Compréhension automatique de la parole sans données de référence

Emmanuel Ferreira Bassam Jabaian Fabrice Lefèvre
Université d'Avignon, CERI-LIA, France
{*prénom.nom*}@univ-avignon.fr

Résumé. La majorité des méthodes état de l'art en compréhension automatique de la parole ont en commun de devoir être apprises sur une grande quantité de données annotées. Cette dépendance aux données constitue un réel obstacle lors du développement d'un système pour une nouvelle tâche/langue. Aussi, dans cette étude, nous présentons une méthode visant à limiter ce besoin par un mécanisme d'apprentissage sans données de référence (zero-shot learning). Cette méthode combine une description ontologique minimale de la tâche visée avec l'utilisation d'un espace sémantique continu appris par des approches à base de réseaux de neurones à partir de données génériques non-annotées. Nous montrons que le modèle simple et peu coûteux obtenu peut atteindre, dès le démarrage, des performances comparables à celles des systèmes état de l'art reposant sur des règles expertes ou sur des approches probabilistes sur des tâches de compréhension de la parole de référence (tests des Dialog State Tracking Challenges, DSTC2 et DSTC3). Nous proposons ensuite une stratégie d'adaptation en ligne permettant d'améliorer encore les performances de notre approche à l'aide d'une supervision faible et ajustable par l'utilisateur.

Abstract.
Spoken language understanding without reference data
Most recent state-of-the-art spoken language understanding models have in common to be trained on a potentially large amount of data. However, the required annotated corpora are not available for a variety of tasks and languages of interest. In this work, we present a novel zero-shot learning method for spoken language understanding which alleviate the need of any annotated or in-context data. Instead, it combines an ontological description of the target domain and the use of a continuous semantic space trained on large amounts of unannotated and unstructured found data with neural network algorithms. We show that this very low cost model can reach instantly performance comparable to those obtained by either state-of-the-art carefully hand crafted rule-based or trained statistical models on reference spoken language understanding tasks (test sets of the second and the third Dialog State Tracking Challenge, DSTC2,DSTC3). Eventually we extend the approach with an online adaptative strategy allowing to refine progressively the initial model with only a light and adjustable supervision.

Mots-clés : Compréhension automatique de la parole, espace sémantique continu, apprentissage sans données de référence, données d'apprentissage hors domaine.

Keywords: Spoken language understanding, continuous semantic space, zero-shot learning, out-of-domain training data.

1 Introduction

Dans un système de dialogue homme-machine, le module de compréhension automatique de la parole (Spoken Language Understanding, SLU) joue un rôle intermédiaire entre le module de reconnaissance de la parole et le gestionnaire de dialogue. Son rôle est d'extraire une liste d'hypothèses de séquence d'étiquettes sémantiques à partir d'une transcription automatique de la requête de l'utilisateur. Actuellement, les systèmes état de l'art pour la compréhension sont basés sur des approches probabilistes et sont appris grâce à différentes méthodes d'apprentissage automatique afin de pouvoir attribuer des étiquettes sémantiques aux entrées des utilisateurs.

Les techniques d'apprentissage supervisé nécessitent un grand nombre de phrases annotées sémantiquement par des utilisateurs experts. Ces corpus annotés sont coûteux (en expertises humaines et en temps de construction) et sont dépendants du domaine d'application (souvent restreint) et de la langue utilisée.

Plusieurs études ont comparé les différentes approches probabilistes pour la compréhension de la parole, e.g. (Hahn *et al.*, 2010; Lefèvre, 2007; Deoras & Sarikaya, 2013). Les approches état de l'art utilisent le plus souvent des modèles statistiques discriminants, tels que les champs aléatoires conditionnels de Markov (conditional random fields, CRF) (Wang & Acero, 2006) ou les réseaux de neurones profonds (Deep Neural Networks, DNN) (Deoras & Sarikaya, 2013). Malgré leurs bonnes performances, ces approches sont très dépendantes des données et sont donc difficilement généralisables.

Pour faire face à cette limitation, plusieurs recherches ont proposé un processus d'annotation non-supervisé, e.g. en se basant sur des allocations latentes de Dirichlet (Camelin *et al.*, 2011). D'autres travaux ont porté sur l'utilisation d'algorithmes d'apprentissage non-supervisé (Tur *et al.*, 2011; Lorenzo *et al.*, 2013) ou semi-supervisé (Celikyilmaz *et al.*, 2011; Hakkani-Tur *et al.*, 2011) pour palier à l'absence de ressources annotées en exploitant notamment le web sémantique pour permettre une recherche de données d'apprentissage supplémentaires afin d'améliorer les performances des classifieurs employés.

Un autre groupe d'études s'est intéressé à proposer des techniques visant à réduire le temps de collecte, de transcription et d'annotation de nouveaux corpus. Par exemple, dans (Gao *et al.*, 2005) ou encore dans (Sarikaya, 2008), il a été proposé de construire en premier lieu un petit corpus pour apprendre un système pilote et d'utiliser ce système pour poursuivre la collecte de nouvelles données. D'autres travaux ont employé des techniques issues de l'apprentissage actif (active learning) pour réduire le temps nécessaire à l'annotation et à la vérification d'un corpus, e.g. (Tur *et al.*, 2003) et (Tur *et al.*, 2005). Plus récemment, plusieurs recherches ont été conduites pour diminuer le coût et l'effort de collecte de données par l'étude de portabilité de systèmes à travers les langues (Lefèvre *et al.*, 2010; Jabaian *et al.*, 2013), et les domaines (Lefèvre *et al.*, 2012).

En outre, (Dauphin *et al.*, 2014) proposent l'adoption d'un algorithme d'apprentissage dit sans données de référence (zero-shot learning) pour une classification sémantique d'énoncés. Cette méthode tente de trouver un lien entre les catégories et les énoncés dans un espace sémantique. Ce dernier est appris par un réseau de neurones profond sur une grande quantité de données non-annotées et non-structurées.

