Représentation sémantique distributionnelle et alignement de conversations par chat

Tom Bourgeade Philippe Muller
IRIT, Université de Toulouse
tom.bourgeade@irit.fr, philippe.muller@irit.fr

RÉSUMÉ

Les mesures de similarité textuelle ont une place importante en TAL, du fait de leurs nombreuses applications, en recherche d'information et en classification notamment. En revanche, le dialogue fait moins l'objet d'attention sur cette question. Nous nous intéressons ici à la production d'une similarité dans le contexte d'un corpus de conversations par *chat* à l'aide de méthodes non-supervisées, exploitant à différents niveaux la notion de sémantique distributionnelle, sous forme d'*embeddings*. Dans un même temps, pour enrichir la mesure, et permettre une meilleure interprétation des résultats, nous établissons des alignements explicites des tours de parole dans les conversations, en exploitant la distance de Wasserstein, qui permet de prendre en compte leur dimension structurelle. Enfin, nous évaluons notre approche à l'aide d'une tâche externe sur la petite partie annotée du corpus, et observons qu'elle donne de meilleurs résultats qu'une variante plus naïve à base de moyennes.

ABSTRACT

Distributional semantic representation and alignment of online chat conversations

Textual similarity measures have an important place in NLP, because of their many applications, particularly in information retrieval and classification. However, dialog receives less attention on this issue. We are interested here in the production of a similarity measure in the context of a corpus of online chat conversations using non-supervised methods, exploiting at different levels the notion of distributional semantics, in the form of embeddings. At the same time, to enrich the measure, and to allow a better interpretation of the results, we establish explicit alignments of speaker turns in these conversations, using Wasserstein's distance, which allows us to take into account their structural dimension. Finally, we evaluate our approach using an external task on the small annotated part of the corpus, and observe that it yields better results than a naive variant based on averages.

MOTS-CLÉS : similarité textuelle, analyse de conversations, représentations sémantiques, sémantique distributionnelle, distance de Wasserstein.

KEYWORDS: textual similarity, dialog analysis, semantic representation, distributional semantics, Wasserstein distance.

1 Introduction

Les problèmes de similarité textuelle ont connu un essor rapide en TAL, en passant de questions au niveau lexical avec des relations de proximité sémantique puis rapidement au niveau phrastique, avec la question de la paraphrase ou l'inférence (Cer *et al.*, 2017), et même au-delà sur la similarité de passages textuels plus grands (Kusner *et al.*, 2015; Le & Mikolov, 2014). Les productions qui ne sont pas dans un cadre mono-locuteur sont relativement négligées, à l'exception de l'*embedding* de tours de parole isolés (Yang *et al.*, 2018), voire du cas particulier de la similarité de questions (Nakov *et al.*,

2017), mais ces modèles ne prennent pas non plus en compte la structure au-delà de la phrase. Nous nous intéressons ici au cas de la similarité de conversation, où l'échange est structuré en tours de parole produits par deux interlocuteurs, dans le cadre de dialogues orientés par une tâche, où chacun des interlocuteurs joue un rôle différent. C'est la similarité globale de conversations entières qui est ciblée. La popularité croissante d'échanges écrits sous la forme de dialogues textuels (*chat*, forum, micro-blogging) soulève des questions intéressantes de mises en rapport de conversations similaires, avec des applications directes à la fouille de conversation, souvent motivées par des problématiques de gestion de la relation client (*Customer Relationship Management*), qui nous sert de cas d'étude.

Dans ce contexte, les applications potentielles d'une définition de similarité de conversation sont nombreuses : par exemple, la possibilité de construire des marches à suivre "types", en regroupant des conversations portant sur un problème technique similaire et/ou dont les étapes de résolution suivies sont similaires ; ou bien encore, la possibilité d'effectuer une recherche rapide d'une conversation similaire pendant qu'une autre se déroule, afin de guider le conseiller, sans passer par des méthodes de recherche classique, telle que la recherche par mots-clés, qui ne renvoient pas toujours un petit nombre de résultats pertinents. Pour établir cette mesure de similarité conversationnelle, on ne s'intéresse pas ici qu'à la dimension sémantique des messages envoyés par chacun des participants, mais également à la structure globale de la conversation elle-même. De plus, le corpus utilisé ici ne disposant que d'une très faible proportion de données annotées (<1%), on n'emploie ici que des méthodes non-supervisées.

