Comparaison et combinaison d'approches pour la portabilité vers une nouvelle langue d'un système de compréhension de l'oral

Bassam Jabaian [1,2], Laurent Besacier [1], Fabrice Lefèvre [2]

(1) LIG, University Joseph Fourier, Grenoble - France
(2) LIA, University of Avignon, Avignon - France

{bassam.jabaian,laurent.besacier}@imag.fr , fabrice.lefevre@univ-avignon.fr

Résumé

Dans cet article, nous proposons plusieurs approches pour la portabilité du module de compréhension de la parole (SLU) d'un système de dialogue d'une langue vers une autre. On montre que l'utilisation des traductions automatiques statistiques (SMT) aide à réduire le temps et le cout de la portabilité d'un tel système d'une langue source vers une langue cible. Pour la tache d'étiquetage sémantique on propose d'utiliser soit les champs aléatoires conditionnels (CRF), soit l'approche à base de séquences (PH-SMT). Les résultats expérimentaux montrent l'efficacité des méthodes proposées pour une portabilité rapide du SLU vers une nouvelle langue. On propose aussi deux méthodes pour accroître la robustesse du SLU aux erreurs de traduction. Enfin on montre que la combinaison de ces approches réduit les erreurs du système. Ces travaux sont motivés par la disponibilité du corpus MEDIA français et de la traduction manuelle vers l'italien d'une sous partie de ce corpus.

Abstract

In this paper we investigate several approaches for language portability of the spoken language understanding (SLU) module of a dialogue system. We show that the use of statistical machine translation (SMT) can reduce the time and the cost of porting a system from a source to a target language. For conceptual decoding we propose to use even conditional random fields (CRF) or phrase based statistical machine translation PB-SMT). The experimental results show the efficiency of the proposed methods for a fast and low cost SLU language portability. Also we proposed two methods to increase SLU robustness to translation errors. Overall we show that the combination of all these approaches reduce the concept error rate. This work was motivated by the availability of the MEDIA French corpus and the manual translation of a subset of this corpus into Italian.

Mots-clés : Système de dialogue, compréhension de la parole, portabilité à travers les langues, traduction automatique statistique

Keywords: Spoken Dialogue Systems, Spoken Language Understanding, Language Portability, Statistical Machine Translation.

1 Introduction

La portabilité d'un système de dialogue d'une langue vers une autre est une tâche difficile qui a fait l'objet, récemment, de plusieurs recherches. Certains composants d'un système de dialogue, tel que le gestionnaire de dialogue, sont relativement indépendants de la langue, ce qui n'affectent pas le processus de portabilité. Cependant, d'autres modules tel que le module de compréhension automatique de la parole (*Spoken Language Understanding, SLU*), doivent être réadaptés pour chaque nouvelle langue cible considérée. Dans cette étude, nous nous intéressons particulièrement à la portabilité d'un système de compréhension automatique de la parole vers une nouvelle langue.

Des travaux récents ont proposé d'utiliser des méthodes stochastiques pour la compréhension automatique de la parole. Ces méthodes sont des alternatives efficaces aux méthodes à base de règles; elles réduisent les besoins en expertise humaine tout en ayant la capacité de produire efficacement des réseaux d'hypothèses ou des listes de N-meilleures (N-best) (Suenderman, Liscombe, 2009) (Hahn, Lehnen, Raymond, Ney, 2008) (Raymond, Riccardi, 2007) (Wang, Acero, 2006) (Schwartz, Miller, Stallard, Makhoul, 1996). L'apprentissage de tels modèles nécessite un corpus annoté qui représente une couverture complète de la sémantique du domaine. Le portage d'un tel modèle vers une nouvelle langue consiste à transférer les connaissances présentes dans le corpus annoté en langue source vers une nouvelle langue avec un minimum de temps et d'effort humain. Par la suite, nous nommerons «langue source » la langue d'origine du système NLU et « langue cible », la langue vers laquelle le système doit être porté.

Récemment, quelques études ont montré que l'utilisation de la traduction automatique à différents niveaux du processus de compréhension peut aider au portage d'un système SLU vers une nouvelle langue (Suenderman, Liscombe, 2009) (Servan, Camelin, Raymond, Bechet, De Mori, 2010) (Lefèvre, Mairesse, Young, 2010) (Jabaian, Besacier, Lefèvre, 2010). Par exemple, dans (Suenderman, Liscombe, 2009) les auteurs proposent de traduire automatiquement les données de la langue source vers la langue cible, puis de re-apprendre une grammaire stochastique pour effectuer l'interprétation dans la langue cible. Une autre possibilité est de considérer que la sémantique d'un domaine est indépendante de la langue. Dans ce cas, une solution est de traduire le corpus d'apprentissage vers la langue cible et d'inférer les balises sémantiques associées au corpus traduit. Un système SLU stochastique peut ensuite être re-entraîné sur ce nouveau corpus annoté en langue cible. Comme décrit dans (Servan, Camelin, Raymond, Bechet, De Mori, 2010), les phrases du corpus d'apprentissage sont composées d'un ou plusieurs segments annotés sémantiquement. Traduire le corpus d'apprentissage en conservant l'information de segmentation en segments permet alors un appariement direct des segments traduits avec les étiquettes sémantiques. Les auteurs de (Servan, Camelin, Raymond, Bechet, De Mori, 2010) montrent qu'un portage du français vers l'italien est possible en utilisant cette approche, avec une traduction manuelle ou automatique. Le portage de l'annotation est encore moins difficile lorsqu'on utilise des méthodes qui n'ont pas besoin d'annotation sémantique au niveau mot ou segment. Par exemple, dans (Lefèvre, Mairesse, Young, 2010), le modèle proposé ne nécessite pas d'informations d'alignement.

