Évaluation automatique de la satisfaction client à partir de conversations de type "chat" par réseaux de neurones récurrents avec mécanisme d'attention

Jeremy Auguste[1] Delphine Charlet[2] Géraldine Damnati[2] Benoit Favre[1]
Frédéric Béchet[1]
(1) Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
(2) Orange Labs, Lannion
(1) {jeremy.auguste,benoit.favre,frederic.bechet}@lis-lab.fr
(2) {delphine.charlet,geraldine.damnati}@orange.fr

RÉSUMÉ

Cet article présente des méthodes permettant l'évaluation de la satisfaction client à partir de très vastes corpus de conversation de type "chat" entre des clients et des opérateurs. Extraire des connaissances dans ce contexte demeure un défi pour les méthodes de traitement automatique des langues de par la dimension interactive et les propriétés de ce nouveau type de langage à l'intersection du langage écrit et parlé. Nous présentons une étude utilisant des réponses à des sondages utilisateurs comme supervision faible permettant de prédire la satisfaction des usagers d'un service en ligne d'assistance technique et commerciale.

ABSTRACT

Customer satisfaction prediction with attention-based RNNs from a chat contact center corpus

This paper presents methods to perform knowledge extraction from very large databases of WEB chat conversations between operators and clients in customer contact centers. Extracting knowledge from chat corpus is a challenging research issue. Simply applying traditional text mining tools is clearly sub-optimal as it takes into account neither the interaction dimension nor the particular nature of this language which shares properties of both spoken and written language. We present a method predicting users satisfaction in a chat-based service trained on answers from users to satisfaction surveys.

MOTS-CLÉS : Réseaux de neurones récurrents, Attention, Satisfaction client, Conversations.

KEYWORDS: Recurrent Neural Networks, Attention-based RNNs, chat, satisfaction prediction.

1 Introduction

L'analyse automatique d'enregistrements de conversations, écrites ou orales, représente un défi pour les méthodes de Traitement Automatique de la Langue à cause d'une part de la nature spontanée du langage employé, et d'autre part de la construction interactive du discours au fur et à mesure de l'échange entre les participants. En dehors des tâches d'extraction d'informations telles que la classification de conversations (Koço *et al.*, 2012) ou le résumé automatique (Trione *et al.*, 2016),

l'une des tâches les plus étudiées dans le cadre des conversations *avec but* est celle de l'évaluation de la *satisfaction* des intervenants, en particulier dans le contexte applicatif des centres de contact clientèle des entreprises ou des administrations.

En effet les services techniques, commerciaux ou juridiques de la majorité des grands groupes peuvent être contactés par l'intermédiaire de centres d'appels ou grâce à des systèmes de conversation textuels hébergés sur leurs sites web. Ce dernier type de conversations, dites conversations *médiées*, sont faciles à collecter et une étude poussée de leur contenu permet de contrôler et d'améliorer la qualité des services rendus. Le champ de l'*analytics* s'intéresse à ces aspects avec par exemple l'étude manuelle par des experts d'échantillons de conversations, l'extraction de statistiques à partir de l'analyse automatique du contenu. Cependant une grande partie des informations sont le plus souvent issues de l'analyse des sondages de satisfaction, sous forme de questionnaires soumis aux usagers suite aux conversations. La réponse aux questions de type "*Votre problème a-t-il été résolu ?*" ou "*Recommanderiez-vous le service à vos proches ?*" sont des indicateurs de performance importants pour les services concernés. Une des limites de ce type d'étude est dû au fait que de nombreux usagers ne prennent pas la peine de répondre aux questionnaires de satisfaction. Se limiter aux seuls sondages renseignés donne une vue tronquée de la réalité d'un service.

Nous nous intéressons dans cette étude à la problématique de la prédiction automatique des réponses aux enquêtes de satisfaction à partir des seules transcriptions de conversations. Nous décrivons un système de prédiction automatique d'indicateurs de qualité entraîné sur des corpus de questionnaires post-conversations. Cette tâche ouvre plusieurs perspectives applicatives comme la prédiction de la satisfaction pour les clients ne répondant pas aux enquêtes. Elle constitue également un travail préalable à la mise en œuvre de monitoring en temps réel de la satisfaction client.