Un autre aspect important dans ce cadre d'assistance est l'explicitation des liens entre conversations qui permet de décomposer la similitude des cas rencontrés dans une tâche de support, et de comprendre la pertinence des conversations mises en relation.

La contribution du travail porte donc ici sur les aspects suivants : une définition de similarité entre conversations, avec un alignement explicite des parties de conversations qui sous-tendent cette similarité, et une méthode non-supervisée pour calculer les représentations utilisées pour calculer la similarité. Pour montrer l'intérêt de l'alignement, nous comparons l'utilisation de cette méthode avec d'autres plus simples pour prédire le résultat d'une tâche de classification annexe par plus proches voisins.

2 Représentation de conversations

Pour cette tâche, on a choisi une approche plus modulaire que monolithique, afin de pouvoir évaluer indépendamment différentes méthodes et architectures existantes. L'objectif final étant de pouvoir comparer et mesurer la similarité entre des conversations, on choisit de définir une conversation, dans le contexte du corpus utilisé, comme étant constituée d'une séquence de messages accompagnés de l'identité de leurs auteurs respectifs. De ce fait, la tâche s'organise assez naturellement autour de deux sous-objectifs principaux : dans un premier temps, on va chercher à construire des représentations sémantiques numériques, sous la forme de vecteurs dans un espace d'*embedding*, pour chacun des tours de parole qui font partie de la conversation. Puis, dans un second temps, en s'appuyant sur ces représentations, on va construire une représentation de la conversation qui prenne en compte sa structure, et dont on pourra mesurer une distance avec les représentations d'autres conversations, que l'on assimilera à une mesure de similarité.

En pratique ici, on a choisi d'utiliser un modèle encodeur-décodeur type *seq2seq* (Sutskever *et al.*, 2014) pour construire des représentations sémantiques vectorielles des tours de parole. En particulier, on a choisi une implémentation du modèle non-supervisé *Skip-Thought Vectors* (Kiros *et al.*, 2015), en travaillant sur des séquences de *word embeddings* produit par un modèle *FastText* (Bojanowski

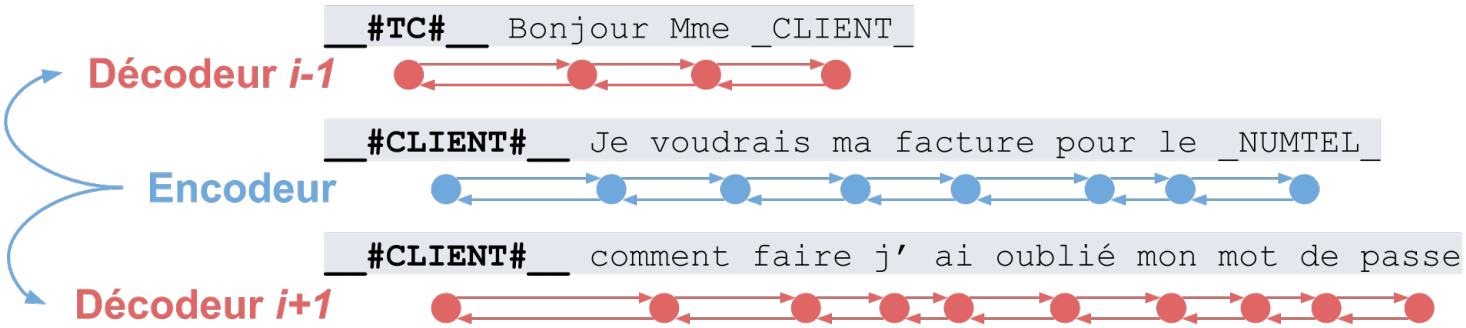


FIGURE 1 – Schéma du modèle *Skip-Thought* utilisé. Le modèle encodeur produit une représentation du i-ème message (concaténation des états cachés finaux des deux directions du modèle récurrent) qui est ensuite utilisée pour conditionner deux décodeurs récurrents respectivement sur les $(i-1)$ et $(i+1)$-èmes messages, en espérant ainsi forcer cette représentation à capturer des informations sémantiques conversationnelles.

