Traitement quantique des langues : état de l'art

Sabrina Campano — Tahar Nabil — Meryl Bothua

EDF Lab Paris-Saclay, boulevard Gaspard Monge, 91120 Palaiseau

RÉSUMÉ. *Cet article propose un état de l'art des travaux d'informatique quantique pour le traitement automatique des langues (TAL). Leur objectif est d'améliorer les performances des modèles actuels, et de mieux représenter des phénomènes linguistiques comme l'ambiguïté ou les dépendances à longue distance. Plusieurs familles d'approches sont présentées, dont les modèles symboliques diagrammatiques, et les réseaux de neurones hybrides. Ces travaux montrent que des expérimentations sont déjà possibles, et ouvrent des perspectives de recherche concernant le développement de nouveaux modèles et leur évaluation.*

MOTS-CLÉS : *informatique quantique, traitement quantique des langues.*

TITLE. *Quantum Natural Language Processing : a review*

ABSTRACT. *This article presents a review of quantum computing research works for Natural Language Processing (NLP). Their goal is to improve the performance of current models, and to provide a better representation of several linguistic phenomena, such as ambiguity and long range dependencies. Several families of approaches are presented, including symbolic diagrammatic approaches, and hybrid neural networks. These works show that experimental studies are already feasible, and open research perspectives on the conception of new models and their evaluation.*

KEYWORDS: *quantum computing, quantum natural language processing.*

1. Introduction

Depuis que l'idée a été émise notamment par le physicien Richard Feynman en 1981, l'informatique quantique intrigue, générant de nombreux efforts de recherche pour comprendre et réaliser son potentiel (Feynman, 1982). En tirant profit de la superposition d'états et de l'intrication, deux propriétés fondamentales des systèmes quantiques, la promesse des ordinateurs quantiques est d'obtenir des avantages calculatoires par rapport aux machines dites « classiques », en réduisant la complexité algorithmique de résolution de certains problèmes. Par exemple, les algorithmes de Shor (1994) pour la factorisation des nombres entiers et de Harrow-Hassidim-Lloyd (HHL) pour la résolution d'un système linéaire (Harrow *et al.*, 2009) garantissent respectivement, par la preuve théorique, des accélérations exponentielles par rapport à toute contrepartie classique. L'algorithme de recherche d'information de Grover apporte quant à lui un gain quadratique (Grover, 1996). Le développement de méthodes algorithmiques précède ainsi l'avènement de l'ordinateur quantique, la première réalisation expérimentale d'un algorithme quantique – celui de Grover – datant de 1998 (Chuang *et al.*, 1998). Si la notion d'avantage quantique est multiforme et reste difficile à mesurer (Rønnow *et al.*, 2014), les progrès significatifs récents des constructeurs d'infrastructures matérielles permettent dorénavant d'élargir l'informatique quantique expérimentale à plusieurs disciplines. À titre illustratif, Madsen *et al.* (2022) ont ainsi mené des essais démontrant un avantage sur une tâche d'échantillonnage de bosons.

Dans le domaine du traitement automatique des langues (TAL), l'informatique quantique a fait émerger un nouveau champ disciplinaire, le traitement quantique des langues (TQL, ou *QNLP, Quantum Natural Language Processing* en anglais). Les origines du TQL remontent aux travaux de Coecke *et al.* (2010), où le sens d'une phrase est calculé en utilisant la composition de produits tensoriels, et la première conférence dédiée a eu lieu en 2019, accueillant à la fois des contributions théoriques et expérimentales. Si les méthodes classiques de TAL connaissent des succès spectaculaires sur de nombreuses tâches, p. ex. *via* les modèles d'apprentissage profond, la perspective du TQL tient avant tout aux fondements théoriques de la discipline.

À terme, il s'agit en effet de trouver une façon de représenter et de traiter la langue pouvant dépasser les limites actuelles des approches classiques. Au-delà de l'accélération des calculs, certains travaux en TQL sont motivés par des approches symboliques, qui pourraient permettre de mieux prendre en compte des aspects comme l'ambiguïté, et la façon dont le sens d'une phrase est composé à partir de ses constituants. Meichanetzidis *et al.* (2023) voient ces approches symboliques plus transparentes que celles fondées sur les réseaux de neurones, qualifiés de boîtes noires. Selon Correia *et al.* (2022), les approches à base de règles en TAL pourraient être réinstituées, car les calculs requis deviendraient plus efficaces avec des algorithmes quantiques.

Dans ce contexte, l'objectif de cette revue de l'état de l'art en TQL est de familiariser le lecteur de la communauté TAL avec cette discipline naissante. Nous définissons ainsi les concepts clés de l'informatique quantique en section 2. Puis la section 3 présente de nouveaux modèles classiques de langue s'appuyant sur ces notions. Les

modèles quantiques sont décrits en section 4, avec d'une part les approches diagrammatiques et symboliques, et d'autre part les hybridations quantiques des réseaux de neurones classiques. L'inclusion de ces derniers modèles, la structuration fondée sur les approches plutôt que sur les cas d'application, et les explications détaillées fournies dans ce document complètent ainsi la première revue de l'état de l'art sur le sujet proposée par Wu *et al.* (2021). Nous détaillons également quelques expérimentations, rendues possibles par les progrès technologiques. La comparaison rigoureuse de ces méthodes entre elles ou bien à des algorithmes classiques n'entre pas toutefois dans le champ de cette revue, le domaine étant encore jeune. Cela nous amène à discuter sur les limites actuelles et sur les développements futurs du TQL en section 5. Enfin, nous concluons en section 6.

2. Concepts

Introduisons tout d'abord les concepts clés nécessaires à la compréhension des travaux de recherche en TQL. D'après le premier postulat de l'informatique quantique, un système physique isolé (fermé) est décrit par un vecteur *ket* $|\psi\rangle$ appartenant à un espace de Hilbert[1] à valeurs complexes, appelé espace d'état (Nielsen et Chuang, 2010). Le transconjugué complexe $|\psi\rangle^\dagger$ de $|\psi\rangle$ est noté *bra* $\langle\psi|$, aussi appelé *effet*, ainsi la quantité $\langle\phi|\psi\rangle$ est un produit scalaire tandis que l'opérateur $|\psi\rangle\langle\psi|$ est une matrice de projection sur la direction $|\psi\rangle$. Un qubit $|\psi\rangle$, ou « bit quantique », est un vecteur dans un espace d'état de dimension 2, il s'écrit comme une combinaison linéaire $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, où $\alpha, \beta \in \mathbb{C}$ et $(|0\rangle, |1\rangle)$ est une base de l'espace, $|0\rangle = (1, 0)^\intercal$ et $|1\rangle = (0, 1)^\intercal$. Il s'agit de la plus petite quantité d'information d'un système quantique. Contrairement à son équivalent classique qui ne prend que deux valeurs, 0 ou 1, le qubit est dans une *superposition* de valeurs entre les états $|0\rangle$ et $|1\rangle$. Par suite, l'espace d'état d'un système à n qubits a pour dimension $N = 2^n$ et est le produit tensoriel des n espaces à un qubit, avec pour base canonique l'ensemble des $\{|j_1\rangle \otimes \cdots \otimes |j_n\rangle \,|\, j_k \in \{0, 1\}, k = 1, \ldots, n\}$; le *i*-ème vecteur de la base est $|e_i\rangle = (\delta_{1i}, \ldots, \delta_{Ni})^\intercal$ où $\delta_{ij} = 1$ si et seulement si $i = j$, 0 sinon. Toutefois, bien qu'un $n-$qubit encode 2^n coefficients, seuls n bits d'information classique peuvent en être extraits. En effet, tout qubit décomposé dans la base des $|e_i\rangle$ s'écrit $|\psi\rangle = \sum_{i=1}^{N} \alpha_i |e_i\rangle$, $\alpha_i \in \mathbb{C}$, et le postulat de la mesure indique qu'une observation du système aura pour résultat de le figer dans l'état $|e_i\rangle$ avec la probabilité $|\langle e_i|\psi\rangle|^2 = |\alpha_i|^2$. Cela implique en particulier la condition $\sum_i |\alpha_i|^2 = 1$: les qubits sont des vecteurs unitaires.

Outre la superposition, une seconde propriété des systèmes quantiques donne l'intuition qu'ils pourraient « mieux » traiter l'information que les systèmes classiques : c'est l'*intrication*. Un n-qubit est intriqué s'il n'est pas séparable, c.-à-d. s'il ne peut

1. Espace vectoriel muni d'un produit scalaire.

pas s'écrire comme produit tensoriel de qubits de plus petite dimension[2]. Dans un système intriqué, les propriétés des qubits sont fortement corrélées, plus que ne peuvent l'être des particules classiques, ce qui ouvre *a priori* des possibilités algorithmiques inacessibles sur ordinateur classique (Nielsen et Chuang, 2010).

Enfin, l'évolution d'un système de n-qubits entre deux mesures est décrite par un opérateur linéaire U, de taille $2^n \times 2^n$, tel que $|\psi(t_0)\rangle \rightarrow |\psi(t)\rangle = U(t,t_0)\,|\psi(t_0)\rangle$. U conserve les angles et les normes (opérateur unitaire) : $U^\dagger U = UU^\dagger = I$, toute matrice unitaire de taille 2×2 applique donc une transformation par rotation d'un qubit – voir Nielsen et Chuang (2010) pour une liste des opérateurs les plus fréquents (portes NON ou contrôlée NON, matrices de Pauli, porte d'échange, etc.).

