Traduction de dialogue: résultats du projet NESPOLE! et pistes pour le domaine

Hervé Blanchon, Laurent Besacier

Laboratoire CLIPS
BP 53
38041 Grenoble Cedex 9
{prenom.nom}@imag.fr

Résumé

Dans cet article, nous détaillons les résultats de la seconde évaluation du projet européen NESPOLE! auquel nous avons pris part pour le français. Dans ce projet, ainsi que dans ceux qui l'ont précédé, des techniques d'évaluation subjectives — réalisées par des évaluateurs humains — ont été mises en œuvre. Nous présentons aussi les nouvelles techniques objectives — automatiques — proposées en traduction de l'écrit et mises en œuvre dans le projet C-STAR III. Nous conclurons en proposant quelques idées et perspectives pour le domaine.

Mots Clés

Traduction de dialogue, évaluation subjective et objective de composants de TALN

Introduction

Le projet NESPOLE! [Lazzari G., 2000] visait à capitaliser les efforts des partenaires européens et américains du consortium C-STAR II et aller plus loin en termes scientifiques. Les langues impliquées sont l'italien, le français, l'allemand et l'anglais.

Les démonstrateurs NESPOLE! mettent en situation de dialogue un agent touristique italophone et un client parlant anglais, français ou allemand. Une importance particulière a été donnée à l'évaluation des deux démonstrateurs produits (2001 et 2002). Ces évaluations ont été conduites avec des évaluateurs humains qui jugent la qualité de traduction du système en comparant l'énoncé source et l'énoncé cible produit. On parle d'évaluation subjective. Ces évaluations ont permis de mesurer à la fois des performances brutes, ainsi que les progrès accomplis.

Dans le domaine de la traduction de l'écrit, afin de diminuer le coût de l'évaluation, la méthode BLEU [Papineni K., et al., 2002] a d'abord été proposée puis raffinée ensuite sous d'autres noms. Il s'agit ici de comparer automatiquement la sortie du système à une traduction étalon éventuellement complétée par un ensemble de paraphrases. On parle d'évaluation objective. Le domaine de la traduction de dialogue a adopté récemment les techniques d'évaluation objectives qui obligent à produire plusieurs paraphrases d'une traduction étalon.

Dans cet article, nous détaillons d'abord les résultats de la seconde évaluation du projet NESPOLE!. Nous présentons ensuite, en formulant quelques remarques, les nouvelles techniques objectives proposées en traduction de l'écrit. Nous commentons aussi les résultats d'une première expérience pilote en évaluation objective réalisée dans le cadre de C-STAR III. Nous concluons avec quelques idées et perspectives plus générales pour le domaine de la traduction de parole.

1 Évaluation du Démonstrateur en « tourisme étendu » (2002)

Les résultats de l'évaluation du premier démonstrateur en « tourisme restreint » ont été présentés en détail dans [Rossato S., et al., 2002].

1.1 Données et protocole

Deux dialogues extraits de la seconde collecte NESPOLE! [Mana N., et al., 2003] ont été utilisés. Ces dialogues couvrent des scénarios complexes non couverts par le premier démonstrateur, événements culturels, châteaux, lacs et forfaits. Pour l'italien, il s'agit de tours de paroles d'un agent de voyage, pour les trois autres langues, ce sont ceux d'un client.

Les signaux recueillis servent d'entrée aux modules de reconnaissance vocale. Les transcriptions manuelles de ces signaux servent de référence pour la traduction (elles simulent une reconnaissance sans erreur). Les tours de parole sont segmentés en unités sémantiques de dialogue (SDU[1]). Après avoir appliqué les modules de reconnaissance et/ou de traduction sur ces données, des évaluateurs humains jugent, pour chaque tour de parole transcrit, la qualité de la traduction de chaque SDU au sein du tour. L'évaluation est faite au niveau des SDU car elles représentent un élément atomique de la tâche. Ne pas faire la traduction correcte de l'une d'entre elles au sein d'un tour de parole ne signifie pas que tout le tour de parole soit mal traduit et que la tâche ne converge pas vers son objectif.

Les différentes classes d'évaluation réalisées sont les mêmes que celles de la première campagne soit : évaluation de la reconnaissance de la parole (Word Accuracy Rate et hypothèse comme paraphrase[2]), et évaluation des traductions monolingues et bilingues sur les références et sur les hypothèses du module de reconnaissances. Afin de mesurer les progrès accomplis, nous avons aussi utilisé, sur ces même données, les analyseurs et générateurs développés pour le premier démonstrateur pour les configurations monolingues et bilingues sur les transcriptions uniquement (*Sur Refs (01)* dans la Table 1).

