Conférence TALN 1999, Cargèse, 12-17 juillet 1999

Classification multicritère floue d'analyses robustes pour les systèmes de dialogue parlé

David Roussel Jean Caelen Michel Grabisch

Laboratoire CLIPS
Univ. Joseph Fourier
BP 53, 38041 Grenoble Cedex
David.Roussel@imag.fr Jean.Caelen@imag.fr

THOMSON/CSF-LCR
Domaine de Corbeville
91404 Orsay Cedex
grabisch@thomson-lcr.fr

Résumé

Nous présentons une méthode de classification d'analyses robustes sur des hypothèses concurrentes d'un système de reconnaissance de la parole. Pour réaliser cette classification, différents critères hétérogènes sont combinés, comme le score de reconnaissance, diverses caractéristiques syntaxiques et sémantiques propres à l'analyse robuste effectuée ou encore des estimations de la cohérence pragmatique. L'analyse est fondée sur une variante des LTAG (Lexicalized Tree Adjoining Grammars). La classification proposée est évaluée à partir d'un corpus d'analyses robustes d'hypothèses de reconnaissance.

1. Introduction

Dans un système de dialogue dont l'interface est un système de reconnaissance, le degré d'élaboration d'une analyse robuste est limité par des impératifs d'interactions simples avec l'utilisateur. Un compromis doit être recherché entre le nombre d'analyses qu'il est possible de considérer (en entrée ou en sortie d'un analyseur), le coût des connaissances linguistiques mobilisées, et l'intérêt de produire ou non des analyses robustes dans certains contextes de dialogue. Nous étudions pour cela une classification dynamique des analyses en fonction de la capacité des couches de traitement supérieures à prendre en compte certaines alternatives.

L'objectif est double : limiter le nombre d'analyses à considérer en intégrant des connaissances de haut niveau, et contrôler, en cas de résultats multiples, que le système puisse générer une interaction pertinente. Dans le cas inverse, les conséquences peuvent varier d'une situation presque transparente pour l'utilisateur à une situation d'incommunicabilité qui amène cet utilisateur à sous-utiliser le système. Il s'agit donc de mettre au point des stratégies de dialogue robustes qui jonglent avec les limites d'un système. En voici une :

i. Si le contexte de dialogue permet de valider une analyse, alors privilégier cette analyse.
ii. Si le contexte de dialogue ne permet pas de prédire un ensemble fini d'énoncés possibles (c'est notamment le cas lorsque le système réalise un acte assertif ou répond à l'utilisateur), et si les analyses concurrentes sont très divergentes ou incompatibles avec le contexte, justifier une incapacité du système et demander une reformulation.
iii. Si une analyse potentielle correspond à une demande de la part de l'utilisateur, alors privilégier cette analyse. Demander une confirmation sur cette demande avant de la satisfaire (le système de dialogue ne peut pas ignorer une demande de l'utilisateur).
iv. Dans le cas contraire, si l'interaction personne-système précédente n'est pas déjà un métadialogue, prendre en compte différentes variations lexicales autour d'un même focus *a priori* valide et initier un dialogue de désambiguïsation explicite.
v. Si aucune autre alternative n'est possible, maintenir plusieurs analyses pour une remise en cause ou une rétrovalidation ultérieure.

(i) et (ii) sont des heuristiques de recherche qui limitent le nombre d'analyse à manipuler simultanément lorsque le contexte ne permet pas de décider. (iii) est un exemple de cas particulier que l'on doit pouvoir intégrer dans toute stratégie. (iv) est un principe rhétorique (l'utilisateur ne doit pas avoir le sentiment de perdre le contrôle du dialogue). On permet au système d'engager un métadialogue avec l'utilisateur (ex : demande de confirmation, mise au point) si celui-ci n'a pas déjà été embarqué récemment dans une interaction qui ne faisait pas avancer la tâche. Enfin (v) est une stratégie générale qui implique des mécanismes de révision des connaissances pour la résolution des conflits.