Le choix d'une approche dépend de considérations techniques et également des caractéristiques du domaine ainsi que des données disponibles. Disposer de données manuellement traduites ou annotées, disposer d'annotateurs ou d'outils spécifiques pour la langue cible, peut faire la différence quant au choix de l'approche. Dans cet article, nous proposons plusieurs approches pour le portage d'un système de compréhension automatique de la parole vers une nouvelle langue. La langue source est le français étant donné que nous travaillons sur le corpus MEDIA (Bonneau -Maynard, Rosset, Ayache, Kuhn, Mostefa, 2005) et la langue cible considérée est l'italien puisque nous disposons également, au départ, d'une partie du corpus MEDIA traduite en italien. Nous sommes conscients de la proximité des langues source et cibles dans cette étude, mais ce choix est guidé par les données disponibles au départ. Le portage vers l'arabe est aussi envisagé et fait l'objet de travaux en cours. Les approches proposées dans cet article sont complètement automatiques et sans aucune supervision humaine lors du processus de portage.

Plus précisément, le but de cet article est de :

- proposer et évaluer différentes approches utilisant la traduction automatique pour porter un système SLU vers une nouvelle langue,

- considérant ensuite la meilleure approche obtenue, accroître sa robustesse aux erreurs de traduction.

Dans ces travaux, les modèles utilisés sont les champs aléatoires conditionnels (*Conditional Random Fields, CRF*) pour la compréhension automatique et l'approche à base de séquences (*phrase-based statistical machine*

translation) pour la traduction automatique. Les CRF (Lafferty, McCallum, Pereira, 2001) sont connus pour être très performants sur des tâches d'étiquetage temporel (annotation en entités nommées, étiquetage syntaxique, etc). Ils nécessitent un corpus annoté au niveau mots. Pour porter notre système vers une nouvelle langue, nous avons proposé plusieurs méthodes qui diffèrent selon le moment où est utilisé le module de traduction. En effet, un système de compréhension peut être porté soit au niveau du test (*TestOnSource*), en conservant le système SLU en langue source et en traduisant simplement les données de test en langue cible vers la langue source. La seconde possibilité consiste à porter le système au niveau de l'apprentissage *(TrainOnTarget)* en construisant un nouvel étiqueteur sémantique dans la langue cible. Pour cela, nous traduisons automatiquement l'intégralité du corpus d'apprentissage de la langue source vers la langue cible, puis nous inférons l'annotation sémantique de ce corpus pour les données traduites. Pour ce faire, nous proposons deux méthodes différentes. La première consiste à l'aide d'un système de traduction probabiliste source-cible, à traduire chaque phrase source annotée, segmentée en segments, et d'utiliser les segments traduits, associés aux étiquettes sémantiques, pour construire un nouveau système SLU en langue cible. La seconde approche utilise les informations extraites des alignements mot-à-mot pour inférer une relation mot_cible-étiquette_sémantique (plus de détails seront données dans la section 2).

Une autre approche, complètement différente, consiste à voir le processus de compréhension comme une tâche de traduction automatique d'une chaîne de mots vers une chaîne d'étiquettes sémantiques. Dans ce cas, le modèle de compréhension est une table de traduction mots-concepts. L'approche à base de séquences (*phrase-based*) (Koehn, Och, Marcu, 2003) nécessite des données alignées au niveau phrase avant le processus d'apprentissage (dont la première étape sera un alignement automatique en mots utilisant les modèles IBM). Dans ce cas, la portabilité vers une nouvelle langue consiste simplement à traduire en langue cible la partie "mots" du corpus d'apprentissage mots-concept sans modifier les concepts.

Pour répondre au deuxième point (robustesse), et puisque nous allons voir que la méthode *TestOnSource* donne les meilleures performances, nous proposons quelques méthodes pour augmenter la robustesse aux erreurs de traduction du système porté. Pour cela, une première approche présentée consiste à re-entrainer le système de compréhension, fondé sur les CRF, sur des données bruitées représentant les erreurs potentielles de traduction. La deuxième approche consiste à utiliser une post-édition automatique statistique (*Statistical Post Edition, SPE*) dans la langue-source pour tenter de corriger automatiquement les sorties issues du système de traduction automatique, avant de les envoyer à l'étiqueteur sémantique. Pour finir, nous proposons aussi dans cet article de combiner toutes les approches proposées dans cette étude, afin de réduire le taux d'erreurs de compréhension.

Cet article est structuré de la façon suivante : la section 2 présente en détail les approches que nous proposons pour porter un système de compréhension vers une nouvelle langue. La section 3 décrit deux solutions pour améliorer la robustesse de notre meilleur système porté, aux erreurs de traduction automatique. Le corpus MEDIA et les outils utilisés sont décrits dans la section 4 tandis que la section 5 présente les résultats expérimentaux obtenus. Finalement, conclusion et perspectives sont présentées dans la section 6.

2 Différentes méthodes pour porter un système de compréhension d'une langue vers une autre

Dans un système de dialogue, le rôle du processus de compréhension est d'extraire une liste d'hypothèses d'étiquettes de concepts à partir d'une phrase en entrée. Ces concepts représentent la sémantique de l'information existant dans la phrase en entrée. Les modèles de compréhension développés dans cette étude sont entraînés sur le corpus MEDIA, annoté en concepts sémantiques (voir section 4).