Plus généralement, un système physique peut être soit dans l'état $|\psi\rangle$ (état pur, connaissance certaine du système), soit dans un mélange statistique classique d'états quantiques, c.-à-d. dans l'état $|\psi_i\rangle$ avec la probabilité $\theta_i \in [0,1]$, $\sum_i \theta_i = 1$ (connaissance incomplète du système) – chaque $|\psi_i\rangle$ pouvant lui-même être dans une superposition d'états quantiques. Ces systèmes sont décrits par le formalisme des *matrices de densité*, plutôt que par des vecteurs d'état : pour un état pur, il s'agit de la projection $\rho = |\psi\rangle\langle\psi|$, tandis qu'elle s'écrit $\rho = \sum_i \theta_i\,|\psi_i\rangle\langle\psi_i|$ pour un mélange. On montre que tout opérateur ρ hermitien, positif[3], de trace $\mathrm{tr}(\rho) = 1$ est un opérateur de densité d'un système $\{\theta_i, |\psi_i\rangle\}$, et réciproquement – cette équivalence définit l'opérateur de densité et permet de reformuler les postulats quantiques précédemment décrits (Nielsen et Chuang, 2010). L'entropie de Von Neumann $S(\rho) := -\mathrm{tr}(\rho \log \rho)$ mesure l'incertitude associée au mélange, elle est nulle pour un état pur. Enfin, ρ attribue à tout événement $|v\rangle$ la probabilité $\mathrm{tr}(\rho\,|v\rangle\langle v|)$, et généralise les distributions de probabilité classiques, qui correspondent elles à ρ diagonale : pour une loi classique de probabilité θ_i associée à l'événement i, il suffit d'associer le vecteur $|e_i\rangle$ de la base canonique à i et de considérer $\rho_\theta = \sum_i \theta_i\,|e_i\rangle\langle e_i|$. ρ_θ est alors diagonale, et on vérifie que $\mathrm{tr}(\rho_\theta\,|e_i\rangle\langle e_i|) = \theta_i$. Nous verrons en section 3 que de nombreux travaux de TQL s'appuient sur le formalisme général des matrices de densité pour encoder des relations de dépendance entre mots et pour élaborer de nouveaux modèles.

Les notions élémentaires étant établies, décrivons finalement le *circuit quantique*, composant au cœur des ordinateurs quantiques. Un circuit quantique est un système fermé composé (i) d'un registre de n qubits, typiquement préparés dans l'état $|0\rangle \otimes \cdots \otimes |0\rangle$, (ii) d'une séquence d'opérateurs unitaires, les portes quantiques, modifiant l'état du système (iii) et enfin d'une action de mesure de l'état final du système (DiVincenzo, 2000). Pour illustration, le circuit réalisant l'algorithme de Grover est représenté en figure 1. La mesure de l'état final étant par nature aléatoire, l'exécution du circuit est répétée plusieurs fois – on parle de « tirs » – pour obtenir la probabilité du résultat. Parmi les circuits, un *circuit variationnel* (VQC, *Variational Quantum Circuit*) est un circuit quantique paramétrique, c.-à-d. contenant des portes paramétrisées

2. C'est le cas fameux de l'état de Bell $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, que l'on ne peut écrire comme produit de deux qubits – où $|00\rangle = |0\rangle \otimes |0\rangle$ et similairement pour $|11\rangle$.
3. Opérateur hermitien : $\rho^\dagger = \rho$, positivité de ρ : pour tout $|\psi\rangle$, $\langle\psi|\,\rho\,|\psi\rangle \geq 0$.

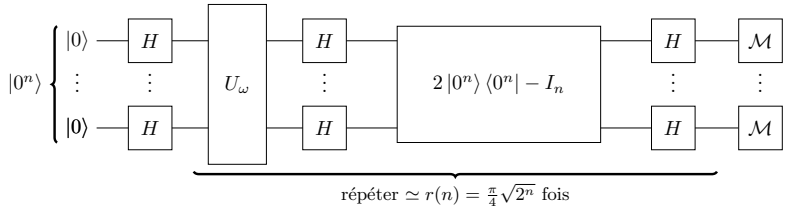

FIGURE 1. *Circuit quantique réalisant l'algorithme de Grover (1996) pour la recherche non structurée d'information : trouver l'unique $\omega \in \{0,1\}^n$ tel que $f_\omega(x) = \delta_{x\omega}$ est non nulle, $x \in \{0,1\}^n$. Un registre de n qubits dans l'état initial $|0^n\rangle = |0\rangle \otimes \cdots \otimes |0\rangle$ est mis dans la superposition de tous les qubits de la base canonique par la porte H de Hadamard. Une série d'opérateurs est répétée $r(n)$ fois, et la mesure finale donne l'état $|\omega\rangle$ avec probabilité au moins $1 - 4/2^n$.*

$U(\theta)$. Les paramètres libres θ des portes du VQC peuvent être optimisés pour minimiser une fonction objectif : typiquement, un algorithme hybride optimisera de façon classique les paramètres du VQC, p. ex. par descente de gradient, avec une fonction coût fournie par la mesure du circuit. Enfin, un *ansatz* désigne une sous-séquence d'un VQC correspondant à une certaine architecture et remplissant une certaine fonctionnalité : préparation de l'état initial, combinaisons spécifiques de portes paramétriques, etc. Il s'agit de blocs élémentaires paramétriques à utiliser pour configurer le circuit, à la manière des blocs constitutifs des réseaux neuronaux classiques (type de couche neuronale, profondeur, etc.). La configuration de l'*ansatz* est un hyperparamètre du VQC et reste donc fixe lors de l'optimisation.

Notons cependant que tout circuit n'est pas implémentable en pratique. En effet, le développement technologique actuel des ordinateurs quantiques reste contraint par le bruit quantique ; c'est l'ère du NISQ, *Noisy Intermediate Scale Quantum*, qui ne garantit pas que le système reste fermé à l'environnement extérieur. Le NISQ induit une double limitation des circuits quantiques, la première en espace, en restreignant le nombre de qubits simultanément contrôlables (de l'ordre de 100 sur les machines actuelles), la seconde en temps, en réduisant la longueur des circuits pour maintenir la cohérence des états. En outre, les machines quantiques actuelles ne sont pas capables d'implémenter toutes les portes, bien que des résultats d'universalité exacte ou approchée permettent de décomposer toute opération unitaire à n qubits en un circuit composé d'un nombre fini de portes universelles (Nielsen et Chuang, 2010). Cela ajoute ainsi une étape de compilation des circuits afin de les convertir en portes universelles admises par la machine, ce qui rallonge leur longueur effective.

3. Modèles de langue d'inspiration quantique

La théorie quantique a d'abord été employée pour produire de meilleurs modèles classiques de résolution de tâches fondamentales de TAL, sans recourir à du matériel

quantique : l'objectif est soit de mieux représenter les dépendances entre mots, soit de mieux représenter certaines relations lexicales comme l'homonymie ou l'hyponymie.

3.1. *Modéliser les dépendances entre termes : le modèle de langue quantique*

Le modèle de langue quantique (QLM, *Quantum Language Model*) de Sordoni *et al.* (2013) répond au premier objectif. Soit un vocabulaire $\mathcal{V}$ de taille n, le QLM associe un vecteur *one-hot* $|e_w\rangle$ à tout mot $w \in \mathcal{V}$, par la projection $\Pi_w = |e_w\rangle \langle e_w|$ sur le vecteur $|e_w\rangle$ de la base canonique de l'espace d'état de dimension n. Les relations de cooccurrence entre K termes $\{w_k\}_{k=1}^K$, ici, toute apparition des $\{w_k\}$ dans un ordre quelconque dans une fenêtre de taille fixe, sont modélisées par la superposition $\Pi_\kappa = |\kappa\rangle \langle\kappa|$ (état pur), $|\kappa\rangle = \sum_{k=1}^K \lambda_k |e_{w_k}\rangle$, $\lambda_k \in \mathbb{R}$, $\sum_k \lambda_k^2 = 1$. Les λ_k sont fixés, de façon uniforme ou d'après la fréquence inverse de document. Prenons par exemple $\mathcal{V}$ = *{traitement, automatique, langues}*, avec une cooccurrence des termes *traitement* et *langues* modélisée par la superposition $|\kappa_{tl}\rangle = \frac{1}{\sqrt{5}} |e_{traitement}\rangle + \frac{2}{\sqrt{5}} |e_{langues}\rangle$. Les projections $\{\Pi_i\}$ individuelles et de cooccurrence sont alors respectivement :

$$\Pi_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \Pi_a = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \Pi_l = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \Pi_{tl} = \begin{bmatrix} \frac{1}{5} & 0 & \frac{2}{5} \\ 0 & 0 & 0 \\ \frac{2}{5} & 0 & \frac{4}{5} \end{bmatrix}.$$

Un document d est ainsi décrit par l'ensemble des projections $\{\Pi_i\}$ associées à ses termes individuels et aux sous-ensembles de mots cooccurrents – sans ces dépendances, le QLM reviendrait à une loi classique sur les termes individuels (densité diagonale, section 2). L'ajout des événements de superposition permet de contrôler à quel point l'observation de la cooccurrence informe sur les constituants individuels, *via* λ_k. Enfin, la matrice de densité ρ_d de taille $n \times n$ représentant d est apprise à partir des Π_i par maximum de vraisemblance : $\rho_d = \arg\max_\rho \sum_i \log \text{tr}(\rho \Pi_i)$, telle que ρ est positive et de trace unitaire. Ce problème difficile d'optimisation est résolu par l'algorithme itératif EM (*expectation-maximization*), sans garantir la convergence rapide vers un optimum global.

Sordoni *et al.* (2013) appliquent le QLM à des tâches de recherche d'information, consistant à trouver la réponse qui correspond le mieux à une requête utilisateur. Après modélisation des documents du corpus, le modèle ρ_q d'une requête est comparé au modèle ρ_d de chaque document d'après l'entropie relative quantique $-\text{tr}(\rho_q(\log \rho_q - \log \rho_d))$, qui généralise la divergence de Kullback-Leibler et classe donc les documents par degré de proximité avec la requête. Pour conserver un temps d'exécution raisonnable[4], le nombre de mots contenus dans la requête est limité à 3. Les expérimentations sont effectuées avec le moteur de recherche *open source* Indri sur 4 jeux de données de la collection TREC, contenant de 90 257 à 50 220 423 documents. Le modèle QLM atteint des performances équivalentes ou supérieures aux

4. Il y a $2^{|Q|}$ sous-ensembles de mots possibles pour une requête de taille $|Q|$.

contreparties classiques – à savoir, une représentation en sacs de mots ou un modèle à champ aléatoire de Markov –, une première lors de la publication en 2013.