Dans le même esprit, dans cet article, nous présentons une méthode visant à limiter la dépendance aux données annotées par l'utilisation d'un mécanisme similaire. En effet, notre méthode repose sur une description ontologique minimale de la tâche visée et sur l'utilisation d'un espace sémantique continu appris par des approches à base de réseaux de neurones sur des données génériques non-annotées (facilement disponible sur web). Notre étude expérimentale a été menée sur une tâche de compréhension de la parole en utilisant les données de la seconde et de la troisième campagne d'évaluation Dialog State Tracking Challenge [1] (DSTC2 and DSTC3) (Henderson *et al.*, 2014a,b). Nous montrons que la technique proposée offre des performances comparables à celles obtenues par des systèmes à base de règles expertes d'une part et appris sur des données annotées d'autre part.

Cependant, une telle approche est dépendante de la qualité de la description ontologique fournie mais aussi de l'espace sémantique continu considéré (sa capacité à modéliser la richesse sémantique du domaine cible). Pour faire face à ces limites, nous proposons l'ajout d'une stratégie d'adaptation « en ligne ». Cette approche a pour objectif d'introduire une faible supervision dans l'optique de raffiner de façon incrémentale la définition de notre connaissance ontologique et de mieux exploiter l'espace sémantique considéré.

Cet article est organisé comme suit : dans la section 2 nous décrivons la tâche de compréhension de la parole. La section 3 présente les approches proposées pour l'apprentissage sans données de référence puis pour l'adaptation en ligne du système, suivie d'une présentation de quelques travaux connexes. Nous présentons notre étude expérimentale dans la section 5 et nous concluons enfin par quelques remarques et perspectives.

2 Compréhension automatique de la parole

Le rôle du module de compréhension de la parole est d'extraire une séquence de m étiquettes sémantiques, également appelées concepts, $C = c_1, c_2, ..., c_m$ d'une phrase d'utilisateur de n mots, $W = w_1, w_2, ..., w_n$. Si classiquement chaque étiquette sémantique c_i est définie par un couple champ/valeur, comme par exemple *food=Italian* ou encore *destination=Boston*, dans cet article, nous adopterons le standard d'annotation sémantique employé dans les corpus en langue anglaise des campagnes d'évaluation DSTC2 et DSTC3 (Henderson *et al.*, 2014a).

Dans ces corpus, les étiquettes sémantiques correspondent à des actes de dialogue de la forme `acttype(champ=valeur)`

1. http ://camdial.org/ mh521/dstc/

où acttype représente le nature de l'acte de dialogue considéré, à savoir son intention dialogique (e.g. la confirmation ou la réfutation). Par exemple, la phrase utilisateur « hello i am looking for a french restaurant in the south part of town » sera associée à la séquence d'actes de dialogue suivante « `hello(), inform(food=french), inform(area=south)` ».

Les combinaisons possibles de `acttype(champ=valeur)` sont déterminées sur la base d'un inventaire ontologique des différents types d'actes de dialogue, des champs ainsi que de leurs valeurs respectives.

Les différents types d'actes de dialogue sont en grande partie indépendants de la tâche visée. Ils peuvent se diviser en quatre grands groupes : ceux ayant pour but de transmettre de l'information (`inform`), ceux représentant différents types de requêtes (`request, reqalts, reqmore`), ceux relatifs aux confirmations (`confirm, affirm, negate, deny`) et les formules de politesse (`hello, thankyou, bye`).

L'ensemble des couples champs/valeurs est quant à lui très lié à la tâche de dialogue, chaque couple correspond généralement à une entrée spécifique dans la base de données utilisée pour répondre aux requêtes des utilisateurs (e.g. contraintes de recherche).

3 Apprentissage sans données de référence pour la compréhension automatique de la parole

L'apprentissage sans données de référence (zero-shot learning), proposé pour la première fois dans (Palatucci *et al.*, 2009), correspond à un cas particulier d'apprentissage où certaines valeurs de l'ensemble des sorties possibles, Y, ne sont pas présentes dans l'ensemble d'exemples du corpus d'apprentissage.

Dans cette étude, nous examinons le problème de prédire la séquence d'actes de dialogue d'une phrase utilisateur sans avoir vu au préalable un exemple de phrase utilisateur dans le contexte de l'interaction et donc sans avoir vu un exemple d'actes de dialogue dans ce dit contexte.

Pour ce faire, une source de connaissance sémantique doit être exploitée pour extrapoler ces sorties à partir de leur définition. Notre méthode se base donc sur trois composants principaux :

— un espace sémantique continu noté F qui peut être défini comme un espace de dimension d à même de coder les différentes propriétés des étiquettes sémantiques ;
— une base de connaissances K qui peut être vue comme un dictionnaire d'exemples dans F utilisé pour relier l'espace sémantique à l'espace de sortie du système ;
— l'analyseur sémantique qui extrait une liste ordonnée des meilleures hypothèses de séquence d'étiquettes sémantiques à partir d'un transducteur à états finis représentant l'ensemble des hypothèses pour une phrase utilisateur (scorées par des informations issues de F et de K).

Dans la suite de cette section nous décrivons plus en détails ces différents composants. Cependant, les choix faits quant à leurs implémentations concrètes pour la tâche visée seront donnés dans la partie expérimentale.

3.1 Espace sémantique continu

De récentes avancées sur les réseaux de neurones, ont permis d'envisager l'apprentissage de diverses représentations vectorielles compactes de mots (word embedding) présentant des régularités notables avec les propriétés syntaxiques et sémantiques des mots qu'elles modélisent (Mikolov *et al.*, 2013a; Bian *et al.*, 2014). Des travaux ont déjà pu montrer l'intérêt de considérer ce type de représentation sur différentes tâches de traitement automatique des langues naturelles (Bengio & Heigold, 2014; Clinchant & Perronnin, 2013).

L'objectif du module de compréhension étant d'extraire des informations sémantiques à partir d'entrées utilisateur en langage naturel, l'utilisation d'une telle représentation pour définir l'espace sémantique continu offre des possibilités de généralisation d'un grand intérêt.