2. Application de connaissances hétérogènes

L'exemple suivant illustre l'intérêt d'intégrer des connaissances hétérogènes. Cet exemple est issu d'une application d'aménagement d'intérieur virtuel qui dispose d'une interface vocale (en anglais). Cette application a été développée pour le projet ACTS *COllaborative Virtual ENvironment* (COVEN). La figure 1 en est une visualisation. La figure 2 est un exemple d'hypothèses de reconnaissance en situation.

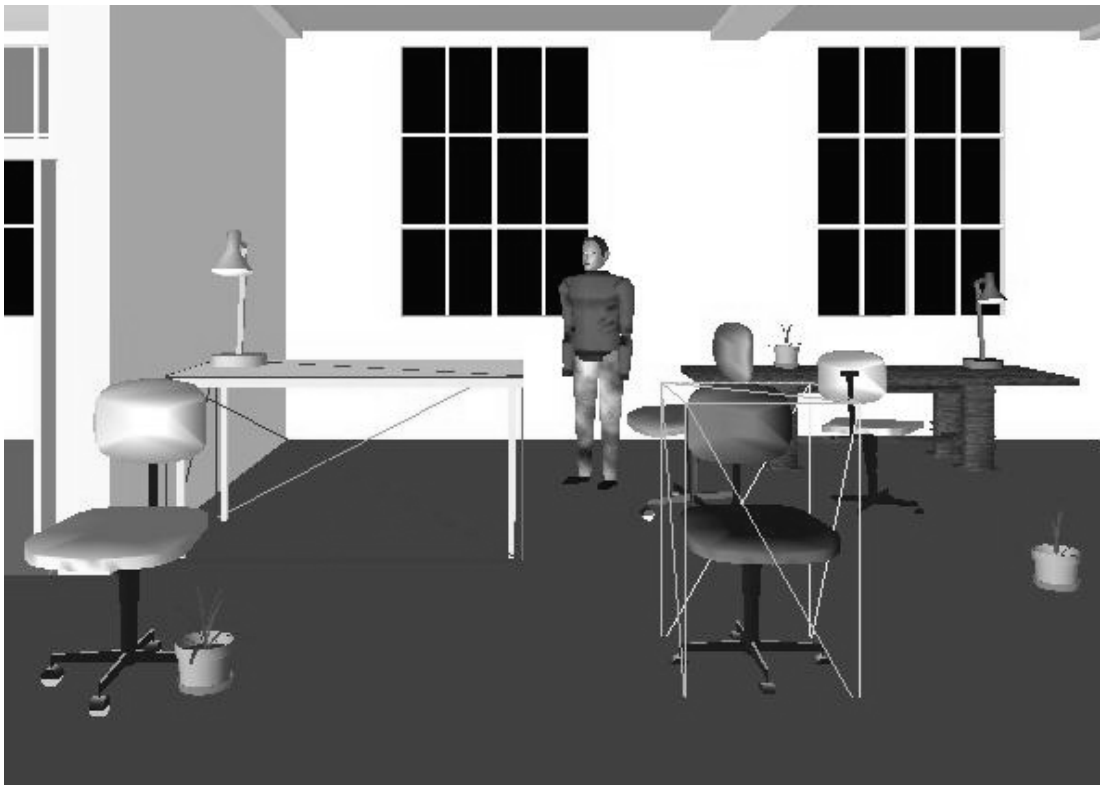

Figure 1: vue d'une scène virtuelle d'aménagement d'intérieur. Une chaise et une table sont déjà sélectionnées, la table par un autre participant (qui apparaît dans la scène).

```
énoncé: I WANT TO CHANGE THE TABLE FOR A SMALLER ONE
```

```
H1:   COULD YOU CHANGE A TABLE FOUR SMALL ONE
H2:   COULD YOU CHANGE THE TABLE FOUR SMALL WALL
H3:   COULD YOU CHANGE A TABLE FOR A SMALL WALL
H4:   COULD YOU CHANGE A TABLE FOR A SMALL METAL ONE
H5:   COULD YOU CHANGE THE TABLE NOT THE SMALL ONE
H6:   WOULD YOU CHANGE THE TABLE FOR A SMALL BLUE ONE
```

Figure 2: hypothèses observées avec le système de reconnaissance ABBOTT™.