3.2. *Extensions du QLM*

Plusieurs extensions du modèle QLM ont été étudiées. Xie *et al.* (2015) reprennent le QLM en modifiant la notion de dépendance. Plutôt que la simple cooccurrence, il s'agit des ensembles de termes statistiquement corrélés entre eux, au sens d'une équivalence à la propriété d'intrication quantique. Cela permet en pratique de mieux traiter les requêtes plus longues. Une explication avancée est que ce modèle élimine les cooccurrences redondantes avec les projecteurs individuels, qui introduisent un bruit inutile pour la requête. Li *et al.* (2018) exploitent un algorithme de convergence globale du maximum de vraisemblance pour étendre la requête avec du vocabulaire hors requête. Zhang *et al.* (2018c) adaptent le modèle QLM sur la tâche de classification de sentiments. Pour cela, ils créent artificiellement deux « dictionnaires de sentiments », et représentent ces dictionnaires et les documents à classer par des matrices de densité. L'entropie quantique relative est ensuite utilisée pour mesurer la similarité entre les matrices de densité des documents et des dictionnaires. La faisabilité de l'approche est démontrée sur deux jeux de données provenant de Twitter en langue anglaise : *Obama-McCain Debate (OMD)* et *Sentiment Strength Twitter Dataset (SS-Tweet)*.

Le modèle *Neural Network based Quantum-like Language Model* (NNQLM) de Zhang *et al.* (2018a) étend quant à lui le QLM par deux aspects, sur une tâche de questions-réponses. Premièrement, afin de prendre en compte des informations sémantiques globales, la matrice de densité correspondant à une phrase est construite à partir de vecteurs de plongement de mots $\mathbf{h}_w$ issus du modèle contextuel Word2Vec *skip-gram* (Mikolov *et al.*, 2013). Un vecteur $\mathbf{h}_w$ est traité, après normalisation, comme l'observation d'un système quantique en état de superposition $|w\rangle = \mathbf{h}_w / ||\mathbf{h}_w||_2$. Une phrase $p = \{w_i\}_{i=1}^n$ est alors un mélange statistique de matrices de densité $\rho_p = \sum_i \theta_i |w_i\rangle \langle w_i|$, $\sum_i \theta_i = 1$ – p. opp. à un encodage *one-hot*, moins riche, et un état pur dans le QLM. Deuxièment, un réseau convolutionnel en deux dimensions prend pour entrée la représentation jointe $\rho_q \rho_a$ d'une question ρ_q et d'une réponse ρ_a, pour extraire automatiquement des caractéristiques de similarité entre chaque paire de documents à comparer. Une couche neuronale dense avec activation *softmax* est enfin ajoutée pour prédire si la réponse est correcte ou non. Ainsi, NNQLM intègre QLM dans un réseau de neurones, et tire profit de données annotées grâce à une mise à jour par rétropropagation. Par la suite, afin de répondre à une limite commune des modèles QLM et NNQLM, dont la représentation des phrases ne prend pas en compte l'ordre ou la position des mots (modèles en « sacs de mots »), Zhang *et al.* (2022) ont étendu NNQLM avec un réseau de neurones quantique à valeurs complexes (C-NNQLM). Le modèle est évalué sur des tâches (i) de questions-réponses (jeux de données TREC-QA, WIKIQA et YahooQA), (ii) de recherche de documents (données MS-MARCO), et (iii) de classification de textes (6 jeux de données contenant de 2 à 6 classes, des corpus de taille jusqu'à 120 k et 45 k mots de vocabulaire).

Les résultats sont supérieurs au modèle QLM, mais inférieurs aux réseaux de neurones classiques à l'état de l'art.

3.3. *Autres modèles de langue quantiques*

Un modèle de langue alternatif, qui n'utilise pas les matrices de densité, est étudié par Zhang *et al.* (2018b). Dénommé QMWF-LM, *Quantum Many-body Wave Function Language Model*, il repose sur une analogie avec le problème physique à n corps et aboutit à un modèle de réseau de neurones convolutionnel, créant ainsi un lien théorique entre formalisme quantique et certaines architectures classiques.

Dans ce modèle, chaque mot d'un document est assimilé à une particule, qui vit dans son propre espace de Hilbert de dimension m. En associant chacun des m vecteurs de base à l'un des sens du mot, cette configuration modélise plus finement la polysémie : un mot w est représenté par $|w\rangle = \sum_{i=1}^{m} \alpha_i^{(w)} |e_i^{(w)}\rangle$. La base $\{|e_i^{(w)}\rangle\}_i$ de l'espace associé à w est soit la base canonique (encodage *one-hot*), soit construite à partir de plongements de mots normalisés, option retenue par les auteurs avec la représentation GloVe (Pennington *et al.*, 2014). L'espace global – p. ex. associé à une phrase $p = \{w_i\}_{i=1}^{n}$ – est alors le produit tensoriel des n espaces de mots. Il a pour dimension m^n et est décrit par un vecteur d'état $|\psi_S\rangle = |w_1\rangle \otimes \cdots \otimes |w_n\rangle$, appelé fonction d'onde. Au-delà de cette représentation *locale* d'une phrase, les auteurs cherchent une représentation *globale* $|\psi\rangle$, qui dépend du corpus entier. En projetant la représentation globale sur la représentation locale, *via* le produit scalaire $\langle\psi_S|\psi\rangle$, ils obtiennent un score utilisable pour des tâches telles que la réponse à des questions. Toutefois, la décomposition de la fonction d'onde globale $|\psi\rangle$ dépend d'un tenseur de très grande dimension, intractable. La contribution de Zhang *et al.* (2018b) est de montrer que ce tenseur se décompose en tenseurs de faible rang, et que cette décomposition correspond à une certaine architecture de réseaux de neurones convolutionnels, donnant un nouvel éclairage sur les connexions entre théorie quantique et apprentissage classique. Des performances supérieures à QLM et à NNQLM sont obtenues sur le problème de questions-réponses sur les données TREC-QA, WIKIQA et YahooQA. La comparaison à C-NNQLM est moins nette : QMWF-LM est au moins aussi bon sur les deux premiers jeux de données, mais nettement moins performant sur YahooQA, jeu de données en plus grande dimension (près de 60 k questions et 300 k paires questions-réponses contre environ 1 k questions pour les deux autres) (Zhang *et al.*, 2022).

À noter que la décomposition tensorielle est un outil utilisé p. ex. par Ma *et al.* (2019) pour proposer une architecture moins gourmande en paramètres du mécanisme d'auto-attention des réseaux Transformers (Vaswani *et al.*, 2017), avec en prime de meilleures performances en modélisation de langue sur les jeux de données en grande dimension Penn Treebank et WikiTest-2[5]. De même Panahi *et al.* (2020) ont proposé un modèle de compression de vecteurs de plongement de mots inspiré par l'intrication quantique, réduisant ainsi le nombre de paramètres entrainables. Ce formalisme

5. Code disponible : https ://github.com/szhangtju/The-compression-of-Transformer.

appartient à la théorie des *réseaux de tenseurs*, dont les circuits quantiques sont un cas particulier (Biamonte et Bergholm, 2017). Zhang *et al.* (2019) ont ainsi étendu le modèle QMWF-LM avec le modèle TSLM, *Tensor Space Language Model*, qui apprend la probabilité conditionnelle $\mathbb{P}(w_t|w_{1:t-1})$ d'un mot w_t sachant l'historique passé $w_1, \ldots, w_{t-1}$ dans une séquence, à la façon des réseaux récurrents RNNs – cf. section 4.2. Les auteurs montrent que TSLM généralise les équations des RNNs, et obtient de bons résultats par rapport à ces derniers sur Penn Treebank et WikiTest-2. Cela illustre une nouvelle fois comment la théorie et les outils de l'informatique quantique permettent de revisiter certaines architectures classiques de réseaux de neurones.

3.4. *Modélisation de relations lexicales*

Une motivation de certains modèles de langue d'inspiration quantique est de représenter des relations lexicales, comme l'ambiguïté ou l'hyponymie. En effet, d'après Meyer et Lewis (2020), des modèles classiques contextuels représentant les mots par des vecteurs seuls peuvent gérer des cas de polysémie mais pas d'ambiguïté, telle l'homonymie. Selon Piedeleu *et al.* (2015), un mot polysémique a plusieurs sens reliés par un concept commun, tandis que deux homonymes ont des sens complètement distincts. Ainsi, le mot *bank* en anglais est polysémique, car il peut désigner une banque en tant qu'institution financière, ou le bâtiment où ces services sont délivrés. Il est aussi homonyme, puisqu'il peut désigner les berges d'une rivière.

Piedeleu *et al.* (2015) encodent un mot polysémique comme un état pur, et un homonyme comme un mélange statistique, avec une distribution de probabilité pour chaque sens. Dans les deux cas, le mot est représenté par la matrice de densité associée ρ, et l'entropie de Von Neumann $\mathrm{tr}(\rho \log \rho)$ est calculée pour mesurer l'ambiguïté, depuis un mot unique jusqu'à un constituant plus large, p. ex. de *bank* à *river bank*. Chaque mot est initialisé avec un vecteur sémantique ayant pour base les 2000 mots pleins W les plus fréquents. Les poids $v_i(t)$ du vecteur d'un mot cible t dépendent des probabilités de chaque mot contextuel $c_i \in W$ d'apparaître dans le voisinage de t. Chaque vecteur est associé à un couple $(\text{lemme}, \text{catégorie grammaticale})$. Ainsi, le verbe et le nom anglais *book* sont représentés par deux vecteurs différents. Sur 5 mots ambigüs, les auteurs montrent que chaque mot seul a une mesure d'ambiguïté plus élevée que s'il est accompagné d'autres mots, ce qui valide la modélisation.

S'appuyant à la fois sur cette approche et sur le modèle contextuel Word2Vec *skip-gram* avec échantillonage négatif (SGNS, Mikolov *et al.* (2013)), Meyer et Lewis (2020) développent le modèle de matrice de densité Word2DM. L'apprentissage par SGNS modifie les vecteurs de mots en maximisant la similarité des vecteurs des mots cooccurrents, et en minimisant celle de mots n'apparaissant pas dans le même contexte. L'algorithme est étendu pour produire des matrices de densité plutôt que des vecteurs, avec une fonction objectif spécifique pour préserver en particulier la propriété de positivité d'une matrice de densité (Meyer et Lewis, 2020). Il repose sur l'apprentissage d'une matrice intermédiaire B en calculant la matrice de densité $A = BB^\intercal$, exploitant la propriété que pour toute matrice B, le produit $BB^\intercal$ est

semi-défini positif. La fonction objectif SGNS est modifiée de la façon suivante : $J(\theta) = \log \sigma(tr(A_t A_c)) + \sum_{k=1}^{K} \log \sigma(-tr(A_t A_{w,k})))$, avec A_t et A_c les matrices de densité des mots cibles et de contexte, $A_{w,k}$ les matrices de densité de K échantillons négatifs, et θ l'ensemble des poids des matrices intermédiaires B_t, B_c et $B_{w,k}$.