De plus, ce type de représentation présente l'avantage de ne pas reposer sur l'exploitation de données liées à la tâche, mais au contraire sur un apprentissage réalisé sur une très grande quantité de données (de large couverture) souvent plus facilement accessible (e.g. dump wikipedia). Cependant différentes techniques, comme celle présentée dans (Zou *et al.*, 2013) par exemple, permettent d'adapter/de transférer le modèle ainsi appris pour une tâche spécifique ou encore pour

une autre langue.

3.2 Base de connaissance sémantique

La base de connaissance sémantique K est définie comme la matrice d'affectation représentant les informations ontologiques du domaine visé, qui dans cette étude se limitent à la liste des étiquettes sémantiques et aux exemples de formes de surface qui leurs sont associées. Dans cette matrice (illustrée Figure 1), chaque ligne correspond à un vecteur d'exemple de dimension d dans F et chaque colonne à une étiquette sémantique. Ainsi la valeur de chaque cellule de la matrice (notée $c_{i,j}$ et appelée valeur d'affectation par la suite) indique s'il existe une éventuelle affectation entre le $i^{\text{ème}}$ vecteur dans l'espace sémantique F et la $j^{\text{ème}}$ étiquette sémantique.

Les exemples (entrées de la matrice) sont obtenus en projetant dans F un certain nombre de formes de surface associées à la description ontologique du domaine. Ces formes de surface peuvent être composées d'un ou plusieurs mots. Par exemple, « what food is served ? » pour `request(food)`, « yes » pour `affirm()` ou encore « french food » pour `inform(food=french)`.

Elles peuvent être facilement obtenues automatiquement en se basant sur l'ontologie du domaine (e.g. guide d'annotation), la base de données associée à la tâche (e.g. extraction des valeurs possibles pour chaque champ) ainsi que sur un certain nombre d'exemples illustrant les différents types d'actes de dialogue (e.g. données par un expert). Il est à noter que la méthode employée ne nécessite en aucun cas d'être exhaustive lors de la définition de ces exemples (contrairement à une approche à base de règles expertes - grammaire), en effet le recours à l'espace sémantique F permettra leur généralisation après coup.

Sur la Figure 1, les lignes et colonnes de K sont respectivement étiquetées par les formes de surface et les étiquette sémantiques pour en faciliter la lecture. Les valeurs d'affection sont d'abord initialisées par des valeurs binaires, 1 si affectation et 0 sinon. Il est important de noter que nous ne contraignons pas la représentation actuelle de K par une correspondance unique entre une forme de surface et une étiquette sémantique. Ainsi, plusieurs valeurs d'affectation peuvent être mise à 1 sur une même ligne. Par exemple la forme de surface « Paris » pourrait très bien être à la fois affectée à l'étiquette sémantique `inform(location=Paris)` et à `inform(name=Paris)` si un établissement présent dans la base de données avait pour nom Paris.

3.3 Analyseur sémantique

En phase de décodage, pour chaque nouvelle phrase utilisateur que l'on cherche à étiqueter, toutes les séquences de mots contigües (formes de surface) sont considérées. Par exemple pour la phrase « yeah downtown », trois formes de surface différentes sont extraites : « yeah », « downtown » et « yeah downtown ». Ces formes de surface sont ensuite projetées dans l'espace sémantique F (cercles bleus dans la Fig. 1) pour être comparées aux vecteurs associés aux exemples de la base de connaissance K (croix noires dans la Fig. 1). Pour ce faire un critère de similarité (e.g. similarité cosinus) entre ces vecteurs est employé.

L'algorithme des k plus proches voisins est ensuite utilisé pour associer à chaque forme de surface extraite de la phrase utilisateur une liste ordonnée d'hypothèses sémantiques. Ces dernières sont ensuite utilisées pour construire un transducteur à états finis dans lequel les formes de surface et leurs hypothèses sémantiques sont respectivement les entrées et les sorties des arcs, eux-même pondérées par les distances (déduites des similarités).

Un processus de repondération (e.g. penalité appliquée pour chaque mot présent sur un arc) permets de régler l'influence de la longueur des formes de surface considérées. L'algorithme du plus court chemin est appliqué sur l'automate à états finis obtenu pour générer des hypothèses ordonnées de séquences d'étiquettes sémantiques (le plus court chemin étant mis en gras sur la Figure 1).

3.4 Adaptation en ligne

La méthode proposée pour une adaptation en ligne permet de mettre à jour les valeurs d'affectation de K en fonction des retours utilisateurs suite à l'interrogation du module de compréhension (en ligne).

En effet, une association directe entre mot (ou séquence de mots) et étiquette sémantique peut être retrouvée dans le

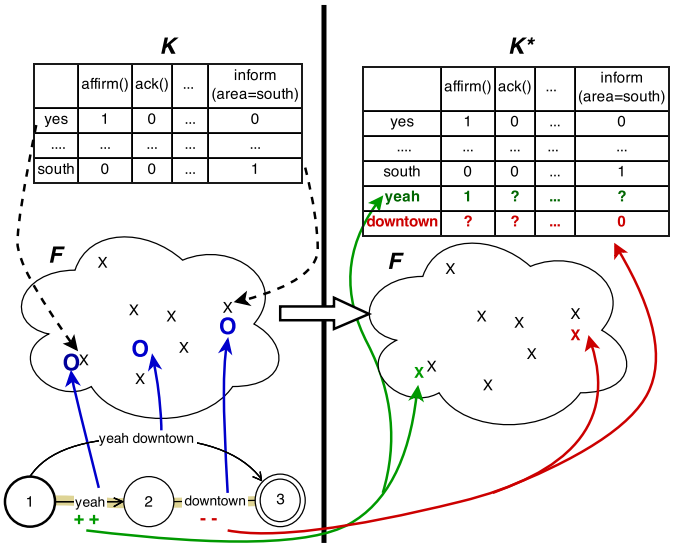

FIGURE 1 – Illustration d'un décodage sémantique basé sur une technique d'apprentissage sans données de référence

transducteur produit par l'analyseur sémantique. Ceci offre la possibilité d'exploiter cette information pour adapter le modèle dynamiquement.

