Orsay, le 12 novembre 2008

Rapport sur le document de thèse présenté par Christophe Servan

Christophe Servan présente dans ce document son travail portant sur des modèles statistiques pour le décodage conceptuel de la parole dans le cadre de systèmes de dialogue dans des domaines limités (réservation d'hôtels pour une première partie du travail, puis routage d'appels dans une seconde partie).

Après une (brève) introduction qui présente le contexte du travail de thèse ainsi que l'organisation du document, Christophe Servan présente dans un premier temps (chapitre 2) le domaine du dialogue oral homme-machine en insistant plus particulièrement sur la Reconnaissance Automatique de la Parole et la Compréhension Automatique de la Parole qui constituent le coeur de son travail. Ensuite, dans les chapitres 3 et 4, Christophe Servan décrit plus précisément son travail, dans le cadre du projet MEDIA, en particulier ses propositions de modélisations sont évaluées sérieusement sur le corpus obtenu lors de ce projet. Puis, une quatrième partie (chapitre 5) décrit une expérience, fondée en partie sur les résultats précédents, portant sur le routage d'appel et la réduction du coût lié à l'apprentissage (développement de corpus) par la réutilisation de corpus existants. Là aussi, une série d'expériences a été menée et est décrite.

Dans la suite de ce rapport je présente les différentes partie de ce mémoire que j'associe à quelques commentaires. Je termine par une appréciation globale du document et du travail réalisé.

Dialogue Homme-Machine et compréhension de la parole

Dans ce chapitre, Christophe Servan commence par une présentation du domaine du dialogue homme-machine et des différents types de dialogue possible, existant aujourd'hui dans le grand public ou en laboratoire. Il est dommage que ses exemples fassent l'impasse sur les systèmes intermédiaires (et donc les dialogues intermédiaires) entre ses deux derniers exemples (je pense en particulier au

système Let's Go abondamment décrit dans la littérature et accessible au grand public). Cette présentation s'achève sur un schéma d'architecture classique de système de dialogue oral homme-machine qui met bien en évidence les points ou modules sur lesquels se concentre ce travail de thèse.

Suit ensuite une explication sur les systèmes de reconnaissance de la parole, les différents modèles utilisés (modèles acoustiques et modèles de langage) ainsi que la combinaison de ces modèles dans un système. Même si on peut parfois être surpris de certains raccourcis d'annotation qui peuvent donner lieu à des interprétations surprenantes, la présentation des systèmes de reconnaissance est claire et permet de bien comprendre comment va s'articuler le travail présenté plus loin.

La section suivante présente la compréhension automatique de la parole, en terme d'objectif et de définition (qu'est-ce que comprendre la parole, sur quelle représentation peut-on s'appuyer) puis un état de l'art des différentes approches est présenté et discuté. Lors de la présentation de ce qu'est la compréhension automatique de la parole, Christophe Servan insiste particulièrement sur les difficultés inhérentes à la parole (versus l'écrit). En particulier les phénomènes typiques de l'oral comme les disfluences sont décrits. Il est également noté que que l'oral *n'est pas aussi bien structuré que l'écrit* et que *l'élocution est particulièrement riche en erreurs de grammaire, de conjugaison, d'accord, etc.*. Ici j'aurais aimé voir une discussion plus poussée et notamment des arguments permettant d'affirmer ceci. Ne peut-on pas considérer (et de nombreux travaux vont dans ce sens) que l'oral a une syntaxe (on y retrouve bien des régularités, dont l'absence par ailleurs ne permettrait pas le développement de modèles statistiques !) qui est simplement différente de celle de l'écrit ? Est-il raisonnable de considérer que le respect de la grammaire scolaire est ce qui caractérise l'écrit ? Et que l'absence de ce respect caractérise l'oral ? Cela parait difficile à croire en particulier si on regarde les travaux réalisés sur des documents écrits tout venant (du Web par exemple) et pour lesquels les problèmes sont finalement assez similaires à ceux présentés ici comme étant spécifiques de l'oral. Par ailleurs, la plupart des difficultés présentes au niveau d'un énoncé inhérentes à l'oral et/ou à la situation de dialogue, avec la présence d'un interlocuteur humain donc, sont plutot bien présentées et décrites. Christophe Servan présente également la notion de représentation informatique de la sémantique, en particulier la représentation FrameNet qu'il illustre par différents exemples judicieusement choisis.

