Un modèle stochastique de compréhension de la parole à 2+1 niveaux

Hélène Bonneau-Maynard

LIMSI/CNRS
Groupe Traitement du Langage Parlé
helene.maynard@limsi.fr

Fabrice Lefèvre

LIA/Université d'Avignon
Equipe Dialogue Homme-Machine
fabrice.lefevre@univ-avignon.fr

ABSTRACT

In this paper an extension is presented for the 2-level stochastic speech understanding model, previously introduced in the context of the ARISE corpus [6]. In the new model, the additional stochastic level is in charge of the attribute value normalisation. Due to data sparseness, the full (3 level) model is not applicable straightforwardly and a variant is introduced where the conceptual decoding and value normalisation phases are decoupled.

The proposed approach is evaluated on the French MEDIA task (hotel booking and tourist information). This recent corpus has the advantage to be semantically annotated with conceptual segments, which allows for a direct training of the 2-level model. We also present some further model improvements such as the modality propagation or the 2-step hierarchical recomposition. On the whole, the various proposed techniques reduce the understanding error rate from 37.6% to 28.8% on the development set (24% relative improvement). This model has been engaged in the 2005 MEDIA evaluation campaign where it achieved the best results amoung the 5 participants with an error rate of 29%.

1. INTRODUCTION

Les approches stochastiques pour la compréhension de la parole offrent une alternative efficace aux approches par règles en réduisant le recours à l'expertise humaine et, ainsi, le coût global de développement du modèle [1, 2, 3, 4, 5]. Dans un précédent article [6], le développement d'un modèle stochastique a été présenté sur la tâche de dialogue ARISE (renseignements sur des horaires de trains et réservation de billets). Le présent article décrit les deux principales améliorations de notre modèle de compréhension par rapport au modèle de base : l'utilisation de corpus d'apprentissage segmenté sémantiquement (par rapport aux précédents travaux dans lesquels l'annotation sémantique portait sur des mots clés uniquement) et le recours aux modèles stochastique pour la normalisation des valeurs des attributs.

Avec une annotation sémantique par mots clés telle qu'elle était réalisée dans le corpus ARISE, la modélisation stochastique à 2 niveaux était fondée sur des segments sémantiques de taille fixe. Ces segments, centrés sur les mots clés précisés par l'annotation, étaient déterminés artificiellement a posteriori. Dans le nouveau schéma d'annotation, l'annotation sémantique est alignée sur des séquences de mots qui sont déterminés par les annotateurs humains. Les segments ont donc des tailles variables, ajustées en fonction des situations, et doivent permettre un meilleur apprentissage des modèles à 2 niveaux. Parallèlement, le modèle de compréhension est transformé en un modèle stochastique complet : la normalisation des valeurs est intégrée au processus stochastique, alors qu'elle était précédemment obtenue par le biais des règles semi-manuelles.

Notre participation au projet d'évaluation Technolangue EVALDA-MEDIA nous a permis de mettre au point et d'évaluer les améliorations proposées sur une nouvelle tâche. La tâche MEDIA concerne la réservation de chambres d'hôtel accompagnées de demande d'informations touristiques en France, les informations provenant d'une base de données disponible sur Internet.

L'organisation de l'article est la suivante : la prochaine section décrit la représentation sémantique. La modélisation stochastique intégrée est décrite dans la section 3. Finalement, après les descriptions du corpus MEDIA et des conditions expérimentales, la dernière section présente les résultats obtenus sur l'ensemble de développement et lors du test.

2. REPRÉSENTATION SÉMANTIQUE

La représentation sémantique du projet MÉDIA, décrite en détail dans [8], est fondée sur des structures d'attributs-valeurs dans lesquelles les relations hiérarchiques entre les concepts sont implicitement représentées par les noms et l'ordre des attributs. Chaque tour de parole est segmenté en un ou plusieurs segments sémantiques alignés sur les séquences de mots. Pour la compréhension littérale, un énoncé est représenté par une suite de segments sémantiques, chaque segment étant représenté par un triplet qui contient :

– le mode : affirmatif '+', négatif '-', interrogatif ' ?' ou optionnel '~' ;
– le nom de l'attribut représentant le sens de la séquence de mots ;
– la valeur de l'attribut.

