Évaluation de l'adaptation par renforcement d'un générateur en langage naturel neuronal pour le dialogue homme-machine

Matthieu Riou, Bassam Jabaian, Stéphane Huet, Fabrice Lefèvre

Évaluation de l'adaptation par renforcement d'un générateur en langage naturel neuronal pour le dialogue homme-machine

Matthieu Riou Bassam Jabaian Stéphane Huet Fabrice Lefèvre
CERI-LIA, Université d'Avignon, Avignon, France
matthieu.riou@alumni.univ-avignon.fr, {bassam.jabaian, stephane.huet, fabrice.lefevre}@univ-avignon.fr

RÉSUMÉ

Jusqu'à récemment, la génération en langage naturel dans les systèmes de dialogue utilisait des systèmes à base de règles et de patrons, mais de nouveaux modèles à base de réseaux de neurones récurrents ont été proposés (Wen *et al.*, 2016a). Cependant ces modèles nécessitent une grande quantité de données d'apprentissage qui peuvent être compliquées à collecter et à annoter. Pour répondre à cette problématique, nous avons proposé un protocole d'apprentissage en ligne utilisant un algorithme de bandit contre un adversaire, permettant d'améliorer l'utilisation d'un modèle initial appris sur un corpus plus restreint généré par patrons (Riou *et al.*, 2017). Dans cet article, nous étudions l'intérêt pratique de notre approche en utilisant des données réelles obtenues par reconnaissance automatique de la parole des propositions des utilisateurs et en faisant évaluer les sorties du système par des humains.

ABSTRACT

Evaluation of the reinforcement adaptation of a neural natural language generation system for human-machine dialogue.

Traditional systems for natural language generation in spoken dialogue systems use patterns and rules to generate system answers. Recently, systems based on recurrent neural network models have been proposed (Wen *et al.*, 2016a). Those systems require a large amount of data to be learned, which can be difficult to collect and annotate. Therefore we proposed a framework to adapt the NLG module online through direct interactions with the users (Riou *et al.*, 2017). In this paper, we study the practical interest of the approach with real data collected as automatic speech recognition of users' suggestions and having humans assessing the system's outputs.

MOTS-CLÉS : génération en langage naturel, réseau de neurones récurrent, bandit contre un adversaire, apprentissage en ligne, adaptation à un utilisateur, reconnaissance automatique de la parole.

KEYWORDS: natural language generation, recurrent neural network, adversarial bandit, online learning, user adaptation, automatic speech recognition.

1 Introduction

Le composant de génération en langage naturel (NLG pour *Natural Language Generation*) d'un système de dialogue oral a pour rôle de transformer la réponse du gestionnaire de dialogue (qui est sous forme d'actes de dialogue) en une forme textuelle exploitable par le module de synthèse

vocale. Par exemple, l'acte $inform(type = restaurant, count = 4, food = pizza)$ peut générer les phrases « *There are 4 restaurants serving pizzas* » ou encore « *To eat pizza I can propose you 4 restaurants* ».

Les approches à base de patrons rédigés par des experts, traditionnellement utilisées pour cette tâche (Rambow *et al.*, 2001), produisent des phrases de bonne qualité pour des tâches bien spécifiées mais ces phrases restent assez répétitives et souvent peu naturelles. Les modèles statistiques permettent de palier ces limites avec des capacités de généralisation plus grandes, permettant l'introduction d'une plus grande variabilité lors de la génération mais aussi un plus grand risque d'erreurs de syntaxe. Les premiers modèles de ce type utilisaient des modèles de langue n-grammes de mots (Oh & Rudnicky, 2002) et ont été étendus par la suite par l'utilisation de facteurs intégrant les étiquettes sémantiques dans le calcul des probabilités de séquences produites (Mairesse & Young, 2014). Plus récemment, avec l'expansion du *deep learning* dans différentes tâches de traitement du langage naturel, des modèles à base de réseaux de neurones récurrents ont été proposés pour le module de génération (Wen *et al.*, 2015a; Riou *et al.*, 2017). En parallèle, des nouvelles approches ont été étudiées pour apprendre des systèmes de bout en bout (Serban *et al.*, 2016). Ces études sont toutefois effectuées dans un contexte différent du nôtre car elles considèrent des systèmes conversationnels non dirigés par le but et pour lesquels de grandes ressources de données sont disponibles.