Christophe Servan présente ensuite différentes modélisations utilisées pour l'analyse sémantique (la compréhension automatique). Il décrit les systèmes à base de grammaires manuelles, les modèles de compréhension dépendant de la tâche, les modèles statistiques de compréhension et les modèles de classification. Les différents modèles sont bien présentés même si certains arguments ne peuvent que surprendre le lecteur. Ces présentations sont claires et dans l'ensemble bien discutées. Une remarque dans cette présentation est surprenante : *La modélisation stochastique du langage est privilégiée dans le cadre d'applications de dialogue oral, en particulier lorsque celles-ci permettent à l'utilisateur de s'exprimer librement. Cependant dans le cadre d'applications ad-hoc les modèles utilisés sont à base de règles manuelles.* Que sont ces applications ad-hoc ?

Pour développer des systèmes de compréhension de la parole, des corpus annotés selon la définition de la tâche et le mode de représentation choisis sont necessaires. Ces corpus servent en outre à l'évaluation des systèmes. Cette problématique est bien présentée et argumentée. Plus surprenante est la remarque concernant des résultats d'évaluation obtenus par différentes approches, sachant que quelques sections plus tard, une campagne d'évaluation est clairement présentée et ses résultats discutés.

Christophe Servan présente ensuite la problématique même de la compréhension de la parole et op-

pose deux approches : une approche séquentielle, approche la plus intuitive mais non dénuée de problèmes, et une approche intégrée, approche dans laquelle se situe son travail. Un bref état de l'art de ces approches et proposé et discuté.

Une partie du travail de thèse de Christophe Servan se rapporte à la campagne d'évaluation MEDIA. Il utilise le corpus collecté et annoté dans le cadre de ce projet ainsi que les outils d'évaluation. Il s'appuie également sur les premiers résultats obtenus lors de la campagne officielle. Une section est entièrement consacrée à cette campagne et à ses données. La présentation des différents aspects (corpus, annotation, ontologie) est très claire et permett de bien mettre en évidence la difficulté à définir une ontologie et faire annoter de manière aussi fiable que possible les corpus en grande quantité. Les expériences portant sur l'accord inter-annotateurs qui ont eu lieu sont clairement présentées. Cette section s'achève avec une brève présentation des différents systèmes ayant participé à cette campagne d'évaluation et les résultats qu'ils ont obtenu. Une discussion est amorcée, se fondant sur la typologie des différentes approches utilisées (logique, syntaxique, statistiques, mixtes).

Dans sa discussion sur les résultats des différents systèmes, Christophe Servan souligne que les systèmes à base de règles manuelles se heurtent à un problème de traduction de leur représentation dans le format MEDIA. Ceci laisse penser que peut-être ces systèmes ont été évalués davantage finalement sur leur (in)capacité à projeter leur sortie dans une représentation différente de la leur. Malheureusement, Christophe Servan ne discute pas ce point au-delà de sa seule évocation.

Décodage conceptuel et apprentissage automatique

Après avoir présenté le contexte et le domaine (chapitre 2), Christophe Servan s'attache à décrire son travail. Il faut remarquer ici que cette organisation permet au lecteur de clairement voir et comprendre l'apport propre du travail du candidat. Cette partie se décompose en deux chapitres. Le premier qui décrit le modèle proposé et le deuxième qui décrit les expériences menées et les résultats obtenus.