L'ordre des triplets dans la représentation suit l'ordre des segments dans l'énoncé. Les valeurs des attributs sont des nombres, des noms propres ou des classes sémantiques qui regroupent des unités lexicales qui sont équivalentes pour la tâche. Des segments d'un même énoncé peuvent porter des modes différents.

La hiérarchie des attributs de base de la tâche est définie dans un dictionnaire sémantique. Différentes classes d'attributs y apparaissent. Certains attributs, dits **attributs BD**, (ex : `nom-hotel`) sont directement issus de la base de données liée à la tâche. Les attributs dits **mo-**

difieurs (e.g. `comparatif`), associés aux attributs BD, permettent d'en modifier le sens. Les attributs dits **généraux** correspondent aux commandes relatives à la tâche (`reservation`), ou au dialogue (`reponse`). Un des attributs généraux est utilisé pour représenter les références linguistiques (`lien-ref`). Le dictionnaire sémantique définit également pour chaque attribut l'ensemble des valeurs normalisées qui lui sont associées. Trois types de définitions de valeurs sont utilisées : par liste de valeurs (ex : `comparatif` avec les valeurs `inferieur`, `superieur`...), par expressions régulières (ex : dates), ou enfin sans restriction (ex : noms de clients).

Une représentation hiérarchique des connaissances, permettrait d'exprimer les relations complexes entre les constituants explicitement. Cependant, dans une approche orientée corpus, une représentation hiérarchique complique grandement l'annotation manuelle des données d'apprentissage. Pour tenir compte de cette difficulté, la représentation MEDIA, qui repose sur une représentation à un seul niveau, a été enrichie d'un ensemble de **spécifieurs** qui, combinés avec les noms des attributs BD et les modifieurs, permettent d'établir la relation au constituant principal. En reproduisant les relations représentées par les spécifieurs et en utilisant l'ordre des segments, il est alors possible de reconstruire une représentation arborescente à partir de la représentation à plat.

Dans nos travaux précédents, l'annotation sémantique était fondée sur des mots-clés : les attributs étaient associés uniquement aux mots qui déterminaient leur valeur. Dans le nouveau schéma d'annotation, la requête est découpée en segments sémantiques : les attributs sont maintenant associés à des *séquences de mots* - les segments - qui en désambiguïsent le mieux leur sens.

3. MODÉLISATION STOCHASTIQUE INTÉGRÉE

Le but de la compréhension stochastique est de déterminer la séquence de concepts $C = c_1 c_2 \dots c_N$ qui va représenter le sens de l'énoncé en posant l'hypothèse qu'il existe une correspondance séquentielle entre les concepts et les séquences de mots [1]. Soit $W = w_1 w_2 \dots w_N$ la séquence de mots de la phrase, le processus de compréhension recherche la séquence de concepts qui maximise la probabilité *a posteriori*, qui peut être écrite selon la formule de Bayes :

$$\hat{C} = \arg\max_C \Pr(C|W) = \arg\max_C \Pr(W|C)\Pr(C)$$

$\Pr(W|C)$ est estimé au moyen de probabilités n-grammes de mots connaissant le concept associé au mot i :