Dans le but de minimiser l'effort de supervision, un scénario dans lequel la supervision est limitée à un ensemble de retours binaires (validation/rejet) sur les étiquettes sémantiques produites est proposé. Compte tenu du fait que ce scénario ne nécessite pas une correction manuelle des étiquettes de la part des utilisateurs, il peut être facilement intégré au sein d'une plate-forme de dialogue existante en utilisant des retours simples de l'utilisateur (oui/non).

Un ratio coût/amélioration peut ainsi être contrôlé en déterminant une politique de demande de retours aux utilisateurs. La définition d'une stratégie optimale pour gérer ce ratio fera l'objet de travaux ultérieurs. Dans l'étude actuelle, nous considérons un utilisateur (simulé) donnant à chaque tour un retour sur chaque étiquette sémantique produite par le système. Ces retours sont ensuite utilisés pour mettre à jour K en K^*.

Un exemple de ce processus est donné dans la Figure 1. Ce dernier illustre un cas où les véritables étiquettes sémantiques de la phrase utilisateur sont mal reconnues par l'analyseur sémantique : ici la phrase « yeah downtown » est étiquetée comme `affirm(), inform(area=south)` au lieu de `affirm(), inform(area=centre)`.

Les m retours utilisateurs constituent un jeu de m tuples $U = ((c_k, T_k, f_k))_{1 \leq k \leq m}$, où (c_k, T_k) est le couple forme-de-surface/étiquette-sémantique proposé à l'utilisateur et f_k est son retour (1 si positif, 0 si négatif). L'algorithme 1 (partiellement illustré sur la Figure 1) est utilisé pour mettre à jour K en K^* en fonction de K et U.

Chaque cellule (i, j) dans K est constituée de 4 valeurs distinctes : $p_{i,j}$ et $n_{i,j}$ représentant respectivement le nombre de retours positifs et négatifs observés jusqu'alors, $knn_{i,j}$ la valeur obtenue par une addition par élément des k plus proches lignes voisines (repondérées par un produit scalaire des similarités normalisées, voir Algorithme 1.16) et $c_{i,j}$ la valeur d'affectation présentée ci-dessus qui est aussi la valeur utilisée par notre analyseur sémantique. Lors d'une mise à jour l'ensemble des ces valeurs peut être impacté. Ainsi, l'algorithme 1 montre les conditions et la nature de leurs mises à jour mais aussi comment K peut être étendue (ajout d'une nouvelle ligne en présence d'une séquence de mots inconnus, par exemple).

Dans un premier temps, K^* est initialisé avec une copie de K. Puis toutes les nouvelles formes de surface c de U qui ne figurent pas parmi nos exemples connus sont ajoutées dans K^* (cf. Algorithme 1.4-6). Ensuite tous les comptes sur les retours sont mis à jour en se basant sur les informations contenues dans U (cf. Algorithme 1.8-9). Pour ce faire deux facteurs d'échelles α_p et α_n on été définis afin de permettre d'ajuster l'importance d'une nouvelle observation par rapport aux connaissances courantes au regard de sa valence (on pourrait par exemple choisir de faire plus confiance au retours positifs). Pour les couples forme-de-surface/étiquette-sémantique initiaux (issus de notre définition ontologique initiale

Algorithm 1 Mise à jour de la base de connaissance K

Sachant : K et U Sortie : K^*
$K^* \leftarrow K$
for all $(c, T, f) \in U$ **do**
    if $c \notin K^*$ **then**
        ajouter une nouvelle ligne pour c dans K^* avec valeurs de cellule initialisées par défaut
        $m_{last} = 1$
    $i \leftarrow$ identifiant ligne c, $j \leftarrow$ identifiant colonne T
    $p_{i,j} \leftarrow p_{i,j} + f \times \alpha_p$
    $n_{i,j} \leftarrow n_{i,j} + (1 - f) \times \alpha_n$
    if $p_{i,j} + n_{i,j} > 0$ **then**
        $old_c \leftarrow c_{i,j}$
        $c_{i,j} \leftarrow \frac{p_{i,j}}{p_{i,j}+n_{i,j}}$
        if $c_{i,j} - old_c < 0$ **then** $m_i \leftarrow 1$
    else $c_{i,j} \leftarrow 0$
for all $c_{i,j} \in K^*$ **do**
    calculer $knn_{i,j}$
for all $c_{i,j} \in K^*$ **do**
    if $p_{i,j} + n_{i,j} = 0$ et $m_i = 1$ **then** $c_{i,j} \leftarrow knn_{i,j}$

du domaine), les valeurs $p_{i,j}$ sont initialisées avec une valeur a priori p_0.

Dans le cas général, la valeur d'affectation à une étiquette sémantique est obtenue par un simple ratio entre les retours positifs et négatifs associés à la cellule concernée (voir Algorithme 1.12).

Pour chaque modification de ligne, un marqueur m_i est employé afin de détecter si une connaissance à priori (affectation positive) est remise en question par de nouvelles observations (détecté par une baisse de la valeur d'affection $c_{i,j}$, cf. Algorithme 1.13). Dans ce cas, les affections pour lesquelles il n'y a eu aucune observation pour cette forme de surface (autres cellules sur la même ligne) ont leurs valeurs d'affectation correspondant à la valeur knn à la place de 0. Ainsi, de nouvelles propositions pourront être testées et évaluées par l'utilisateur si la forme de surface venait à se représenter (processus d'exploration de l'espace d'affectation).

4 Travaux connexes

Le problème de l'apprentissage sans données de référence a été déjà abordé par la communauté de l'apprentissage automatique. On peut notamment citer les premiers travaux de Larochelle et al. (Larochelle *et al.*, 2008) qui ont introduit ce type spécifique d'apprentissage pour résoudre une tâche de reconnaissance optique de caractères.