Si l'on applique un critère uniquement syntaxique sur les deux premières hypothèses de la figure 2, on est amené à supposer une **absence d'interprétation** du fait de l'agrammaticalité

syntaxique du groupe nominal final. C'est peut être une chance concernant la deuxième hypothèses mais pas pour la première, qui est très proche de l'énoncé de l'utilisateur. Une analyse syntaxique partielle qui ne prend en compte que l'îlot *could you change a/the table* n'est pas non plus une bonne idée : celle-ci produirait une **mauvaise interprétation**.

Si l'on considère la troisième hypothèse et les suivantes, on remarque que celles-ci sont syntaxiquement correctes. Différents critères doivent être appliqués pour détecter qu'elles ne correspondent pas à l'énoncé de l'utilisateur. Une contrainte propre à l'application permet de rejeter la troisième hypothèse : le fait que la position des murs est prédéfinie dans l'application. Dans ce cas, le système observe une contradiction entre des connaissances générales et le contenu de l'énoncé. Si l'on emprunte la classification de (McRoy, 98), il s'agit d'un cas particulier d'**incohérence de l'interprétation** qui s'applique pour toute situation interprétative où les représentations du système ne semblent pas alignées avec celles de l'utilisateur. D'autres contraintes contextuelles —dynamiques cette fois— permettent de rejeter la quatrième hypothèse du fait d'une incohérence d'interprétation : si une table de la scène est déjà sélectionnée, l'usage d'un déterminant indéfini est suspect. Si plusieurs référents, usuellement désignés par le terme *table*, existent dans la scène, le fait que l'expression référentielle ne permette pas de sélectionner une table particulière peut également être suspecté.

Il est plus difficile d'écarter la cinquième et sixième hypothèse. On risque du même coup d'interpréter l'interaction de l'utilisateur comme une requête contenant une précision (H5), ou comme une question (H6). Il faudrait ici considérer que le rang de ces hypothèses les rend incertaines. La stratégie proposée en introduction gère cette situation en (iv) ou (v).

Premier constat : une analyse renseignée est nécessaire pour identifier les différents types d'erreurs et mettre en concurrence les meilleures hypothèses d'un système de reconnaissance de la parole. **Deuxième constat** : l'intégration de ces connaissances est d'autant plus souhaitable qu'une analyse robuste est nécessaire ; or cette analyse produit potentiellement beaucoup d'hypothèses, et nécessite un contrôle par des connaissances de haut niveau.

Pour intégrer différentes sources de connaissances, des approches empiriques comme celle de (Van Noord 98) ne nous semble pas viable[1]. Un modèle stochastique ne nous semble pas non plus facile à mettre en œuvre[2]. Dans la section suivante, nous introduisons la décision multicritère floue en tant que cadre théorique pour la détermination automatique des interactions entre les critères les plus importants.

3. Notions basiques de décision multicritère floue

Un problème de décision multicritère est représenté par un vecteur d'attributs (ou critères) $X = X_1 \times X_2 \times \ldots \times X_n$ et une relation de préférence notée $\succ$. $\succ$ dénote une relation de préférence binaire, connexe et transitive sur l'espace multidimensionnel X. Le problème est de remplacer l'ordre faible $(X, \succ)$ par un ordre total, par exemple $(\mathbb{R},\geq)$, c'est-à-dire déterminer une fonction d'utilité $u : X \rightarrow \mathbb{R}$ telle que $x \succ y \Leftrightarrow u(x_1 \ldots x_n) > u(y_1 \ldots y_n)$. Ceci consiste à déterminer une fonction multidimensionnelle $u(x_1 \ldots x_n) = H(u_1(x_1),\ldots, u_n(x_n))$ où H est un opérateur d'agrégation qui agit sur des fonctions utilitaires monodimensionnelles $x_i \succ y_j \Leftrightarrow u_k(x_i) > u_k(y_j)$. Les travaux comme ceux de (Grabish, Nguyen, Walker, 95) ont prouvé que de

[1] La formule appliquée dans (Van Noord *et al.*, 98) pour la sélection des meilleures analyses nécessite la détermination de deux constantes empiriques pour mettre en relation quatre critères de bas niveau. L'intégration de connaissances sémantiques est évoquée mais reste une perspective.