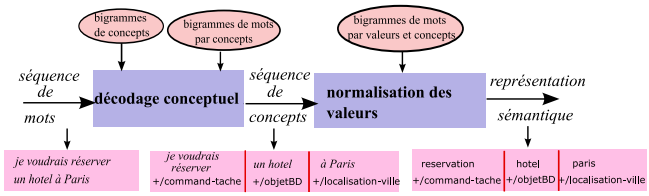
$$\Pr(W|C) \simeq \prod_{i=1}^{N} \Pr(w_i|w_{i-1}, \dots, w_{i-n}, c_i)$$

et $\Pr(C)$ est estimé par des probabilités m-grammes de concepts :

$$\Pr(C) \simeq \prod_{i=1}^{N} \Pr(c_i|c_{i-1}, \dots, c_{i-m})$$

A partir de cette formulation, plusieurs approches peuvent être considérées selon l'ordre des modèles utilisés pour produire l'estimation de $\Pr(W|C)$ et de $\Pr(C)$. Généralement, des bigrammes de concepts $\Pr(c_i|c_{i-1})$ ($m = 1$)

FIG. 1: Compréhension stochastique : modélisation à 2+1 niveaux.



sont suffisants pour modéliser les séquences de concepts. Lorsque l'on dispose d'une annotation sémantique segmentale pour le corpus d'apprentissage, on peut envisager d'utiliser des bigrammes de mots conditionnés au concept $\Pr(w_i|w_{i-1}, c_i)$ ($n = 1$). On parle alors d'une modélisation stochastique à 2 niveaux [6]. Afin d'améliorer la généralisation des modèles, il est possible d'utiliser un ensemble de classes lexicales.

La figure 1 représente les étapes du processus de compréhension. Dans notre modélisation, un concept est constitué de la combinaison du nom de l'attribut et de sa modalité. La première phase (*décodage conceptuel*) cherche à déterminer le concept le plus probable pour chaque sous-séquence de mots de l'énoncé. La seconde phase (*normalisation des valeurs*) consiste à déterminer, pour chaque attribut associé à chaque séquence de mots, la forme normalisée de la valeur attendue selon la représentation sémantique. Dans le schéma de la figure 1 par exemple, la séquence de mots *je voudrais réserver* associée à l'attribut `command-tache` lors de la phase de décodage conceptuel, doit être transformée en sa forme normalisée : `reservation`. Une même forme normalisée peut être produite par différentes séquences de mots : *je voudrais réserver, pour ma réservation....*

La normalisation est classiquement obtenue au moyen d'un ensemble de règles. Dans la modélisation décrite ici, nous proposons d'étendre la modélisation stochastique y compris à la phase de normalisation. Dans ce contexte, le modèle de compréhension peut être considéré comme un modèle intégrant 3 niveaux : mot, concept et valeur, comme indiqué dans les équations :

$$\begin{aligned} \hat{C}, \hat{V} &= \arg\max_{C,V} \Pr(C, V|W) \\ &= \arg\max_{C,V} \Pr(W|C, V)\Pr(C, V) \end{aligned}$$

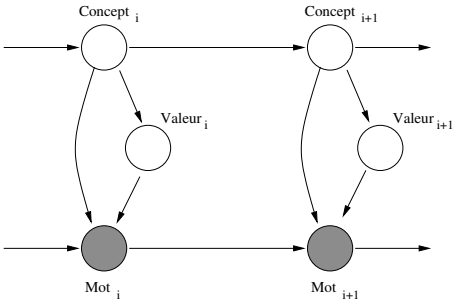
Cette modélisation conduit à des probabilités conditionnées par les valeurs normalisées et par conséquent à prendre en compte un nombre d'états considérablement augmenté. De plus, une telle solution est peu adaptée dans le cas où la liste des valeurs associées à un concept est ouverte (cas par exemple des nombres ou des noms de clients). Pour ces raisons, dans notre modèle, la normalisation n'a pas été totalement intégrée à la phase de décodage mais découplée de celui-ci pour être effectuée dans un second processus stochastique :

$$\hat{C} = \arg\max_C \sum_V \Pr(W|C, V)\Pr(C, V) \quad (1)$$

$$\begin{aligned} \hat{V} &= \arg\max_V \Pr(\hat{C}, V|W) \\ &= \arg\max_V \Pr(W|\hat{C}, V)\Pr(\hat{C}, V) \end{aligned} \quad (2)$$

