Vers l'évaluation de systèmes de dialogue homme-machine : de l'oral au multimodal

Frédéric Landragin

TALN 2008, Avignon

Constats préalables

1. Beaucoup de propositions récentes sur l'évaluation, même dans le domaine relativement restreint qu'est le dialogue homme-machine en langage naturel
   - propositions sur l'évaluation de la compréhension de la parole
   - propositions concernant la compréhension de la multimodalité (pas encore dans la communauté française)
   - propositions de métriques, de tests utilisateurs, de méthodes de constitution et d'analyse de questionnaires, etc.
2. Cependant, les systèmes de dialogue restent toujours au niveau de prototypes de recherche
   - sauf quelques systèmes plus ludiques qu'opérationnels
   - sauf quelques systèmes tellement finalisés qu'ils ne s'adressent qu'à un nombre extrêmement réduit d'utilisateurs

Constats préalables (suite)

3. Il risque d'exister bientôt autant de méthodologies d'évaluation que de systèmes proprement dits
   - attention à ne pas proposer une nouvelle méthode juste en vue de pouvoir évaluer les avancées d'un système (exemple classique en multimodal : nouvelle modalité à prendre en compte)
   - attention à bien séparer la conception et l'évaluation d'un système (normalement : pas les mêmes personnes)
4. L'évaluation sert non seulement à améliorer le développement d'un système (diagnostic) mais aussi à comparer les systèmes les uns par rapport aux autres
   - plusieurs campagnes lancées, mais plus difficiles à mettre en œuvre que pour les autres domaines du TAL
   - les systèmes étant très différents les uns des autres (ils n'ont jamais exactement la même finalité), peut-on vraiment comparer ? évaluer ?

Méthodologies existantes pour le dialogue oral

Stock dans lequel chaque évaluateur peut piocher

- méthodes à visée qualitative, méthodes visant un diagnostic, méthodes « boîte noire », méthodes « boîte transparente », etc.
- un seul type de test devient insuffisant

Quelques exemples maintenant classiques

- MADCOW (Hirschman, 1992) = apporte la notion de **gabarit** qui caractérise les solutions minimales et maximales à une requête
- PARADISE (Walker et al., 2001) = maximisation de la satisfaction de l'utilisateur avec la **satisfaction de la tâche** comme référence
- (López-Cózar et al., 1997) = propose d'évaluer en **générant automatiquement des énoncés** utilisateurs de test, donc en modélisant des utilisateurs et en incluant des modèles d'erreurs
- DQR (Zeiliger et al., 1997) puis DCR (Antoine, Caelen, 1999) = principe original consistant à questionner le système sur le point à évaluer
- PEACE (Devillers et al., 2002) = paraphrase pour historique du dialogue

Méthodologies existantes pour le dialogue multimodal

Quelques propositions, globalement moins abouties :

- PROMISE (Beringer et al., 2002) = extension de PARADISE à la multimodalité, avec affectation de scores aux différentes entrées et sorties multimodales. Problème : proposition approximative, bien en-deçà de la variété des phénomènes multimodaux
- (Bernsen, Dybkjaer, 2004) = questionnaire après test utilisateurs. Problème : questions superficielles, par exemple : « avez-vous utilisé la souris ou avez-vous pointé sur l'écran ? »
- (Dybkjaer et al., 2004) = malgré le titre de l'article, c'est plus une revue de méthodes qu'une proposition...
- µEVAL (Vuurpijl et al., 2004) = outil pour la transcription des données multimodales et l'évaluation d'un système. Problème : l'évaluation ne concerne que les tours de parole, et peu la multimodalité

Il y a donc un retard du multimodal sur l'oral...

De DQR vers DQR multimodal

1. Principes de DQR
   - méthodologie de type « boîte noire » permettant de faire un diagnostic du système, qui repose sur des tests génériques pour l'évaluation de la compréhension d'un énoncé isolé
   - prix à payer : aspects contextuels négligés (cf. PEACE)
   - l'évaluation prend la forme d'une question Q adressée au système :
     - D = « vous prenez la rue à droite et vous la suivez »
     - Q = « suivre rue à droite ? »
     - R = « oui » (donc l'anaphore a bien été comprise)
   - sept niveaux caractérisent la portée des questions Q posées
2. Approche pour son extension à la multimodalité
   - prise en compte dans un premier temps du geste de désignation dans les questions Q, puis d'autres aspects de la multimodalité
   - l'extension se fait niveau par niveau

