Le Contexte en Traitement Automatique des Langues

Le Contexte dans le TAL Neuronal

Juan Luis Gastaldi Christian Retoré

1 Le contexte dans le TAL neuronal

Au début des années 2010, le développement de nouvelles techniques neuronales dans l'analyse des données a entraîné une profonde transformation du domaine du traitement automatique des langues (TAL). Cette transformation, qualifiée de véritable "tsunami" par l'un des principaux chercheurs du domaine (Manning, 2015), a donné un sens renouvelé à la notion de contexte en linguistique computationnelle qui, bien que décisif dans ce cadre, n'est pas forcément bien spécifié.

Plusieurs traits singularisent l'usage de la notion de contexte dans les développements récents du TAL, justifiant un traitement relativement indépendant des usages classiques dans notre présentation. À commencer par le fait qu'il ne s'agit pas tant ici de prendre en compte le contexte pour enrichir l'analyse des mots dont le sens est censé être déterminé par d'autres moyens que de *définir le sens des mots par rien d'autre que par leurs contextes linguistiques*, suivant les principes de l'école *distributionnelle* inaugurée par le linguiste américain Zelig Harris (1960). Ensuite, par la façon dont cette approche est mise en œuvre dans les modèles neuronaux actuels, *la distinction entre différents niveaux d'analyse de la langue, et donc de niveaux contextuels, est inopérante*, sinon en principe, du moins en pratique. En effet, l'apprentissage profond fait une force du fait d'analyser une quantité extraordinaire de textes pratiquement bruts et laisser à la boîte noire du modèle le travail d'identifier les éléments de tous les niveaux nécessaires à l'accomplissement de la tâche pour laquelle le modèle est entraîné. Cette stratégie s'est trouvée empiriquement validée par les résultats souvent surprenants exhibés par ces modèles dans le traitement des tâches linguistiques de la vie réelle. Pourtant, le prix à payer pour un gain de performance si drastique comparé aux modèles précédents est la perte presque aussi drastique de l'intelligibilité théorique et épistémologique des procédures d'analyse. En effet,

To be published in *Les concepts fondateurs de la philosophie du langage: Contexte*. Ed. by Béatrice Godart-Wendling. ISTE, volume series, 2023.
This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 839730.
File version: 0.0.20 février 2023.

–c'est le dernier trait par lequel on voudra distinguer ces modèles neuronaux des approches précédentes– alors que la nature des méthodes et de leur efficacité relative ne soulevait guère de doutes dans les modèles classiques, le caractère de boîte noire des réseaux de neurones profonds présente un *problème d'interprétabilité fondamental* qu'il est de plus en plus urgent de résoudre, tant du point de vue théorique que sociétal. Cela vaut tout particulièrement pour la notion de contexte, d'autant plus obscure qu'elle est censée occuper une place centrale dans la justification des méthodes neuronales. Il en résulte que, à la différence des approches précédentes, le contexte est ici plus une notion fondamentale à élucider qu'un problème circonscrit à résoudre.

Tenant compte de ces spécificités, dans cette dernière section, nous nous proposons alors de parcourir l'évolution récente des modèles d'apprentissage profond en TAL, pour en dégager la place et le sens de la notion de contexte qui se jouent dans ce cadre.

1.1 Vecteurs de mots, pierre angulaire du TAL neuronal

Comparée au traitement d'autres types de données, comme les images ou les sons, l'utilisation de réseaux de neurones profonds (RNP) [1] pour le TAL était relativement marginale jusqu'au début des années 2010, en partie en raison de la solide implantation des méthodes formelles dans le domaine de la linguistique computationnelle. Cette circonstance a commencé à changer lorsque des chercheurs du domaine ont progressivement réalisé que la première couche des modèles neuronaux pour le TAL (dite parfois couche de projection) était douée d'une signification particulière. En effet, il est apparu que si l'on extrait la première couche d'un RNP entraîné pour une tâche linguistique spécifique et qu'on l'utilise comme première couche d'un autre RNP visant une tâche linguistique différente, on peut constater une amélioration substantielle des performances de ce dernier. Ce phénomène tout à fait remarquable, a permis de faire avancer l'idée que les vecteurs résultant du traitement des mots par la première couche d'un RNP pouvaient être considérés comme des *représentations génériques* de ces mots capturant certaines de leurs caractéristiques linguistiques essentielles, en opposition à une représentation symbolique, purement atomique, des unités linguistiques comme celle des méthodes traditionnelles. En conséquence, l'idée est apparue d'entraîner cette première couche de manière séparée, indépendamment de toute tâche spécifique (*downstream task*) pour laquelle elle pourrait être utilisée par la suite, et de remplacer la représentation atomique des mots par la représentation vectorielle produite par cette couche pour chaque mot.