[2] Il est difficile de procéder à des rétroactions en fonction de l'état du dialogue: comment découper les corpus qui servent à l'apprentissage des HMM en fonction d'hypothétiques contextes ? et surtout, comment conserver une bonne représentation des données dans les sous-corpus nécessaires à l'apprentissage ?

nombreux opérateurs d'agrégation peuvent se représenter par une intégrale floue et que l'addition des mesures floues est indépendante des attributs.

Un point important est que les mesures floues sont facilement interprétables. Soit **I** l'ensemble des critères, une mesure floue μ sur **I** exprime en soi l'importance des coalitions de critères. L'importance d'un critère particulier est donné par la *valeur de Shapley* de ce critère (Shapley, 1953). L'interaction I_{ij} entre deux critères i et j est également mesurable (Murofushi et Soneda, 93). L'indice d'interaction détermine notamment si deux critères de faible importance pris séparément sont importants pris ensemble. Plus généralement :

$I_{ij} > 0$ implique une *complémentarité* entre i et j : l'importance des critères i et j pris ensemble est plus grande que la somme des importances individuelles.

$I_{ij} < 0$ implique une *redondance* entre i et j : l'importance des critères i et j pris ensemble est moindre que la somme des importances individuelles

4. Application à l'analyse robuste d'hypothèses de reconnaissance

Nous avons testé quatre types de critères. Les indications du système de reconnaissance figurent bien entendu parmi eux, en l'occurrence - $\mathbf{c_1}$ - le score attribué à l'hypothèse analysée, et - $\mathbf{c_2}$ - le nombre de mots reconnus dont le score est fiable. Voici les autres types de critères.

4.1 Description de l'analyse produite

L'analyse est fondée sur un formalisme appelé LTFG (*Lexicalized Tree Furcating Grammar*), qui n'utilise pas l'adjonction mais une variante plus simple : la furcation (De Smedt et Kempen, 91). La complexité de l'analyse est celle d'une analyse hors contexte. L'analyse robuste obéit à des contraintes entre traits sémantiques (Roussel et Pernel, 98). Elle se fonde sur ensemble de règles de rattrapage déclaratives comme celles présentées dans (Roussel et Lopez, 99) et des arbres de "jonction" (voir Roussel et Halber, 97). La souplesse du formalisme permet également de reconstituer des structures syntaxiques par fusion, mode d'analyse présenté dans (De Smedt et Kempen, 91).

Schématiquement, l'analyse étend progressivement les arbres élémentaires puis dérivés qui s'appliquent à la séquence à analyser. Les règles dédiées au rattrapage d'erreurs de reconnaissance ou d'extragrammaticalités sont, le cas échéant, mises à contribution suivant les contraintes et possibilités d'extension des îlots d'analyse (représentés dans un *chart*).

Nous décrivons le résultat de l'analyse par deux critères :

($\mathbf{c_3}$) nombre d'îlots complets par rapport au nombre total de mots couverts
($\mathbf{c_4}$) nombre d'opérations de rattrapage utilisées par rapport à l'ensemble des opérations

Ce sont les opérations que les arbres élémentaires prévoient ou rendent possibles qui donnent lieu à des hypothèses de rattrapage. Ces hypothèses doivent respecter des contraintes sémantiques entre les différents sèmes (traits sémantiques élémentaires) associées aux arbres d'analyse. Pour forcer une analyse, les contraintes sont relâchées. Les sèmes sont alors propagés avec un statut particulier. Nous considérons deux types de contraintes sémiques : des restrictions de sélection entre constituants recteurs et régis (elles permettent le contrôle des relations fonctionnelles ou sémantiques à créer entre les constituants) et des contraintes plus générales de compatibilité entre traits sémantiques. Ces dernières indiquent si un arbre auxiliaire a_i peut être ancré sur un arbre a_j[3]. Étant donné cette stratégie, nous ajoutons quatre critères pour décrire le processus d'analyse :

[3] Dans ce cas, les traits sémantiques associés à chaque nœud concerné (ou à leur projection directe) doivent dénoter au moins une complémentarité référentielle. Ces contraintes sont moins complexes que l'unification.