2 Travaux connexes

L'évaluation de la satisfaction client au sein de centres d'appels a donné lieu à de nombreuses études, la plupart sur des conversations orales téléphoniques. Les critères utilisés pour mesurer cette satisfaction peuvent être multiples, de critères objectifs tels que la réalisation effective de la tâche ayant motivé l'appel ou le temps d'attente, jusqu'aux critères subjectifs relatifs à la perception de l'efficacité ou la capacité d'écoute du téléconseiller ayant géré l'appel, ou encore la volonté de l'utilisateur de recommander ou pas le service qu'il vient d'utiliser, ce dernier critère étant l'un des plus importants pour les entreprises concernées (Reichheld, 2003).

Les systèmes qui ont été développés pour prédire automatiquement ces critères peuvent utiliser deux types de supervisions pour entraîner leurs modèles : une supervision directe sous la forme de sondages demandant aux utilisateurs de répondre à une enquête de satisfaction immédiatement après une conversation ; une supervision indirecte en demandant à des experts d'évaluer la satisfaction des appelants perçue à partir des transcriptions. C'est généralement la supervision indirecte qui est utilisée en raison des difficultés à obtenir les enquêtes de satisfaction des clients. Par exemple le système QA^{RT} décrit dans (Roy *et al.*, 2016) permet de prédire directement la qualité d'une conversation, au fur et à mesure de son déroulement, à l'aide d'indices et de classifieurs entraînés sur des avis d'experts. Cette prédiction de satisfaction à base d'avis d'experts a aussi été utilisée pour évaluer le ressenti d'utilisateurs d'interfaces de dialogue homme-machine, comme récemment dans (Stoyanchev *et al.*, 2017) où des classifieurs à base de Support Vector Machine (SVM) sont utilisés, ou encore dans (Pragst *et al.*, 2017) où des approches à base de réseaux de neurones récurrents (RNN) permettent de

prendre en compte la séquentialité des tours de parole dans une conversation.

L'utilisation d'une supervision indirecte, sous forme d'avis d'experts, pose problème pour évaluer des critères aussi subjectifs que la perception d'une qualité d'écoute, ou la volonté de recommander un service. Ce point est discuté dans (Ultes *et al.*, 2013) où il est montré qu'il existait une bonne corrélation entre les enquêtes d'opinion et ces avis d'experts. Cependant l'une des principales limitations des études précédentes est la faible taille des corpus utilisés. En effet, même pour les avis d'experts, obtenir des annotations sur de grands corpus reste une tâche difficile et coûteuse.

L'une des principales originalités de notre étude est d'avoir pu utiliser de très grands corpus de dialogue avec une *supervision directe* sous la forme de sondages de satisfaction effectués à l'issue des conversations. En effet le format des conversations "*chat*", ainsi que le volume des conversations disponibles font que nous pouvons disposer d'un très grand ensemble de sondages, sur lesquels des modèles de prédiction sont entraînés. Dans (Hara *et al.*, 2010), les auteurs exploitent des enquêtes de satisfaction renseignées par les utilisateurs eux-même afin d'estimer la satisfaction face à un système de dialogue homme-machine. Cependant, ces questionnaires étaient dans leur cas directement liés à la qualité perçue du système de dialogue, avec des questions orientées dans ce sens. Ici, nous disposons d'un volume important de conversations "*chat*" avec une annotation directe de satisfaction selon plusieurs dimensions.

Disposer d'une supervision directe est bien évidemment un atout majeur, cependant cela pose des questions sur la difficulté de la tâche de prédiction : en effet, contrairement à la supervision indirecte effectuée par des experts se basant uniquement sur des transcriptions de conversation, nous ne savons pas dans quelle mesure les notes données par les utilisateurs ont des traces *objectives* dans la conversation elle-même ou bien proviennent d'un ressenti et d'une expérience utilisateur prenant en compte l'historique des rapports entre le client et le service. Cette étude se propose d'essayer de prédire cette supervision directe de la satisfaction des utilisateurs d'un service, en présentant tout d'abord dans le paragraphe suivant le type de données et de sondage auxquels nous avons eu accès.