La génération automatique de ces concepts à partir d'une séquence de mots par des méthodes stochastiques telle que décrite dans (Raymond, Riccardi, 2007), peut être résumée de la façon suivante :

Soit C= c_1,...,c_n une séquence d'étiquettes sémantiques qui peut être associée initialement à la séquence de mots W= w_1,…,w_n ; pour chaque concept, une séquence de mots de W est associée et une étiquette est attribuée à chaque mot. Cette étiquette correspond au concept sémantique c_i et à la position de w_i.

Plusieurs études ont proposé de comparer différentes méthodes pour entraîner un modèle de compréhension de la parole (Hahn, Lehnen, Raymond, Ney, 2008) (Raymond, Riccardi, 2007). Dans cet article, nous proposons d'utiliser et d'évaluer deux approches état-de-l'art.

La première est fondée sur les champs aléatoires conditionnels (*Conditional Random Fields, CRF*), qui ont besoin d'un corpus annoté au niveau mot pour être entraînés. La seconde utilise une approche de traduction

probabiliste fondée sur les séquences (*Phrase-Based Statistical Machine Translation, PB-SMT*) et nécessite un corpus d'apprentissage annoté au niveau des phrases.

2.1 Champs aléatoires conditionnels (CRF)

Les CRF ("Conditional Random Fields" ou "Champs Aléatoires (Markoviens) Conditionnels") sont une famille de modèles graphiques introduits récemment (Lafferty, McCallum, Pereira, 2001). Ils permettent d'apprendre à annoter des données, en se basant sur un ensemble d'exemples déjà annotés. Les CRF ont le plus souvent été utilisés dans le domaine du TAL, pour étiqueter des séquences d'unités linguistiques. Ces modèles possèdent les avantages des modèles génératifs et discriminants. En effet, comme les classifieurs discriminants, ils peuvent manipuler un grand nombre de descripteurs, et comme les modèles génératifs, ils intègrent des dépendances entre les étiquettes de sortie et prennent une décision globale sur la séquence. Par rapport aux modèles de Markov Cachés (HMMs), les CRF ont par ailleurs l'avantage de relâcher certaines hypothèses d'indépendance.

Dans notre cas, pour apprendre notre modèle CRF, les données d'apprentissage doivent être représentées selon le formalisme BIO décrit dans (Raymond, Riccardi, 2007), qui indique les frontières entre les concepts sémantiques selon l'exemple ci-dessous :

"Je voudrais réserver un hôtel à Paris "

Sera représenté par la séquence de couples (w,c):

(je, B_command-tache) (voudrais, I_command-tache) (réserver, I_command-tache) (un, B_Objet) (hôtel, I_objet) (à, B_loc-ville) (Paris, I_loc-ville)

La probabilité d'une sequence de concepts, étant donnée une séquence de mots est alors calculée par :

$$p(c_1^N|w_1^N) = \frac{1}{Z}\prod_{n=1}^{N} H\,(c_{n-1}, c_n, w_{n-2}^{n+2})$$

avec

$$H(c_{n-1}, c_n, w_{n-2}^{n+2}) = \sum_{m=1}^{M} \lambda_m . h_m\,(c_{n-1}, c_n, w_{n-2}^{n+2})$$

H est un modèle log-linéaire fondé sur des fonctions caractéristiques $h_m\,(c_{n-1}, c_n, w_{n-2}^{n+2})$ qui représentent l'information extraite du corpus d'apprentissage ; les poids λ du modèle log-linéaire sont estimés lors de l'apprentissage et *Z* est un terme de normalisation définit tel que :

$$Z = \sum_{m=1}^{M}\prod_{n=1}^{N} H\,(c_{n-1}, c_n, w_{n-2}^{n+2})$$

Afin de porter notre système de compréhension fondé sur les CRF (note SLU/CRF par la suite) d'une langue à une autre, nous avons proposé plusieurs approches qui diffèrent selon le moment où est appliqué le processus de transfert entre les langues.

2.1.1 Portage au niveau du test (*Test On Source*)

Dans cette approche, nous supposons qu'un système SLU est disponible en langue source, et nous utilisons un système de traduction automatique probabiliste pour traduire les phrases de test en langue cible vers la langue source. Ces traductions sont ensuite les entrées du système SLU original. En d'autres termes, nous portons le système « au niveau du test » sans modifier le processus d'apprentissage du système SLU. Cette technique sera dénommée *TestOnSource* dans la suite de cet article. Elle a l'avantage d'être très simple mais ses performances dépendront, bien évidemment, des performances du système de traduction automatique utilisé pour revenir de la langue cible à la langue source.