En parallèle les auteurs de (Palatucci *et al.*, 2009) ont proposé une approche similaire pour apprendre un classifieur qui prédit des classes omises dans l'ensemble d'apprentissage. Cet algorithme utilise également une base de connaissances de propriétés sémantiques des classes connues afin d'explorer de nouvelles classes (généralisation).

En tant qu'application de cette technique dans le domaine du traitement naturel de la langue, notre proposition s'inscrit dans une même ligne que la proposition faite dans les travaux de Dauphin et al. (Dauphin *et al.*, 2014). Cependant, la nature et la manière dont nous définissons notre représentation sémantique se distinguent de ces travaux. En effet dans notre cas nous n'utilisons pas de données reliées à notre domaine mais au contraire nous utilisons une représentation généraliste. De plus, la tâche considérée n'est pas identique puisque notre objectif est d'avoir une annotation sémantique complète d'une phrase utilisateur (séquence d'étiquettes) et non pas une simple classification globale de la phrase en catégorie.

Ayant toujours le même objectif de minimiser le besoin de données d'apprentissage coûteuses en temps et en expertises humaines, différentes approches ont déjà été appliquées pour exploiter le web sémantique pour des tâches de classification d'énoncés.

Par exemple, les auteurs de (Heck & Hakkani-Tur, 2012) ont proposé une approche d'apprentissage non-supervisé pour la

compréhension de la parole basées sur l'utilisation des connaissances sémantiques du Web sémantique. Ces propositions reposent sur une combinaison d'un système de recherche d'information du web et d'un analyseur de dépendance basé sur des informations syntaxiques.

Anastasakos et Deoras (Anastasakos & Deoras, 2014) ont également proposé d'exploiter un espace continu pour modéliser les mots. Ils ont proposé une approche pour obtenir des représentations vectorielles spécifiques à des tâches et des domaines précis afin d'apprendre un système de compréhension en utilisant un algorithme d'apprentissage non-supervisé. Ils ont également proposé de transférer ces représentations d'une langue à une autre permettant l'apprentissage d'un système de compréhension multilingue.

Notre technique proposée pour l'adaptation en ligne rejoint également celles de récents travaux ayant pour but d'adapter des modèles pour la tâche de compréhension. Par exemple, dans (Bayer & Riccardi, 2013) une approche basée sur les exemples est proposée pour l'adaptation en ligne du modèle sémantique. Une autre solution présentée dans (Gotab *et al.*, 2010) utilise une méthode supervisée qui permet de mettre à jours les modèles avec une supervision limitée effectuée par les utilisateurs du système.

5 Expérimentations et résultats

5.1 Description de données

Toutes les expériences présentées dans cet article sont basées sur les corpus DSTC2 et DSTC3 (Henderson *et al.*, 2014a,b). Ces corpus ont été construits pour un défi de recherche dédié à la détection du but de l'utilisateur tout au long d'un dialogue oral (et non pas uniquement l'étiquetage sémantique des énoncés de l'utilisateur au fur et à mesure). Cependant, dans notre étude expérimentale, nous exploitons ces données (transcriptions, annotation sémantique, etc.) comme un ensemble de test pour évaluer notre approche d'apprentissage sans données de référence pour l'étiquetage sémantique sur deux configurations de dialogues réalistes.

Le défi DSTC2 couvre le domaine de la recherche d'informations sur des restaurants alors que DSTC3 étend le domaine et couvre également la recherche d'informations touristiques plus générale en incluant notamment des nouveaux types d'établissement (pubs, coffee shops) mais aussi de nouveaux champs et valeurs. Dans notre expérience, seules les données de test de ces deux corpus sont utilisées (9890 énoncés d'utilisateurs pour DSTC2 et 18715 pour DSTC3). Chaque ensemble est évalué en deux modes différents : transcriptions manuelles et n-meilleures transcriptions automatiques des entrées de l'utilisateur.

5.2 Évaluation de l'approche proposée

Afin de constituer notre espace sémantique, un modèle word2vec (Mikolov *et al.*, 2013a) a été utilisé pour apprendre une représentation vectorielle des mots sur 300 dimensions. Ce modèle a été appris avec l'algorithme *Skip-gram* (avec une fenêtre de 10 mots) sur une grande quantité de données [2] en langue anglaise disponibles librement et présentant une grande couverture thématique.

Ce type de représentation présente certaines régularités avec les propriétés syntaxiques et sémantiques des mots comme celles montrées dans Mikolov (Mikolov *et al.*, 2013b) ainsi qu'une structure linéaire permettant la combinaison des représentations des mots par une simple addition vectorielle élément par élément. Cette technique est donc utilisée pour projeter nos formes de surface vers leur représentation sémantique vectorielle de type word2vec vue comme une somme des représentations individuelles de chaque mot les constituant.

Plusieurs travaux état-de-l'art ont montré que la similarité cosinus est une métrique pertinente pour comparer différents vecteurs de mots word2vec (Mikolov *et al.*, 2013a,b). De ce fait, nous avons également utilisé cette métrique dans l'algorithme de type k plus proche voisins pour la prédiction sur les formes de surface et l'adaptation de la base de connaissance. Ainsi, dans les expériences considérées, $k = 1$ pour l'analyse sémantique et 20 pour les valeurs knn dans la matrice d'affectation. De plus, nous avons utilisé l'algorithme du plus court chemin pour parcourir le graphe sémantique (transducteur) avec une métrique de distance cosinus (voir la section 3.3).