et al., 2017), tout deux entraînés sur les données non-annotées du corpus. Due à la nature du corpus, il est nécessaire de prendre en compte un certain nombre de contraintes sur le contenu textuel des conversations : en particulier, on a choisi un modèle *FastText* car celui-ci peut obtenir des *embeddings* pour des mots hors-vocabulaire à l'aide d'*embeddings* de *n-grammes*, ce qui nous permet de mieux gérer la présence de fautes de frappe et d'orthographe. Le modèle *Skip-Thought Vectors* produit des encodages de séquences de mots, en prenant en compte leurs contextes immédiats durant l'apprentissage (ici, le tour de parole directement suivant et précédent, voir figure 1), ceci permettant en théorie d'enrichir les informations sémantiques extraites par cet encodeur. En sortie, on obtient donc des représentations vectorielles des messages d'une conversation résultants de la concaténation des états cachés finaux du modèle récurrent bidirectionnel, dont les hyper-paramètres utilisés dans nos expériences sont donnés en table 1.

Une fois les messages des conversations encodés, et afin de prendre en compte la dimension structurelle de celles-ci, on a choisi de s'inspirer de la mesure *Word Mover's Distance (WMD)* présentée dans Kusner *et al.* (2015), elle-même proche de Wan (2007) : cette méthode consiste normalement à employer la distance de Wasserstein, qui mesure le "coût" nécessaire pour transformer une distribution de probabilité en une autre, sur des phrases sous la forme de séquences de *words embeddings*, en résolvant le problème de transport optimal *Earth Mover's Distance (EMD)* associé : on assimile chaque vecteur-mot de la phrase A à un tas de terre et chaque vecteur-mot de la phrase B à un trou, tous de même capacité et définis dans le même espace muni d'une mesure de distance (ou coût) donnée, le but étant de trouver l'ensemble de déplacements de coût total minimum qui permette de remplir tous les trous (ou d'épuiser tous les tas de terre ainsi). Cette approche fournit à la fois une mesure de distance, assimilable à une mesure de similarité sémantique ici, et un alignement optimal (au sens du problème *EMD*) entre deux phrases données. On peut adapter cette méthode, à un niveau d'abstraction plus élevé, en travaillant sur des séquences d'*embeddings* de phrases aux seins de conversations, comme c'est notre cas ici. La résolution d'instances du problème *EMD* peut se faire à l'aide d'un solveur dédié (tel que *POT*[1]), et ne nécessite que de calculer la matrice des coûts entre les vecteurs, à l'aide de la similarité cosinus ou d'une autre norme vectorielle par exemple. L'avantage principal de cette approche est qu'en plus d'obtenir une mesure de similarité conversationnelle structurelle, on obtient également un alignement optimal des messages des conversations traitées, qui peut être utilisé comme entrée pour une autre tâche, ou pour permettre une interprétation qualitative des résultats : en effet on

1. Source : `https://github.com/rflamary/POT`

pourra alors observer comment se transpose l'enchaînement d'actes de dialogue d'une conversation à une autre qui lui est similaire (d'après cette mesure) et ainsi pouvoir vérifier quelles parties de celles-ci ont été jugées similaires, et de quelle manière elle le sont (dans la forme globale que prennent certaines sections, ou dans la thématique d'un ou plusieurs messages en particulier, par exemple). Avec cette approche, une conversation sera donc représentée par la distribution des messages qui la composent dans leur espace d'*embedding*, ceci s'opposant aux méthodes qui agrège ces vecteurs en une représentation simple, qui ont pour avantage d'être plus économes en mémoire et en temps de calcul (pour effectuer des mesures de similarité), mais ont pour désavantage de détruire au moins en partie l'information structurelle de la conversation.

3 Données utilisées

Les données utilisées ici proviennent d'un corpus constitué de journaux de conversations par *chat*, entre des téléconseillers et des clients, provenant de la plateforme d'assistance technique et commerciale en-ligne de l'opérateur téléphonique Orange [2]. Chaque conversation est constituée d'une suite de messages horodatés et munis d'un marqueur d'identité anonymisé, avec éventuellement des méta-données associées au contexte technique de la conversation. Une petite partie du corpus a été annotée par les utilisateurs télé-conseillers de la plateforme, avec des labels indiquant l'état de résolution du problème technique du client à la fin de la conversation (par exemple : `PbTechResolu`, si le problème du client a été résolu en fin de dialogue, `InfoFournie` s'il s'agissait d'une question posée par le client et qu'une réponse satisfaisante a été fournie, `SuspenduClient` si le client à mis fin à la conversation abruptement, etc.).