Afin de réduire le coût lié à une annotation d'un nouveau corpus de génération pour un nouveau domaine, Wen *et al.* ont proposé une approche incrémentale pour gérer l'adaptation de domaine d'un modèle de génération à base de RNN (Wen *et al.*, 2016b). Ils recourent à des données contrefaites synthétisées à partir d'un ensemble hors-domaine pour ajuster leur modèle sur un ensemble réduit de phrases du domaine. Il est aussi envisageable d'employer des méthodes d'extension de corpus (Manishina *et al.*, 2016) mais elles ne permettent pas une adaptation simultanée aux préférences de l'utilisateur.

Dans un travail précédent (Riou *et al.*, 2017), nous avons proposé de réduire le coût de la production de nouvelles données, en adaptant un modèle initial afin de générer des phrases avec une plus grande diversité. Dans cette logique, une approche par renforcement basée sur un algorithme de bandit contre un adversaire a été appliquée (Auer *et al.*, 2002) pour adapter un modèle RNN aux nouvelles phrases, différentes de l'ensemble d'apprentissage, en prenant en compte le coût imposé à l'utilisateur pour les fournir au système. Cette proposition a été évaluée sur des données simulant les erreurs de transcription automatique des retours des utilisateurs.

Dans cet article, nous étendons cette études en évaluant notre proposition sur des données réelles extraites d'un système de reconnaissance de la parole et en complétant les comparaisons préliminaires effectuées à l'aide de métriques automatiques par des mesures faites par des annotateurs humains sur trois critères : l'informativité, le naturel et la grammaticalité.

2 Le système de génération

Pour modèle de génération, nous avons proposé d'utiliser le LSTM (*Long short-term memory*) à contexte combiné décrit plus en détail dans (Riou *et al.*, 2017). Ce modèle combine deux modèles neuronaux proposés par Wen *et al.*, le modèle LSTM conditionné sémantiquement (SCLSTM) et l'encodeur-décodeur RNN avec attention. Chacun de ces systèmes traite différemment l'information sémantique représentée par un acte de dialogue (AD) pour produire la phrase. La cellule de contrôle (*reading gate*) du SCLSTM permet de filtrer l'AD en ne conservant à chaque étape que l'information restante, qui n'a pas été encore traitée. Au contraire, le mécanisme d'attention permet de sélectionner

dans l'AD au complet l'information à considérer spécifiquement à l'étape suivante, mais ne modélise pas explicitement la progression de l'information déjà traitée. L'objectif de notre système est de combiner les avantages des deux systèmes, en utilisant une cellule de contrôle similaire au SCLSTM et un mécanisme d'attention pour traiter séquentiellement l'AD restant à générer.

3 Un modèle pour l'adaptation en ligne

Afin de réduire le coût d'annotations complémentaires pour rajouter plus de variabilité dans le système, nous avons proposé un protocole d'apprentissage en ligne en deux étapes : tout d'abord, un modèle est appris sur un corpus constitué de références générées par patrons, puis le modèle est utilisé pour générer des phrases, en interaction vocale avec l'utilisateur, auquel il peut demander de produire un énoncé meilleur ou différent (Riou *et al.*, 2017).

On rappelle brièvement qu'à chaque tour de parole et après la génération de l'énoncé par le système, ce dernier doit décider de la meilleure action à suivre (à partir d'une distribution de probabilité et en tenant compte du coût $\phi(\text{Action})$ estimé) parmi :
• **Skip** : n'appliquer aucune mise à jour au modèle. Le coût de cette action est nul ($\phi(\text{Skip}) = 0$).
• **AskDictation** : affiner le modèle en considérant un énoncé alternatif proposé par l'utilisateur et transcrit automatiquement par un système de reconnaissance ($\phi(\text{AskDictation}) = 1$).
• **AskTranscription** : demander à l'utilisateur de transcrire la correction ou l'énoncé alternatif ($\phi(\text{AskTranscription}) = 1 + l$, avec l la taille de l'énoncé proposé).

Nous avons estimé le **gain** de chaque action $g(i) \in [0, 1]$ comme décrit dans (Riou *et al.*, 2017) et nous avons défini une fonction de perte $l(i) \in [0, 1]$ qui permet de maximiser ce gain $g(i)$ tout en minimisant l'effort de l'utilisateur $\phi(i)$:

$$l(i) = \underbrace{\alpha(1 - g(i))}_{\text{amélioration du système}} + \underbrace{(1 - \alpha)\frac{\phi(i)}{\phi_{max}}}_{\text{effort de l'utilisateur}} \quad (1)$$

avec α un scalaire qui pondère le gain par rapport au coût, permettant au système de s'adapter aux préférences de l'utilisateur et ϕ_{max} correspond à l'effort maximal pour normaliser l'effort de l'utilisateur entre 0 et 1.

