnaire de dialogue.

Remerciements

Nous tenons à remercier chaleureusement les annotatrices, Christelle AYACHE et Anne KUHN, pour la qualité de leur travail et leur participation enthousiaste et constructive à la définition du manuel d'annotation. Nous tenons également à remercier les relecteurs pour la pertinence de leurs commentaires.

Références

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFEVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C. & VILLANEAU J. (2006). Results of the French Evalda-Media evaluation campaign for literal understanding. In *LREC'06*, Genoa.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal.

CHINCHOR N. & HIRSCHMANN L. (1997). MUC-7 coreference task definition, version 3.0. In *Actes de MUC-7*.

CORBLIN F. (1987). *Indéfini, Défini et Démonstratif*. Genève : Droz.

DENIS A., PITEL G. & QUIGNARD M. (2006a). A deep-parsing approach to natural language understanding in dialogue system : Results from a corpus-based evaluation. In *LREC 2006*, Genoa, Italy.

DENIS A., PITEL G. & QUIGNARD M. (2006b). Resolution of reference grouping in practical dialogues. In *SIGDial 2006*, Sydney, Australia.

GROSZ B. & SIDNER C. (1986). Attention, intention and the structure of discourse. *Computational Linguistics*, **12**, 175–204.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, p. 282–289 : Morgan Kaufmann, San Francisco, CA.

LANDRAGIN F. & ROMARY L. (2004). Dialogue history modelling for multimodal human-computer interaction. In *Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*.

MAYNARD H. & DEVILLERS L. (2000). A framework for evaluating contextual understanding. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Beijing, Chine.

POPESCU-BELIS A., RIGOUSTE L., SALMON-ALT S. & ROMARY L. (2004). Online evaluation of coreference resolution. In *Proceedings of LREC 2004*.

REBOUL A., BALKANSKI C., BRIFFAULT X., GAIFFE B., POPESCU-BELIS A., ROBBA I., ROMARY L. & G. G. S. (1997). *Le projet CERVICAL : Représentations mentales, référence aux objets et aux événements*. Rapport interne, Loria-CNRS/Limsi, France.

REBOUL A. & GAIFFE B. (1999). Représentations mentales et référence.

REBOUL A. & MOESCHLER J. (1994). *Dictionnaire encyclopédique de la pragmatique*. Editions du Seuil.

SALMON-ALT S. (2001). *Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel*. PhD thesis, Université Henri Poincaré, Nancy.

SALMON-ALT S. & ROMARY L. (2004). Towards a reference annotation framework. In *Proceedings of LREC 2004*.

SERVAN C. & BECHET F. (2006). Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media. In *TALN*, Leuven.

VAN DEEMTER K. & KIBBLE R. (2000). On coreferring : Coreference in MUC and related annotation schemes. *Computational Linguistics*, **26**(4), 629–637.