2.1.2 Portage au niveau de l'apprentissage (*Train On Target*)

Cette approche consiste à re-entraîner un système SLU en langue cible. L'idée générale est de traduire le corpus d'apprentissage de la langue source vers la langue cible et d'inférer les étiquettes sémantiques associées. Pour inférer l'annotation sémantique, nous proposons deux approches différentes :

1. Traduire avec des tags XML (*Tagged Translation*):
   Dans cette approche, le corpus d'apprentissage est traduit en prenant en compte la segmentation en « segments sémantiques » (un « segment » est composé potentiellement de plusieurs mots mais correspond à une et une seule étiquette sémantique). Pour cela, nous utilisons une option (*-xml-input*) du décodeur MOSES (Koehn, Hoang, Birch, Callisonburch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, Herbst, 2007), qui force la segmentation d'une phrase à traduire, cette segmentation étant décrite par des tags XML. Ainsi, en sortie, nous obtenons chaque phrase du corpus d'apprentissage traduite, ainsi qu'une projection des tags XML de la source vers la cible. L'exemple donné précédemment peut alors être représenté sous la forme suivante :

   <tag c=command_tache > Je voudrais réserver </tag> <tag c=objet > un hôtel </tag> <tag c=localisation_ville> à Paris </tag>

   En utilisant l'option de MOSES qui prend en compte les tags XML comme information de segmentation, nous obtenons la sortie traduite suivante :

   <tag c= command_tache > vorrei prenotare </tag> <tag c=objet> un hotel </tag> <tag c= localisation_ville> a Parigi </tag>

   Tout le corpus d'apprentissage est traduit de cette façon puis re-formaté au format BIO avant un nouvel apprentissage du modèle CRF de compréhension en langue cible.

2. Projections des concepts sémantiques d'une langue à l'autre en utilisant un alignement en mots (*Alignment*):

   L'alignement automatique en mots est une étape importante dans le processus de construction d'un modèle de traduction probabiliste. Plusieurs boites à outils existent pour cette tâche telles que GIZA++ (Och, Ney, 2000) qui utilise les modèles IBM et HMM, ou Berkeley aligner (Liang, Taskar, Klein, 2006) qui repose sur une méthode d'alignement par consensus (*alignment by agreement*).

   Pour projeter les concepts sémantiques d'une langue à l'autre, on peut utiliser les informations d'alignement bilingue en mots. Plus précisément, la première phase consiste à aligner automatiquement le corpus parallèle source-cible. Ensuite, comme le corpus source est déjà annoté sémantiquement, il est possible d'apparier les étiquettes sémantiques aux mots en langue cible en utilisant l'information d'alignement. Certains cas ambigus demeurent cependant, comme illustré dans la figure 2 (alors que la figure 1 présente un cas où la projection est évidente). Dans cet article, l'aligneur utilisé est *Berkeley Aligner*.

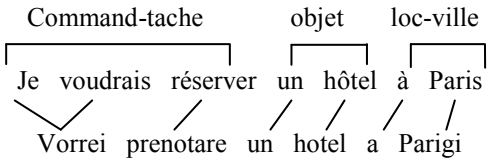

Figure 1 : Exemple de projection des tags sémantiques du français vers l'italien

Pour faire cette projection, nous avons développé un algorithme qui parcourt la phrase en langue cible et associe aux mots la bonne étiquette sémantique. Pour les cas ambigus où un mot cible est aligné avec plusieurs mots sources correspondent à deux concepts différents (voir figure 2), nous devons prendre une décision sur quel concept doit être associé au mot cible. Pour cela, notre proposition est de simplement associer le mot cible au premier concept rencontré. Par exemple, sur la figure 2, le mot italien *alla* sera associé au concept *loc-dis* et pas au concept *loc-lieu*. Cette décision, bien qu'arbitraire, a l'avantage d'être cohérente d'un bout à l'autre du corpus si le même cas est rencontré à nouveau.

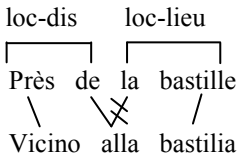

Figure 2 : Exemple de cas ambigu pour la projection des concepts sur les mots cible

2.2 Compréhension par une approche de traduction (PB-SMT)

Afin d'éviter de recourir à un corpus d'entrainement annoté au niveau mot, nous proposons d'utiliser l'approche PB-SMP qui ne nécessite qu'un alignement au niveau des phrases complètes. Dans cette approche, nous considérons que les séquences de concepts sont les traductions des séquences de mots initiales. Ainsi l'étiquetage sémantique est vu comme une tâche de traduction : la meilleure séquence C à partir des mots W est définie par :

$$\hat{C} = argmax_c\, P(C|W) = argmax_c\, P(W|C).P(C)$$

Pour résoudre cette équation sont requis : un modèle de langage de concepts *P(C)* (qui peut être appris à l'aide de SRILM (Stolcke, 2002) sur le corpus de concepts) et d'un model de traduction *P(W|C)* (qui peut être un modèle PB-SMT par exemple). Nous avons utilisé MOSES pour entraîner un tel modèle PB-SMT à partir d'un corpus « parallèle ». Les poids associés à ce modèle sont optimisés par un apprentissage à taux d'erreur minimum (MERT) qui est traditionnellement utilisé pour optimiser le score BLEU. Puis les performances de cette approche PB-SMT de base ont été améliorées en utilisant des caractéristiques de la tâche de compréhension sémantique.

D'abord, en suivant l'hypothèse raisonnable que la sémantique d'une phrase respecte l'ordre dans lequel les mots sont émis, la table de segments est re-entrainée en utilisant une contrainte de monotonie durant l'alignement automatique en mots. Puis, dans la mesure où une difficulté majeure du processus de traduction est l'alignement automatique correct d'un mot du langage source avec le mot correspondant dans le langage cible, nous avons tenté d'aider le processus d'alignement par l'utilisation du formalisme BIO. De cette façon, l'extraction de la table de segments a été obtenue sur un corpus avec un alignement de meilleure qualité. Enfin, la mesure d'évaluation du SLU étant le CER et non le score BLEU, nous avons modifié l'optimisation MERT pour optimiser le CER directement. Finalement, pour éviter les mots hors-vocabulaire (venant principalement de noms de ville absents des données d'entrainement), une liste de villes est ajoutée aux données d'apprentissage et le système *SLU/PB-SMT* est re-entrainé.

