Modèle de langage sémantique pour la reconnaissance automatique de parole dans un contexte de traduction

Quang VU-MINH quang.vu-minh@imag.fr
Laurent BESACIER laurent.besacier@imag.fr
Hervé BLANCHON herve.blanchon@imag.fr
Brigitte BIGI brigitte.bigi@imag.fr

CLIPS-IMAG Lab. UJF, BP53, 38041 Grenoble cedex 9, France

Résumé

Cet article présente une méthode pour intégrer des informations sémantiques dans le modèle statistique de langage (ML) d'un système de Reconnaissance Automatique de Parole (RAP). Ce travail a été réalisé dans le cadre d'un projet global de traduction automatique de parole. L'approche de traduction est fondée sur un langage pivot ou *Interchange Format (IF),* qui représente le sens de la phrase indépendamment de la langue. Notre méthode consiste à introduire des informations sémantiques dans le modèle de langage utilisé pour la RAP, sous forme de classes qui correspondent directement à des concepts du langage pivot (IF) utilisé en traduction. Des résultats préliminaires montrent qu'avec cette approche, le système de reconnaissance peut analyser directement en IF un volume important de données de dialogues, sans faire appel au système de traduction (35% des mots ; 58% des tours de parole) et sans dégrader le système global.

1 Introduction

Dans un système de traduction ou compréhension automatique de parole, le rôle du module de reconnaissance automatique de la parole (RAP) est d'obtenir une hypothèse textuelle a partir du signal tandis que, généralement, cette hypothèse est ensuite traitée séparément par un autre module de compréhension ou d'analyse qui transforme le texte en une représentation sémantique. Tous ces modules utilisent des ressources linguistiques comme des dictionnaires, modèles de langage et/ou grammaires, mais ils sont souvent indépendants l'un de l'autre. Bien qu'il y ait quelques travaux (voir Vermobil [1] ou SLT [2]) qui proposent une utilisation intelligente des ressources communes entres module de RAP et module d'analyse, à notre connaissance, très peu de travaux expérimentaux proposent d'introduire des informations sémantiques directement dans un module de RAP pour une meilleure intégration du système complet.

Cet article présente une méthode pour intégrer des informations sémantiques dans le modèle de langage statistique du système de reconnaissance. Ce travail a été réalisé dans le cadre d'un projet global de traduction de parole intitulé NESPOLE1 [3]. Au sein du projet, une approche de traduction fondée sur un langage pivot (appelé IF pour *Interchange Format*), qui représente le sens de la phrase indépendamment de la langue, est utilisée. L'architecture de ce système de traduction utilisant l'approche pivot est décrite dans la *figure 1*.

---

[1] see http://nespole.itc.it/

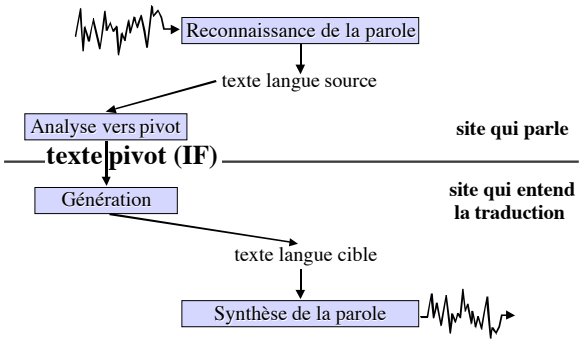

Figure 1: Interaction entre les modules de traduction de parole dans l'architecture IF.

L'avantage le plus évident de l'approche par pivot est la réduction du nombre de modules à réaliser. Si n langues sont impliquées, deux modules, d'analyse et de génération vers/depuis l'IF, pour chaque langue permettent la traduction pour toutes les paires de langues possibles. Si une nouvelle langue vient s'ajouter au projet, il suffit alors de développer les modules d'analyse et de génération vers/depuis l'IF pour cette langue afin qu'elle puisse être intégrée avec les n autres. Le défaut de cette approche réside dans la difficulté à définir le langage pivot, les concepts qui sont couverts, ainsi que sa syntaxe. Cela est vrai même lorsque le domaine est limité à une tâche particulière comme l'information touristique.