3 Conversations et métadonnées

Les conversations utilisées sont issues des logs de conversations de type "chats" provenant du service client de l'entreprise Orange. Les différentes conversations portent sur plusieurs sujets, à la fois techniques sur les problèmes rencontrés avec les services proposés, ou encore des questions à propos d'une offre commerciale. Ces conversations textuelles étant directement issues des "*chat*" entre clients et téléconseillers, il est important de noter qu'il y a une présence assez importante de fautes d'orthographe et autres types de *bruits*. Lorsqu'une conversation avec un agent est terminée, le client a la possibilité de remplir un questionnaire contenant les 5 questions suivantes :

Question	Alias
J'ai été accompagné(e) et j'ai eu les explications pour faire par moi-même	Accompagnement
J'ai été écouté(e) et ma demande a été prise en charge	Ecoute
J'ai été bien conseillé(e)	Conseil
La solution proposée par Orange me convient	Solution
Suite à votre contact avec le Service Clients, recommanderiez-vous Orange à vos proches ?	Recommander

Si certaines questions portent directement sur l'interaction en elle même ("*Accompagnement*",

"*Ecoute*", "*Conseil*"), d'autres ne sont qu'indirectement liées. Ainsi "*Solution*" peut être liée à l'expérience du client à l'issue de la conversation. Enfin, la question "*Recommander*" relève également d'une appréciation générale pour lesquels les clients peuvent exprimer un ressenti plus large que celui qui résulte de la simple conversation. Pour cette question "*Recommander*", le client doit répondre sur une échelle allant de 0 à 10. Suivant les conventions du domaine de l'analyse de la relation client, nous avons réalisé des regroupements pour définir trois catégories : *détracteurs* (note de 0 à 6), *passifs* (7 ou 8) et *promoteurs* (9 ou 10). Pour les autres questions, le client doit répondre sur une échelle à 5 niveaux allant de "Pas du tout satisfait(e)" à "Très satisfait(e)". Les réponses à ces questions sont des indicateurs importants pour juger de la qualité de service.

Les données ont été collectées sur une période d'un mois et nous avons sélectionné le sous-ensemble de conversations pour lesquelles nous avions une réponse à toutes les questions. Les corpus d'entraînement, de développement et de test sont respectivement constitués de 47685, 15899 et 15892 conversations. Le corpus d'entraînement est composé de 140000 tokens différents. Comme précisé dans le paragraphe précédent, l'originalité de l'étude est que la supervision de l'annotation est faite directement par le client. Il y a ainsi autant d'annotateurs que de conversations. La quantité très importante de données d'apprentissage et de test disponibles avec cette supervision directe (près de 80K conversations), est aussi très inhabituelle pour ce type d'étude ou ce sont généralement de petits volumes qui sont considérés.

4 Prédiction automatique de la satisfaction client

L'objectif est d'évaluer dans quelle mesure il est possible, à partir de l'analyse du contenu des conversations, de prédire automatiquement les réponses aux 5 questions posées à l'issue de ces conversations. Dans une première approche, nous considérons ce problème comme une tâche de classification où pour chaque dimension considérée, un classifieur doit prédire la réponse à la question posée. Les classifieurs diffèrent entre autres par le mode de représentation du texte des conversations traitées : simples sacs de mots (1) contenant toute la conversation, découpage en blocs contigus (2) ou séquences de mots ordonnées (3).

Pour évaluer le premier mode en sac de mots (1), nous utilisons un classifieur SVM avec un modèle pour chaque tâche, dans l'implémentation des SVM à noyau linéaire de Pedregosa *et al.* (2011).

Pour la représentation en bloc (2), nous utilisons un réseau de neurones convolutionnels (CNN) basé sur le réseau décrit par Kim (2014). Nous créons un modèle par tâche ayant des filtres de tailles 3, 4 et 5 avec 100 filtres pour chaque taille.