3 Accroître la robustesse du SLU aux erreurs de traduction

Nos expériences, ainsi que d'autres travaux (Lefèvre, Mairesse, Young, 2010) (Jabaian, Besacier, Lefèvre, 2010), ont montré que la meilleure méthode pour la portabilité SLU est aussi la plus simple, le *TestOnSource*. La faiblesse principale de cette méthode est que la qualité de l'étiquetage dépend grandement de la qualité de la traduction préalable. Ainsi, le système SLU doit prendre en compte des entrées bruitées par les erreurs de traduction.

De sorte à améliorer la robustesse de l'approche, nous proposons deux méthodes dans ce papier. La première prend en compte le bruit venant de la traduction durant le processus d'apprentissage des modèles SLU ; la seconde corrige automatiquement la sortie du système de traduction avant de la transférer au système SLU. Il est notable que, bien que pas encore évaluées dans ce cadre (par manque de données audio dans la langue cible), les deux méthodes seront tout à fait adaptées pour traiter aussi les erreurs dues à la reconnaissance de la parole dans une tache de compréhension réelle.

3.1 Apprentissage sur des données bruité

Le principe de cette méthode est d'entrainer un modèle SLU (dans le langage source) avec des données additionnelles provenant de la sortie d'un système de traduction automatique. En pratique, nous traduisons les données d'apprentissage disponibles entre les langues cible et source et nous inférons les concepts associés aux données bruitées (en suivant la même méthode que *TrainOnTarget*). Puis nous ajoutons les données corrompues (maintenant annotées sémantiquement) aux données originales et l'ensemble est utilisé pour entrainer le nouveau modèle SLU (dans la langue source) qui alors intégrera le bruit présent dans les données traduites.

3.2 Post-édition statistique

Plusieurs travaux récents en traduction automatique comme (Simard, Goutte, Isabelle, 2007) (Diaz de Ilarraza, Labaka, Sarasola, 2008) ont utilisé une approche basée sur un système de traduction pour post-éditer les sorties d'un autre système de traduction. Un tel système, malgré une démarche qui peut paraitre contre-intuitive, a été proposé pour améliorer la qualité des données traduites avant leur envoi à des post-éditeurs humains. Pour entrainer un tel post-éditeur, (Simard, Goutte, Isabelle, 2007) (Diaz de Ilarraza, Labaka, Sarasola, 2008) utilisent les sorties d'un système SMT avec comme données parallèles leur post-édition manuelle.

Dans notre cas, dans la mesure où la sortie du système SMT sera utilisée comme entrée du système SLU entrainé sur les données du langage source, nous proposons de post-éditer cette sortie afin de diminuer le bruit du à la traduction des entrées utilisateurs.

Pour apprendre un SPE, notre choix a été de traduire automatiquement l'ensemble de données disponibles pour la langue cible, puis d'utiliser les sorties traduites avec les parties correspondantes transcrites manuellement, comme corpus parallèle. Nous pensons que le module de post-édition permettra ainsi de réordonner quelques mots ou de retrouver des mots manquants dans un certain nombre de phrases.

4 Description du corpus et d'outils

Toutes les expériences décrites dans le papier ont été réalisées sur le corpus français MEDIA. Ce travail a été motive par la disponibilité d'une traduction manuelle en italien d'une sous-partie de ce corpus.

4.1 Le corpus MEDIA

Comme décrit dans (Bonneau -Maynard, Rosset, Ayache, Kuhn, Mostefa, 2005), ce corpus couvre un domaine lié aux réservations de chambres d'hôtels et aux informations touristiques. Le corpus est constitué de 1257 dialogues enregistrés par 250 locuteurs, collectés en situation de Wizard-of-Oz (un humain simule le système de dialogue).

Les dialogues sont regroupés en 3 parties : un ensemble d'apprentissage (environ 13k phrases), un ensemble de développement (1,3k phrases) et un ensemble d'évaluation (3k phrases). Dans nos expériences, nous ne prenons en compte que les phrases utilisateurs.

Le corpus est étiqueté avec 99 concepts différents. Ces étiquettes peuvent être simples comme les dates ou les noms de ville ou peuvent être plus complexes comme les coréférences. A titre d'illustration, voici une phrase de MEDIA :

Je voudrais une chambre double à Marseille

L'annotation sémantique de cette phrase aura la forme :

Je voudrais [null], une [nombre-chambre], chambre double [chambre-type], à Paris [localization-ville].

Cette annotation sémantique découpe chaque phrase en plusieurs segments. Chaque segment est non seulement annoté avec le nom du concept mais aussi par une valeur, une modalité et un specifieur. Les expériences présentées dans le papier prennent en compte uniquement le nom du concept et la modalité du segment.

Un sous-ensemble de l'apprentissage (environ 5.6k phrases), de même que les ensembles de test et de développement, ont été manuellement traduit en italien dans le contexte du projet européen LUNA (Servan, Camelin, Raymond, Bechet, De Mori, 2010).

4.2 Les SMTs apprit

Dans cette étude, nous utilisons deux systèmes de traduction automatique pour obtenir les traductions du français vers l'italien et de l'italien vers le français. Pour réaliser ces traductions, la boite à outils Moses (Koehn, Hoang, Birch, Callisonburch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, Herbst, 2007) est utilisée. Moses implémente l'état-de-l'art des systèmes de traduction par segments utilisant des modèles log-linéaires.