Afin de réduire l'effort demandé à l'utilisateur et éviter des actions inutiles, un algorithme de bandit contre un adversaire à été adopté. A chaque itération, le système produit une phrase puis choisit une action $i_t \in \mathcal{I}$. Une fois que l'action i_t est effectuée, le système calcule l'estimation du gain $g(i_t)$, l'effort de l'utilisateur $\phi(i_t)$ et la perte $l(i_t)$. Le rôle du bandit est donc de trouver $i_1, i_2, \ldots$, afin que pour chaque t, le système minimise la perte $l(i_t)$.

4 Expériences

4.1 Cadre expérimental

Les expériences ont été menées sur le corpus *SF restaurant* décrit dans (Wen *et al.*, 2015b) et librement accessible en ligne [1]. Ces données contiennent 5 191 phrases, pour 271 AD distincts. Le

1. https://www.repository.cam.ac.uk/handle/1810/251304

corpus associe à chaque acte une phrase générée par patron et plusieurs phrases proposées par des annotateurs humains, chaque phrase étant délexicalisée [2].

Le LSTM à contexte combiné a été implémenté en utilisant la bibliothèque Tensorflow [3]. Ce système a ensuite été entraîné sur le corpus séparé en 3 parties suivant un ratio 3 :1 :1 : apprentissage, validation et test, en utilisant uniquement les références proposées par les annotateurs humains.

4.2 Comparaison des systèmes de génération

Une évaluation a été conduite avec deux métriques objectives calculées à partir des générations de référence, le score BLEU-4 (Papineni *et al.*, 2002) et le taux d'erreur en concepts SER. Le BLEU-4 valide la génération de phrases, notamment la grammaticalité, tandis que le SER se concentre spécifiquement sur le contenu sémantique. Pour chaque exemple, nous produisons 20 hypothèses et ne gardons pour l'évaluation que les 5 meilleures selon le score donné par le modèle NLG. Des références multiples pour chaque AD sont obtenues en groupant les phrases délexicalisées du même AD et en les relexicalisant ensuite.

Le LSTM à contexte combiné obtient un score BLEU-4 de 71,1%, voisin des deux autres systèmes initiaux (72,2% pour le SCLTSM et 69,7% pour l'encodeur-décodeur), mais le taux d'erreur SER est divisé par trois par rapport aux autres systèmes, en passant de 0,77% et 0,65% pour le SCLSTM et l'encodeur-décodeur, à 0,24% pour le LSTM à contexte combiné. Cela veut dire qu'il propose une meilleure couverture des concepts à exprimer et donc moins d'omissions ou d'erreurs de concepts, ce qui est le principal but recherché pour un module NLG.

4.3 Évaluation de l'apprentissage en ligne

Nous avons utilisé le même corpus, mais avec un découpage en apprentissage, validation et test suivant un ratio 2 :1 :1. Le modèle NLG utilisé est encore le LSTM à contexte combiné. Pour initialiser le modèle, nous l'entraînons en utilisant les références générées par patron du corpus d'apprentissage. Le corpus de validation a permis de décider l'arrêt de la phase d'apprentissage (*early stopping*). Ensuite, nous avons simulé un apprentissage en ligne en réutilisant le corpus d'apprentissage, mais cette fois en apprenant sur les références proposées par des annotateurs humains. Le modèle ainsi que la politique de bandit ont été mis à jour toutes les 400 phrases. Dans cette série d'expériences, le WER a été simulé en insérant de manière aléatoire des erreurs (substitution, insertion et suppression) dans les exemples du corpus, jusqu'à atteindre un taux global de WER prédéfini.

Le modèle initial, entraîné sur les références générées par patron, atteint un haut score BLEU-4 de 80,2% quand celui-ci est calculé à partir de références générées par patrons, mais qui est réduit à seulement 39,7% en refaisant les calculs à partir des seules références proposées par des annotateurs humains. Cela montre la grande diversité possible des réponses dans une conversation.

4.4 Évaluation humaine

Les évaluations basées sur les métriques automatiques, tel BLEU-4, ne reflètent pas nécessairement les vraies préférences utilisateurs (Callison-Burch *et al.*, 2006). En particulier la dimension naturelle est très complexe à formaliser. Dans le but de mieux comparer les modèles initiaux et adaptés, nous

2. La délexicalisation remplace les formes de surface des concepts par des variables, *inform(name=la mimosa, food=mediterranean)* devient « *$name sert des plats $food* ».