Pour le dernier mode prenant en compte l'ordre des mots (3), nous avons implémenté un réseau de neurones récurrent (RNN) de type *Long Short Term Memory* (LSTM). Etant donné la forte variabilité inhérente à ce mode de représentation où chaque conversation est représentée comme une unique séquence de tokens (environ 500 mots en moyenne par conversation), nous avons ajouté en complément au RNN, un *mécanisme d'attention* (Bahdanau *et al.*, 2014; Xu *et al.*, 2015) permettant au système de se focaliser sur les mots importants d'une conversation par rapport à la tâche visée. Dans cette configuration, les interventions de l'usager et du téléconseiller sont concaténées en ajoutant un token <EOT> à la fin de chaque tour de parole. En sortie, le réseau donne, pour chaque conversation, une distribution de probabilité sur l'ensemble des classes de la tâche. La classe sélectionnée est celle qui a la plus haute probabilité.

Le mécanisme d'attention utilisé est le suivant :

$$u_t = v^\top \tanh(W_a h_t + b_a) \qquad \alpha_t = \frac{\exp(u_t)}{\sum_{i=1}^{n} \exp(u_i)} \qquad \text{SelfAttn}(h) = \sum_{i=1}^{n} \alpha_i h_i$$

où W_a et b_a sont des paramètres de la fonction calculant le score d'attention et v est le vecteur de contexte qui est aléatoirement initialisé et qui est également entraîné lors de la phase d'apprentissage.

Soit $W = w_1 \dots w_n$ une conversation. Les distributions de probabilité p sont obtenus avec :

$$x_t = \text{Embedding}(w_t) \quad (1) \qquad \overrightarrow{h}_t = \overrightarrow{\text{LSTM}}(x_t, \overrightarrow{h}_{t-1}) \quad (2) \qquad \overleftarrow{h}_t = \overleftarrow{\text{LSTM}}(x_t, \overleftarrow{h}_{t-1}) \quad (3)$$

$$h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \quad (4) \qquad h = \{h_t \mid t \in [1,n]\} \quad (5) \qquad c = \text{SelfAttn}(h) \quad (6) \qquad p = \text{softmax}(W_d c + b_d) \quad (7)$$

Dans les équations précédentes, LSTM est une couche de Long Short-Term Memory units (Hochreiter & Schmidhuber, 1997). Le sens de la flèche indique le sens de lecture des séquences par la couche LSTM. W_d et b_d sont les paramètres de la couche de décision.

Dans le réseau implémenté, les couches cachées des LSTMs sont de taille 128. Les *embeddings* de mots sont de dimensions 100. L'information indiquant qui est le scripteur du tour de parole (client, téléconseiller, système) est également intégrée sous forme d'*embeddings* de taille 3 concaténés aux *embeddings* de mots. Nous utilisons la fonction d'entropie croisée pour la fonction de coût. Les poids du modèle sont initialisés uniformément dans $[-0.1,0.1]$, et optimisés avec l'algorithme ADAM. Un dropout de 0.5 est appliqué après la couche de LSTMs. Pour des contraintes techniques, les tailles des conversations sont normalisées à une taille de 1200 mots. Pour les conversations plus courtes, un symbole de *padding* est utilisé pour compléter les conversations ; pour celles qui sont plus longues, les 1200 derniers tokens sont pris en compte. Dans l'ensemble du corpus, seules 4% des conversations sont partiellement coupées.

5 Expérimentations

La prédiction de la satisfaction est abordée dans cette étude comme un problème de classification, on mesure donc la performance de la prédiction par le taux de labels correctement prédits, appelée *accuracy* dans la suite de l'étude. Dans ce type d'évaluation, il n'est pas plus grave de faire une confusion entre le label "0" et le label "4" qu'entre le label "0" et le label "1". Cependant, d'un point de vue applicatif, ces confusions n'ont pas la même valeur. Les labels étant des notes sur une échelle de satisfaction graduée, il est plus grave de considérer un client pas du tout satisfait comme très satisfait, plutôt que comme peu satisfait.