Les résultats pour la version optimisée de Word2DM indiquent un bon encodage de l'ambiguïté, telle qu'évaluée d'après la corrélation entre le nombre de sens associés à ce mot (nombre de *synsets* du mot sur l'ontologie WordNet (Miller, 1995)) et son entropie de Von Neumann, définie section 2.

La représentation d'un mot par un opérateur positif – en relâchant la contrainte de trace unitaire des matrices de densité – est aussi employée par Lewis (2019) pour prendre en compte l'hyponymie, en introduisant une relation hiérarchique entre les mots. En effet, l'ensemble des opérateurs positifs d'un espace vectoriel est muni d'une relation d'ordre : pour deux opérateurs A et B, $A \sqsubseteq B \Leftrightarrow B - A$ est positif. Soit un opérateur positif $[\![mammal]\!]$ représentant le mot ***mammifère*** et un opérateur positif $[\![dog]\!]$ représentant le mot ***chien***, alors on obtient : $[\![dog]\!] \sqsubseteq [\![mammal]\!]$. Un mot w est alors vu comme une collection de vecteurs d'état $|w_i\rangle$ représentant chacun une instance du concept – un hyponyme – exprimé par w. $|w_i\rangle$ est obtenu par un plongement de mots GloVe (Pennington *et al.*, 2014) ; les relations sont construites à l'aide de la base WordNet (Miller, 1995), qui contient des relations d'hyponymie et d'hyperonymie entre mots de la langue anglaise. Ainsi, pour chaque mot w, tous les hyponymes w_i de w à chaque niveau i pour lequel il existe un vecteur GloVe sont collectés sur WordNet. Finalement, $[\![w]\!] = \sum_i p_i \, |w_i\rangle \langle w_i|$, où les poids p_i sont dérivés du texte, sans normalisation. Les représentations obtenues sont testées sur trois jeux de données en anglais composés de phrases simples contenant ou non des implications (Sadrzadeh *et al.*, 2018), ce qui est indiqué par un label positif ou négatif. Par exemple, « l'été se termine, la saison prend fin » porte un label positif, mais « la saison prend fin, l'été se termine » porte un label négatif. Les auteurs rapportent une bonne performance de leur modèle sur la tâche de détection d'implication.

4. Modèles quantiques des langues

Au-delà des approches décrites en section 3, des efforts ont été produits pour faire émerger des modèles spécifiquement quantiques de TAL. Nous revenons en détail sur l'approche diagrammatique, visant à convertir les mots et leurs relations en états quantiques au moyen de symboles et de règles, et sur l'approche par réseaux de neurones hybrides, visant à transformer certaines parties des réseaux de neurones classiques en circuits quantiques.

4.1. *L'approche diagrammatique*

4.1.1. *Principe*

L'approche diagrammatique tire ses origines des travaux théoriques de Coecke *et al.* (2010) d'inspiration quantique, définissant une représentation sémantique distributionnelle et compositionnelle du sens d'une phrase. Ils seront repris par Zeng et Coecke (2016) utilisant pour la première fois le calcul quantique appliqué au TAL, constituant les fondations du TQL. Le sens d'une phrase est calculé avec des principes mathématiques compatibles avec la physique quantique, à partir d'une représentation symbolique du lexique et de la syntaxe. Le modèle général est appelé DisCoCat d'après l'anglais *DIStributional COmpositional CATegorical*.

DisCoCat emploie une grammaire catégorielle de prégroupe (Lambek, 1997). Dans cette grammaire, un type est associé à chaque mot, et des règles de réduction s'appliquent sur les types. Un type donné p a un adjoint gauche p^l et un adjoint droit p^r, et il existe deux règles de réduction : $p^l \cdot p \to 1$ et $p \cdot p^r \to 1$. Le type n correspond aux noms et aux propositions nominales, le type s aux phrases. Le type d'un verbe transitif est alors $n^r \cdot s \cdot n^l$, signifiant qu'un type n est attendu à gauche ainsi qu'un autre à droite. La phrase est valide si les réductions successives, supprimant un type et son adjoint, aboutissent au type s.

Cette représentation grammaticale est ensuite convertie en diagramme de cordes, décrivant ici les relations grammaticales entre les mots sous forme de compositions séquentielles, admettant des entrées et produisant des sorties. De façon plus générale, les diagrammes de cordes correspondent à un langage graphique permettant de représenter et de manipuler les états quantiques. Les diagrammes expriment des calculs avec des catégories monoïdales, utilisées pour représenter des processus dans des systèmes de types variés, dont un ordinateur quantique. Une catégorie monoïdale est munie d'un bifoncteur généralisant la notion de produit tensoriel de deux structures algébriques. Cette notion est détaillée par Coecke *et al.* (2010). Un exemple de diagramme est donné sur la figure 2.

Afin de réaliser des expérimentations, telles que celles présentées dans la section suivante 4.1.2, ces diagrammes sont convertis en circuits quantiques. Pour cela, il est nécessaire de fixer certains hyper paramètres définissant l'architecture du circuit : les nombres q_n et q_s de qubits associés à chaque corde respectivement de type n et s, ainsi que les états quantiques avec lesquels tous les mots sont remplacés. Ces choix déterminent un *ansatz* – section 2 –, et sont fixés avant la phase d'entrainement. Les circuits sont entrainables : ils contiennent des paramètres libres, pouvant être optimisés pour la réalisation d'une tâche.

En complément, des travaux théoriques, considérés comme des extensions de DisCoCat par leurs auteurs, ont été proposés pour l'hyponymie (Lewis, 2019) – section 3.4 –, et l'ambiguïté syntaxique (Correia *et al.*, 2022). Pour illustrer ce type d'ambiguïté, les auteurs prennent pour exemple le groupe nominal « *rigorous mathematicians and physicists* » (« des physiciens et des mathématiciens rigoureux »). L'adjectif

« rigorous » peut porter soit sur « mathematicians and physicists », soit sur « mathematicians » uniquement. Afin de gérer ces deux représentations simultanément, elles sont placées sur un circuit commun, dans lequel des portes d'échange de deux qubits contrôlent la portée syntaxique. Un qubit $|c\rangle = c_1|1\rangle + c_2|0\rangle$ est ajouté afin de contrôler ces portes d'échange, avec les probabilités respectives $|c_1^2|$ et $|c_2^2|$ que le premier et le second sens se matérialisent.

4.1.2. *Travaux expérimentaux*

Meichanetzidis *et al.* (2023) réalisent la première expérimentation en TQL sur une machine NISQ. Ils instancient les phrases sur des VQCs au moyen du modèle DisCoCat. Le sens des mots est encodé par des états quantiques, et leurs relations de dépendance par des intrications. Les auteurs considèrent les types de la grammaire de prégroupe n et s, et un nombre de qubits q_n et q_s est associé à chaque type. Ce nombre détermine l'arité k de chaque mot w, c.-à-d. la largeur du circuit requise pour préparer l'état du mot. Les circuits correspondant au type s sont des scalaires ($q_s = 0$). Les mots unaires (1 qubit) sont représentés avec une paramétrisation d'Euler, c.-à-d. un circuit de décomposition des trois angles d'Euler $R_z(\theta_1) \circ R_x(\theta_2) \circ R_z(\theta_3)$. Ces angles peuvent décrire l'orientation d'un référentiel par rapport à un repère cartésien. Les mots avec une arité supérieure à 1 sont représentés par des circuits quantiques de type instantané à temps polynômial, en anglais *Instantaneous Quantum Polynomial-time* (IQP) (Bremner *et al.*, 2011), comportant d couches. Les portes d'un tel circuit sont commutatives car non dépendantes du temps, d'où le terme « instantané ». Un circuit IQP est constitué de portes d'Hadamard suivies d'une couche de portes de Z-rotation contrôlées $CR_z(\theta_i)$ telles que $i \in \{1, 2, ..., d(k-1)\}$. Le pronom relatif *who* en anglais (pronom relatif sujet désignant une personne) est associé à un circuit GHZ, qui n'a pas de paramètres. Ce circuit est choisi d'après les travaux de Sadrzadeh *et al.* (2013) sur les pronoms relatifs. Un état GHZ est une généralisation de l'état de Bell à trois qubits, représentant une intrication entre les qubits.

L'ensemble des θ est la paramétrisation du VQC associé à une phrase et son diagramme : ils sont optimisés de façon à résoudre le problème d'apprentissage formulé à partir de la mesure finale du circuit – ici une classification binaire.

Ces circuits sont entrainés pour une tâche de questions-réponses, sur des données synthétiques créées pour l'expérimentation. L'emploi de données synthétiques permet d'obtenir des phrases courtes, en limitant les structures syntaxiques et le vocabulaire utilisé. Les paramètres du circuit sont entrainés pour prédire une réponse correcte (0 ou 1) à une question. Par exemple la question « *Romeo loves Juliet* » (Roméo aime Juliette) porte le label de réponse 1, et « *Romeo loves Romeo* » (Roméo aime Roméo) le label 0. Les auteurs réalisent une simulation classique, et une expérimentation sur machine NISQ. Pour la simulation classique, le corpus est de 30 phrases, avec un vocabulaire de 7 mots. Sur machine NISQ, le corpus est de 16 phrases et 6 mots de vocabulaire, soit le maximum des capacités du hardware utilisé. Les entrainements utilisent l'algorithme d'optimisation SPSA, un algorithme d'approximation du gradient (Spall, 1998). Les auteurs testent 3 machines NISQ, et reportent les résultats de l'er-

reur d'entrainement et de test. L'erreur de test la plus basse est obtenue sur la machine `ibmq_montreal` avec 0 en erreur d'entrainement, et 0,375 en erreur de test.