Niveau 1 = information explicite

Principe

- il s'agit du repérage d'une information explicitée dans l'énoncé, l'intérêt étant de tester la bonne compréhension de l'énoncé littéral compte tenu de la grande variabilité du langage spontané
- Q reprend une partie de l'énoncé et demande confirmation

Liste des paramètres liés à la multimodalité

- D = « mets ça ici » + geste en (x_1, y_1) + geste en (x_2, y_2)
- test de la référence à l'objet : **Q = « ça ? » + geste en (x_1, y_1)**
- test du positionnement : **Q = « mettre ici ? » + geste en (x_2, y_2)**
- test de l'appariement des gestes aux expressions référentielles : **« mettre ça ? » + geste en (x_2, y_2)**
- test de la gestion de la succession temporelle : **« mettre ça ici ? » + geste en (x_2, y_2) + geste en (x_1, y_1)**
- test de la gestion de la synchronisation temporelle : génération de Q avec différents temps de latence entre gestes et expressions

Niveau 2 = information implicite

Principe

- concerne la résolution des anaphores, des ellipses, des incomplétudes et autres informations implicites mais récupérables aux niveaux syntaxique et sémantique
- la résolution de la référence étant concernée, la multimodalité aussi

Eléments liés à la multimodalité

- test de la résolution de la référence à un objet en introduisant des caractéristiques de l'objet, jusqu'à son identifiant unique :
  - **Q = « mettre cet objet ? » + geste en (x_1, y_1)**
  - **Q = « mettre ce fichier LaTeX ? » + geste en (x_1, y_1)**
  - **Q = « mettre 'submis.tex' ? » (sans geste)**
  - **Q = « mettre obj_{4353} ? » (sans geste)**
- exemple plus complexe : D = « déplace ça ici » + geste en forme de courbe qui illustre une trajectoire de déplacement
- test du geste illustratif : **Q = « suivre cette trajectoire ? » + geste...**

Niveau 3 = inférence

Principe

- il s'agit ici de la construction du sens complet de l'énoncé, la difficulté étant l'identification des sous-entendus
- appel à des raisonnements de sens commun et à des inférences pragmatiques
- exemple donné : D = « je voudrais un aller-retour pour Paris »
  Q = « vouloir billet ? »

Extension à la multimodalité

- cet aspect est indépendant des modalités et reste valable dans l'état
- néanmoins, des inférences peuvent être identifiées lors de l'utilisation consécutive de plusieurs gestes

Niv.4 = interprétation du type d'acte illocutoire

Principe

- on entre ici dans les niveaux de dialogue : capacité du système à identifier le bon type d'acte de langage, même en cas d'acte de langage indirect
- exemple donné : D = « un billet pour Paris »
  Q = « est-ce une demande ? »

1er cas de dialogue multimodal : dans lequel les gestes et les autres modalités de communication restent co-verbaux :

- l'information qu'ils apportent s'ajoute à celle portée par l'énoncé oral en langage naturel
- acte de l'énoncé multimodal = acte de l'énoncé oral
- donc pas de changement par rapport à DQR

Niv.4 = interprétation du type d'acte illocutoire (suite)

2ème cas de dialogue multimodal :

- le dialogue multimodal peut inclure des quasi-linguistiques, et, d'une manière générale, des messages effectués avec des modalités autres que le langage naturel et portant en eux-mêmes un acte de communication similaire à un acte de langage
- exemple : interface graphique autorisant un lexique prédéfini de gestes quasi-linguistiques (croix pour supprimer, flèche pour déplacer) :
  - test de l'acte de l'énoncé oral :
    Q = « cet énoncé est-il une demande ? »
  - test de l'acte du geste :
    Q = « ce geste est-il une demande ? »
    Q = « ce geste accompagne-t-il la parole ? »
  - test de l'acte de l'énoncé multimodal :
    Q = « l'énoncé dans son ensemble est-il une demande ? »

Niv.5 = reconnaissance des intentions

Principe

- il s'agit ici de déterminer les intentions ou les buts sous-jacents aux énoncés, donc à un niveau plus profond que le niveau 4
- on interroge explicitement les états intentionnels, avec des questions **Q telles que « l'utilisateur sait-il que [...] ? »**