L'équation 1 permet une meilleure généralisation du modèle conceptuel. Par ailleurs, l'hypothèse que la normali-

FIG. 2: Représentation du modèle stochastique de compréhension selon le formaliste des réseaux Bayesiens dynamiques.



sation des valeurs n'influence pas ou très peu la détermination des attributs paraît acceptable. Selon cette hypothèse, l'estimation des modèles stochastiques de normalisation peut être effectuée indépendemment de celle des modèles de concepts. Dans notre cas, un bigramme de mots est construit pour chaque couple (concept, valeur). Les indépendances conditionnelles sur les probabilités sont représentées dans le diagramme de la figure 2 dans le formalisme des réseaux Bayesiens dynamiques. Ainsi, le modèle envisagé à 3 niveaux est finalement assimilé à un modèle à 2+1 niveaux et le décodage depuis la séquence de mots jusqu'à la séquence de concepts associés à leurs valeurs normalisées s'effectue en deux temps, comme indiqué dans la figure 1. La forme finale obtenue est une suite de triplets [`mode`, `attribut`, `valeur`] comme attendu dans la représentation sémantique.

4. EXPÉRIENCES ET RÉSULTATS

Pour la compréhension littérale, le corpus MEDIA (tableau 1) consiste en une portion d'apprentissage de 10965 requêtes client, une portion de développement de 1009 requêtes, et un corpus de test de 3003 requêtes. Tous sont transcrits et annotés sémantiquement. Les 676 noms propres apparaissant dans le corpus correspondent essentiellement aux noms de villes (201) et d'hôtels (548) et sont très ambigus pour la tâche.

L'outil d'évaluation développé pour le projet MEDIA effectue un alignement entre deux représentations sémantiques afin de les comparer en terme de suppression, insertion et substitution. En mode *complet*, l'alignement est effectué sur tout le triplet [`mode`, `attribut`, `valeur`]. En mode *simplifié*, seuls deux modes (affirmatif et négatif) sont distingués (les modes ? et ~ étant projetés sur le mode +). Enfin, en mode *relâché*, les spécifieurs ne sont pas pris en compte dans l'identification des noms d'attributs.

4.1. Normalisation par règles (référence)

Les résultats décrits dans cette partie reposent sur la modélisation décrite dans [6]. Le modèle conceptuel à deux niveaux est appris sur les 11k énoncés du corpus d'apprentissage. Grâce à l'annotation segmentale, aucune transformation des annotations n'est nécessaire - comme l'utilisation de marqueurs de concepts - pour estimer les bigrammes de concepts. Un ensemble de classes lexicales est utilisé pour généraliser l'estimation des bigrammes. Les classes sont dérivées des attributs liés à la base de données et se limitent à des mots qui sont syntaxiquement et sémantiquement équivalents pour la tâche (par exemple `aéroport charles de gaulle`) peut apparaître sous différentes formes de surface (*Aéroport Charles de Gaulle*, *aéroport de Gaulle*...). Afin de résoudre les ambiguïtés liées au fait qu'une même forme de surface peut correspondre à différentes classes lexicales (par exemple un nom de ville correspond souvent aussi à un nom d'hôtel), les classes utilisées sont déterminées sélectivement selon le concept indiqué dans l'annotation.

TAB. 1: Caractéristiques principales des énoncés client des corpus d'apprentissage et de développement.

	appr.	dév.
nombre d'énoncés	10965	1009
nombre moyen de mots par énoncé	4.8	5.4
nombre de mots différents	2115	794
nombre d'attributs observés	29980	3125
nombre moyen d'attributs par énoncé	2.7	3.1
nombre d'attributs différents	144	106

L'annotation segmentale permet également de dériver du corpus d'adaptation un ensemble de règles de réécritures qui sont utilisées pour la phase de normalisation des valeurs. Une règle de réécriture est dérivée pour chaque observation d'un concept dans le corpus d'adaptation. Afin d'améliorer leur généralisation, les règles sont regroupées indépendemment du mode, ainsi que pour tous les attributs dont les noms ne diffèrent que par les spécifieurs.