Avec la bibliothèque Python DisCoPy (De Felice *et al.*, 2020), permettant de compiler des diagrammes en code, et donc d'utiliser le modèle DisCoCat, Lorenz *et al.* (2021) proposent une expérimentation de classification de phrases sur des machines NISQ [6]. Deux tâches sont abordées. La première est une tâche de classification binaire sur les thèmes de la nourriture et de l'informatique. Le jeu de données, appelé MC pour *Meaning Classification*, contient 130 phrases d'au plus 5 mots. C'est un jeu synthétique généré automatiquement avec une grammaire indépendante du contexte (*Context-Free Grammar, CFG* en anglais) comportant 4 règles de réécriture simples, et un vocabulaire fixe de 17 mots. Les données contiennent p. ex. les phrases « *skillful programmer creates software* » (le programmeur talentueux développe un logiciel) et « *chef prepares delicious meal* » (le chef prépare un délicieux repas).

La seconde tâche est une classification visant à prédire si une proposition nominale contient une proposition relative référant à un sujet ou à un objet. Le jeu de données, appelé RP, contient 150 propositions nominales issues du corpus RELPRON (Rimell *et al.*, 2016). La taille du vocabulaire est limitée à 115 mots, et chaque mot apparaît au moins 3 fois. Les données contiennent p. ex. les phrases : « *device that detects planets* » (« instrument qui détecte les planètes ») et « *device that observatory has* » (« instrument que l'observatoire possède »).

Ces phrases sont transformées en diagrammes avec DisCoCat. Les représentations sont similaires à celles de l'expérimentation de Meichanetzidis *et al.* (2023) décrites plus haut. Un exemple fourni par les auteurs d'un diagramme DisCoCat et de sa transformation en VQC est montré en figure 2.

Pour le paramétrage des circuits, le nombre de qubits fixé est le plus faible possible pour que l'expérimentation soit réalisable sur machine NISQ. Différentes configurations sont testées, en faisant varier les types d'effets représentant les noms, et la profondeur du circuit. Ces choix déterminent l'architecture du circuit, c.-à-d. l'*ansatz*, et donc son nombre de paramètres.

Les auteurs rapportent en résultat un score F-Mesure de 0,85 et de 0,78 pour les tâches MC et RP respectivement, selon eux conforme à ce qui pourrait être attendu pour un jeu de données de cette taille. Ils ouvrent des perspectives sur l'utilisation de jeux de données plus larges (taille de vocabulaire, nombre de phrases), comptant sur une amélioration des capacités des machines NISQ.

Afin de faciliter les expérimentations en TQL reposant sur l'approche diagrammatique, la bibliothèque Python `lambeq` (Kartsaklis *et al.*, 2021) a été développée. Elle est maintenant intégrée à la bibliothèque PennyLane pour l'informatique quantique. Le lecteur intéressé pourra trouver des tutoriels sur les sites des bibliothèques concernées. `lambeq` permet d'encoder une phrase vers un circuit quantique selon le

6. Le code source des auteurs est à disposition sur https ://github.com/CQCL/qnlp_lorenz_etal_2021_resources.

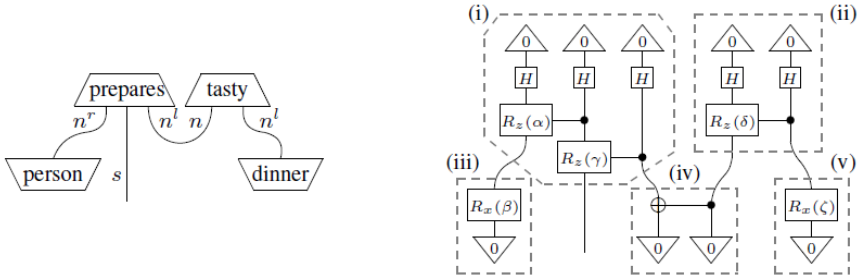

FIGURE 2. *Représentation de la phrase « person prepares tasty dinner » (la personne prépare un diner savoureux) par un diagramme DisCoCat à gauche, et par un circuit quantique à droite (Lorenz* et al.*, 2021). Les mots « prepares » et « tasty » sont remplacés respectivement par les états (i) et (ii), « person » et « dinner » respectivement par les effets (iii) et (v). La corde en forme de coupe entre « prepares » et « tasty » est traduite en intrication par effet de Bell dans le composant (iv).*

modèle DisCoCat, et d'optimiser les paramètres du circuit pour une tâche donnée. La bibliothèque DisCoPy (De Felice *et al.*, 2020) est utilisée pour le chargement des diagrammes et pour leur manipulation. Pour l'encodage, `lambeq` fournit notamment pour l'anglais un parseur à l'état de l'art obtenu par apprentissage automatique (Yoshikawa *et al.*, 2017). Il est fondé sur une grammaire catégorielle combinatoire (GCC, ou CCG en anglais pour *Combinatory Categorial Grammar*) (Steedman, 2001). Celle-ci a été préférée à la grammaire de prégroupe pour son fort pouvoir d'expressivité, et parce qu'il existait déjà des parseurs CCG automatisés. Un schéma de la chaîne de traitement d'une phrase par la bibliothèque `lambeq` est donné en figure 3.

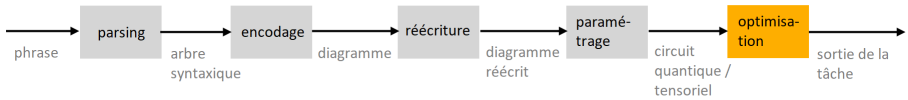

FIGURE 3. *Traitement d'une phrase avec `lambeq` (Kartsaklis* et al.*, 2021)*

4.2. *Hybridation des réseaux de neurones*

Nous avons abordé section 3.2 les approches d'inspiration quantique, avec un premier travail en 2018 (Zhang *et al.*, 2018a) sur une extension du modèle QLM. Il existe un second type d'approche fondé sur des réseaux de neurones hybrides, conçus pour être exécutés sur un ordinateur quantique ou sur une machine NISQ. Il utilise à la fois des VQCs et des étapes de calcul classique au sein d'une même architecture. Son

objectif est de tirer parti de la puissance des réseaux de neurones profonds classiques déjà démontrés pour le TAL, et des avantages théoriques de l'informatique quantique.

Les architectures à longue mémoire à court terme (en anglais *Long Short-Term Memory*, LSTM) (Hochreiter et Schmidhuber, 1997) et Transformer (Vaswani *et al.*, 2017) ont ainsi été adaptées en réseaux de neurones hybrides.

Un réseau de neurones LSTM est une extension d'un réseau de neurones récurrent (RNN), mentionné section 3.2. Un RNN possède un état caché h_t étant à la fois une entrée et une sortie du réseau qui l'actualise à chaque temps t. Cet état représente la mémoire permettant de transmettre les informations des étapes précédentes. Une limite des RNN est qu'ils ne peuvent pas transmettre des informations sur une longue distance en raison du pénomène de disparition du gradient. L'architecture LSTM pallie cette limite en intégrant une cellule de mémoire supplémentaire c_t, actualisée au temps t, permettant au gradient de persister dans le réseau.

Chen *et al.* (2022) conçoivent un modèle LSTM hybride quantique classique appelé QLSTM. Le modèle est capable d'apprendre sur plusieurs types de données temporelles sur des problèmes physiques. Dans certains cas, les performances du QLSTM sont supérieures à sa contrepartie classique, ou convergent plus rapidement. Abbaszade *et al.* (2021) se fondent sur cet avantage pour proposer un algorithme pouvant traduire des phrases de l'anglais vers le persan avec un QLSTM. Les exemples traités sont des phrases simples avec une structure grammaticale comprenant un sujet, un verbe transitif, et un objet. Les phrases sont encodées en circuit avec DisCoCat selon la méthode décrite par Lorenz *et al.* (2021). Les caractéristiques du circuit de la phrase à traduire sont passées en entrée à un QLSTM, qui les encode vers un vecteur caché appris, et les décode vers un vecteur de sortie correspondant à la phrase suivante, c.-à-d. la phrase traduite. Un QLSTM contient six VQCs, qui remplacent certaines opérations du LSTM classique. Chaque VQC est composé de trois couches : une non paramétrique d'encodage des données sous forme de qubits, une pour l'optimisation des paramètres (apprentissage des *ansätze*), une pour la mesure. Nous détaillons ci-dessous pour comparaison les équations du LSTM classique (colonne gauche) et du QLSTM (colonne droite), avec $v_t = [h_{t-1}, x^t]$, la concaténation des états cachés h au temps $t-1$ et des entrées x au temps t, $*$ l'opérateur de multiplication terme à terme et σ la fonction sigmoïde.

$$f_t = \sigma(W_f \cdot v_t + b_f) \qquad f_t = \sigma(VQC_1(v_t)) \qquad [1]$$

$$i_t = \sigma(W_i \cdot v_t + b_i) \qquad i_t = \sigma(VQC_2(v_t)) \qquad [2]$$

$$\tilde{C}_t = tanh(W_c \cdot v_t + b_C) \qquad \tilde{C}_t = tanh(VQC_3(v_t)) \qquad [3]$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{C}_t \qquad c_t = f_t * c_{t-1} + i_t * \tilde{C}_t \qquad [4]$$

$$o_t = \sigma(W_o \cdot v_t + b_o) \qquad o_t = \sigma(VQC_4(v_t)) \qquad [5]$$

$$h_t = o_t * tanh(c_t) \qquad h_t = VQC_5(o_t * tanh(c_t)) \qquad [6]$$

$$y_t = VQC_6(o_t * tanh(c_t)) \qquad [7]$$

La porte d'oubli $f_t \in [0, 1]$ (Eq. [1]) détermine dans quelle mesure les éléments correspondants dans l'état de la cellule c_{t-1} doivent être oubliés ou mémorisés (par application de $f_t * c_{t-1}$ Eq. [4]). La porte d'entrée i_t détermine quelles valeurs seront ajoutées à l'état de la cellule (Eq.[2]). $\tilde{C}_t$ est l'état de la cellule candidate (Eq. [3]) utilisé pour la mise à jour de l'état de la cellule c_t (Eq. [4]). Après la mise à jour de c_t, les sorties peuvent être calculées. o_t représente les valeurs de l'état de la cellule c_t pertinentes pour la sortie (Eq. [5]). Pour le QLSTM, o_t et c_t sont utilisés pour calculer l'état caché h_t avec le VQC_5, et la sortie y_t avec le VQC_6 (Eq. [6] et [7]).