Nous utilisons la partie manuellement traduite en italien de l'ensemble d'apprentissage du corpus MEDIA comme corpus parallèle pour Moses dans les deux directions pour entrainer les modèles. Chacune des parties séparément permet l'apprentissage d'un modèle de langage. Aussi l'ensemble de développement avec sa traduction est utilisé comme corpus parallèle pour ajuster les poids du modèle log-linéaire des systèmes SMT. Finalement, nous obtenons un système de français vers l'italien avec un score BLEU de 43,62 et de l'italien vers

le français avec un score de 47,18. Ces scores sont mesurés sur l'ensemble de test de MEDIA manuellement traduit. Dans la mesure où une seule référence par phrase est utilisée pour évaluer le score BLEU, de même que l'ensemble d'apprentissage est réduit (5,6k), ces performances peuvent être considérées comme très acceptables.

Nous avons aussi re-traduit automatiquement en français la version manuelle en italien du corpus, de sorte à utiliser cette traduction en parallèle de la partie originale pour entrainer le SPE et fournir les données bruitées pour le SCTD. Du point de vue de la performance de traduction exclusivement, l'utilisation de la post-édition automatique améliore le score BLEU du système de 47,18 à 49,25.

4.3 Compléter la traduction de MEDIA

Le système de traduction français-italien est utilisé pour obtenir une traduction automatique de la partie restante (non traduite manuellement) du corpus d'apprentissage, ainsi une traduction intégrale (manuelle + automatique) est disponible. Le système italien-français est utilisé pour traduire le test italien en français, qui sera utilisé pour l'approche *TestOnSource*. Le tableau 1 donne un aperçu des ensembles disponibles pour les expériences.

MEDIA data	Train	Dev	Test
French MEDIA	13K	1,3K	3,5K
Italian manual	5,6K	1,3K	3,5K
Italian automatic	7,4K	-	-

Tableau 1: aperçu du corpus MEDIA et de sa traduction vers l'italien (# phrases).

5 Expériences et résultats

Afin d'évaluer les performances des approches proposés, la traduction manuelle en italien des données de test est utilisée. Le CER est le critère d'évaluation retenu pour cette étude. Le CER est l'équivalent du taux d'erreur en mots (WER), et peut être défini comme le ratio de la somme des concepts omis, insérés et subtilisés sur le nombre de concepts dans la référence. Premièrement nous évaluons et comparons SLU/CRF et SLU/PB-SMT, nous évaluons de même notre proposition de robustesse, puis les systèmes sont combinés. Pour finir, nous validons nos approches en utilisant les traductions obtenues par un système de traduction en ligne (ie sans utiliser de traduction manuelle).

5.1 Les stratégies de portabilité SLU/CRF

La totalité de l'ensemble d'apprentissage de MEDIA est utilisé pour apprendre un étiqueteur français de base utilisant des uni- et bi-grammes. Cette base atteint de bonnes performances (12,9% CER) et peut être considérée comme une référence pour les méthodes proposées. Pour évaluer les performances de l'approche *TestOnSource* la traduction automatique en français du test italien (comme décrit en 4) est fournie à l'étiqueteur de base.

La méthode *TrainOnTarget* décrite dans 2.1.2 a été appliquée. Nous utilisons le système français-italien (décrit en 4) pour traduire le corpus d'entrainement intégrant des balises XML correspondant aux segments conceptuels afin d'évaluer la méthode *TaggedTranslation*, et la totalité des traductions en italien de MEDIA (manuelle et automatique) avec la version française comme corpus parallèle pour obtenir l'alignement mot-a-mot. L'information d'alignement telle que défini en 2.1.2 est utilisée, pour évaluer la performance de la méthode *Alignement*. Toutes les expériences utilisent l'outil CRF++ (*https://crfpp.sourceforge.net)*. L'ensemble des résultats est regroupés dans le tableau 2.

Il apparait clairement à la lecture des résultants que la méthode *Alignement* est meilleure que *TaggedTranslation*. Ceci peut être expliqué par le fait que *Alignement* est seulement influencé par les erreurs d'alignement, tandis que *TaggedTranslation* est influencé par les erreurs de traduction automatique qui sont plus importantes. On note aussi que la méthode *TestOnSource* est plus performante que les méthodes *TrainOnTarget.*les performances de toutes les méthodes sont considérées come bonne en comparaison avec la référence française.

Model	Sub	Del	Ins	CER
FR	3,1	8,1	1,8	12,9
SLU/CRF(TestOnSource)	5,2	12,1	2,6	19,9
SLU/CRF(TaggedTranslation)	3,7	16,9	2,1	22,7
SLU/CRF(Alignment)	3,1	15,0	2,3	20,5

Tableau 2 : Evaluation (CER %) de différentes stratégies de portabilité du SLU utilisant les méthodes SLU/CRF

5.2 SLU/PB-SMT

Nos premières tentatives pour construire le modèle PB-SMT pour le SLU italien ont clairement montré des performances inférieures aux CRF (CER=28,1% après réglage MERT pour le PB-SMT comparé aux ~20% pour les CRF). Les améliorations progressives du modèle proposées en Section 2.2 sont évaluées dans le tableau 3. L'utilisation de la contrainte de monotonie durant l'alignement en mot permet une réduction de 0,6% absolu. Convertir les données selon le formalisme BIO avant la phase d'apprentissage réduit le CER de façon significative de 2,8%. Enfin optimiser le CER à la place du BLEU réduit le CER de 0,3% supplémentaire. L'ajout d'une liste de villes à l'ensemble d'apprentissage avant réapprentissage du modèle PB-SMT permet une réduction finale de 0,5%.