Nous avons changé la procédure d'évaluation sur plusieurs points. Nous avons abandonné l'échelle à trois valeurs utilisée lors de la première évaluation. Les évaluateurs choisis pour cette évaluation sont des élèves de dernière année d'école de traduction (DESS pour le français). Lors de la première évaluation, les évaluateurs n'avaient pas de formation spécifique en traduction et les groupes n'étaient pas homogènes en terme de niveau en seconde langue.

La première échelle de notation comportait les valeurs BAD, OK et PERFECT. Les évaluateurs devaient d'abord vérifier que le sens était préservé (note BAD sinon). Puis, lorsque le sens était préservé, ils devaient dire si la traduction était grammaticale ou non (PERFECT vs. OK). Or, l'environnement d'évaluation présentait aux évaluateurs les trois options en même temps, il est possible qu'ils aient été poussés à choisir la note médiane (OK). La première question était aussi sévèrement interprétée (toute sorte de perte de sens rangeait la SDU dans la catégorie BAD).

Nous avons donc utilisé une échelle de quatre valeurs fondée uniquement sur la préservation du sens : VERY GOOD (toutes les informations sont présentes et faciles à comprendre), GOOD (toutes les informations importantes sont présentes), BAD (une ou plusieurs informations importantes ont été omises), VERY BAD (les informations importantes sont presque toutes absentes).

---

[1] Une SDU est un segment de tour de parole de longueur maximale qui peut être codé par une seule représentation dans le pivot IF que nous utilisons dans NESPOLE! Voici, un exemple de découpage en SDU, séparées par des # d'un tour de parole : « bonjour madame # j aimerais organiser une semaine de vacances dans un parc # et je voudrais aussi une chambre simple à cavalese du quinze au vingt septembre ».

[2] Le WAR ne prend pas en compte le fait que certaines erreurs de reconnaissance peuvent avoir des conséquences plus ou moins importantes sur la qualité de la traduction produite par le système. Ainsi, nous avons aussi vérifié si la sortie du module de reconnaissance peut être, ou non, considérée comme une paraphrase de la transcription manuelle du signal.

Finalement, pour pouvoir comparer les résultats de cette évaluation à la précédente, les notes VERY GOOD et GOOD ont été additionnées comme ACCEPTABLE. Les résultats complets de cette seconde évaluation sont donnés Table 1.

1.2 Résultats

Reconnaissance WAR	58	56	51	76
Hypos comme paraphase	60	67	62	76
Traduction monolingue (ACCEPTABLE**)**	*FRA-FRA*	*ENG-ENG*	*GER-GER*	*ITA-ITA*
Sur Refs (01) / sur Refs (02) / Hypos (02)	*69* / 77 / 58	*68* / 68 / 50	*45* / 61 / 51	*36* / 51 / 42
Traduction Bilingue (ACCEPTABLE**)**	*FRA-ITA*	*ENG-ITA*	*GER-ITA*	
Sur Refs (01) / sur Refs (02) / Hypos (02)	*72* / 78 / 58	*64* / 70 / 50	*44* / x / x	
	ITA-FRA	*ITA-ENG*	*ITA-GER*	
Sur Refs (01) / sur Refs (02) / Hypos (02)	*19* / 37 / 33	*33* / 33 / 30	*38* / 45 /38	

Table 1 : Résultats de la seconde campagne d'évaluation Nespole!

1.3 Commentaires

Pour le français, on atteint, en monolingue sur les hypothèses, un taux de 58% de traductions acceptables alors que l'on pouvait espérer 60%. Vers l'italien, les résultats sont cette fois meilleurs que depuis l'anglais. Enfin, depuis l'italien, le générateur vers le français affiche maintenant des performances comparables aux générateurs vers l'anglais et l'allemand.

Les résultats produits dans Verbmobil lors de l'évaluation de masse [Tessiore L. and Hahn W., 2000] pour un taux de reconnaissance inférieur à 75% sont de 66% en allemand-anglais et de 58% en anglais allemand. Les résultats obtenus avec notre second démonstrateur sont du même ordre.