Di Sipio *et al.* (2022) proposent une architecture alternative de LSTM hybride, dans laquelle le VQC est positionné entre deux couches classiques. Un VQC ne pouvant pas changer les dimensions des données d'entrée, la première couche adapte la dimensionalité au nombre de qubits du VQC, et la seconde adapte la dimensionalité de la sortie du VQC pour les vecteurs cachés. Le modèle est implémenté sur une tâche d'étiquetage morpho-syntaxique (*Part-Of-Speech tagging*)[7], exécutée en simulation quantique, c.-à-d. sur machine classique. Le jeu de données constitué à la main est composé de deux phrases simples en anglais : « *The dog ate the apple* » (le chien a mangé la pomme) et « *Everybody read that book* » (tout le monde a lu ce livre). Le jeu de test est identique au jeu d'entrainement. Les résultats des modèles hybrides et classiques sont similaires, avec 100 % de résultats corrects. Le LSTM hybride doit être entrainé plus longtemps en simulation que sa contrepartie classique (15 minutes contre 8 secondes), mais utilise deux fois moins de paramètres (199 contre 477).

Les auteurs proposent également une architecture de Transformer hybride. Un modèle Transformer classique (Vaswani *et al.*, 2017) gère les dépendances longues dans les textes grâce à un mécanisme d'auto-attention prenant en entrée des représentations de plongements de mots. Une couche d'auto-attention possède trois types de vecteurs : un vecteur requête Q, un vecteur clé K, et un vecteur valeur V, chacun associé à une matrice de poids entrainable. Chaque vecteur est calculé en multipliant les vecteurs de plongement de mots en entrée par la matrice entrainée correspondante. Soit d_k la taille des requêtes et des clés, les sorties de la couche d'attention sont calculées par la formule : $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$. Pour concevoir un modèle Transformer hybride, Di Sipio *et al.* (2022) remplacent les transformations linéaires du mécanisme d'attention par des VQCs. Ils réalisent une expérimentation quantique simulée avec cette architecture sur une tâche de classification de sentiments sur le corpus IMDB[8]. Les scores de performance ne sont pas rapportés. L'entrainement d'une seule itération sur tout le jeu de données – environ cent heures – s'est révélé trop long pour optimiser pleinement les paramètres du modèle.

Une approche similaire fondée sur un Transformer a été développée par Li *et al.* (2022), avec une architecture de réseau de neurones à attention quantique *quantum self-attention neural network* (QSANN) utilisant une projection gaussienne de l'auto-attention. D'après les auteurs, les travaux de Di Sipio *et al.* (2022) utilisent directement

7. Le code source des auteurs est à disposition sur https ://github.com/rdisipio/qlstm.
8. Code source à disposition sur https ://github.com/rdisipio/qtransformer.

le produit scalaire dans l'auto-attention, ce qui ne permet pas d'établir des corrélations entre des éléments lointains. La projection gaussienne utilisée dans un QSANN exploite un espace de Hilbert exponentiellement large, à même de prendre en compte des corrélations cachées entre les mots, ce qui pourrait permettre de dépasser les performances des architectures classiques.

Dans ce modèle, la couche d'auto-attention d'un Transformer classique est remplacée par une couche d'auto-attention quantique, dénommée *Quantum Self Attention Layer* (QSAL). Le schéma de cette couche est présenté sur la figure 4. Sur machine quantique, les entrées $\{y_s^{(l-1)}\}$ de la couche l sont utilisées comme angles de rotation d'*ansatz* (boîtes en ligne pointillée rouge) pour les encoder dans leur état quantique $\{|\psi_s\rangle\}$. Ensuite pour chaque état, trois classes d'*ansatz* sont exécutées : les deux premières correspondent à la requête et à la clé, la troisième correspond à la valeur. Les sorties des requêtes $\langle Z_q\rangle_s$ et des clés $\langle Z_k\rangle_j$ sont mesurées et calculées sur machine classique par une fonction gaussienne pour obtenir les coefficients d'auto-attention $\alpha_{s,j}$ (cercles verts). Toujours sur machine classique, la somme pondérée par $\alpha_{s,j}$ de la partie correspondant aux valeurs (troisième *ansatz*, petits carrés colorés) est calculée, puis les entrées y sont ajoutées pour obtenir les sorties $\{y_s^{(l)}\}$.

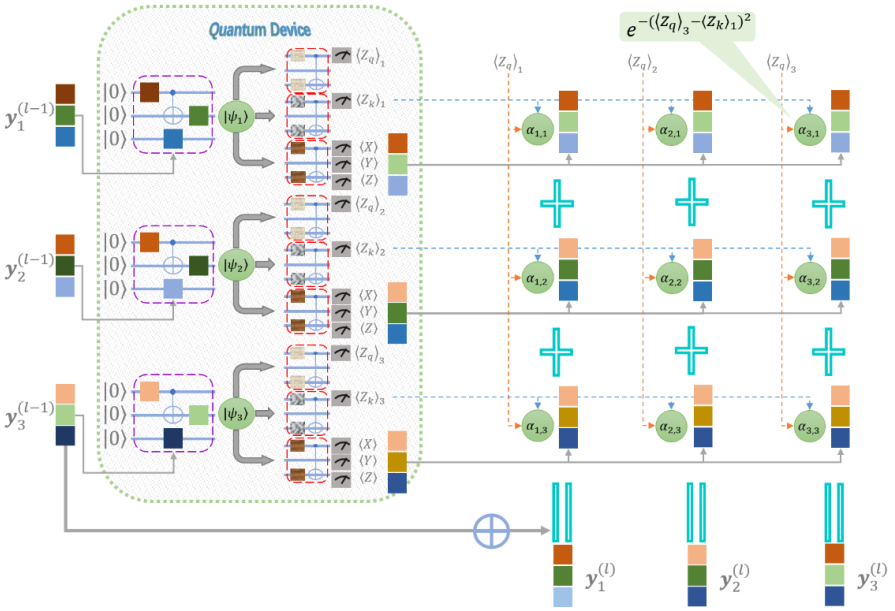

FIGURE 4. *Schéma d'une couche d'auto-attention quantique par Li* et al. *(2022)*

Les auteurs évaluent leur modèle sur deux tâches. La première est une classification thématique identique à celle réalisée par Lorenz *et al.* (2021) avec l'outil DisCoCat au moyen d'une approche diagrammatique. Les résultats obtenus sont équivalents

entre le QSANN et le modèle diagrammatique. La seconde tâche est une classification binaire de sentiments sur trois jeux de données issus du monde réel, contenant des commentaires sur des restaurants, des films et des produits, sélectionnés respectivement dans Yelp, IMDb et Amazon. Chaque jeu contient 1000 séquences avec autant d'avis négatifs que positifs, dont la longueur varie de quelques mots à plusieurs dizaines de mots. QSANN se révèle un peu plus performant que sa contrepartie classique sur les trois jeux de données, avec moins de paramètres (785 contre 49).

Yang *et al.* (2022) proposent le premier modèle hybride avec un Transformer BERT (Devlin *et al.*, 2019) pour une tâche de classification d'intention. BERT est un modèle de langue général pouvant être adapté à une tâche spécifique, entrainé en masquant partiellement des entrées que le modèle doit apprendre à prédire. Les auteurs proposent une architecture composée d'un modèle BERT préentrainé, suivi d'un décodeur quantique, qui prend en entrée les plongements de mots issus du modèle BERT. Les données classiques sont transférées sur circuit quantique au moyen d'une convolution temporelle quantique, utilisée dans des travaux sur le traitement de l'image (Henderson *et al.*, 2020) et de la voix (Yang *et al.*, 2021). Cette technique découpe le signal en fenêtres glissantes traitées par un filtre convolutionnel temporel. Les représentations latentes ainsi obtenues après filtrage sont fournies au VQC. La taille des vecteurs de plongement de mots du modèle BERT est limitée à 50, et la taille de la fenêtre glissante est de 4, ce qui détermine le nombre de qubits du circuit.

Les expérimentations sont réalisées en simulation quantique et sur machine NISQ avec deux jeux de données sur une tâche de classification d'intention : Snips (Coucke *et al.*, 2018), contenant des phrases prononcées dans un contexte d'interaction vocale avec un assistant personnel, et ATIS (AirLine Travel Information Systems) (Hemphill *et al.*, 1990), contenant des phrases prononcées correspondant à des réservations de vol en avion. Le modèle est testé sur les données texte avec différentes configurations sur le nombre et la taille des filtres. Les meilleurs scores pour les jeux Snips et ATIS sont respectivement de 96,62 % et de 96,98 % d'exemples correctement classés. Ces scores sont inférieurs à ceux pouvant être obtenus actuellement avec un modèle classique fondé sur une architecture Transformer, comme le montrent les résultats reportés par Rafiepour et Sartakhti (2023), avec respectivement 99,42 % et 98,07 % d'intentions correctement détectées pour Snips et ATIS. Ces travaux classiques incorporent le modèle BERT large, dont la taille des vecteurs de plongement de mots est de 1024. En comparaison, pour l'expérimentation quantique de Yang *et al.* (2022) la taille du modèle BERT est limitée à 50.

5. Synthèse des travaux et discussion

Nous avons présenté dans cet état de l'art des modèles d'inspiration quantique destinés à être exécutés sur machine classique, ainsi que des modèles quantiques conçus pour des simulations quantiques ou des machines NISQ. Le tableau 1 liste ces modèles quantiques, ainsi que les tâches et les données sur lesquelles ils ont été testés. Les expérimentations réalisées uniquement avec les approches diagrammatiques sont

actuellement limitées à des phrases synthétiques, tandis qu'il est déjà possible d'expérimenter sur des données naturelles avec des réseaux de neurones hybrides.