2. enwik9, One Billion Word Language Modelling Benchmark, Brown corpus, English GigaWord de 1 à 5 - soit plus de 4 milliards de mots en contexte

Tâche	Modèle	Entrée	F-score	P	R
DSTC2	**S-règles**	n-meilleures	0,782	0,900	0,691
	S-appris	n-meilleures	0,802	0,846	0,762
	ZSSP	manuelle	0,919	0,898	0,942
		n-meilleures	0,794	0,796	0,792
DSTC3	**S-règles**	n-meilleures	0,824	0,852	0,797
	ZSSP	manuelle	0,899	0,873	0,928
		n-meilleures	0,826	0,806	0,849

TABLE 1 – Evaluation des performances de l'analyseur sémantique basé sur l'apprentissage sans données de référence en termes de F-score, **P**récision et **R**appel.

Les bases de connaissances liées à la tâche utilisées dans les expériences sont extraites de la description ontologique du domaine fournie dans le challenge (e.g. listes des champs/valeurs) ainsi que d'un ensemble d'informations de dialogue générique en suivant la procédure automatique décrite dans la section 3.2.

La sémantique du domaine est représentée par 8 champs et 215 valeurs dans DSTC2 et par 13 champs et 279 valeurs dans DSTC3. Pour les deux tâches 16 acttype différents sont considérés, en résultent 663 différentes étiquettes sémantiques pour DSTC2 et 855 pour DSTC3.

Nous avons définis manuellement 53 formes de surface pour modéliser les types d'actes de dialogue, par exemple « say again » pour l'acte de demande de répétition. Cet effort est commun aux deux tâches cibles. Dans les deux descriptions ontologiques considérées, les champs et les valeurs ont des noms significatifs (lexicalisés) et ils peuvent être directement utilisés dans les formes de surface (par exemple « address », « french », « has tv »). Au total, 4160 formes de surface ont été ainsi générées complètement automatiquement et sont utilisées pour DSTC2, 6555 pour DSTC3.

Pour évaluer nos propositions, nos résultats sont comparés avec deux systèmes état de l'art : le premier est un système à base de règles expertes utilisé dans le défi DSTC et le second est un système présenté dans (Williams, 2014), appris sur les données d'apprentissage du DSTC2 (nommé SLU1 dans l'article de Williams). Ces deux systèmes sont respectivement référencés par « S-règles » et « S-appris » dans la suite de cet article.

Les résultats de nos expériences (présentés dans le tableau 1) montrent que l'approche proposée, nommé ZSSP (pour Zero-Shot Semantic Parser) par la suite, atteint un niveau de performance (en termes de F-score) légèrement meilleur que celui de l'approche à base de règles ($0,794$ contre $0,782$ sur DSTC2 et $0,826$ contre $0,824$ sur DSTC3) et comparable à celui d'un modèle appris ($0,794$ contre $0,802$ sur DSTC2). Ainsi le modèle proposé atteint au démarrage des performances état-de-l'art sans utilisation de nombreuses règles spécifiques manuellement établies (coût d'experts humains) ni de données d'apprentissage (coûts de collecte et d'annotation).

Cependant, afin de mesurer l'impact de la représentation sémantique choisie sur la performance globale de l'approche, un système qui n'utilise pas ce type de représentation a été construit. Un F-score de $0,839$ (contre $0,919$ en configuration normale) est obtenu sur les transcriptions manuelles du DSTC2 par une simple stratégie de détection de patrons de mots à partir des exemples de la même base de connaissances K. Cette dernière observation confirme l'avantage d'avoir recours à une représentation sémantique riche apprise sur une grande quantité de données non annotées. En effet, cette dernière permet une meilleure généralisation des connaissances lexicales initiales (qui elles peuvent être assez limitées).

5.3 Adaptation en ligne

Comme mentionné précédemment, un mécanisme l'adaptation en ligne a également été proposé dans la section 3.4 pour améliorer les performances de l'approche ZSSP. Ainsi, les énoncés transcrits du corpus d'apprentissage du DSTC2 sont utilisés pour simuler des retours de validation utilisateurs et donc pour adapter notre base de connaissance K dynamiquement (en évitant le bruit dû à des erreurs de transcription automatique). Nous utiliserons l'ensemble de test DSTC2 (comme précédemment) pour juger de l'évolution des performances de l'approche ZSSP avec cette mise à jour. Pour positionner notre approche par rapport à l'état de l'art, les mêmes systèmes de référence que précédemment sont utilisés.

Ainsi, pour rendre possible la phase d'adaptation, les évaluations/retours des utilisateurs sont simulées en comparant la meilleure hypothèse du modèle avec l'étiquette sémantique de référence des phrases utilisateurs dans le corpus d'apprentissage DSTC2. Toutes les formes de surface de notre meilleure hypothèse ayant une étiquette sémantique présente dans

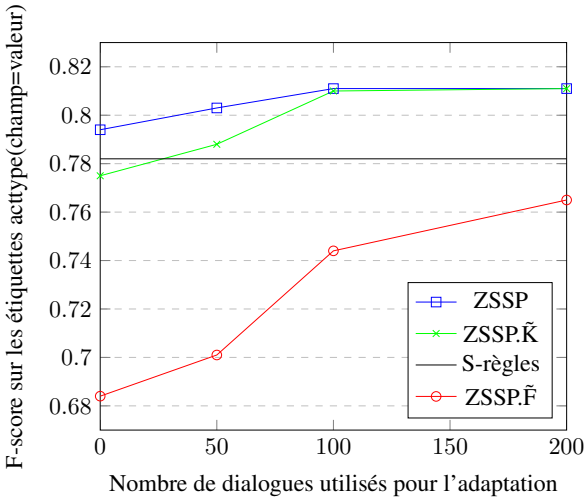

FIGURE 2 – Performances de diverses configurations de la méthode ZSSP en termes de F-score en fonction du nombre de dialogues utilisées pour l'adaptation.

l'annotation de référence sont considérées comme positives et toutes les autres comme négatives. K est mise à jour à la fin de chaque tour en suivant l'algorithme présenté dans la section 3.4 (avec $\alpha_p = \alpha_n = 1$).