Le processus de compréhension est effectué après une transformation des énoncés dans laquelle les séquences de mots qui correspondent à des classes lexicales déterminées lors de l'apprentissage sont remplacées par le nom de la classe correspondante. Les mots vides de sens comme *euh, ah* sont également retirés de l'énoncé avant le décodage conceptuel. Les taux d'erreur de compréhension pour ce système de référence sont donnés dans la première ligne du tableau 2 : de 37,6% en mode complet à 23,0% en mode relâché et 20,8% pour les valeurs seules.

4.2. Normalisation stochastique des valeurs

La modélisation 2+1 avec une normalisation stochastique découplée, décrite dans la section 3, est ensuite substituée à la normalisation par règles du système de référence. La normalisation stochastique permet une amélioration relative des résultats de 6% par rapport à une normalisation par règles (deuxième ligne du tableau 2, `norm. stoch.`). Le gain relatif passe à 7.6% si on complète la normalisation stochastique avec un système de pénalités (troisième ligne du tableau 2, `norm. stoch.+`). Les pénalités sont appliquées en distinguant trois cas :

1. aucun mot de la séquence traitée n'a jamais été observé pour la valeur considérée ;
2. la forme normalisée apparaît telle quelle parmi les mots de la séquence traitée ;
3. tous les autres cas.

Les probabilités fournies par le module de normalisation stochastique reçoivent un malus dans le premier cas, un bonus dans le deuxième et ne sont pas modifiées dans le dernier.

Les résultats tendent à montrer que la normalisation stochastique représente une alternative efficace à la normalisation par règles. Elle permet une bonne généralisation

TAB. 2: Taux d'erreur de compréhension (%) sur le corpus de développement : complet, relâché (2 modes, sans spécifieurs) et valeurs uniquement. La colonne *#cpt* donne le nombre de concepts dans le modèle.

	#cpt	complet	relâché	valeurs
référence	390	37,6	23,0	20,8
norm. stoch.	390	36,9	21,8	19,6
norm. stoch.+	390	36,9	21,6	19,2
modes+	393	35,6	21,3	19,2
spécifieurs+	344	28,8	21,5	19,7
test (*officiel*)	344	29,0	21,6	19,7
test (*corrigé*)	344	27,7	20,4	18,5

en dépit du très petit nombre d'observations disponible pour chaque valeur de chaque concept. Afin d'augmenter la quantité de données par valeur, une normalisation partagée a été évaluée, comme dans le cas de la normalisation par règles. Si nos expériences ont montré que cette méthode permettait un petit gain avec un partage indépendant du spécifieur, ce gain disparaît après l'intégration des deux techniques présentées ci-après. Elle n'a donc pas été retenue dans le système complet.

4.3. Identification des modes

La différence importante entre les résultats en mode complet et en mode relâché indique que les confusions sur les modes sont la source de nombreuses erreurs. Une difficulté vient de ce qu'au cours de l'annotation manuelle du corpus, un mode positif a été affecté à tous les segments associés à l'attribut `null`, introduisant ainsi des discontinuités artificielles qui bloquent la propagation des modes lors de la phase de décodage conceptuel (limitée aux successions de 2 concepts par la modélisation en bigrammes). Une modification automatique du mode de tous les segments `null` situés entre deux segments de même mode non affirmatif a donc été réalisée. Comme le montre la troisième ligne du tableau 2 (`modes+`), cette simple propagation des modes a permis - en introduisant 3 concepts supplémentaires : `-/null`, `?/null` et `~/null` - d'améliorer les résultats d'un score relatif de 3,6% sur le taux de compréhension en mode *complet*.