Modèle	Tâches	Données	Moyen
DisCoCat (Meichanetzidis *et al.*, 2023)	Questions-réponses	Synth.	Simulation NISQ
DisCoCat (Lorenz *et al.*, 2021)	Classification thématique	Synth. (MC)	NISQ
	Résolution d'anaphore	Synth. (RP)	
LSTM hybride (Di Sipio *et al.*, 2022)	Etiquetage morpho-syntaxique	Synth.	Simulation
Transformer hybride (Di Sipio *et al.*, 2022)	Classification de sentiment	Naturelles IMDb	Simulation
Transformer hybride QSANN (Li *et al.*, 2022)	Classification thématique	Synth. (MC)	Simulation
	Résolution d'anaphores	Synth. (RP)	
	Classification de sentiment	Naturelles Yelp IMDb Amazon	
Transformer hybride BERT-QTC (Yang *et al.*, 2022)	Classification d'intention	Naturelles. Snips, ATIS	Simulation NISQ

TABLEAU 1. *Principaux modèles en TQL exécutables sur machine NISQ et / ou en simulation quantique. La colonne « Données » indique si les données utilisées sont synthétiques, ou naturelles.*

Le recours aux données synthétiques pour les approches diagrammatiques est nécessaire afin de construire des phrases courtes en limitant leur vocabulaire et leur syntaxe. Lorenz *et al.* (2021) reportent qu'au moment de leur expérimentation, les capacités des machines NISQ n'ont pas permis d'augmenter davantage la longueur des phrases et la taille du jeu de données. Ces approches nécessitent également des étapes de prétraitement pouvant s'avérer coûteuses en temps, comme l'étiquetage morpho-syntaxique des phrases, et l'ajustement des circuits quantiques, afin que ces derniers puissent représenter le vocabulaire et la syntaxe d'une phrase tout en étant exécutables sur une machine NISQ. En pratique, des formes d'*ansatz* sont choisies de par leur disponibilité sur la machine quantique utilisée (Lorenz *et al.*, 2021). À l'inverse, les approches par réseaux de neurones hybrides utilisent les représentations par plongements de mots, permettant de prendre en compte des phrases complexes issues de jeux de données naturels. Pour autant, ces réseaux de neurones hybrides se confrontent actuellement à des temps d'exécution élevés.

Certains chercheurs (Li *et al.*, 2022) défendent l'approche par réseaux de neurones hybrides car elle permet de prendre en compte toute la complexité de la langue sans recourir à un prétraitement des données. D'autres (Correia *et al.*, 2022) souhaitent au contraire développer une approche fondée sur les règles de grammaire en TAL, pour privilégier des développements purement quantiques. Les travaux présentés dans

cet état de l'art montrent que ces deux approches peuvent être complémentaires. En effet, l'approche diagrammatique telle qu'implémentée par la bibliothèque `lambeq` (Kartsaklis *et al.*, 2021) utilise un étiquetage grammatical CCG dont la catégorie lexicale est prédite grâce à un réseau de neurones BERT (Devlin *et al.*, 2019), et le modèle LSTM hybride proposé par Abbaszade *et al.* (2021) encode les phrases avec DisCoCat, un formalisme appartenant à l'approche diagrammatique.

La diversité des modèles présentés montre que le dévelopement actuel des machines quantiques permet de réaliser des premières expérimentations, et de travailler sur de nouvelles architectures et de nouveaux algorithmes sur la base de ces résultats. Ces expérimentations restent limitées par les capacités actuelles des machines quantiques. Les gains espérés par les modèles quantiques en TQL, comme une meilleure gestion de l'ambiguïté (Meyer et Lewis, 2020), la meilleure prise en compte des dépendances longues (Li *et al.*, 2022), ou une accélération du temps de traitement, sont théoriquement fondés, mais ces avantages restent à confirmer.

Concernant les gains espérés sur la complexité algorithmique, Wiebe *et al.* (2015) présentent une variante quantique de l'algorithme du plus proche voisin, démontrant théoriquement un gain quadratique pour ce problème sous certaines conditions. Zeng et Coecke (2016) proposent une adaptation de cet algorithme dans le cadre du modèle DisCoCat, qui conserverait cet avantage. Wiebe *et al.* (2019) développent une représentation des structures syntaxiques pouvant conduire à un avantage exponentiel sur l'estimation de la validité grammaticale d'une phrase.

Toutefois, en raison du faible nombre de qubits disponibles, les performances en simulation quantique ou sur machine NISQ demeurent inférieures à celles des modèles classiques. Par exemple, pour l'expérimentation quantique de Yang *et al.* (2022), la taille des vecteurs de plongement de mots du modèle BERT est limitée à 50, alors qu'elle est de 1024 dans les expérimentations classiques, dont les scores surpassent ceux du modèle quantique. Au sujet de leur expérimentation sur machine NISQ, Lorenz *et al.* (2021) ont pour objectif explicite de rendre compte de leur démarche et des résultats obtenus afin de les partager à la communauté TAL, mais sans chercher à prouver un avantage quantique sur le temps de calcul en raison des capacités limitées des machines quantiques. Des ordinateurs quantiques de plus grande capacité seraient nécessaires pour évaluer pleinement le potentiel des modèles quantiques.

En outre, à l'ère du NISQ, une autre limitation pratique apparaît lors de l'optimisation des VQCs, même pour des problèmes de petite taille. Typiquement, ces circuits sont optimisés *via* un algorithme de descente de gradient, p. ex. SPSA (Spall, 1998), ce qui reste un défi pour deux raisons.

Tout d'abord, des chercheurs ont mis en évidence le phénomène de *plateau aride* (en anglais : *barren plateau*), consistant en l'effacement exponentiel du gradient dans toutes les directions respectivement à la taille du système quantique (McClean *et al.*, 2018). De même, Holmes *et al.* (2022) montrent qu'augmenter la capacité des *ansätze* peut réduire le gradient. Ainsi, le plateau aride aboutit au paradoxe que l'augmentation du nombre de qubits des machines NISQ peut rendre le problème d'optimisation plus

compliqué et fournir de moins bonnes solutions. Ce phénomène est observé en TQL par Niroula *et al.* (2022) sur un problème de résumé extractif de texte, avec de moins bonnes performances de l'algorithme quantique L-VQE lorsque le nombre de qubits augmente de 14 à 20. La mise au point de stratégies pour contrer les plateaux arides est un pan important de la recherche sur les VQCs, p. ex. avec de meilleures initialisations des paramètres (Grant *et al.*, 2019).

Le second défi pratique du NISQ porte sur le calcul même du gradient. Les auteurs ont recours soit comme Chen *et al.* (2022), à une expression analytique du gradient disponible pour certaines classes d'*ansatz* (Schuld *et al.*, 2019 ; Crooks, 2019), soit comme Meichanetzidis *et al.* (2023), à un schéma numérique d'approximation du gradient. Les deux cas requièrent d'évaluer au moins deux fois la fonction objectif, c.-à-d. le circuit quantique. La mise au point d'algorithmes d'optimisation robustes, au meilleur compromis entre la qualité et le coût de l'estimation, et adaptés à ces environnements stochastiques est donc une condition supplémentaire pour évaluer pleinement les concepts proposés en TQL.

6. Conclusion et perspectives

Les travaux présentés dans cet état de l'art combinent le TAL et l'informatique quantique avec deux objectifs principaux : réaliser des tâches en obtenant de meilleures performances par rapport aux modèles dits « classiques », et mieux modéliser des phénomènes linguistiques comme l'ambiguïté et les dépendances longues. Les méthodes utilisées spécifiques à l'informatique quantique, telles que les VQCs, ont motivé la conception de nouvelles approches. L'approche symbolique, dite « diagrammatique », permet de convertir un texte en circuit quantique, et l'approche par réseaux de neurones hybrides remplace certaines parties de réseaux classiques par des circuits quantiques. Les phénomènes quantiques de superposition et d'intrication sont aussi exploités par certains modèles. La spécificité de ces méthodes est à l'origine d'un nouveau champ disciplinaire : le traitement quantique des langues.

Les machines quantiques ont actuellement une capacité de calcul limitée en raison du faible nombre de qubits disponibles. Malgré cela, des expérimentations sur machines NISQ ont pu être réalisées pour quelques modèles. Même si les gains de performance sont encore difficiles à démontrer dans ces conditions contraintes, cela montre qu'il est déjà possible de concevoir des modèles quantiques en TAL, et de travailler sur l'architecture et les opérations de modèles quantiques. Les futurs travaux de recherche pourraient porter sur des extensions de modèles existants, et sur l'étude de nouvelles tâches. Par exemple Di Sipio *et al.* (2022) mentionnent que tous les blocs d'un réseau de neurones Transformer pourraient être remplacés par des VQCs, et que le modèle hybride pourrait être utilisé pour la génération automatique des langues. Il existe également des perspectives concernant le développement du TQL pour la langue française. Nous n'avons pas recensé à ce jour de travaux en TQL réalisés sur des données en français, et certaines ressources ne sont pas disponibles pour le français. Par exemple, le parseur BobCat utilisé dans la bibliothèque `lambeq` (Kartsaklis

et al., 2021) conçue pour convertir des phrases en circuits quantiques gère uniquement l'anglais et l'allemand. Enfin, il n'existe pas à notre connaissance de jeu de données commun, ou de procédure d'évaluation commune, pour comparer les modèles quantiques. Construire un tel jeu de données serait bénéfique pour les futurs développements dans ce domaine.

7. Bibliographie

Abbaszade M., Salari V., Mousavi S. S., Zomorodi M., Zhou X., « Application of quantum natural language processing for language translation », *IEEE Access*, vol. 9, p. 130434-130448, 2021.

Biamonte J., Bergholm V., « Tensor networks in a nutshell », *arXiv preprint arXiv :1708.00006*, 2017.

Bremner M. J., Jozsa R., Shepherd D. J., « Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy », *Proc. of the Royal Society A : Mathematical, Physical and Engineering Sciences*, vol. 467, n° 2126, p. 459-472, 2011.

Chen S. Y.-C., Yoo S., Fang Y.-L. L., « Quantum long short-term memory », *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 8622-8626, 2022.

Chuang I. L., Gershenfeld N., Kubinec M., « Experimental Implementation of Fast Quantum Searching », *Physical Review Letters*, vol. 80, p. 3408-3411, Apr, 1998.

Coecke B., Sadrzadeh M., Clark S., « Mathematical foundations for a compositional distributional model of meaning », *arXiv preprint arXiv :1003.4394*, 2010.

Correia A., Moortgat M., Stoof H., « Quantum computations for disambiguation and question answering », *Quantum Information Processing*, vol. 21, n° 4, p. 1-25, 2022.