Les résultats montrent qu'en dépit de réglages fins de l'approche SMT, les approches à base de CRF obtiennent toujours les meilleures performances. De plus, dans une expérience parallèle, un modèle PB-SMT a été construit pour le SLU Français afin de le tester dans l'approche *TestOnSource*. Mais les performances de l'approche sont décevantes et bien en-deça de celles des autres méthodes. Elle a donc été écartée pour le reste de l'étude.

A partir d'une analyse rapide du type d'erreurs de chaque modèle, nous pouvons observer que les méthodes utilisant des CRF ont un haut niveau de suppressions comparativement aux autres types d'erreurs, tandis que la méthode PB-SMT présente un meilleur compromis entre les erreurs de suppression et d'insertion, et ce bien qu'elle abouti à un CER plus élevé.

SLU/PB-SMT	Sub	Del	Ins	CER
Initial	6,5	4,0	18,6	29,1
+ MERT (BLEU)	6,3	9,3	12,5	28,1
+ Monotone align	7,4	8,4	11,8	27,5
+ BIO format	6,5	10,6	7,7	24,7
+ MERT (CER)	6,4	10,9	7,2	24,4
+ City list	7,2	10,5	6,1	23,9

Tableau 3 : améliorations itératives de la méthode SLU/PB-SMT sur le test italien de MEDIA (CER%)

5.3 SLU/CRF *TestOnSource* robuste

Nous avons tenté d'améliorer les performances de la méthode *TestOnSource* SLU/CRF en renforçant sa robustesse aux erreurs de traduction. Premièrement nous traduisons automatiquement la partie manuelle en italien. Ensuite nous apprenons un nouvel étiqueteur CRF simultanément sur les données d'apprentissage en français et traduites (approche +SCTD, décrite en 3.1). La méthode de la section 3.2 (SPE) a aussi été évaluée, dans laquelle le test traduit post-édité a été transmis aux CRF de base (+SPE) ou aux CRF appris sur les données corrompues (+SCTD+SPE).

L'évaluation des performances de ces approches sont rapportées dans le tableau 4. Les deux méthodes, d'apprentissage sur données bruitées et SPE, améliorent les performances de l'étiqueteur sémantique. Leur mise en série donne les meilleures performances.

SLU/CRF	Sub	Del	Ins	CER
TestOnSource	5,2	12,1	2,6	19,9
+SCTD	5,9	11,4	2,3	19,6
+SPE	6,5	10,6	2,5	19,7
+SCTD +SPE	6,4	9,9	2,9	19,3

Tableau 4 : Evaluation (CER %) des approches proposées pour la robustesse des systèmes au bruit de traduction

5.4 Combinaison de systèmes

Nous proposons de combiner les trois approches principales (*TestOnTarget* et *TrainOnTarget* SLU/CRF, et SLU/PB-SMT) afin de bénéficier de leurs caractéristiques respectives pour améliorer la performance globale. La combinaison (dénotée BASIC dans le tableau 5) est simple : un réseau de confusion est construit à partir des trois hypothèses et la séquence de concept correspondant à la plus grande probabilité a posteriori est calculée. La performance est améliorée de façon significative (-1,3% CER) ce qui confirme la complémentarité des méthodes.

Finalement nous combinons toutes les méthodes proposées dans ce papier (SLU/CRF *TrainOnTarge,* SLU/CRF *TestOnSource*, +SCTD, +SPE, +SCTD+SPE, SLU/PB-SMT). Ce qui permet d'atteindre les meilleures performances rapportées sur ce test (18,2%). Afin de mesurer l'influence de la méthode SLU/PB-SMT sur les performances de la combinaison, nous avons aussi évalué les performances de la combinaison diminuée de SLU/PB-SMT. Cette expérience a montré qu'en dépit de ses mauvais résultats individuels, la méthode PB-SMT a une influence importante sur la combinaison.

Model	Sub	Del	Ins	CER
BASIC	6,2	9,7	2,7	18,6
ALL	5,4	10,5	2,3	18,2
ALL – SLU/PB-SMT	6,6	10,2	2,7	19,4

Tableau 5 : combinaison de systèmes avec et sans l'approche PB-SMT

5.5 Validation des stratégies de portabilité SLU/CRF en utilisant des traductions en ligne uniquement

Les expériences présentées dans cet article ne sont pas totalement non-supervisées, dans tous les cas nous avons utilisé des données traduites manuellement pour obtenir le système de traduction pour *TestOnSource* ou pour compléter la traduction de l'apprentissage et obtenir les informations d'alignement pour la méthode *TrainOnTraget*.

Le coût associé à cette traduction manuelle est relativement bas comparé à celui de collecter et d'annoter un nouveau corpus d'apprentissage, mais il reste non négligeable. Nous voulons vérifier qu'un tel coût est justifié par comparaison à une approche totalement non-supervisée et donc (potentiellement) « gratuite ».

Afin de répondre à cette interrogation nous proposons de reproduire les expériences en utilisant un système de traduction gratuit en ligne à la place de notre système de traduction appris sur la tâche.