Christophe Servan propose, expose et argumente une approche intégrée de décodage conceptuel fondé sur des modèles markoviens. Dans un premier temps il revient sur une distinction en deux familles des différentes approches possibles : l'une qui considère le décodage comme la conséquence d'un processus d'analyse et l'autre qui le considère comme le résultat d'une traduction automatique d'une suite de symboles en une autre suite de symboles. Son travail se situe clairement dans cette dernière. Ensuite il présente rapidement l'approche classique et séquentielle de la compréhension de la parole, approche dans laquelle le processus de reconnaissance de la parole et le processus de compréhension sont nettement dissociés. Arguant de la rupture du lien de dépendance entre le signal de parole et l'interprétation du message, Christophe Servan qualifie cette approche de sous-optimale. Ensuite l'approche intégrée est présentée et discutée. Le choix d'utiliser une approche probabiliste pour la compréhension (approche proche de ce qui est fait en reconnaissance de la parole) est argumenté notamment par le fait que cela permet simplement une bonne intégration entre les deux processus (reconnaissance de la parole et compréhension de la parole). Deux autres arguments sont présentés, d'une part la nécessité de passer par des graphes de mots du fait de la fenêtre courte utilisée par les modèles de langage et d'autre part le fait que même si les analyses syntaxiques peuvent travailler sur des graphes, la nature même de l'oral et la présence des disfluences les réduit le plus souvent à une succession d'analyses partielles qui ne permet pas d'atteindre, ou difficilement, une analyse fonctionnelle et complète du message.

Enfin, Christophe Servan présente le modèle théorique sur lequel il s'appuie. Il s'agit d'un modèle de Markov caché (HMM) à deux niveaux où sur le premier niveau les états sont les concepts et les symboles générés les mots et sur le deuxième niveau les états sont les mots et les symboles générés les séquences d'observations acoustiques. Dans ce modèle, le choix de la meilleure interprétation se fait avec un maximum a posteriori. Trois différentes sources de probabilités sont utilisées : la probabilité P(A|W) donnée par les modèles acoustiques à la suite de mot ; la probabilité P(W) donnée par le modèle de langage et enfin la probabilité P(W|I) donnée par le modèle de décodage conceptuel et qui représente la probabilité d'une suite de concepts étant donnée la séquence de mots W. Deux facteurs, estimés sur les corpus de développement, viennent pondérer le poids des différents modèles. Pour estimer la probabilité P(W|I) Christophe Servan s'inspire des travaux de Ramshaw et Marcus (1995) concernant le chunking. Dans cette approche, à chaque mot une étiquette est associée qui précise le type de segment dans lequel se trouve le mot. Une information binaire (est-ce le premier mot du segment ou un mot à l'intérieur du segment) est également associée. Dans ce cadre, l'estimation de P(W|I) revient donc à une tâche d'étiquetage où chaque mot reçoit un label correspondant au concept qu'il représente et à sa position à l'intérieur de celui-ci. Dans la mesure où le module de compréhension de la parole doit prendre en entrée un graphe de mots, il faut pouvoir structurer l'espace de recherche et surtout réévaluer les scores des chemins du graphe. C'est pourquoi le module repose certes sur l'utilisation de l'étiqueteur stochastique mais aussi sur des grammaires de concepts. Ces grammaires de concepts représentent pour chaque concept la totalité de leurs réalisations lexicales. Elles peuvent être apprises sur les corpus ou bien écrites manuellement. Le formalisme des grammaires régulières est utilisé et la représentation choisie est celle des automates à états finis. Ce choix, restrictif, est judicieusement argumenté.

Christophe Servan fait remarquer que le transducteur résultant n'étant pas évalué, ajouter des grammaires manuelles est simple. Ainsi pour l'application MEDIA, il a ajouté les grammaires manuelles pour coder les dates, les expressions numériques comme les prix, les entités de la base de données comme les villes, les noms d'hôtels et les restaurants. Partant de là, Christophe Servan décrit précisément et clairement les opérations permettant de passer d'un graphe de mots (l'entrée du module de compréhension) à une liste des N-meilleures solutions structurées. Pour cela, il s'appuie sur un exemple clair qui est filé tout au long de l'explication. Il termine son explication par une présentation de l'architecture générale de ce système qui illustre clairement l'ensemble du travail présenté.