L'IF [4] est fondé sur de actes de dialogue (DA) qui sont constitués d'un acte de parole (SA) éventuellement complété par des concepts. L'acte de parole exprime ce que veut ou ce que fait celui qui parle. Les concepts sont sub-divisés en attitudes, prédicats principaux et participants du prédicat. Ils expriment le focus informationnel de ce qui est dit. Actes de parole et concepts peuvent admettre des arguments qui instancient les variables du discours. Les arguments admis par les actes de parole et les concepts sont des arguments supérieurs (top-level arguments). Il existe aussi des arguments dominés (embedded arguments) qui raffinent les arguments supérieurs.

Pour une phrase signifiant "*et je voudrais une chambre simple à 100 euros à Cavalese du 10 au 15 septembre*" prononcée par un client, l'IF est :

```
c:give-information+disposition+price+room(
  conjunction=discourse,disposition=(desire, who=i),
  room-spec=(identifiability=no,single_room),
  price=(quantity=100,currency=euro),
  location=name-cavalese,
  time=(start-time=(md=10),end-time=(md=15, month=9))
)
```

où :

- `c:` indique que c'est le client qui parle
- `give-information+disposition+price+room` est l'acte de dialogue constitué de
  - `give-information` acte de parole qui permet d'instancier :
    - `conjunction=` un argument supérieur de type rhétorique
  - `disposition` un concept de type attitude qui permet d'instancier :
    - `disposition=` un argument supérieur qui est réalisé par

- `desire` qui est la valeur et `who=` qui est un argument dominé dont la valeur est :
- `price` un concept qui joue le rôle de prédicat principal qui permet d'instancier :
  - `price=` un argument supérieur réalisé par `quantity=100` et `currency=euro`
- `room` un concept participant du prédicat qui permet d'instancier :
  - `room-spec=` un argument supérieur réalisé par `identifiability=no` et `single_room`
  - `location=` un argument supérieur réalisé par `name-cavalese`
- `time` un argument supérieur réalisé par :
  - `start-time=(md=10)` et
  - `end-time=(md=15, month=9)`

Le travail présenté ici se situe à l'interface entre le module de reconnaissance et le module d'analyse en IF (voir *figure 1*). Plus précisément, le module de reconnaissance de la parole a été adapté afin d'être capable de délivrer une chaîne partiellement ou complètement analysée en IF. Cela a été achevé par l'usage d'un modèle de langage utilisant des classes qui correspondent directement à des concepts de l'IF. La méthodologie utilisée pour obtenir ce modèle de langage « sémantique » est détaillée dans la section 2. La section 3 présente quelques résultats expérimentaux obtenus tandis que la section 4 propose une première conclusion à ce travail.

2 Construction du modèle de langage

Le modèle de langage statistique constitue un élément important dans un système de Reconnaissance Automatique de la Parole. Il a pour but de donner la probabilité d'existence d'une chaîne de n'importe quelle suite de trois mots (dans le cas du modèle trigramme). Pendant l'apprentissage du modèle, certains mots peuvent être regroupés en classes, si on considère que chaque mot, à l'intérieur d'une classe, a la même probabilité d'occurence. Une classe n'est donc qu'une liste de mots qui sont tous équiprobables dans cette classe. Lorsqu'on a définit des classes, on peut alors directement remplacer les mots du corpus d'apprentissage par leur classe avant l'apprentissage du modèle de langage.

Pour introduire des connaissances sémantiques dans le ML, notre idée est alors la suivante : regrouper tous les mots (par exemple : « bien », « d'accord », « okay », etc.) en des classes correspondant à des entités sémantiques de l'IF, (ces mots correspondent par exemple à la classe « *c:acknowledgment* »). Plusieurs articles [5] [6] ont montré l'intérêt de l'utilisation de classes dans diverses tâches de Traitement Automatique de Langue Naturelle. La plupart des méthodes pour constituer automatiquement des classes (donc des ensembles de mots) utilisent des critères statistiques permettant, par exemple, de diminuer la perplexité d'un modèle de langage. Dans notre cas, notre critère de choix de classes est guidé par la définition du langage pivot et par les concepts les plus utilisés dans l'IF. Notre approche consiste en deux étapes : (1) la sélection des IFs les plus fréquentes à intégrer comme classes dans le nouveau modèle de langage (2) la calcul du modèle de langage proprement dit. Ces étapes toutes automatiques sont détaillées dans le paragraphe suivant.