Pour évaluer la méthode *TestOnSource* nous traduisons le test MEDIA en italien à l'aide d'une solution gratuite en ligne puis nous utilisons cette traduction comme entrée de l'étiqueteur CRF de base. Pour la méthode *TrainOnTarget* deux approches ont été testées. Pour permettre la comparaison avec la méthode

TaggedTranslation, nous proposons de traduire les données d'entrainement de MEDIA, segment par segment, au moyen du traducteur en ligne, puis ces traductions sont associées aux étiquettes sémantiques initiales. Dans une seconde version, les données sont traduites intégralement puis utilisées comme corpus parallèle pour l'approche *Alignement*.

Pour choisir un traducteur automatique dans le cadre de nos expériences nous avons comparé les performances de deux traducteurs réputés. Le test MEDIA et sa traduction manuelle ont été utilisé comme couple test/référence dans chacune des directions de traduction. Pour l'Italien vers le français un traducteur de score BLEU 42,58 a été sélectionné (à comparer à 47,18 obtenu par le système SMT appris sur les traductions manuelles), et pour le français vers l'italien un traducteur de score 39,75 a été retenu (à comparer avec 43,62). Les résultats de cette expérience sont reportés dans le tableau 6.

De manière attendue les performances des systèmes obtenus par cette approche non-supervisée sont inferieures à celle des systèmes semi-supervisés. Toutefois malgré la dégradation du CER pour toutes les approches son niveau absolu reste tout à fait acceptable considérant les besoins de la tache et la réduction substantielle du coût de développement.

La méthode *Alignement* est la plus perturbée et devient presque équivalente à *TaggedTranslation* en version non-supervisée. Le CER augmente de 22,7% à 26,6% (+3,9% absolu) pour *TaggedTranslation* et de 20,5% à 26,5% (+6%) pour *Alignement*. *TestOnSource* perd 3,2% mais reste la plus performante. Ces résultats nous engagent à tester de nouvelles langues pour lesquelles nous ne disposons pas de traductions manuelles.

Model	**Sub**	**Del**	**Ins**	**CER**
Semi supervisé				
SLU/CRF(TestOnSource)	5,2	12,1	2,6	19,9
SLU/CRF(TaggedTranslation)	3,7	16,9	2,1	22,7
SLU/CRF(Alignment)	3,1	15,0	2,3	20,5
Non supervisé				
SLU/CRF(TestOnSource)	6,1	14,5	2,5	23,1
SLU/CRF(TaggedTranslation)	5,5	15,4	5,7	26,6
SLU/CRF(Alignment)	6,3	14,8	5,4	26,5

Tableau 6 : Evaluation (CER %) des stratégies de portabilité SLU/CRF en utilisant des traductions en ligne

6 Conclusion

Dans cet article on a proposé et comparé plusieurs approches pour la portabilité d'un SLU a travers les langues. Les CRFs et le PB-SMT on été utilisé pour cette tache et les résultats montrent que l'utilisation d'un étiqueteur a base de CRF avec des données de test traduites donne la meilleur performance. On a aussi montré l'intérêt de l'utilisation de deux méthodes différentes pour accroitre la robustesse du SLU aux erreurs de traduction. Enfin on a montré que la combinaison de toutes les méthodes proposées augmente la performance du système.

Remerciements

Ce travail est supporté par le projet ANR PORT-MEDIA (ANR 08 CORD 026 01). Plus d'information disponible sur le site du projet : www.port-media.org

Références

Suenderman K., Liscombe J. (2009). From rule-based to statistical grammars: Continuous improvement of large-scale spoken dialog system. Actes de *ICASSP*.

Hahn S., Lehnen S., Raymond C., Ney H. (2008). A comparison of various methods for concept tagging for spoken language understanding. Actes de *LREC*.

Raymond C., Riccardi G. (2007). Generative and discriminative algorithms for spoken language understanding. Actes de *Interspeech*.

Wang Y., Acero A. (2006). Discriminative models for spoken language understanding. Actes de *ICSLP*.

Schwartz R., Miller S., Stallard D., Makhoul J. (1996). Language understanding using hidden understanding models. Actes de *ICSLP*.

Suenderman K., Liscombe J. (2009). Localization of speech recognition in spoken dialog systems: How machine translation can make our lives. Actes de *Interspeech*.

Servan C., Camelin N., Raymond C., Bechet F., De Mori R. (2010). On the use of machine translation for spoken language understanding portability. Actes de *ICASSP*.

Lefèvre F., Mairesse F., Young S. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. Actes de *Interspeech*.

Jabaian B., Besacier L., Lefèvre F. (2010). Investigating multiple approaches for SLU portability to a new language. Actes de *Interspeech*.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D. (2005). Semantic annotation of the French media dialog corpus. Actes de *Eurospeech.*

Lafferty J., McCallum A., Pereira F. (2001). Conditional random fields: Probabilistic models for segmenting and labelling sequence data. Actes de *ICML.*

Koehn P., Och F., Marcu D. (2003). Statistical phrase_based translation. Actes de *HLT/NAACL*.

Koehn P., Hoang H., Birch A., Callisonburch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E. (2007). Moses: Open source toolkit for statistical machine translation. Actes de *ACL.*

Och F., Ney H. (2000). Improved Statistical Alignment Models. Actes de *ACL*.

Liang P., Taskar B., Klein D. (2006). Alignment by agreement. Actes de *HLT*.

Stolcke A. (2002). SRILM an extensible language modeling toolkit. Actes de *SLP*.

Simard M., Goutte C., Isabelle P. (2007). Statistical phrase-based post-editing. Actes de *NAACL*.

Diaz de Ilarraza A., Labaka G., Sarasola K. (2008). Statistical post-editing: A valuable method in domain adaptation of RBMT systems for less-resourced languages. Actes de *MATMT*.