C'est pourquoi nous évaluons également les classifieurs avec des mesures qui exploitent la gradation des labels, qui sont ici des valeurs numériques ordonnées (que l'on peut qualifier de notes). Nous utilisons pour cela le coefficient de corrélation de Spearman entre l'ensemble des notes prédites et l'ensemble des notes réelles car il permet d'évaluer si les notes prédites conservent l'ordonnancement des notes réelles. En effet, ce coefficient de corrélation vaut 1 quand il existe une fonction monotone croissante entre labels prédits et labels réels, c'est-à-dire si les ordonnancements des labels prédits sont identiques aux ordonnancements des labels réels. Afin d'avoir une mesure d'évaluation plus interprétable, nous utilisons également la mesure Δ_{abs}, définie comme la moyenne de la valeur absolue de la différence entre le label réel (score de satisfaction de la vérité-terrain) et le label prédit. Cette mesure doit être la moins élevée possible, un classifieur parfait rendant une mesure de 0.

satisfaction (taille de l'échelle)	approche	accuracy	spearman	Δ_{abs}
Accompagnement (5)	Majorité	48.48	-	0.974
	SVM	55.28	0.542	0.729
	CNN	**56.85**	0.549	0.64
	RNN	55.68	0.553	0.644
	RNN+Attn	56.82	**0.57**	**0.633**
Conseil (5)	Majorité	53.24	-	0.867
	SVM	59.84	0.517	0.613
	CNN	61.22	0.568	0.562
	RNN	60.56	0.565	0.569
	RNN+Attn	**61.43**	**0.594**	**0.556**
Solution (5)	Majorité	44.38	-	1.195
	SVM	54.62	0.544	0.788
	CNN	55.82	0.587	0.724
	RNN	54.12	0.574	0.767
	RNN+Attn	**56.21**	**0.611**	**0.713**
Ecoute (5)	Majorité	54.57	-	0.833
	SVM	61.26	0.517	0.613
	CNN	62.77	0.554	0.54
	RNN	61.91	0.556	0.554
	RNN+Attn	**63.10**	**0.570**	**0.532**
Recommander (3)	Majorité	42.71	-	0.882
	SVM	56.31	0.441	0.593
	CNN	56.27	0.468	**0.562**
	RNN	56.01	0.435	0.605
	RNN+Attn	**57.53**	**0.478**	0.581

TABLE 1 – Résultats des classifieurs pour la prédiction des indicateurs de qualité. Accuracy et Spearman doivent être les plus élevés possibles et Δ_{abs} doit être idéalement proche de 0.

Le tableau 1 présente les résultats des différents classifieurs, pour prédire les réponses aux différentes questions sur la satisfaction. On reporte également les résultats obtenus en attribuant simplement la classe majoritaire à tous les exemples (approche *Majorité* dans le tableau). Cette *baseline* permet de vérifier que le déséquilibre entre les classes n'est pas trop important (au maximum une classe couvre environ 50% des exemples).

On peut constater dans la table 1 qu'en utilisant un RNN simple sans attention, on obtient des scores d'accuracy équivalents à $\pm 0{,}5$ points près, aux scores obtenus à l'aide du SVM. Cependant, lorsque les mécanismes d'attention sont utilisés, on observe des gains permettant de gagner entre $1{,}2$ et $1{,}8$ points par rapport au SVM. Ces meilleurs résultats indiquent que la présence ou non de certains mots dans une conversation sont de forts indicateurs pour la prédiction de la satisfaction. Au contraire, l'ordre des mots de la conversation l'est moins comme on peut le voir avec les résultats obtenus par les réseaux récurrents sans attention. En comparant les résultats obtenus par le CNN avec ceux obtenus par le RNN avec attention, nous pouvons constater que le CNN obtient des scores inférieurs ou égaux obtenant jusqu'à $1{,}26$ points de moins sur la tâche "**Recommander**".

En regardant de plus près les mots qui obtiennent le plus souvent le plus haut score d'attention, on remarque que les mots en lien avec les remerciements reviennent le plus souvent pour la majorité des tâches. Pour la tâche "**Ecoute**", on constate également qu'il y a plusieurs mots portant sur l'agent comme "*efficacité*" ou "*gentillesse*". Pour la tâche "**Solution**", on observe la présence des mots "*aide*", "*satisfait*" et "*navrée*" faisant probablement référence à la résolution ou non du problème.