2.1 Sélection des classes-IF les plus fréquentes

Utiliser toutes les unités sémantiques présentes dans la définition de l'IF, comme classes dans notre modèle de langage, conduirait à un modèle inutilisable pour la reconnaissance automatique de la parole. En effet, le nombre de classes doit être limité et surtout, le nombre d'occurrences de mots

d'une même classe doit être suffisamment important pour que les probabilités apprises soient correctes. Nous avons donc choisi de nous limiter à la sélection de classes-IF les plus fréquemment rencontrées dans les dialogues du projet NESPOLE, correspondant à la tâche que nous voulons traiter. Dans cette étape, nous identifions donc ces IFs les plus fréquentes et les regroupons dans des classes. Une classe correspond alors à un ensemble de mots conduisant à une même représentation IF (on trouve par exemple dans ces classes des actes de dialogue tels que *acknowledge*, *affirm*, *negate* …). La *figure 2* illustre comment cette sélection des IFs les plus fréquentes est réalisée automatiquement à partir d'un corpus textuel brut non annoté manuellement en IF.

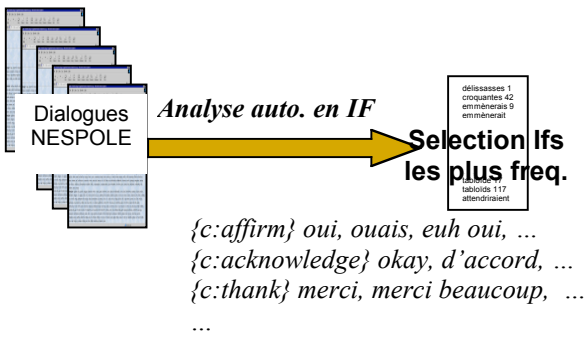

Figure 2: Sélection des IFs les plus fréquentes

L'analyseur en IF automatique du CLIPS [7] a été utilisé pour analyser en IF automatiquement un corpus qui comprend 46 transcriptions de dialogues collectés lors du projet NESPOLE [8]. Ce corpus représente des dialogues possibles entre un client et un agent de voyage concernant l'organisation de vacances, la réservation d'hôtels et les activités sportives ou culturelles, dans la région de Trente en Italie. L'analyseur transforme automatiquement tous ces dialogues en une représentation de langage IF. Nous avons par conséquent un corpus aligné français-IF. Bien sûr, ce corpus n'est pas parfait car l'analyseur fait éventuellement des erreurs, mais nous supposons que, malgré ces erreurs, la distribution des différentes IFs est correctement respectée. Ensuite, nous regroupons ces données alignées par IF et listons toutes les unités sémantiques de dialogue correspondant à une même IF, et obtenons enfin nos classes « sémantiques » dont certaines sont présentées dans la *table 1*. Par exemple, la classe la plus fréquente *affirm* contient les variantes représentant le même sens (*affirmer*) en français.

Le nombre de classes sémantiques obtenues de cette façon étant important, nous avons retenu seulement 41 classes en tenant compte de la taille (le nombre de mots ou variantes dans une même classe) et de la fréquence d'apparition dans les dialogues des classes.

CLASSES IFs	Variantes d'unités sémantiques associées	Pourcentage dans un total de 3194 unités
{c:affirm}	Oui, ouais, mouais…	22%
{c:acknowledge}	d'accord, entendu, ok…	19%
{c:exclamation (exclamation=oh)}	Oh, ah, ha….	4%
…	…	…

Table 1: Exemples de classes IF obtenues automatiquement

2.2 Calcul du modèle de langage

Après avoir obtenu la liste des classes sémantiques, comme illustré dans la *figure 2*, nous nous en servons en combinaison avec les données de l'apprentissage du modèle de langage afin de construire notre nouveau modèle « sémantique ». Ce processus est illustré dans la *figure 3*.