4.4. Recomposition hiérarchique

Le modèle stochastique ne permet pas de gérer les dépendances hiérarchiques à long terme. Afin de prendre en compte les limites du modèle, l'identification des spécifieurs de concepts - qui portent les dépendances à long terme dans la représentation MEDIA - est transformée en une procédure en deux étapes.

La plupart des spécifieurs sont activés dans des contextes sémantiques particuliers portant les dépendances à long terme, mais peuvent être décrits en terme de présence de concepts de base dans l'énoncé. Par exemple, le spécifieur `reservation` apparaît presque exclusivement dans le cas où le concept `command-tache` avec la valeur `reservation` a été identifié pour un autre segment de l'énoncé. Ces contextes peuvent donc être retrouvés *après* la phase de décodage conceptuel. C'est ainsi que les modèles sont *dorénavant* appris avec des concepts sans de spécifieurs - réduisant ainsi le nombre de concepts du modèle de 390 à 344. Les spécifieurs sont déterminés dans un second temps par un ensemble de règles. Cette procédure en deux temps a permis une amélioration relative des résultats de 19% (ligne `specifieurs+` du tableau 2), avec une très faible détérioration dans la normalisation des valeurs (de 19,2 à 19,7%), qui peut s'expliquer par l'augmentation du nombre possible de valeurs par concept.

CONCLUSION

Dans cet article, nous avons proposé et évalué un modèle de compréhension de la parole stochastique à 2+1 niveaux. Par rapport au modèle de référence à 2 niveaux, une amélioration relative de 24% du taux d'erreur de compréhension a été obtenue avec un modèle comptant 344 concepts (incluant des constituants hiérarchiques). Une part importante de l'amélioration des performances provient d'une technique simple mais efficace de traiter la composition hiérarchique en 2 étapes. Toutefois, les erreurs sur les spécifieurs représentent toujours 25% du total des erreurs. Une amélioration de la méthode actuelle par l'introduction de classifieurs statistiques devrait permettre une meilleure couverture des contextes sémantiques pour une recomposition hiérarchique plus précise des concepts.

Le système développé a participé à l'évaluation MEDIA 2005. Les résultats sur l'ensemble de test (3003 énoncés utilisateurs) sont donnés dans la dernière partie du tableau 2. La première ligne correspond aux résultats officiels après adjudication, la seconde ligne correspond aux résultats obtenus après correction d'une erreur de manipulation lors de l'évaluation (échange de 2 fichiers). Avec un taux d'erreur de 29.0% en mode complet, le système se classe 1er parmi les 5 participants.

RÉFÉRENCES

[1] E. Levin and R. Pieraccini, "Concept-based Spontaneous Speech Understanding System," in ESCA Eurospeech, Madrid, 1995.

[2] R. Schwartz, S. Miller *et al.*, "Hidden Understanding Models for Statistical Sentence Understanding," in IEEE ICASSP, Munich, 1997.

[3] F. Pla , A. Molin *et al.*, "Language Understanding using Two-level Stochastic Models with POS and Semantic Units," LNCS series, vol. 2166, 2001.

[4] Y. He and S. Young, "Hidden Vector State Models for Hierarchical Semantic Parsing," in IEEE ICASSP, Hong Kong, 2003.

[5] C. Raymond, F. Bechet *et al.*, "On the use of finite state transducers for semantic interpretation," Speech Communication, vol. 48 :3-4, pp 288-304, 2006.

[6] F. Lefevre and H. Bonneau-Maynard, "Issues in the development of a stochastic speech understanding system," in ICSLP, Denver, 2002.

[7] L. Devillers, H. Maynard *et al.*, "The French MEDIA/EVALDA project : the evaluation of the understanding capability of Spoken Language Dialogue Systems," in LREC, Lisbon, 2004.

[8] H. Bonneau-Maynard, S. Rosset *et al.*, "Semantic annotation of the MEDIA corpus for spoken dialog," in ISCA Eurospeech, Lisbon, 2005.