Pourtant, de par la nature même des méthodes d'apprentissage, l'entraînement d'un modèle produisant ces représentations vectorielles ne peut pas se faire sans recours à une tâche prédictive, aussi générique soit-elle. La question se pose, alors, de savoir quel objectif d'apprentissage est capable de produire des représentations de mots parfaitement génériques, utilisables dans le cadre d'une grande variété de tâches linguistiques spécifiques. La réponse est venue,

1. Nous assumons ici une connaissance élémentaire des RNP. Une grande quantité de présentations de ces modèles peuvent de nos jours être trouvés sur Internet (eg. Wikipedia). Le lecteur voulant obtenir une présentation plus détaillé pourra consulter Goodfellow et al. (2016) ou Brunton and Kutz (2022). Pour un livre de référence sur le TAL neuronal, cf. Goldberg (2017), et pour les vecteurs de mots, on pourra consulter Pilehvar and Camacho-Collados (2020).

précisément, d'un recours à la notion de *contexte* : il s'agira donc d'entraîner ces modèles sur une tâche prédictive rapportant chaque mot à l'ensemble de contextes linguistiques dans lequel il est susceptible d'apparaître dans un corpus donné (parfois appelé sa "distribution")

Un pas décisif dans cette direction a été franchi en 2013 avec la publication de *word2vec*, un modèle pré-entraîné de représentations vectorielles résultant d'une implémentation particulièrement efficace de cette idée (inc., 2013; Mikolov et al., 2013c,b). Word2vec comprend en réalité deux modèles, Skip-gram and CBOW, [2] représentant deux manières symétriques de traiter le rapport entre un mot et son contexte. Dans le premier cas, étant donné un mot, le modèle neuronal est entraîné à prédire les mots l'entourant dans une fenêtre de longueur donnée (typiquement 5 mots à droite et à gauche), tandis que dans le second, il s'agit pour le modèle de prédire le mot central étant donné les mots l'entourant.

Prenons, par exemple, le modèle Skip-gram. Des représentations vectorielles "one-hot" [3] sont utilisées à la fois pour saisir le mot central choisi dans le corpus et pour évaluer à la sortie les mots contextuels prédits. Une seule couche intermédiaire ou cachée est entraînée. Le vecteur d'entrée est alors multiplié par une matrice (représentation une transformation linéaire) initialisée de manière aléatoire, produisant un vecteur de faible dimension (typiquement 300 dimensions). [4] Ce vecteur est à son tour transformé de manière similaire en un vecteur de sortie d'autant de dimensions que la taille du vocabulaire, et enfin normalisé de manière à ce que ses composantes puissent être interprétées comme une distribution de probabilité sur le vocabulaire. L'erreur entre ce vecteur de sortie et chacun des vecteurs "one-hot" correspondant aux mots du contexte est ensuite utilisée pour ajuster les paramètres du modèle par un algorithme de descente du gradient connu sous le nom de rétropropagation (*backpropagation*). Une fois ce processus terminé pour une occurrence donnée d'un mot dans son contexte, l'entraînement se poursuit de la même façon avec le mot suivant du corpus, en essayant de prédire à son tour son propre contexte. Le processus est ainsi répété pour chaque mot du corpus jusqu'à ce que l'erreur atteigne un minimum stable, en recommençant depuis le début, si nécessaire, lorsque le dernier mot du corpus est atteint.

Une fois le processus d'apprentissage terminé, l'ensemble des vecteurs intermédiaires de basse dimension correspondant à chacun des mots d'entrée fournit les représentations vectorielles recherchées. Ainsi, alors qu'un mot était précédemment représenté par un vecteur "one-hot" de (très) grande dimension indexant sa place dans un vocabulaire, ce même mot pourra maintenant être représenté par un vecteur dense de basse dimension donné par la couche cachée de ce réseau.

Bien que son objectif principal ait été de fournir un algorithme efficace pour entraîner des modèles neuronaux pour le TAL, le succès et la popularité de word2vec ont marqué le triomphe des représentations vectorielles distribuées sur les méthodes traditionnelles. En effet, les modèles neuronaux entraînés pour des tâches spécifiques sur la base de vecteurs de mots construits de cette façon

2. De l'anglais *Continuous Bag Of Words*, Sac de mots continu.

3. C'est-à-dire, au moyen de vecteurs de dimension égale à la taille du vocabulaire en question, comportant des zéros partout, sauf à la place correspondant à l'indice dans le vocabulaire du mot représenté.