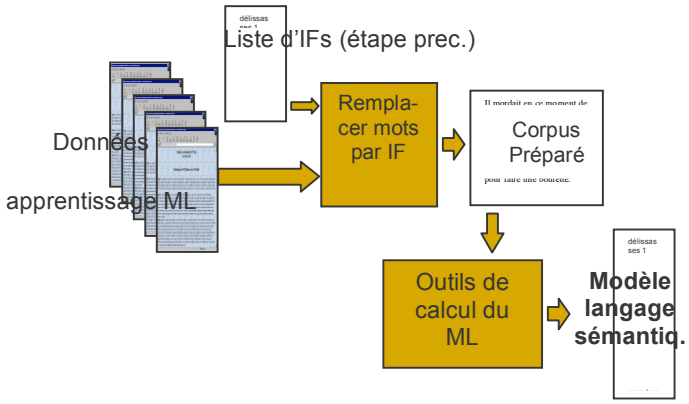

Figure 3: Méthode d'apprentissage du modèle de langage utilisant les classes issues de l'IF

Dans le corpus d'apprentissage du ML qui comprend les 46 dialogues NESPOLE, nous remplaçons tous les mots (ou séquences de mots) qui sont éléments de nos nouvelles classes sémantiques, par le nom de la classe. Il en résulte alors un corpus "préparé" qui contient à la fois des mots français et des IF. Ensuite les outils traditionnels de calcul de ML sont utilisés pour obtenir notre nouveau modèle de langage « sémantique ».

Après avoir eu le nouveau ML proprement construit, nous l'avons intégré et testé dans le système de reconnaissance. La section suivante présente quelques résultats expérimentaux.

3 Résultats expérimentaux

3.1 Description du système de RAP

Notre système RAPHAEL de reconnaissance de parole continue utilise la boite à outils Janus-III du CMU [9]. Le modèle acoustique dépendant du contexte a été appris sur un corpus qui contient 12 heures de parole continue prononcée par 72 locuteurs, issues de la base BREF80. Le vocabulaire contient approximativement 20000 formes lexicales parmi lesquelles quelques-unes sont spécifiques

au domaine de la réservation touristique. Plus de détail sur le système RAP du français utilisé dans NESPOLE se trouvent dans [3].

Un test contradictoire a donc été conduit pour comparer un même système de RAP utilisant d'une part l'ancien modèle de langage, et d'autre part le nouveau modèle de langage « sémantique » obtenu par la méthodologie présentée dans la section 2.

3.2 Corpus de test

Les signaux de test sont 216 tours de parole extraits du corpus de dialogues du projet NESPOLE. La *table 2* illustre quelques exemples de ces tours de parole de test. Dans la deuxième colonne sont présentées les hypothèses textuelles obtenues à la sortie du module de RAP utilisant notre modèle de langage sémantique. Nous remarquons que des tours de parole simples sont déjà complètement analysés en IF. D'autres tours de parole plus complexes sont, eux aussi, analysés partiellement ou complètement en IF.

Phrases de référence	Sortie de RAP avec nouveau ML
oui je vous entends	c:affirm c:dialog-hear(who=i, to-whom=you)
euh je vous entends pas très fort mais c'est correct	euh c:dialog-hear(who=i, to-whom=you) pas très forme ce_qu on est
oh oui c'est bon	c:exclamation (exclamation=wow) c:affirm c:acknowledge
Oui	c:affirm
d'accord	c:acknowledge

Table 2: Exemple d'hypothèses obtenues à la sortie du système de RAP avec notre nouveau ML sémantique

3.3 Analyse des résultats

3.3.1 Comparaison de taux d'erreur

Premièrement, dans le but de vérifier que ces changements dans le modèle de langage, permettant une analyse partielle en IF, ne dégradent pas la performance intrinsèque du système de reconnaissance initial, nous avons comparé le taux d'erreur du système initial avec celui de notre nouveau système. Le taux d'erreur de mots (Word Error Rate -WER) du système initial qui utilise des classes construites manuellement est de 31.9% alors que le taux d'erreur du système utilisant le nouveau modèle, après avoir reconstitué les mots français à partir de la classe IF, est 32.9%. Ainsi, nous pouvons constater que le nouveau modèle n'impose pas une dégradation très importante de la performance du système initial.