La nature de ce corpus implique quelques spécificités par rapport à des corpus conversationnels plus classiques. Ne s'agissant pas de transcriptions de dialogues oraux, le contenu textuel est fourni tel qu'il a été saisi par les interlocuteurs du *chat*, ce qui implique la présence en abondance de fautes de frappes, de grammaire, d'orthographe, et de problèmes de structure dans les messages (atténuées ici par l'emploi de *FastText*, moins sensible aux petites différences morphologiques lors de la production de *word embeddings*). De plus, le *chat* en-ligne est un medium dans lequel on trouve des phénomènes linguistiques particuliers, dû, notamment, à la nature asynchrone de la communication (messages de "correction", phrases communiquées "par morceaux", question-réponse en décalage, etc.). De même, il est nécessaire de prendre en compte la présence d'éléments non-linguistiques, comme des hyperliens, des marqueurs de balisage (HTML ou autre) ou encore des pictogrammes de nature diverses (émoticônes, *smileys* ou *emojis*). De ce fait, il est nécessaire d'effectuer plusieurs étapes de pré-traitement avant de pouvoir les utiliser. En s'inspirant de l'anonymisation déjà effectuée (portant principalement sur les données à caractère personnel du client, par exemple, ses numéros de téléphone sont remplacés par des jetons `_NUMTEL_`), on remplace ces différents éléments par des marqueurs simplifiés (`_HTML_` pour des éléments de balisage HTML, etc.) qui permettent aux modèles d'*embedding* de mots et de tours de parole de les prendre en compte d'une manière similaire à des éléments de ponctuation, sans se soucier de leur sémantique particulière. Dans un même temps, on encode l'identité de l'auteur par un marqueur au début de son message (`__#TC#__` et `__#CLIENT#__` pour le télé-conseiller et le client, respectivement), et qui seront prises en compte dans la représentation du tour de parole.

2. Les données étant mises à disposition des participants du projet ANR Datcha.

Nombre de conversations dans le corpus complet	432 768
Nombre de conversations annotées (état résolution)	2775
Nombre de tours de paroles dans le corpus complet	15 682 118
Taille du vocabulaire pour *word embeddings*	120 391
Dimensions des *word embeddings*	100
Dimensions cachées de l'encodeur *Skip-Thoughts*	1024
Structure de l'encodeur *Skip-Thoughts*	biLSTM, 1 couche, 0.5% *dropout*
Structure des décodeurs *Skip-Thoughts* (entraînement)	2 biGRU (suivant, précédent), 1 couche
Paramètres de l'entraînement	10 *epochs*, taille des *batchs* = 32, optimiseur Adam

TABLE 1 – Description des données et des paramètres du modèle présenté.

4 Expériences et résultats

Pour tester notre méthode, au vu du fait que le corpus n'est pas annoté d'une manière directement pertinente pour une mesure de similarité, on opte pour une évaluation indirecte via une tâche externe et des annotations détachées. On fait donc l'hypothèse que la mesure de similarité ainsi produite par notre approche peut potentiellement être corrélée avec l'appartenance à une catégorie d'état de résolution des conversations, qui ont été annotées sur une petite partie du corpus (2775 conversations, contre 432768 pour le corpus d'entraînement total) par l'opérateur de téléphonie. On effectue donc une tâche d'évaluation externe de classification portant sur ces labels, en utilisant un classifieur k-plus proches voisins (avec $k = 5$ ici) auquel on fournit une matrice de distance produite par différentes variantes de notre méthode, et sur lesquelles on effectue une validation croisée sur *10-fold* (le jeu de données étant relativement petit). On dispose de 14 classes d'états de résolution annotées, la classe majoritaire représentant 26.19% des données. Les cinq variantes évaluées et comparées ici (table 2) utilisent toutes les mêmes *embeddings* de messages produit par le modèle *FastText* comme base, les différences portant sur la production des représentations des messages, la phase d'alignement et la mesure de distance inter-conversations : la variante *baseline* utilise la moyenne des vecteurs d'*embedding* de messages produits par le modèle *Skip-Thoughts* comme représentation vectorielle des conversations, puis effectue une simple mesure de similarité cosinus comme entrée du classifieur. Les variantes *SIF (Smooth Inverse Frequency)* implémentent la méthode décrite dans (Arora *et al.*, 2017) pour construire des représentations de messages (une version moyenne comme la *baseline*, l'autre avec l'approche *EMD*). Les deux dernières variantes implémentent l'approche basée sur le problème *EMD* présentée dans cet article (avec vecteurs de messages *Skip-Thoughts*), l'une employant la norme $L2$ des vecteurs d'*embedding* de messages comme matrice de coûts pour le solveur, l'autre la similarité cosinus. La variante *baseline* peut paraître excessivement brutale, mais il semblerait qu'au moins au niveau des phrases, l'approche *CBOW (Continuous Bag Of Words)* consistant à effectuer une moyenne des représentations vectorielles capture déjà une partie suprenante de l'informations sémantiques dans les phrases, comme il est montré dans (Adi *et al.*, 2017; Shen *et al.*, 2018).