1.4 Mesure des progrès accomplis

En traduction vers l'italien, les trois systèmes produisent un taux de plus de 30% de traductions acceptables. Ces taux ne semblent pas montrer de gros progrès par rapport à la première évaluation. Cependant, si on compare les taux d'acceptabilité sur les références italiennes avec les modules des premier et second démonstrateurs, on se rend compte que, pour le français, le générateur progresse de 18%, le générateur allemand de 7%, le générateur anglais restant stable.

En ce qui concerne la traduction monolingue ou bilingue du côté client, on observe que les taux de traduction sont supérieurs ou égaux à 50% sur les hypothèses et supérieurs à 70% sur les références. Ces taux sont en augmentation de six à dix points pour le français par rapport aux modules de la première évaluation.

Les tours de parole de l'agent de voyage (italien) sont devenus plus complexes dans le second démonstrateur. Avec un taux d'hypothèses paraphrasant l'entrée de 76%, le pourcentage de traductions acceptables en italien semble être nettement inférieur aux résultats monolingues des autres langues. Une étude plus fine des données permet d'expliquer ce résultat.

Les développeurs des systèmes ont classé les SDU des dialogues d'évaluation en trois catégories : SDU couvertes par le premier démonstrateur (classe 1), SDU couvertes par le second démonstrateur uniquement (classe 2), SDU hors du domaine (classe 3). Nous avons calculé les performances des systèmes sur ces trois groupes séparément. Pour le français, l'anglais et l'allemand (client), seulement 5% des SDU appartiennent aux classe 2 et 3. Cependant, pour l'italien (agent), 13% sont dans la seconde catégorie et 25% dans la troisième.

Les différences de performance apportent des réponses intéressantes. Sur le premier groupe, on observe une amélioration des performances de 56.6% pour le premier démonstrateur à 63.2% pour le second démonstrateur. Ce qui montre une amélioration du second démonstrateur dans la

couverture du domaine du premier démonstrateur. Sur les données du groupe 2, le premier démonstrateur atteint seulement 14.2% de traduction acceptable alors que le second démonstrateur atteint un score de 38.4%. Le premier démonstrateur n'étant pas préparé pour ce type de tour de parole cela n'est pas surprenant. Bien que le second démonstrateur ait des performances bien supérieures au premier, celui-ci n'atteint cependant pas un niveau comparable aux performances atteintes sur le domaine du premier démonstrateur.

L'analyseur de l'italien ne couvre pas, bien sûr, les SDU hors du domaine. Lorsqu'on les exclue, les performances du système italien passent de 49.3% pour le premier démonstrateur à 58.9 pour le second. Ces résultats sont alors comparables à ceux des autres systèmes monolingues. De même, sans les SDU du groupe 3, les performances des autres systèmes bilingues depuis l'italien sont accrues de 9 points.

2 Vers des évaluations communes sur un même corpus

Nous présentons maintenant les tendances actuelles, et leurs limites, en évaluation objective. Nous évoquons aussi la proposition du consortium C-STAR III pour des évaluations compétitives sur un même corpus.

2.1 Tendances actuelles

Le coût important de l'évaluation subjective a motivé le passage vers des protocoles d'évaluation objectifs (automatiques). Plusieurs métriques ont été proposées. Les plus utilisées sont BLEU [Papineni K., et al., 2002] et NIST [Doddington G., 2002] qui calculent des distances statistiques sur des ensembles de n-grammes entre la traduction produite par le système et des paraphrases d'une traduction de référence.

Le score BLEU est constitué de deux composants une précision modifiée sur les n-grammes (calculée pour chaque traduction) et une pénalité pour les traductions plus courte que les références (calculée sur tout le corpus). Avec NIST, un poids plus important est donné aux n-grammes les plus longs. De plus, le poids d'un n-gramme est calculé en fonction de sa valeur informationnelle. Plus un n-gramme est présent dans les sorties, plus son poids est faible.

La communauté pratique aussi des évaluations en WER (taux d'erreur mesuré sur un alignement entre une sortie et une référence), inspirée directement du domaine de la reconnaissance automatique de la parole, et en MWER (WER sur des références multiples), ou en PER (WER indépendant de la position des mots) ou en MPER (PER sur des références multiples). Le lecteur peut consulter [Sugaya F., et al., 2001] pour plus d'informations.

2.2 Limites

Les promoteurs et les utilisateurs de telles techniques n'oublient pas que, dans l'absolu, les chiffres produits par les méthodes automatiques ne veulent rien dire. De nombreuses études essaient d'établir une corrélation entre les résultats d'une évaluation objective et ceux d'une évaluation subjective [Coughlin D., 2003, Doddington G., 2002, Papineni K., et al., 2002]. Cependant, de notre point de vue, si cette corrélation est utile pour vérifier les progrès réalisés au cours du développement, elle ne répond pas à la question de l'utilisabilité.