3. `https://www.tensorflow.org`

	Système initial	Système adapté
Score global	2.356	**2.425**
Informativité	**2.528**	2.509
Grammaticalité	2.272	**2.383**
Naturel	2.267	**2.383**

TABLE 1 – Moyenne des scores pour chaque système

recourons à une évaluation humaine. 5 annotateurs ont reçu pour consigne d'évaluer les phrases générées automatiquement. Pour chaque exemple, l'évaluateur était confronté avec l'acte de dialogue visé et les 3 meilleures propositions de chaque système. Les hypothèses à juger sont ordonnées aléatoirement, les phrases équivalentes regroupées et aucune indication du système d'origine n'est disponible. Nous avons demandé aux annotateurs de donner à chacune des phrases (6 au maximum) trois scores :

- **Informativité** évalue si l'ensemble des informations présentes dans l'acte de dialogue sont bien toutes transmises dans la phrase générée, et si aucune supplémentaire n'est introduite, sur une échelle de 1 à 3 :
  - 3 : toutes les informations données par l'acte de dialogue (et seulement ces informations) sont présentes.
  - 2 : une information mineure est manquante, où une extra information non contradictoire est présente.
  - 1 : dans les autre cas.
- **Grammaticalité** évalue le niveau de correction syntaxique de la phrase, sur une échelle de 1 à 3 :
  - 3 : la phrase est correcte.
  - 2 : il y a quelques imperfections, mais qui probablement ne sont pas audibles.
  - 1 : il y a des erreurs importantes dans la phrase.
- **Naturel** évalue à quel point la phrase est proche d'une production potentielle humaine, sur une échelle de 1 à 3 :
  - 3 : la phrase aurait pu être prononcée par un humain dans cette situation.
  - 2 : la phrase est correcte mais moins appropriée à la situation, ou semble « automatique ».
  - 1 : même en corrigeant les erreurs grammaticales le cas échéant, la phrase n'aurait jamais pu être prononcée par un humain.

En supplément à l'évaluation de chaque propositions, les annotateurs ont aussi indiqué la phrase qu'ils jugeaient la meilleure de façon globale. Afin de mesure le niveau d'accord entre annotateur avec la métrique Kappa de Fleiss, les 20 premiers exemples étaient communs à tous. Au total, 471 annotations ont été réalisées

L'accord moyen global entre annotateurs présente un κ de 0,55. La tâche qui présente le moins grand agrément est le jugement sur le naturel (κ=0.468), à comparer avec des κ de 0.59 et 0.58 respectivement pour l'informativité et la grammaticalité.

Comme on peut le constater dans le tableau 1 le système adapté obtient un score global moyen plus élevé que le système initial. Plus particulièrement, ses scores sont meilleurs pour le naturel et la grammaticalité mais un peu dégradés pour l'informativité. Dans le tableau suivant 2, on peut observer que les deux systèmes ont tendance à avoir des scores globaux plus élevés pour des actes de dialogues de longueurs moyennes (2 à 3 slots). Au delà le score décroit rapidement du fait d'une plus grande

# slots	Système	Tous	Informativité	Naturel	Grammaticalité
0	Initial	1,74	1,76	1,72	1,74
	Adapté	2,59	2,60	2,60	2,58
1	Initial	2,38	2,59	2,34	2,21
	Adapté	2,38	2,55	2,31	2,27
2	Initial	2,58	2,80	2,50	2,46
	Adapté	2,64	2,75	2,58	2,58
3	Initial	2,47	2,72	2,30	2,39
	Adapté	2,32	2,48	2,26	2,29
4	Initial	2,28	2,35	2,24	2,27
	Adapté	1,71	1,74	1,69	1,70
5	Initial	1,75	2,00	1,67	1,58
	Adapté	1,66	1,50	1,75	1,58

TABLE 2 – Moyennes des scores pour chaque système en fonction du nombre de slots dans l'acte de dialogue

complexité, plus favorable à l'introduction d'erreurs dans le cas de la génération stochastique.

Enfin, le tableau 3 nous permet de constater les variations de scores en fonction des types d'actes de dialogue. On observe, contrairement à notre intuition, que les scores sont assez réguliers selon les types, et ce alors qu'ils représentent bien sûr des complexités très variables (mais qui doivent aussi être liées au nombre moyen de concepts associés en moyenne à chacun des actes, il peut par exemple être assez grand pour l'acte très générique inform, alors qu'il est nul pour goodbye).