Modèle	**Exactitude (*accuracy*)**
Skip-Thoughts + cos + EMD	**52.53%** ($\sigma \approx 3.36\%$)
Skip-Thoughts + L2 + EMD	51.20% ($\sigma \approx 3.14\%$)
SIF + cos + EMD	48.78% ($\sigma \approx 2.66\%$)
SIF + cos + moyenne	44.73% ($\sigma \approx 1.71\%$)
Skip-Thoughts + cos + moyenne (*baseline*)	43.41% ($\sigma \approx 2.62\%$)

TABLE 2 – Résultats de la tâche de classification externe sur différentes variantes (*accuracy* moyenne et écart-type sur la validation croisée.)

On observe que la variante *EMD* avec similarité cosinus affiche les meilleures performances, avec une amélioration moyenne de $+9\%$ sur le modèle *baseline*. La similarité cosinus est en général l'opération de prédilection pour comparer des vecteurs d'*embedding* car elle n'est pas sensible à leurs normes, qui, due à la manière dont ceux-ci sont construits, traduit en général une notion analogue à la fréquence d'apparition de l'élément associé dans le corpus, mesure qui n'est en général pas vraiment pertinente quand on souhaite effectuer des comparaisons sémantiques. Ceci peut expliquer les petites différences de performances entre la variante utilisant la norme $L2$ et celle utilisant la similarité cosinus. La distance de Wasserstein, quant à elle, est beaucoup plus sensible aux détails structurels de la conversation : en effet, avec le modèle *baseline* "sac de mots", on peut imaginer que deux conversations structurellement différentes puissent avoir des vecteurs moyens proches, tandis qu'il serait très improbable de trouver deux conversations différentes ayant des distributions de messages très similaires dans l'espace d'*embedding*.

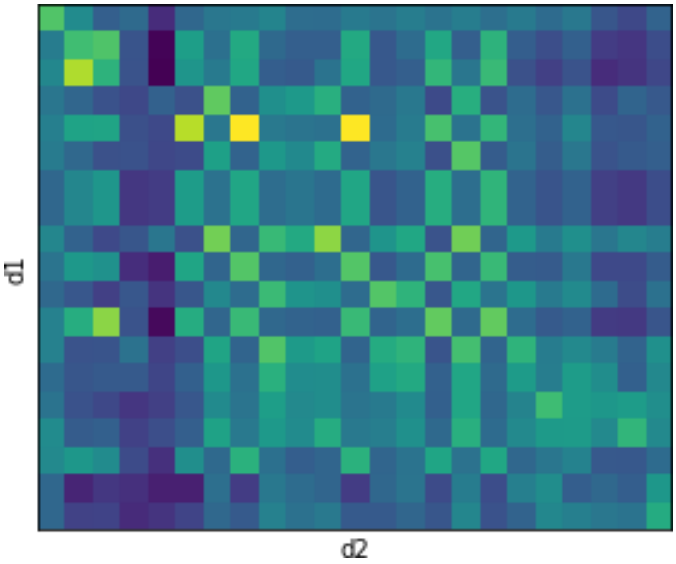
I1	**D1**	**I2**	**D2**
1	__#TC#__ Bonjour Mme _CLIENT_	1	__#TC#__ _HTML_ Bonsoir , je me prénomme _TC1_ et je vais traiter votre demande . En quoi puis -...
2	__#CLIENT#__ Je voudrais ma facture pour le _NUMTEL_	3	__#CLIENT#__ et je veu remplir le formulaire de rembourssement
3	__#CLIENT#__ comment faire j' ai oublié mon mot de passe	2	__#CLIENT#__ bonsoir j ai oublier mon mot de passe svp
4	__#TC#__ Souhaitez - vous récupérer le mot de passe de votre adresse mail afin de consult...	7	__#TC#__ _HTML_ Votre demande consiste t - elle a récupérer le mot de passe de votre adresse de ...
5	__#CLIENT#__ oui	8	__#CLIENT#__ oui

FIGURE 2 – Extrait d'un alignement optimal produit par notre approche, entre tours de parole de deux conversations (`d1` et `d2`). **I1** et **I2** correspondent respectivement aux indices originaux des interventions dans les conversations (la conversation `d2` étant aligné sur `d1`).