En effet, nous savons qu'un taux de 100% de traductions acceptables ou un score de un à une évaluation BLEU ne seront pas atteints avant longtemps, si ce n'est jamais, dans des domaines assez largement couverts. Il serait donc important de savoir à partir de quelles performances un système devient utile et utilisable. On pourrait ainsi essayer de corréler les résultats d'évaluation objective avec l'utilisabilité. Les évaluations actuelles ne traitent pas de ce sujet.

Du point de vue de leur mise en œuvre pratique, les méthodes automatiques induisent aussi quelques questions essentielles. Ainsi, on ne trouve pas dans la littérature d'instructions pour la fabrication des références. On sait bien que la « typologie » des références doit être la plus

proche possible de la « typologie » des sorties du système à évaluer. En conséquence, la comparaison de plusieurs systèmes sur des mêmes données de test peut devenir délicate si la collection de paraphrases de chaque référence n'est pas assez exhaustive.

2.3 Évaluations sur un même corpus (C-STAR III)

Afin de préparer une campagne d'évaluation ouverte sur le corpus BTEC [Takezawa T., et al., 2002], une première expérience pilote interne a été conduite en 2003. Cette évaluation concerne uniquement la traduction textuelle. Les données de développement et de test des différents systèmes utilisaient BTEC et d'autres ressources monolingues. 500 phrases anglaises traduites dans toutes les langues sources ont été utilisées comme données de test.

Nous avons évalué cinq systèmes depuis le chinois, l'italien, et le japonais vers l'anglais en procédant, d'une part, à une évaluation subjective des résultats et d'autre part, à une évaluation objective en utilisant BLEU et NIST. L'évaluation subjective a été conduite selon les recommandations publiées par le Linguistic Data Consortium pour l'évaluation du projet TIDES de la DARPA[3]. Les scores BLEU et NIST ont été calculés sur les sorties brutes des systèmes.

À l'examen des résultats produits, nous avons rencontré une inconsistance entre les scores pour un système classé 3ième par BLEU et 5ième par NIST. Les sorties de ce système étaient significativement plus courtes que les références, ce qui influe fortement le calcul du score. À ce détail près, les classements des systèmes pour chacune des évaluations sont homogènes.

Nous avons aussi observé des différences importantes dans la forme des sorties des différents systèmes : usage de majuscules, rendu des numéraux (lettres, chiffres séparés ou non), abréviations, mots composés (avec ou sans tirets), ponctuation, etc. Nous avons montré que cela induit des différences de ±0,15 points pour BLEU et ±1,8 pour NIST, ce qui n'est pas négligeable. Pour les expériences futures, il faudra donc normaliser les sorties des différents systèmes comme cela est déjà le cas dans la communauté du traitement de la parole.

Afin de mobiliser la communauté du domaine et d'apporter quelques réponses aux questions que nous posons dans les sections 2.2 et 3, le consortium C-STAR III organise un atelier satellite à la conférences ICSLP-2004[4]. Dans ce cadre, le consortium diffusera une partie du corpus BTEC afin que des membres extérieurs puissent évaluer leurs systèmes sur nos données.

3 Pistes pour le futur

Pour aller plus loin, il nous semble bien sûr important et nécessaire de réfléchir à l'évaluation en proposant un cadre qui réponde non seulement aux besoins des développeurs, mais aussi aux besoins des utilisateurs. Il nous paraît aussi nécessaire de réfléchir de nouveau au contexte de nos travaux : nous faisons de la ***traduction de dialogue***. Si dans le cadre du projet Verbmobil ce contexte particulier a été exploité, il nous semble que dans les travaux postérieurs et actuels, les systèmes proposés n'en font plus usage. Chaque tour de parole est traduit pour lui même comme s'il n'était pas énoncé dans un contexte (celui du dialogue). Les architectures mises en œuvre sont exclusivement des architectures pipe-line dans lesquelles les différents composants s'échangent des informations minimales. Cette lacune va se renforcer avec l'utilisation des techniques statistiques qui exploitent des données (suites de caractères constituées de lemmes fléchis) alignées pour lesquelles le contexte se réduit forcément au tour de parole.

Nos propositions plus générales concernant la traduction de dialogue vont donc dans deux directions : l'intégration des composants en entrée et en sortie, et la gestion du dialogue.