(c_5) nombre d'îlots ignorés ou substitués (par une hypothèse concurrente)
(c_6) nombre de fusions (dénote autant de compatibilité sémantiques)
(c_7) nombre d'insertions d'arbres génériques pour obtenir la dérivation maximale
(c_8) nombre d'adaptations sémantiques pour rendre les îlots compatibles

4.2 Estimation de la cohérence pragmatique

Le premier critère pragmatique testé - **c_9** - représente la probabilité de l'acte de dialogue manifesté dans un énoncé. Ces actes de dialogue sont déterminés en fonction des besoins de la tâche. Certains se décomposent en deux actes autonomes de façon à faciliter le calcul des probabilités par un modèle tri-grammes. Ce modèle a été entraîné sur un corpus de dialogues d'aménagement d'intérieur (cf. Mignot, 95) étiqueté en actes de dialogue. Il détermine la probabilité d'apparition d'un nouvel acte de dialogue en fonction du contexte immédiat. Un test sur deux dialogues, l'un sans problème, l'autre au contraire plusieurs fois interrompu, a montré que le taux de prédiction est important si l'on considère les deux actes ayant la plus forte probabilité. Les taux de prédiction respectifs sont de 90 % et 83 %.

Le second critère pragmatique - **c_{10}** - concerne les expressions référentielles. Pour évaluer la confiance d'un locuteur dans l'adéquation d'une expression référentielle, (Edmonds, 93) propose une pondération des caractéristiques des expressions référentielles en fonction des types d'entité en présence. L'estimation que nous avons testée repose aussi sur des coefficients de saillance. Ces coefficients sont associés aux sèmes qui opposent les unes aux autres les définitions lexicales des noms d'entités. Chaque coefficient est défini par le contraste qu'une caractéristique établit dans le contexte d'un ensemble d'objets ou de groupes d'objets. Les disjonctions entre les caractéristiques sont représentées par des transitions (très) pénalisantes entre les états de différents automates (chaque automate contrôle un ensemble de propriétés exclusives). Un tri décroissant des hypothèses suivant la somme des logarithmes des valeurs de saillance permet de renforcer ou déclasser correctement 70 % de ces hypothèses (sur 200 testées). Etant donné l'imprécision de certaines caractéristiques ou relations spatiales dans la représentation virtuelle, nous convertissons le score d'adéquation référentielle en trois degrés de satisfaction : possible, indécidable et impossible.

4.3 Critères subsidiaires

En plus des critères précédents, sont pris en compte - **c_{11}** - le rang relatif de l'hypothèse selon le classement du système de reconnaissance et - **c_{12}** - le nombre d'analyses concurrentes trouvées. *A priori*, ces critères peuvent affaiblir ou renforcer une décision de classification.

5. Premiers Résultats

Le tableau 1 présente les indices d'importance des critères. Sur la base des classes distinguées en introduction, les résultats de classification sont meilleurs avec 9 critères (*Cl_9*) que 12 (*Cl_12*) . Nous obtenons 82 % de classification correcte sur cinquante analyses robustes en situation. La classe (iv) est confondue avec la classe (iii), la classe (ii) avec la classe (iv).

c_i	Cl_12	Cl_9	c_i	Cl_12	Cl_9
1	1.16	0.93	7	1.31	1.16
2	0.83	_	8	0.95	0.99
3	1.11	0.97	9	1.61	2.02
4	1.28	1.16	10	1.61	1.45
5	1.20	_	11	1.48	1.48
6	1.19	_	12	1.21	1.05

Tableau 1 : Valeurs de Shapley avec 12 (Cl_12) et 9 critères (Cl_9).

j\i	4	7	9	10	11
1	-0.09	0.01	0.03	-0.19	-0.23
2	-0.40	-0.20	-0.19	-0.07	-0.14
3	-0.33	-0.20	-0.19	-0.13	**0.18**
5	-0.14	-0.23	-0.07	0.00	-0.06
6	-0.24	-0.24	-0.07	-0.04	-0.06
8	**0.28**	-0.02	0.08	**0.21**	**0.21**

Tableau 2 : Indices d'interaction I_{ij} entre les critères faibles et forts sur les 12 critères.