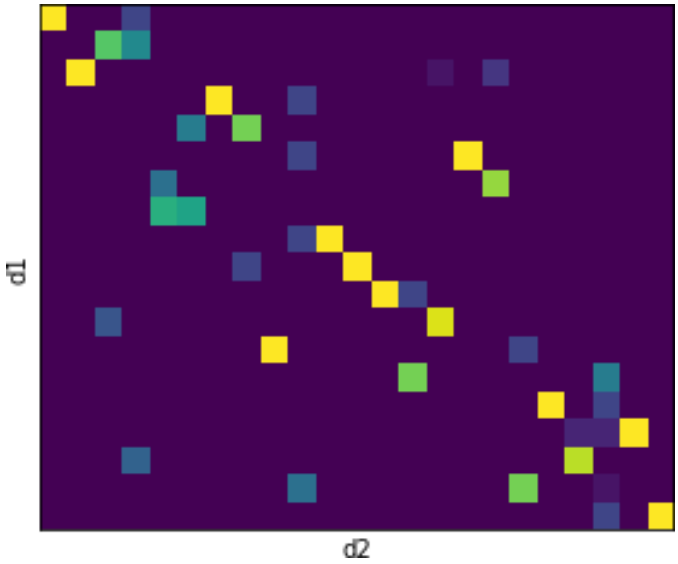
L'approche avec *EMD* est également beaucoup plus interprétable, grâce aux alignements de conversations construits. L'interprétabilité étant un des défis majeurs de la recherche en intelligence artificielle aujourd'hui, la possibilité d'extraire une forme d'explication humainement interprétable, ainsi que les différents objets mathématiques utilisés pour arriver à cette décision, sont des avantages importants de cette approche : dans la figure 2 on peut observer un alignement optimal d'une conversation **D1** avec une conversation similaire **D2**. On peut par exemple voir ici une instance de sémantique structurelle : dans le message de **D1** d'indice 2, le client énonce sa demande principale (récupérer sa facture), suivie en indice 3 du problème instrumental à sa résolution (perte de son mot de passe), tandis que dans **D2**, ces actes de dialogue apparaissent dans l'ordre opposé. Malgré la différence de demande (dans **D2**, le client souhaite remplir un formulaire de remboursement) et l'ordre inversé du déroulement de la conversation, l'alignement produit nous permet qualitativement de juger en quoi celles-ci sont similaires. Une autre manière d'obtenir des interprétations est de directement observer les matrices manipulées par la méthode (figure 3), en particulier la matrice d'alignement solution du problème *EMD* (3b), qui nous permet de rapidement identifier les messages fortement similaires (avec flux important) et les segments de conversation communs mais potentiellement transposés (structures en "diagonales").

L'évaluation présentée ici est bien sûr préliminaire, dans la mesure où le modèle est très simple par rapport à la tâche considérée, et celle-ci ne correspond pas de toutes façons aux applications visées par la similarité. Ces dernières nécessitent une mise en place plus complexe vis à vis de la plate-forme

de conseil, qui n'a pas encore été mise en oeuvre dans le projet global.



(a) Matrice de similarité (en cosinus) entre les représentations sémantiques latentes des messages des deux conversations (jaune : similarité élevée ; bleu : similarité faible).



(b) Matrice solution du problème *EMD* associé, correspondant au meilleur alignement possible entre les deux conversations (jaune : flux important ; bleu : flux faible).

FIGURE 3 – Exemples de matrices associées aux dialogues `d1` et `d2`.