3.3.2 Statistiques sur l'analyse partielle en IF lors de la phase de reconnaissance

Les 216 tours de parole de test comprennent 915 mots. Parmi ces 915 mots, 35% ont été analysés directement en IF dès la phase de reconnaissance.

Au niveau des tours de parole, 125 tours sur un total de 216 (58%) ont été analysés directement en IF, aussi, dès la phase de reconnaissance. Bien sûr, ce sont surtout les tours de parole courts qui sont totalement analysés en IF, mais ce résultat reste encourageant car, désormais, un volume important du travail du module d'analyse peut être réalisé directement par l'usage d'un module de reconnaissance utilisant un modèle de langage sémantique.

Par ailleurs, sur les 58% de tours directement analysés, 84% sont proprement analysés sans erreur. Les 16% d'erreur restant sur cette partie du corpus de test, correspondent essentiellement aux erreurs de reconnaissance faites par le système, avec ou sans modèle sémantique, et qui ne seront jamais récupérées par le module d'analyse.

4 Conclusion

Nous avons présenté une nouvelle méthodologie pour introduire des classes sémantiques dans le modèle de langage statistique d'un système de reconnaissance, dans un contexte de traduction de parole. Ce modèle « sémantique » a été testé dans le cadre du projet de traduction de parole NESPOLE. Avec notre nouveau modèle de langage, le module de reconnaissance peut réaliser directement une partie du travail d'analyse vers la représentation sémantique (IF) : 35% des mots du dialogue test ; 58% des tours de parole du dialogue test. Parmi ces 58% tours analysés directement, 84% sont proprement analysés. Evidemment, la principale limitation de notre approche est que ce sont majoritairement les tours de parole les plus courts, et donc les plus faciles à analyser, qui sont traités dès la phase de reconnaissance. Par ailleurs, le module d'analyse devra être légèrement adapté afin de pouvoir traiter en entrée un mélange de mots français et d'IF. Néanmoins, cette modification reste relativement facile à implémenter.

5 Références

[1] Wolfgang Wahlster (Springer edition) "Verbmobil : Foundations of Speech-to-Speech Translation", 2000.

[2] Manny Rayner, David Carter, Pierrete Bouillon, Vassilis Digalakis, Mats Wirén "Spoken Language Translation" *Cambride Press*, 2000

[3] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz "Speech Translation for French in the NESPOLE! European Project", *Eurospeech 2001*, Aalborg, Danemark, September 2001

[4] Levin L. & al. "An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues". *Proc. ICSLP'98*, 30th November - 4th December 1998, Sydney, Australia, vol.4/7, pp.1155- 1158.

[5] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai anh Robert L. Mercer. "Class-based n-gram Models of Natural Language". Computational Linguistics 18(4): 467-479. 1992.

[6] Reinherd Kneser and Hermann Ney. "Improved Clustering Techniques for class-based Statistical Language Modelling". *In proceeding of the 3rd European Conference on Speech Communication and Technology*, 973-976. 1993.

[7] Blanchon, H. (2002). "A Pattern-Based Analyzer for French in the Context of Spoken Language Translation: First Prototype and Evaluation". *Proc. COLING*. Taipei, Taiwan. Vol. 1/2: pp 92-98. 24 August - 1 September, 2002.

[8] S. Burger, L. Besacier, P. Coletti, F. Metze, C. Morel "The NESPOLE! VoIP Dialogue Database", *Eurospeech 2001*, Aalborg, Danemark, September 2001

[9] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, A. Waibel “Recognition of conversational telephone speech using the Janus speech engine” IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, 1997.