Le système présenté devant être évalué avec un système de reconnaissance, Christophe Servan décrit rapidement la mise en place du système de reconnaissance qu'il a ensuite utilisé pour ses expériences. Dans la mesure où un des objectif affiché est d'évaluer l'impact du taux d'erreur mot (WER) sur le système, Christophe Servan propose et implémente différentes méthodes pour faire varier ce WER dans les graphes obtenus. Il propose deux approches différentes. La première consiste en l'injection des données de test et produit 4 graphes hypothèses différents en faisant varier le coefficient d'interpolation. La deuxième consiste en une combinaison de systèmes. Il utilise pour cela les approches de combinaison de log-linéaire des probabilités postérieures proposées notamment par Barrault (2008). Il utilise trois différents systèmes puis la combinaison de ces trois systèmes. Ceci aboutit à 4 graphes hypothèses différents avec des WER différents. Au total, Christophe Servan dispose pour ses expériences de 8 graphes différents dont le WER varie de 18.5% à 33%.

Le chapitre suivant (chapitre 4) relate les différentes expériences effectuées avec ces différents graphes de mots et le système présenté précédemment.

Christophe Servan présente les résultats qu'il obtient avec le nouveau système complet présenté dans

le troisième chapitre. Les amélioration successives apportées au système durant ces travaux de thèse ont permis un gain sur le Concept Error Rate (CER) de 10% absolu en mode *full* et d'environ 3% en mode *relax* quelque soit le nombre de modes considérés (2 ou 4). Ensuite Christophe Servan présente des évaluations pour mesurer l'intérêt des listes des n-meilleures hypothèses en comparant méthode séquentielle et méthode intégrée. Cette dernière offre les meilleurs résultats. Si la raison principale évoquée parait effectivement la bonne, elle n'est ni argumentée ni étayée ce qui est dommage.

Pour étudier l'impact du WER sur le CER, Christophe Servan utilise les différents graphes de mots présentés au chapitre précédent. Les expériences menées sur les graphes obtenus via l'introduction des données de tests montrent une supériorité systématique de la méthode intégrée sur la méthode séquentielle. D'autre part, elles mettent en évidence une relation linéaire entre le WER et le CER. Il n'a pas été possible à Christophe Servan de mener l'expérience sur les graphes de mots obtenus à partir des combinaisons de systèmes avec la méthode intégrée, le système de reconnaissance automatique de la parole n'ayant pu fournir des graphes exploitables. Il n'en reste pas moins qu'on constate également une forte correlation entre le WER et le CER.

Un des aspects importants à prendre en compte lorsqu'on travaille sur le développement de modèles statistiques est celui des corpus, de la taille minimale nécessaire de ces corpus. Christophe Servan a cherché à évaluer cet aspect en faisant varier la taille du corpus d'apprentissage, de 25 dialogues à 720 dialogues. Il a ainsi constitué 7 corpus d'apprentissage qu'il a utilisé pour apprendre les automates à états finis et le modèle de langage. Il a évalué ensuite le CER en fonction du corpus utilisé pour l'apprentissage. Les résultats montrent qu'à partir de 400 dialogues le CER se stabilise autour de 25%. Christophe Servan met en parallèle ces résultats avec les expériences effectuées par le consortium MEDIA sur les accords inter-annotateurs. On peut conclure de ces observations et expériences qu'il faut environ deux fois plus de dialogues à un système pour stabiliser ses résultats qu'à des humains pour stabiliser leurs annotations. Par ailleurs un certain nombre de connaissances étant communes à différentes applications (typiquement les entités nommées, les expressions numériques, les dates...), il peut être intéressant de les intégrer tout de suite, indépendamment du corpus d'apprentissage. C'est ce qu'a évalué Christophe Servan. Son expérience montre qu'intégrer ces connaissances via des grammaires a apporté un gain au niveau du CER jusqu'à un corpus de 600 dialogues où le gain devient presque nul. Ceci revient à dire qu'il est possible de diminuer la taille minimale du corpus d'apprentissage si on intègre dès le début des connaissances communes à plusieurs applications. Les résultats de ces 2 expériences sont intéressants mais on peut regretter que le lien entre les deux expériences ne soient pas plus explicites et surtout que la différence dans les résultats ne soient pas davantage expliquée.