Trois critères (2, 5 et 6), dont l'importance moyenne n'est pas forcément la plus faible, apportent plus d'incertitude que de capacité discriminatoire.

Les interactions entre les 6 critères les plus importants et les 6 moins importants (pris isolément) sont indiquées tableau 2. On remarque que les critères 3 et 8, bien que d'importance faible pris isolément, présentent tous deux une synergie positive avec $\mathbf{c_{11}}$, ce qui signifie que le rang de l'hypothèse de reconnaissance et le rapport nb_îlots / nb_mots_couverts (ou le nombre d'adaptation sémantique) fournissent un bon verdict. Avec 9 critères, les couples de critères qui présentent la synergie la plus forte sont {8,10}, {8,9}, {4,10}, {9,10} et {10,11}. Ceci confirme que le score acoustique améliore légèrement la classification des analyses robustes, alors que les critères pragmatiques sont de la plus haute importance pour en mesurer l'acceptabilité.

6. Conclusion

L'expérimentation présentée tente de répondre à un problème méthodologique (l'intégration de sources de connaissances difficiles et coûteuses à modéliser sous forme de probabilités) par des techniques de décision multicritère floue. En fonction des critères disponibles, la méthode peut aboutir à une classification avancée destinée à une stratégie fine de dialogue et / ou permettre la mise au point de fonctions d'estimation dans des algorithmes de réduction de l'espace de recherche.

Références

EDMONDS P. G. (1993), *A Computational Model of Collaboration on Reference in Direction-Giving Dialogues*, Ph.D. thesis, University of Toronto, Canada.

GRABISCH M. (1996), The application of fuzzy integrals in multicriteria decision making, *European J. of Operational Research*, vol 89, pp. 445-456.

GRABISCH M., NGUYEN H. T., WALKER E.A. (1995), *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference*, Dordrecht : Kluwer Academic Publisher.

McROY S. (1998), Detecting, repairing and preventing human-machine miscommunication, Int. Journal of Human-Computer Studies, 48(5), May 1998, pp. 547-552.

MUROFUSHI T., SONEDA S. (1993), Techniques for reading fuzzy measures (iii) : interaction index, *Proc. of the 9th Fuzzy System Symposium*, Sapporo, Japan, Mai 1993, pp. 693-696.

ROUSSEL D., HALBER A. (1997), Filtering errors and repairing linguistic anomalies for Spoken Dialogue Systems, *Proc. of the Workshop Interactive Spoken Dialog Systems*, *ACL/EACL*, Madrid, Juillet 1997, pp. 74-81.

ROUSSEL D., LOPEZ P. (1999), Contribution à l'analyse robuste non déterministe pour les systèmes de dialogue parlé, *TALN'99*, ce volume.

ROUSSEL D., PERNEL, D. (1998), Intégration de prédictions linguistiques dans un système de reconnaissance de la parole : une expérience utilisant une grammaire d'arbres lexicalisée, *TALN'98*, France, Juin 1998, pp. 122-132.

SHAPLEY L.S. (1953), A value for n-person games. In H.W. KUHN, A.W. TUCKER, Eds., *Annals of Mathematics Studies* 2(28), Princeton University Press, pp. 307-317.

De SMEDT K., KEMPEN G. (1991), Segment grammar : a formalism for incremental generation. In C. L. PARIS et al., Eds., *Natural language generation and computational linguistics.* Boston/Dordrecht/London: Kluwer Academic Publishers, pp. 329-349.

Van NOORD G., BOUMA G., KOELING R., NEDERHOF M-J. (1998), Robust grammatical analysis for spoken dialogue systems, *Natural Language Engineering* 1(1), Cambridge University Press, pp. 1-48.