Sur les deux autres mesures, on peut observer que le RNN simple obtient de meilleurs performances

que le SVM sur toutes les tâches en excluant la tâche "**Recommander**". On observe que le réseau avec attention obtient les meilleures scores de corrélations sur toutes les tâches avec des gains allant de $0{,}01$ à $0{,}05$ points, ainsi que sur la mesure Δ_{abs}. Dans le cas du CNN, on obtient des scores de corrélations inférieurs aux scores obtenus par le RNN avec attention avec des différences allant de $0{,}01$ à $0{,}02$ points. Ceci est également le cas sur la mesure Δ_{abs} sauf pour la tâche "**Recommander**" où le CNN est meilleur de $0{,}02$ points.

Les tâches obtenant les meilleurs résultats sont "**Conseil**" et "**Ecoute**" ce que l'on pouvait escompter dans la mesure où ce sont les questions les plus directement liées au déroulement de la conversation. "**Accompagnement**" relève d'une appréciation plus subjective et "**Solution**" relève d'une appréciation technique. Quant à la question de savoir si le client recommanderait l'entreprise, il peut y avoir des facteurs subjectifs dépassant le cadre de la simple conversation. Si l'on s'intéresse à l'ordre relatif des jugements, la prédiction de "**Solution**" présente le meilleur coefficient de Spearman, en revanche c'est bien la dimension "**Ecoute**" qui produit les prédictions les plus proches des prédictions réelles en valeur absolue ($\Delta_{abs} = 0.532$).

6 Conclusion

L'évaluation automatique de la satisfaction client à partir d'une conversation n'est pas une tâche facile. Les modèles de type SVMs permettent d'obtenir des résultats raisonnables et les réseaux de neurones convolutionels permettent d'améliorer ces résultats. Cependant, un réseau de neurones récurrents avec un mécanisme d'attention parvient à obtenir de meilleurs résultats que ce soit du point de vue de la classification que de la corrélation de Spearman.

Pour la suite, il serait intéressant de réaliser directement une régression avec le réseau de neurones afin de mieux prendre en compte le fait que les réponses sont données sur une échelle graduée. Il serait aussi intéressant d'essayer d'utiliser des descripteurs structurels en complément des mots.

Remerciements

Ce travail a été partiellement financé par l'Agence Nationale pour la Recherche au sein du projet ANR-15-CE23-0003 (DATCHA).

Références

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

HARA S., KITAOKA N. & TAKEDA K. (2010). Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system. In *LREC*.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780. 04135.

KIM Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1746–1751.

KOÇO S., CAPPONI C. & BÉCHET F. (2012). Applying multiview learning algorithms to human-human conversation classification. In *Thirteenth Annual Conference of the International Speech Communication Association*.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PRAGST L., ULTES S. & MINKER W. (2017). Recurrent neural network interaction quality estimation. In *Dialogues with Social Robots*, p. 381–393. Springer.

REICHHELD F. F. (2003). The one number you need to grow. *Harvard business review*, **81**(12), 46–55.

ROY S., MARIAPPAN R., DANDAPAT S., SRIVASTAVA S., GALHOTRA S. & PEDDAMUTHU B. (2016). Qa rt : A system for real-time holistic quality assurance for contact center dialogues. In *Thirtieth AAAI Conference on Artificial Intelligence*.

STOYANCHEV S., MAITI S. & BANGALORE S. (2017). Predicting interaction quality in customer service dialogs. In *Proceedings of the 2017 INTERNATIONAL WORKSHOP ON SPOKEN DIALOGUE SYSTEMS TECHNOLOGY (IWSD)*.

TRIONE J., FAVRE B. & BÉCHET F. (2016). Beyond utterance extraction : Summary recombination for speech summarization. In *Interspeech*, p. 680–684.

ULTES S., SCHMITT A. & MINKER W. (2013). On quality ratings for spoken dialogue systems–experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 569–578.

XU K., BA J., KIROS R., CHO K., COURVILLE A., SALAKHUDINOV R., ZEMEL R. & BENGIO Y. (2015). Show, attend and tell : Neural image caption generation with visual attention. In *International Conference on Machine Learning*, p. 2048–2057.