Puisque le travail porte sur la compréhension automatique de la parole dans un système de dialogue, Christophe Servan pose la question de savoir si utiliser les informations précédentes, issues du dialogue, ne pourrait pas aider le décodage et donc améliorer la compréhension. Pour vérifier cette hypothèse, il a étudié spécifiquement le cas des entités nommées et de leur apparition dans les énoncés des utilisateurs suivant une première mention par le système (un compère en l'occurrence). Il a étudié plus précisément le cas des noms d'hôtels, entités très présentes et indispensables pour le bon déroulement de la tâche de réservation d'hôtel. Son étude du corpus MEDIA montre qu'une très grande proportion des noms d'hôtels prononcés par un utilisateur suivent en effet une première mention de ce même hôtel par le compère. Il a confirmé que ces proportions se retrouvaient également dans le corpus de test. Il s'agit donc là d'un comportement qu'on peut qualifier de standard et utiliser ce fait pour améliorer la compréhension semble donc prometteur. Afin de valider cette hypothèse, Christophe Servan a vérifié ce qui se produisait et quels étaient les types d'erreur réalisés par son système (RAP + déco-

dage conceptuel) sur les noms d'hôtels. Cette étude montre qu'il y a un gain net à espérer si une telle connaissance est prise en compte par le système. Toutefois, et ceci est bien souligné par Christophe Servan, les exemples dans le corpus sont trop peu nombreux pour pouvoir en tirer des conclusions fiables. Il est dommage que cette étude n'ait pu être poussée plus avant. Néanmoins, à partir de cette étude, Christophe Servan propose des pistes de recherche qui semble intéressantes (comme l'utilisation d'un modèle cache ou de trigger) proches de ce qui est fait en reconnaissance automatique de la parole ou encore en traduction automatique. Cette incursion dans des domaines connexes est un aspect intéressant de la réflexion proposée.

Routage d'appels et décodage conceptuel

Dans ce dernier chapitre, Christophe Servan propose d'appliquer son modèle de décodage conceptuel à une autre tâche, en rapport avec les systèmes de dialogue : le routage d'appels. Ce chapitre lui permet également de mettre en oeuvre des propositions intéressantes pour régler le problème du manque de données d'apprentissage. Sa proposition se base sur la réutilisation de données pré-existantes associées à des connaissances a priori sur la tâche.

Le cadre applicatif dans lequel s'effectue ces travaux est un système de routage d'appels appliqué aux renseignements téléphoniques du Conseil Général du Vaucluse. Concernant le domaine lui-même les seules données dont dispose Christophe Servan sont celles présentes sur le site Web correspondant, en particulier sa FAQ. S'appuyant sur la structure de la FAQ, Christophe Servan a pu déterminer des classes et les concepts associés pour le système de classification. De la meme façon il a pu extraire un vocabulaire spécifique au domaine, des formulations de questions et leurs réponses associées. Il a également utilisé deux autres corpus pour, d'une part des exemples de parole spontanée (corpus EPAC) et d'autre part des exemples de formulations de question (corpus RITEL). Par ailleurs, afin de disposer d'un corpus de test, il a enregistré un corpus contenant 216 requêtes correspondant aux 12 thèmes de la tâche. Il a procédé à une segmentation en trois parties de ce corpus selon le niveau de difficulté des requêtes (77 requêtes faciles, 94 requêtes associées à un niveau de difficulté moyen et 45 requêtes difficiles). Il est dommage que des informations simples (comme le nombre de locuteurs) ne soient pas indiquées.

A l'aide de différentes combinaisons de ses corpus d'apprentissage, 5 différents modèles de langage ont été développés. Des mesures de perplexité sur le corpus de test sont effectuées pour chacun de ces modèles de langage. Le modèle de langage intégrant les 3 corpus (FAQ, EPAC et RITEL) donne la perplexité la plus basse (96). Christophe Servan donne ensuite des taux de WER sur les différentes parties du corpus de test mais n'indique pas le modèle de langage utilisé (on peut supposer qu'il s'agit de celui ayant obtenu la plus faible perplexité). Ce WER est de 54.2% sur l'ensemble du corpus de test et varie de 28.1% (requêtes faciles) à 73.5% (requêtes difficiles).