Extension à la multimodalité

- il y a des intentions sous-jacentes aux gestes aussi bien qu'aux énoncés oraux, mais les intentions elles-mêmes (désirs, intentions, croyances) sont amodales
- ce niveau reste valable dans l'état

Niv.6 = pertinence de la réponse

Principe

- on interroge la pertinence des réponses du système
- aspects couverts très nombreux et très complexes à modéliser : adéquation des réponses par rapport :
  - aux énoncés initiaux de l'utilisateur
  - aux connaissances du système
  - aux moyens de communication
  - au profil de l'utilisateur, etc.
- quelques exemples donnés, qui interrogent à la fois la forme et le contenu de la réponse : « cette question est-elle nécessaire ? », « cette proposition est-elle possible à cet instant ? »

Eléments liés à la multimodalité

- aspects liés à la multimodalité en sortie : **« le choix des modalités de sortie est-il pertinent ? »** ; **« le message est-il surchargé ? »** ; **« le message est-il redondant ? »** ; **« les informations présentées sont-elles synchronisées ? »** ; etc.

Niv.7 = pertinence de la stratégie

Principe

- ce dernier niveau teste la qualité de la stratégie de dialogue, c'est-à-dire si elle a été efficacement menée et si elle est réussie
- la stratégie de gestion de la tâche est également testée
- quelques exemples :
  - **Q = « le client est-il content ? »**
  - **Q = « y a-t-il trop de questions de confirmation indirectes ? »**

Extension à la multimodalité

- aspects indépendants des modalités
- donc ce niveau reste valable dans l'état

De DCR vers DCR multimodal

1. Principe par rapport à DQR
   - la question évaluative Q est remplacée par un contrôle C
   - qui consiste en une simplification ou reformulation de l'énoncé initial
   - ce qui minimise le problème de la capacité du système à répondre à cette question parfois métalinguistique et souvent complexe...
2. Extension à la multimodalité
   - on paraphrase de manière simple et non ambiguë les références multimodales, avec par exemple la description en langage naturel de coordonnées spatiales :
     - **C = « mets 'submis.tex' en (x_1, y_1) »**
     - **C = « déplace obj_{4353} de (x_1, y_1) à (x_2, y_2) »**
     - **C = « déplace obj_{4353} selon les points de passage (x_3, y_3), (x_4, y_4), (x_5, y_5)… »**
   - les autres aspects ne posent pas plus de problème que le passage de DQR à DCR

PEACE multimodal ?

1. Principes
   - reformulation de l'historique du dialogue en une phrase unique
   - utilisation de cette phrase pour l'évaluation contextuelle de l'énoncé
2. Extension à la multimodalité du principe de reformulation
   - concernant la référence : l'historique doit conserver à la fois :
     - l'identifiant des référents *(pour ressortir ceux-ci lors de l'interprétation d'une anaphore)*
     - les mentions utilisées pour y référer *(pour interpréter les ellipses nominales, les références mentionnelles ou métalinguistiques, etc.)*
   - en dialogue multimodal, l'état de la scène visuelle doit également être conservé à chaque étape, ce qui conduit à un historique linguistique, un historique gestuel, un historique visuel, etc. : « l'objet que je viens de désigner », « celui qui était à droite »...
   - mentions multimodales, état de scène visuelle... :
     il semble impossible de paraphraser tous ces aspects

Bilan

1. De l'oral au multimodal
   - l'évaluation de systèmes de dialogue multimodaux s'avère en fin de compte plus complexe que celle des systèmes oraux (déjà délicate)
   - tous les paradigmes prévus pour l'oral ne sont pas extensibles
   - l'exemple classique et terriblement simple « mets ça ici » permet de tester efficacement l'extension à la multimodalité, comme quoi cet exemple n'a pas fini d'être exploité...
   - DQR et DCR multimodaux : les questions soulevées sont aussi des éléments qui peuvent être utiles lors de la conception d'un système
2. Quelques pistes pour la suite
   - la conception des systèmes de dialogue est en train d'évoluer, avec notamment l'exploitation d'environnements de développement capables d'automatiser certaines phases de conception (exemple : approche MDE-MDA par dérivation et génération de modèle). Conséquence : il faudrait évaluer aussi ces environnements
   - méthode d'évaluation par défi : a priori pertinente pour le multimodal