5 Perspectives et conclusion

Nous avons proposé ici un modèle qui permet de définir une similarité entre dialogues en prenant en compte le contenu sémantique tout en respectant la structure du dialogue en tours de parole. Il est évident que l'organisation dialogique peut être modélisé de nombreuses façons plus précises, en prenant en compte le type des actes de dialogue (Bunt *et al.*, 2010) ou l'organisation des liens entre tours en fonction des besoins de la communication avec des relations dialogiques, comme dans (Asher *et al.*, 2016). Sans aller jusqu'à ce dernier niveau complexe à prédire même avec des données annotées, ajouter l'information du type d'acte de dialogue (assertion, question, ...) est une suite naturelle : ce niveau est l'objet de nombreux travaux avec des performances assez élevées, que ce soit sur l'anglais (Kumar *et al.*, 2018) ou le français (Perrotin *et al.*, 2018). On pourrait alors observer si l'ajout de cette information à chaque énoncé permet d'avoir de meilleurs alignements et une meilleure correspondance entre des types de dialogue. À l'inverse, il pourrait être intéressant de voir l'apport de l'appariement de dialogues dans l'apprentissage d'une séquence d'actes de dialogues. De plus, si en théorie le modèle *Skip-Thoughts* utilisé ici permet de capturer le contexte immédiat des messages, comprenant donc en partie les spécificités liées à la nature des conversations par *chat*, une amélioration possible à explorer serait l'augmentation de la taille de la fenêtre de contexte à l'entraînement (ici de taille 1 seulement), ou bien, l'utilisation de modèles avec prédictions contextualisées, comme *ELMo* (Peters *et al.*, 2018) ou *BERT* (Devlin *et al.*, 2018). Au-delà du dialogue par *chat*, d'autres formes de communication impliquent de modéliser des interactions textuelles, y compris dans un contexte orienté-tâche, et il serait intéressant de généraliser l'approche à la modélisation de dialogue sur des forums, où (Wang *et al.*, 2012) a montré l'intérêt d'une analyse (supervisée) de la structure.

Remerciements

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

Références

(2017). *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

ADI Y., KERMANY E., BELINKOV Y., LAVI O. & GOLDBERG Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In (DBL, 2017).

ARORA S., LIANG Y. & MA T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In (DBL, 2017).

ASHER N., HUNTER J., MOREY M., BENAMARA F. & AFANTENOS S. D. (2016). Discourse structure and dialogue acts in multiparty dialogue : the STAC corpus. In *LREC* : European Language Resources Association (ELRA).

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

BUNT H., ALEXANDERSSON J., CARLETTA J., CHOE J., FANG A. C., HASIDA K., LEE K., PETUKHOVA V., POPESCU-BELIS A., ROMARY L., SORIA C. & TRAUM D. R. (2010). Towards an ISO standard for dialogue act annotation. In *LREC* : European Language Resources Association.

CER D., DIAB M., AGIRRE E., LOPEZ-GAZPIO I. & SPECIA L. (2017). Semeval-2017 task 1 : Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 1–14 : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

KIROS R., ZHU Y., SALAKHUTDINOV R., ZEMEL R. S., URTASUN R., TORRALBA A. & FIDLER S. (2015). Skip-thought vectors. In *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 3294–3302.

KUMAR H., AGARWAL A., DASGUPTA R. & JOSHI S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. In *AAAI*, p. 3440–3447 : AAAI Press.

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966 : JMLR.org.

LE Q. V. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, p. 1188–1196.

NAKOV P., HOOGEVEEN D., MÀRQUEZ L., MOSCHITTI A., MUBARAK H., BALDWIN T. & VERSPOOR K. (2017). Semeval-2017 task 3 : Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 27–48 : Association for Computational Linguistics.

PERROTIN R., NASR A. & AUGUSTE J. (2018). Dialog Acts Annotations for Online Chats. In *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France.

PETERS M., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana : Association for Computational Linguistics.

SHEN D., WANG G., WANG W., RENQIANG MIN M., SU Q., ZHANG Y., LI C., HENAO R. & CARIN L. (2018). Baseline needs more love : On simple word-embedding-based models and associated pooling mechanisms. In *ACL*.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*, p. 3104–3112.

WAN X. (2007). A novel document similarity measure based on earth mover's distance. *Information Sciences*, **177**(18), 3718–3730.

WANG L., KIM S. N. & BALDWIN T. (2012). The utility of discourse structure in identifying resolved threads in technical user forums. In *COLING*, p. 2739–2756 : Indian Institute of Technology Bombay.

YANG Y., YUAN S., CER D., KONG S.-Y., CONSTANT N., PILAR P., GE H., SUNG Y.-H., STROPE B. & KURZWEIL R. (2018). Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, p. 164–174 : Association for Computational Linguistics.