Coucke A., Saade A., Ball A., Bluche T., Caulier A., Leroy D., Doumouro C., Gisselbrecht T., Caltagirone F., Lavril T. *et al.*, « Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces », *arXiv preprint arXiv :1805.10190*, 2018.

Crooks G. E., « Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition », *arXiv preprint arXiv :1905.13311*, 2019.

De Felice G., Toumi A., Coecke B., « DisCoPy : Monoidal Categories in Python », Proc. of the 3rd Annual International *Applied Category Theory Conference 2020,* Cambridge, USA, 6-10th July 2020, vol. 333 of *Electronic Proc. in Theoretical Computer Science*, Open Publishing Association, p. 183-197, 2020.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Vol. 1*, p. 4171-4186, 2019.

Di Sipio R., Huang J.-H., Chen S. Y.-C., Mangini S., Worring M., « The dawn of quantum natural language processing », *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 8612-8616, 2022.

DiVincenzo D. P., « The physical implementation of quantum computation », *Fortschritte der Physik : Progress of Physics*, vol. 48, n° 9-11, p. 771-783, 2000.

Feynman R. P., « Simulating physics with computers », *International Journal of Theoretical Physics*, vol. 21, n° 6-7, p. 467-488, 6, 1982.

Grant E., Wossnig L., Ostaszewski M., Benedetti M., « An initialization strategy for addressing barren plateaus in parametrized quantum circuits », *Quantum*, vol. 3, p. 214, 2019.

Grover L. K., « A fast quantum mechanical algorithm for database search », *Proc. of the 28th annual ACM symposium on Theory of computing, Philadelphia, PA, United States*, p. 212-219, 1996.

Harrow A. W., Hassidim A., Lloyd S., « Quantum Algorithm for Linear Systems of Equations », *Physical Review Letters*, vol. 103, n° 15, p. 150502, 2009.

Hemphill C. T., Godfrey J. J., Doddington G. R., « The ATIS spoken language systems pilot corpus », *Speech and Natural Language : Proc. of a Workshop Held at Hidden Valley, Pennsylvania*, 1990.

Henderson M., Shakya S., Pradhan S., Cook T., « Quanvolutional neural networks : powering image recognition with quantum circuits », *Quantum Machine Intelligence*, vol. 2, n° 1, p. 1-9, 2020.

Hochreiter S., Schmidhuber J., « Long short-term memory », *Neural computation*, vol. 9, n° 8, p. 1735-1780, 1997.

Holmes Z., Sharma K., Cerezo M., Coles P. J., « Connecting *Ansatz* Expressibility to Gradient Magnitudes and Barren Plateaus », *PRX Quantum*, vol. 3, p. 010313, Jan, 2022.

Kartsaklis D., Fan I., Yeung R., Pearson A., Lorenz R., Toumi A., de Felice G., Meichanetzidis K., Clark S., Coecke B., « lambeq : An efficient high-level python library for quantum NLP », *arXiv preprint arXiv :2110.04236*, 2021.

Lambek J., « Type grammar revisited », *International conference on logical aspects of computational linguistics*, Springer, p. 1-27, 1997.

Lewis M., « Modelling hyponymy for DisCo-Cat », *Proc. of the Applied Category Theory Conference, Oxford, UK*, 2019.

Li G., Zhao X., Wang X., « Quantum Self-Attention Neural Networks for Text Classification », *arXiv preprint arXiv :2205.05625*, 2022.

Li Q., Melucci M., Tiwari P., « Quantum language model-based query expansion », *Proc. of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, p. 183-186, 2018.

Lorenz R., Pearson A., Meichanetzidis K., Kartsaklis D., Coecke B., « QNLP in practice : Running compositional models of meaning on a quantum computer », *arXiv preprint arXiv :2102.12846*, 2021.

Ma X., Zhang P., Zhang S., Duan N., Hou Y., Zhou M., Song D., « A Tensorized Transformer for Language Modeling », *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (eds), *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

Madsen L. S., Laudenbach F., Askarani M. F., Rortais F., Vincent T., Bulmer J. F., Miatto F. M., Neuhaus L., Helt L. G., Collins M. J. *et al.*, « Quantum computational advantage with a programmable photonic processor », *Nature*, vol. 606, n° 7912, p. 75-81, 2022.

McClean J. R., Boixo S., Smelyianskiy V. D., Babbush R., Neven H., « Barren plateaus in quantum neural network training landscapes », *Nature Communications*, vol. 9, n° 1, p. 4812, 2018.

Meichanetzidis K., Toumi A., de Felice G., Coecke B., « Grammar-aware sentence classification on quantum computers », *Quantum Machine Intelligence*, vol. 5, n° 1, p. 1-16, 2023.

Meyer F., Lewis M., « Modelling lexical ambiguity with density matrices », *arXiv preprint arXiv :2010.05670*, 2020.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., « Distributed representations of words and phrases and their compositionality », *Advances in neural information processing systems*, 2013.

Miller G. A., « WordNet : a lexical database for English », *Communications of the ACM*, vol. 38, n° 11, p. 39-41, 1995.

Nielsen M. A., Chuang I. L., *Quantum Computation and Quantum Information : 10th Anniversary Edition*, Cambridge University Press, 2010.

Niroula P., Shaydulin R., Yalovetzky R., Minssen P., Herman D., Hu S., Pistoia M., « Constrained quantum optimization for extractive summarization on a trapped-ion quantum computer », *Scientific Reports*, vol. 12, p. 17171, 2022.

Panahi A., Saeedi S., Arodz T., « word2ket : Space-efficient Word Embeddings inspired by Quantum Entanglement », *International Conference on Learning Representations*, 2020.

Pennington J., Socher R., Manning C. D., « Glove : Global vectors for word representation », *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532-1543, 2014.

Piedeleu R., Kartsaklis D., Coecke B., Sadrzadeh M., « Open system categorical quantum semantics in natural language processing », *arXiv preprint arXiv :1502.00831*, 2015.

Rafiepour M., Sartakhti J. S., « CTRAN : CNN-Transformer-based Network for Natural Language Understanding », *arXiv preprint arXiv :2303.10606*, 2023.

Rimell L., Maillard J., Polajnar T., Clark S., « RELPRON : A relative clause evaluation data set for compositional distributional semantics », *Computational Linguistics*, vol. 42, n° 4, p. 661-701, 2016.

Rønnow T. F., Wang Z., Job J., Boixo S., Isakov S. V., Wecker D., Martinis J. M., Lidar D. A., Troyer M., « Defining and detecting quantum speedup », *Science*, vol. 345, n° 6195, p. 420-424, 7, 2014.

Sadrzadeh M., Clark S., Coecke B., « The Frobenius anatomy of word meanings I : subject and object relative pronouns », *Journal of Logic and Computation*, vol. 23, n° 6, p. 1293-1317, 2013.

Sadrzadeh M., Kartsaklis D., Balkır E., « Sentence entailment in compositional distributional semantics », *Annals of Mathematics and Artificial Intelligence*, vol. 82, p. 189-218, 2018.

Schuld M., Bergholm V., Gogolin C., Izaac J., Killoran N., « Evaluating analytic gradients on quantum hardware », *Physical Review A*, vol. 99, n° 3, p. 032331, 2019.

Shor P. W., « Algorithms for quantum computation : discrete logarithms and factoring », *Proceedings 35th Annual Symposium on Foundations of Computer Science, Singer Island, FL, United States*, IEEE, p. 124-134, 1994.

Sordoni A., Nie J.-Y., Bengio Y., « Modeling term dependencies with quantum language models for ir », *Proc. of the 36th international ACM SIGIR conference on Research and development in information retrieval*, p. 653-662, 2013.

Spall J. C., « Implementation of the simultaneous perturbation algorithm for stochastic optimization », *IEEE Transactions on aerospace and electronic systems*, vol. 34, nº 3, p. 817-823, 1998.

Steedman M., *The syntactic process*, MIT press, 2001.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., « Attention is all you need », *Advances in neural information processing systems*, 2017.

Wiebe N., Bocharov A., Smolensky P., Troyer M., Svore K. M., « Quantum language processing », *arXiv preprint arXiv :1902.05162*, 2019.

Wiebe N., Kapoor A., Svore K. M., « Quantum nearest-neighbor algorithms for machine learning », *Quantum information and computation*, vol. 15, nº 3-4, p. 318-358, 2015.

Wu S., Li J., Zhang P., Zhang Y., « Natural Language Processing Meets Quantum Physics : A Survey and Categorization », *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 3172-3182, 2021.

Xie M., Hou Y., Zhang P., Li J., Li W., Song D., « Modeling quantum entanglements in quantum language models », *24th International Joint Conference on Artificial Intelligence*, 2015.

Yang C.-H. H., Qi J., Chen S. Y.-C., Chen P.-Y., Siniscalchi S. M., Ma X., Lee C.-H., « Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition », *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6523-6527, 2021.

Yang C.-H. H., Qi J., Chen S. Y.-C., Tsao Y., Chen P.-Y., « When BERT Meets Quantum Temporal Convolution Learning for Text Classification in Heterogeneous Computing », *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8602-8606, 2022.

Yoshikawa M., Noji H., Matsumoto Y., « A* CCG parsing with a supertag and dependency factored model », *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 277-287, 2017.

Zeng W., Coecke B., « Quantum algorithms for compositional natural language processing », *arXiv preprint arXiv :1608.01406*, 2016.

Zhang L., Zhang P., Ma X., Gu S., Su Z., Song D., « A generalized language model in tensor space », *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 7450-7458, 2019.

Zhang P., Hui W., Wang B., Zhao D., Song D., Lioma C., Simonsen J. G., « Complex-valued Neural Network-based Quantum Language Models », *ACM Transactions on Information Systems (TOIS)*, vol. 40, nº 4, p. 1-31, 2022.

Zhang P., Niu J., Su Z., Wang B., Ma L., Song D., « End-to-end quantum-like language models with application to question answering », *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018a.

Zhang P., Su Z., Zhang L., Wang B., Song D., « A quantum many-body wave function inspired language modeling approach », *Proc. of the 27th ACM International Conference on Information and Knowledge Management*, p. 1303-1312, 2018b.

Zhang Y., Song D., Li X., Zhang P., « Unsupervised sentiment analysis of twitter posts using density matrix representation », *Europ. Conf. on Information Retrieval*, Springer, p. 316-329, 2018c.