Dans le but de quantifier l'influence de l'espace sémantique considéré F et de la base de connaissance initiale K sur l'approche proposée dans ce papier, nous avons fait le choix d'étudier trois configurations différentes de cette dernière. Nous distinguerons donc de l'approche ZSSP classique (base de connaissance K de qualité et un espace sémantique reposant sur une représentation word2vec apprise sur une grande quantité de données) deux variantes : la première, notée ZSSP.F̃, utilise une représentation sémantique « dégradée » et réduite à 50 dimensions, à savoir une représentation word2vec apprise avec l'algorithme *Skip-gram* (avec une fenêtre de 5 mots) sur des données non annotées issues du corpus d'apprentissage du DSTC2 (190366 mots en contexte) ; la seconde, notée ZSSP.K̃ utilise une version « dégradée » de K où 10% des formes de surface (exemples de types d'actes de dialogues) ont été retirés.

Les résultats présentés dans la figure 2 montrent l'évolution du F-score en fonction du nombre de dialogues utilisés pour l'adaptation. Même avant l'adaptation ZSSP $(0,794)$ et ZSSP.K̃$(0,775)$ atteignent des performances proches d'un système à base de règle $(0,782)$. Mais un espace sémantique appris sur une petite quantité de données peut avoir un impact significatif sur cette performance (comme montré avec ZSSP.F̃, $0,684$) dû à la fois à des mots hors vocabulaire et des mauvaises propriétés de généralisation de cet espace sémantique.

Néanmoins, dans toutes les configurations de ZSSP, la performance augmente conjointement avec le nombre de dialogues d'adaptation. En effet, à la fois ZSSP et ZSSP.K̃ obtiennent, après seulement 100 dialogues, des performances nettement meilleures que les modèles de références (0.811 contre 0.782 pour S-règles et 0.803 pour S-appris* [3]).

En outre, l'écart entre ZSSP.F̃ et le modèle à base de règles est nettement réduit tout au long du processus d'adaptation en ligne (de $0,098$ à $0,017$ après 200 dialogues). Cette observation montre que la méthode proposée peut aussi fonctionner avec un espace sémantique bruité. Ces résultats confirment l'avantage de la méthode d'adaptation en ligne proposée pour faire face aux limites de la couverture initiale de K et à la robustesse de l'espace sémantique F.

5.4 Généralisation

L'avantage majeur de l'utilisation d'un modèle word2vec par rapport à un simple modèle de détection par mots clés est l'intégration d'une représentation continue des mots dans le processus de décodage. Cette caractéristique confère au système une capacité de généralisation inhérente permettant de couvrir des mots inconnus correspondant à des valeurs non présentes dans l'ontologie définie du domaine ou de la tâche. Par exemple, dans le contexte d'un domaine de recherche de restaurant, il est intéressant pour un système de dialogue de détecter certaines situations où un utilisateur parle d'un type

3. les performances de S-appris n'ont pas été reportées sur la figure. 2 dans le but d'éviter une possible confusion (au regard de l'axe des abscisses) sachant qu'il utilise beaucoup plus de données d'apprentissage

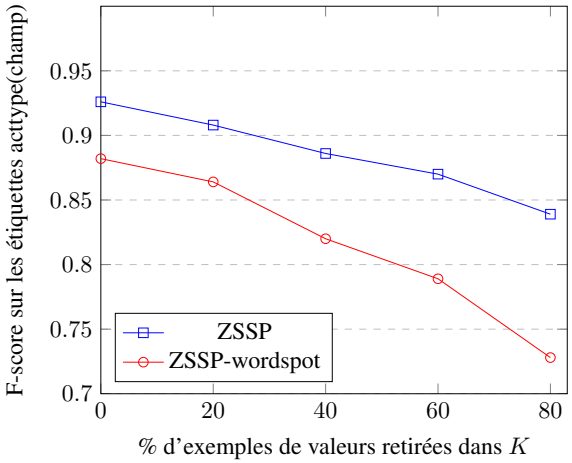

FIGURE 3 – Capacité de généralisation de l'approche ZSSP sur la corpus de test DSTC2 exprimée en termes de F-score sur la détection d 'actes de dialogue génériques (i.e. acttype(champ)) en fonction du pourcentage d'exemples de valeurs retirées dans K

d'aliment inconnu jusqu'alors par le système (si ce dernier n'est pas dans la base de données d'origine) ou au moins être en mesure de proposer une alternative en conséquence (en exploitant par exemple la proximité dans l'espace sémantique).

Afin d'évaluer la capacité de généralisation de notre système, nous avons volontairement supprimé de la base de connaissances de DSTC2 des formes de surface correspondant aux différents pourcentages des valeurs possibles de certains champs spécifiques. Dans cette étude préliminaire, nous avons choisi d'étudier l'impact sur les champs *food*, *area* et *pricerange*. Les performances du modèle sur les transcriptions manuelles ont été évaluées en termes de F-score pour *acttype(champ)* uniquement au lieu de *acttype(champ=valeur)* afin d'évaluer la détection des concepts de haut niveau.

Ainsi, nous comparons la performance de ZSSP avec une autre configuration de l'analyseur, notée ZSSP-wordspot. Ce dernier étiquette uniquement les segments qui atteignent un degré de similarité très élevé (une correspondance quasi parfaite - $0,94$). Vu que ce modèle est capable d'exploiter l'espace sémantique, cette configuration peut être assimilée à une stratégie robuste de détection de mots clés.

Les résultats (présentés dans la figure 3) montrent clairement une légère baisse de performances lorsque le pourcentage de valeurs retirées est grand. La différence entre les deux configurations est de $0,044$ à 0% et de $0,111$ à 80%. Cela confirme que l'approche proposée est tolérante à une faible densité de données dans K. Cette caractéristique peut être utile pour développer un système de dialogue générique permettant une évolution transparente de la base de connaissances contenant une base de données croissante.