Quand les annotateurs devaient voter pour leur phrase favorite, ils ont en majorité voté pour la meilleure proposition de chacun des systèmes (on rappelle que les phrases ne sont pas présentées de façon ordonnée), avec une préférence pour le système adapté (voir le tableau 4). Mais surtout on constate que les phrases proposées en positions 2 et 3 sont aussi beaucoup sélectionnées par les annotateurs dans le cas du système adapté, ce qui participe à confirmer que le système adapté peut générer des phrases satisfaisantes avec une plus grande variabilité que le système initial.

4.5 Évaluation de l'adaptation en ligne avec des vraies données orales

Pour évaluer en pratique le schéma proposé d'adaptation en ligne, et en particulier l'impact du taux d'erreur en mot de la reconnaissance de parole durant les interactions, une collecte de corpus a été réalisée.

Pour chaque phrase nous avons confronté un utilisateur avec son acte de dialogue (et la possibilité de lire quelques exemples de référence de génération de cette phrase, système initial). Puis l'utilisateur devait dicter une alternative correspondant à l'acte de dialogue, qui était automatiquement transcrite. Afin de simplifier le déploiement de l'expérience, les capacités de reconnaissance de la parole offerte par le navigateur Chrome, utilisant l'API RAP de Google, ont été utilisées. Enfin à partir de la sortie automatique l'utilisateur avait la possibilité d'apporter manuellement les corrections nécessaires pour fournir une transcription de référence. Les deux sorties, automatiques et références, ont été collectées, ainsi que le score de confiance des transcriptions automatiques.

Acte de dialogue	Système	Tous	Informativité	Naturel	Grammaticalité
inform	Initial	2.50	2,39	2,39	2,42
	Adapté	2,42	2.52	2,37	2,39
inform_only_match	Initial	2,33	2,50	2,50	2,00
	Adapté	1,94	2,00	2,00	1.83
?inform_no_match	Initial	2,48	2,78	2,33	2,34
	Adapté	2,19	2,23	2,07	2,05
?select	Initial	2,16	2,52	2,01	1,89
	Adapté	2.05	2,52	2,37	2,33
?request	Initial	2,65	2,82	2,65	2.47
	Adapté	2,63	2,77	2,57	2,55
?reqmore	Initial	2,27	2,67	2,07	2,07
	Adapté	2,62	2.47	2,73	2,67
?confirm	Initial	2,02	2,27	1.91	1,88
	Adapté	2,05	2,33	1,94	1,88
goodbye	Initial	1,71	1,70	1,70	1,72
	Adapté	2,59	2,60	2,59	2,57

TABLE 3 – Moyennes des scores pour chaque système selon le type des actes de dialogue

Rang	Système initial	Système adapté
1	111 (22,0%)	143 (28,5%)
2	51 (10,2%)	103 (20,6%)
3	18 (3,6%)	75 (15,0%)
Total	180 (35,9%)	321 (64,1%)

TABLE 4 – Effectif de phrases sélectionnées par les annotateurs selon leur rang.

426 paires de transcriptions (automatiques, manuelles) ont pu être récupérées de cette manière. [4] Le taux d'erreur en mot moyen est de seulement 2,42%, avec un score de confiance moyen de 0,86.

Un nouveau modèle a été appris en suivant le protocole d'adaptation en ligne décrit dans la partie 4.3 (avec $\alpha = 0,5$) en utilisant les données nouvellement collectée, au lieu des références annotées par des humains. Le nouveau corpus est divisé en 300 phrases pour l'apprentissage et 126 pour le test. Nous avons gardé le modèle initial utilisé dans la première expérience, mais il a été cette fois mis à jour, ainsi que le bandit, toutes les 50 phrases du fait de la taille plus réduite du corpus. Pour améliorer l'estimation du gain, l'estimation du WER global a été remplacée par le score de confiance de la transcription :

$$g(\text{AskDictation}) = (1 - \text{BLEU}_{gen/prop}) \times \text{score de confiance} \times (1 - \text{SER})$$

Nous avons testé ce nouveau modèle sur le même corpus que dans la première expérience, en comparant aux références générées par patrons dans un premier temps, et dans un second temps aux références proposées par des annotateurs humains auxquelles nous avons rajouté nos propres références corrigées issues de la collecte de données orales. Les résultats montrent des tendances similaires aux résultats obtenus avec le WER simulé. Le score BLEU-4 par rapport aux références

4. Toutes les données utilisées dans cette étude sont disponibles sur demande.

générées par patrons diminue fortement (10%, de 82,9% à 72,7%), tandis qu'il chute légèrement lorsqu'on le compare aux références humaines (3%, de 48,17% à 45,08%), ce qui s'explique par la forte présence dans le test des références proposées par des annotateurs humains du corpus initial qui n'ont pas été apprises dans cette expérience.