Le routage d'appels est considéré comme une tâche de classification. La classification peut être faite en s'appuyant sur les mots ou sur des concepts. Dans ce dernier cas, un module de compréhension doit être développé. Pour ce faire, Christophe Servan s'est appuyé sur l'ontologie (la partie "générique") de MEDIA et sur le corpus de la FAQ d'où il a extrait 20 concepts spécifiques au domaine. Comme corpus d'apprentissage Christophe Servan utilise le corpus de la FAQ constitué des paires questions/labels et du même corpus annoté sémantiquement. Le paradigme du leave-one-out a été utilisé pour mener à bien les expériences. Le taux de bonne classification en n'utilisant que les mots est de 80% (pour 12 classes) et de 86% en utilisant l'annotation sémantique. Ces résultats ont été

obtenus sur les transcriptions manuelles du corpus de test. Ceci lui a permis de constituer son modèle.

Pour évaluer ce classifieur dans le système complet (incluant la RAP donc), 3 approches sont comparées : la première où la classification est effectuée à partir des mots de la meilleure hypothèse du système de RAP ; la deuxième où la classification est effectuée sur sur la meilleure séquence de mots et de concepts produite par le décodage conceptuel sur le graphe de mots de la RAP ; et enfin la troisième qui s'appuie sur une nouvelle approche, décrite dans ce chapitre. Cette dernière applique un ensemble de filtres supplémentaires sur le transducteur global obtenu à partir du graphe de mots. Chaque filtre correspond à un concept. Une évaluation sur le corpus de test (et chacune de ses parties) est présentée, sur les transcriptions manuelles et sur le système utilisant la RAP. Chacune des approches est ainsi évaluée. Ces résultats montrent clairement d'une part qu'ajouter l'information sémantique permet d'améliorer les résultats et d'autre part que en ce qui concerne la classification sur la RAP, la dernière approche (qui s'applique donc sur le treillis de mots issu du module de RAP) produit des résultats meilleurs. Cette approche semble plus robuste aux erreurs de reconnaissance.

Remarques générales sur le document et le travail

Christophe Servan s'est attaqué au problème de la compréhension automatique de la parole. Il a pleinement intégré la reconnaissance de la parole et a proposé un modèle original de décodage conceptuel, basé sur une approche stochastique qui s'appuie sur des automates à états finis. Ce modèle et son implémentation permettent une grande souplesse d'utilisation, en particulier permettent d'utiliser une méthode intégrée de la reconnaissance de la parole et de la compréhension automatique de la parole. Un travail expérimental sérieux a été effectué qui a montré l'intérêt d'un tel modèle. On peut également souligner la clarté de l'explication concernant le modèle et sa mise en oeuvre, exposé qui s'appuie sur un exemple filé qui permet au lecteur de bien suivre tout le processus de mise en oeuvre du modèle. Une dernière série d'expériences, fondée en partie sur les résultats précédents, portant sur le routage d'appel est exposée. Là encore, on peut constater que le travail expérimental est sérieux. Dans ces deux séries d'expériences, Christophe Servan, qui défend des approches stochastiques coûteuses en corpus, a toujours cherché à proposer des méthodes pour diminuer le coût lié à l'acquisition de ces corpus et a évalué ses propositions de réduction de coûts.

Une critique qui peut être faite concerne l'absence de réelle synthèse et discussion à la fin des différentes parties. La conclusion ne remplit malheureusement pas ce rôle. Ceci dénote peut-être une absence de recul du candidat par rapport à ses travaux de recherche. Il n'en reste pas moins que le travail présenté est sérieux et s'appuie sur des expériences intéressantes et bien menées. Les travaux de Christophe Servan ont fait l'objet de différentes publications nationales (4) et internationales (5) dans des conférences de très haut niveau. Pour ces raisons, je considère que les travaux décrits dans ce mémoire méritent d'être présentés en soutenance de thèse.

Sophie Rosset
chargée de recherche CR1, HDR
groupe Traitement du Langage Parlé
LIMSI-CNRS, Orsay