---

[3] http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf

[4] http://www.slt.atr.co.jp/IWSLT2004/

L'intégration des composants peut être réalisée en entrée et en sortie. En entrée, il faut envisager de transmettre des informations plus riches entre les modules de reconnaissance et d'analyse. Nous proposons même une transmission bidirectionnelle. La reconnaissance peut fournir un treillis de mots (complété éventuellement d'informations sur la prosodie), ou bien une sortie déjà partiellement analysée en utilisant un modèle de langage sémantique [Vu Minh Q., et al., 2004]. Inversement, le module d'analyse peut fournir au module de reconnaissance le thème en cours dans le dialogue (pour utiliser des modèles de langage dynamique), et/ou un cache de mots déjà utilisés dans les tours de parole précédents des différents interlocuteurs (renforcement des mots du cache lors du décodage). En sortie, il s'agit pour l'analyseur de fournir des marques dans le texte produit afin d'obtenir une synthèse plus naturelle.

La gestion du dialogue peut prendre en compte plusieurs contextes [Boitet C., et al., 2000] : le contexte global (type de dialogue, caractéristiques, rôle, localisation des participants), le contexte dialogique (représentation du passé et du présent, prédictions sur le futur), et le contexte linguistique (antécédents possibles d'anaphores et d'ellipses, sélections lexicales).

Conclusion

Nous avons décrit en détail la seconde expérience en évaluation subjective du projet NESPOLE! et montré comment nous avons évalué nos progrès. Nous avons introduit, et critiqué, les pistes actuellement suivies en évaluation dans le domaine. Nous avons enfin évoqué les questions en suspens à propos de l'évaluation objective et fait des propositions afin d'améliorer les systèmes.

Références

Boitet C., Blanchon H. & Guilbaud J.-P. (2000). *A way to integrate context processing in the MT component of spoken, task-oriented translation systems*. *Proc*. MSC-2000. Kyoto, Japan, October 11-13, 2000. vol. 1/1: pp. 83-87.

Coughlin D. (2003). *Correlating Automated and Human Assessments of Machine Translation Quality*. *Proc*. MT Summit IX. September 23-27, 2003: 8 p.

Doddington G. (2002). *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. *Proc*. HLT 2002. San Diego, California, March 24-27, 2002. vol. 1/1: pp. 128-132 (note book proceedings).

Lazzari G. (2000). *Spoken Translation: Challenges and Opportunities*. *Proc*. ICSLP 2000. Beijing, China, Oct. 16-20, 2000. vol. 4/4: pp. 430-435.

Mana N., Burger S., Cattoni R., Besacier L., Maclaren V., Mc Donough J. & Metze F. (2003). *The Nespole! VoIP Corpora in Tourism and Medical Domains*. *Proc*. EUROSPEECH 2003. Geneva, Switzerland, Spetember 1-4, 2003: 4 p.

Papineni K., Roukos S., Ward T. & Zhu V. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. *Proc*. ACL-02. Philadelphia, USA, July 7-12, 2002. vol. 1/1: pp. 311-318.

Rossato S., Blanchon H. & Besacier L. (2002). *Speech-to-Speech Translation System Evaluation: Results for French for the NESPOLE! Project First Showcase*. *Proc*. ICSLP. Denver, USA, 16-20 September, 2002: 4p.

Sugaya F., Yasuda K., Takezawa T. & Yamamoto S. (2001). *Precise Measurement Method of a Speech Translation System's Capabilities with a Paired Comparison Method between the System and Humans*. *Proc*. MT Summit VIII. Santiago de Compostela, Spain, 18-22 September, 2001. vol. 1/1: pp. 345-350.

Takezawa T., Sumita E., Sugaya F., Yamamoto H. & Yamamoto S. (2002). *Towards a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World*. *Proc*. LREC-2002. Las Palmas, Spain, May 29-31, 2002. vol. 1/3: pp. 147-152.

Tessiore L. & Hahn W. (2000). *Functional Evaluation of a Machine Interpretation System: Verbmobil*. *in* Verbmobil: Foundation of Speech-to-Speech Translation. Springer-Verlag. Berlin. pp. 611-631.

Vu Minh Q., Besacier L., Blanchon H. & Bigi B. (2004). *Modèle de langage sémantique pour la reconnaissance automatique de parole dans un contexte de traduction*. *Proc*. TALN 2004. Fès, Maroc, 19-21 avril 2004: dans ce volume 6p.