L'algorithme de bandit permet de ne pas demander constamment des efforts importants à l'utilisateur. Sur la totalité de l'apprentissage il a demandé 53% du temps une transcription, contre 23% pour une alternative à l'oral et 23% aucune alternative. Cela lui permet de diviser presque par deux le coût cumulé sur l'apprentissage sans pour autant trop diminuer les performances par rapport aux références annotées par des humains, soit un coût cumulé de 2430 et un BLEU-4 final de 44,4% contre respectivement 4243 et 45,8% dans le cas où le système demanderait systématiquement des transcriptions. En revanche il fait moins bien que le système qui ne demanderait que des alternatives orales, qui obtiendrait un BLEU-4 équivalent de 45,1% pour un coût cumulé de 300.

Pour permettre au système de mieux évaluer s'il doit ou non risquer de demander une alternative orale, il faudrait pouvoir élargir le contexte de l'exemple traité pris en compte par le bandit (nature de l'acte de dialogue, complexité...), par exemple en gardant le même protocole mais en utilisant un algorithme de bandit contextuel (Auer *et al.*, 2002).

5 Conclusions et perspectives

Dans cet article, nous avons évalué un nouveau protocole pour adapter un modèle initial de génération de langage naturel neuronal à l'aide d'un apprentissage en ligne. Les résultats obtenus par une expérience simulée ont ainsi pu être confirmés et complétés avec des utilisateurs réels, pour fournir des propositions au système et juger des qualités des hypothèses du système adapté. L'algorithme de bandit permet d'équilibrer de manière automatique l'évolution des performances du système avec le coût induit pour l'utilisateur et d'aboutir à un système que les utilisateurs jugent plus varié. Une voie d'amélioration possible du système que nous entrevoyons est l'augmentation des capacités d'apprentissage du bandit par la prise en compte du contexte lors de ses décisions, à l'aide d'un bandit contextuel. Enfin, le générateur de texte doit pouvoir être évalué dans le contexte du système de dialogue complet pour confirmer l'intérêt pratique de l'approche.

Remerciements

Ce travail a été partiellement financé par le Labex BLRI (ANR-11-LABX-0036) et l'ILCB.

Références

AUER P., CESA-BIANCHI N., FREUND Y. & SCHAPIRE R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, **32**(1), 48–77.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the role of bleu in machine translation research. In *EACL*, p. 249–256.

MAIRESSE F. & YOUNG S. (2014). Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, **40**(4), 763–799.

MANISHINA E., JABAIAN B., HUET S. & LEFÈVRE F. (2016). Automatic corpus extension for data-driven natural language generation. In *LREC*.

OH A. H. & RUDNICKY A. I. (2002). Stochastic natural language generation for spoken dialog systems. *Computer Speech & Language*, **16**(3–4), 387–407.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *ACL*.

RAMBOW O., BANGALORE S. & WALKER M. (2001). Natural language generation in dialog systems. In *HLT*.

RIOU M., JABAIAN B., HUET S. & LEFÈVRE F. (2017). Online adaptation of an attention-based neural network for natural language generation. In *INTERSPEECH*.

SERBAN I. V., SORDONI A., BENGIO Y., COURVILLE A. & PINEAU J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., BARAHONA L. M. R., SU P.-H., ULTES S., VANDYKE D. & YOUNG S. (2016a). Conditional generation and snapshot learning in neural dialogue systems. In *EMNLP*.

WEN T.-H., GAŠIĆ M., KIM D., MRKŠIĆ N., SU P.-H., VANDYKE D. & YOUNG S. (2015a). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *SIGDIAL*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., ROJAS-BARAHONA L. M., SU P.-H., VANDYKE D. & YOUNG S. (2016b). Multi-domain neural network language generation for spoken dialogue systems. In *NAACL-HLT*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., SU P.-H., VANDYKE D. & YOUNG S. (2015b). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *EMNLP*.