6 Conclusions et perspectives

Dans cet article nous avons présenté une approche d'apprentissage sans données de référence pour la compréhension de la parole. Cette dernière repose à la fois sur l'utilisation d'une représentation sémantique riche apprise sur des données généralistes et sur une description ontologique minimale décrivant la tâche de compréhension visée. Nous avons montré que cette approche, bien que peu coûteuse, est tout de même comparable en termes de performances à des méthodes statistiques apprises sur de grande quantité de données annotées et aussi à un système à bases de règles expertes. De plus, la méthode proposée montre une meilleur tolérance à des valeurs de concept manquantes et donc offre des propriétés de généralisation pouvant être employées notamment dans l'extension de domaine en ligne.

De plus nous avons montré qu'un processus d'adaptation simple et ajustable en ligne permet de répondre aux deux limites de l'approche, à savoir la qualité de la base de connaissance K et de l'espace sémantique employé F. L'effort de supervision reste acceptable puisque l'utilisateur se contente de confirmer les hypothèses faites par le système et donc n'est pas contraint de corriger explicitement les erreurs du système. La comparaison avec d'autres techniques d'apprentissage active et la généralisation de cette technique par l'adaptation d'une vision plus probabiliste et dynamique sont planifiées pour de futurs travaux, de même que son évaluation dans le contexte d'interactions complètes.

Remerciements

Le travail présenté dans cet article a été partiellement financé par le projet ANR MaRDI (Man Robot Dialogue), ANR-12-CORD-0021. Vous trouvez plus d'informations concernant le projet sur http ://mardi.metz.supelec.fr.

Références

ANASTASAKOS T. & DEORAS A. (2014). Task specific continuous word representations for mono and multi-lingual spoken language understanding. In *ICASSP*.

BAYER A. & RICCARDI G. (2013). On-line adaptation of semantic models for spoken language understanding. In *ASRU*.

BENGIO S. & HEIGOLD G. (2014). Word embeddings for speech recognition. In *INTERSPEECH*.

BIAN J., GAO B. & LIU T. (2014). Knowledge-powered deep learning for word embedding. In *ECML*.

CAMELIN N., DETIENNE B., HUET S., QUADRI D. & LEFÈVRE F. (2011). Unsupervised concept annotation using latent dirichlet allocation and segmental methods. In *EMNLP Workshop on Unsupervised Learning in NLP*.

CELIKYILMAZ A., TUR G. & HAKKANI-TUR D. (2011). Leveraging web query logs to learn user intent via bayesian latent variable model. In *ICML*.

CLINCHANT S. & PERRONNIN F. (2013). Aggregating continuous word embeddings for information retrieval. In *Workshop on Continuous Vector Space Models and their Compositionality*.

DAUPHIN Y., TUR G., HAKKANI-TUR D. & HECK L. (2014). Zero-shot learning and clustering for semantic utterance classification. *arXiv preprint arXiv :1401.0509*.

DEORAS A. & SARIKAYA R. (2013). Deep belief network based semantic taggers for spoken language understanding. In *INTERSPEECH*.

GAO Y., GU L. & KUO H. (2005). Portability challenges in developing interactive dialogue systems. In *ICASSP*.

GOTAB P., DAMNATI G., BÉCHET F. & DELPHIN-POULAT L. (2010). Online slu model adaptation with a partial oracle. In *INTERSPEECH*.

HAHN S., DINARELLI M., RAYMOND C., LEFÈVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2010). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE TASLP*, **19**(6), 1569–1583.

HAKKANI-TUR D., HECK L. & TUR G. (2011). Exploiting query click logs for utterance domain detection in spoken language understanding. In *ICASSP*.

HECK L. & HAKKANI-TUR D. (2012). Exploiting the semantic web for unsupervised spoken language understanding. In *SLT*.

HENDERSON M., THOMSON B. & WILLIAMS J. (2014a). The second dialog state tracking challenge. In *SIGDIAL*.

HENDERSON M., THOMSON B. & WILLIAMS J. (2014b). The third dialog state tracking challenge. In *SLT*.

JABAIAN B., BESACIER L. & LEFÈVRE F. (2013). Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System. *IEEE TASLP*, **21**(3), 636–648.

LAROCHELLE H., ERHAN D. & BENGIO Y. (2008). Zero-data learning of new tasks. In *Conference on Artificial Intelligence*.

LEFÈVRE F. (2007). Dynamic Bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *ICASSP*.

LEFÈVRE F., MAIRESSE F. & YOUNG S. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In *INTERSPEECH*.

LEFÈVRE F., MOSTEFA D., BESACIER L., ESTEVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B. & ROJAS-BARAHONA L. (2012). Robustness and portability of spoken language understanding systems among languages and domains : the PORT-MEDIA project. In *LREC*.

LORENZO A., ROJAS-BARAHONA L. & CERISARA C. (2013). Unsupervised structured semantic inference for spoken dialog reservation tasks. In *SIGDIAL*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MIKOLOV T., YIH W. & ZWEIG G. (2013b). Linguistic regularities in continuous space word representations. In *NAACL-HLT*.

PALATUCCI M., POMERLEAU D., HINTON G. E. & MITCHELL T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22*, p. 1410–1418.

SARIKAYA R. (2008). Rapid bootstrapping of statistical spoken dialogue systems. *Speech Communication*, **50**(7), 580–593.

TUR G., HAKKANI-TUR D., HILLARD D. & CELIKYILMAZ A. (2011). Towards unsupervised spoken language understanding : Exploiting query click logs for slot filling. In *INTERSPEECH*.

TUR G., HAKKANI-TUR D. & SCHAPIRE R. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, **45**(2), 171–186.

TUR G., RAHIM G. & HAKKANI-TUR D. (2003). Active labeling for spoken language understanding. In *EUROSPEECH*.

WANG Y. & ACERO A. (2006). Discriminative models for spoken language understanding. In *ICSLP*.

WILLIAMS J. D. (2014). Web-style ranking and slu combination for dialog state tracking. In *Meeting of the Special Interest Group on Discourse and Dialogue*.

ZOU W., SOCHER R., CER D. & MANNING C. (2013). Bilingual word embeddings for phrase-based machine translation. In *EMNLP 2013*.