4. À la différence des couches classiques des réseaux neuronaux, le modèle original de word2vec ne comporte pas de biais ajouté ni de fonction d'activation.

ont rapidement surpassé de manière significative les performances des modèles existants à travers des tâches linguistiques de diverse nature ((cf. Baroni et al., 2014)). Qui plus est, indépendamment des tâches pour lesquels ils pourraient être utilisés, plusieurs travaux ont montré que ces vecteurs de mots encodent une grande quantité d'information à la fois syntaxique et sémantique correspondant aux mots qu'ils représentent. De manière plus surprenante encore, il est apparu que l'espace défini par l'ensemble de ces vecteurs (l'espace de plongement ou *embedding space*) était aussi doué de propriétés remarquables, exhibant notamment une organisation en sous-espaces selon des directions plus ou moins bien déterminées corrélées à des aspects syntaxiques ou sémantiques de la langue en entier, tels des relations analogiques entre des mots Mikolov et al. (2013c) ou des différences de degré associées à des aspects sémantiques de groupes de mots (Grand et al., 2022).

1.2 Traits généraux du TAL neuronal

Depuis cette arrivée réussie des méthodes d'apprentissage profond dans le domaine du TAL, les modèles basés sur des vecteurs de mots n'ont cessé d'évoluer à une vitesse telle qui fait que modèles comme word2vec paraissent presque trop simples. Pourtant, nous pouvons déjà reconnaître dans ces premiers modèles des traits généraux qui vont caractériser le développement de cette orientation de recherche, jusqu'à l'arrivée des Grands Modèles de Langage (*Large Language Models*) définissant l'état de l'art au moment de l'écriture de ces pages.

Le premier de ces traits est l'existence de composantes *transférables* dans l'apprentissage automatique. En effet, la capacité d'extraire des fragments d'un réseau neuronal déjà entraîné (tels la couche de projection) et de les utiliser comme des composants d'un autre dans le traitement d'une tâche différente n'a rien de trivial. Du point de vue technique, les RNPs ne sont pas directement interprétables en fonction du phénomène analysé et fonctionnent, par rapport à ces phénomènes, comme des boîtes noires (*black boxes*). Ce qui veut dire que, en principe, aucune modularité n'est à attendre dans la structure du modèle. Le fait que l'on puisse, malgré cela, transférer des composantes entre des modèles différents, indique qu'une telle modularité est envisageable. Du point de vue des phénomènes linguistiques traités par ces modèles, la transférabilité des composants suggère l'existence d'une dimension générique du langage qui serait capturée par ces composants, offrant un socle sur la base duquel des tâches linguistiques spécifiques pourraient être effectuées. Bien que les principes de ces transferts restent obscures et que ce manque soit souvent comblé par des interprétations hautement métaphoriques, leur efficacité est avérée et la pratique consistant à "pré-entraîner" un modèle générique pouvant être affiné (*fine-tuned*) par la suite en fonction des tâches ou domaines spécifiques est devenue une orientation majeure des approches neuronales du TAL.

Le second trait qui se dégage est leur *scalabilité*, c'est-à-dire, la capacité de ces modèles d'améliorer leurs performances par l'augmentation pure et simple des données traitées à l'entraînement. En effet, un des facteurs décisifs pour la réussite de word2vec a été la capacité d'implémenter l'entraînement du modèle de façon efficace grâce à des techniques comme le softmax hiérarchique ou l'échantillonnage négatif, mais aussi la parallélisation du calcul (Mikolov et al., 2013b,a). La réduction du coût computationnel résultante s'est ainsi traduite par une capacité d'entraîner des modèles sur une masse accrue de textes, ce

qui s'est révélé avoir un impact décisif sur les performances. Encore faut-il que des quantités de plus en plus significatives de données linguistiques soient disponibles pour l'entraînement. Le fait que l'objectif de l'entraînement soit celui, hautement générique, de prédire des mots étant donné leurs contextes (ou vice-versa) a permis à ces modèles de contourner le besoin de corpus soigneusement annotés à la main propres aux modèles d'apprentissage classique. La stratégie d'entraînement est alors passée de supervisée à non-supervisée, ou plus précisément "auto-supervisée", de sorte que tout texte numérique (typiquement extrait d'Internet) est devenu susceptible d'être utilisé comme donnée pour l'entraînement. Avec l'arrivée de modèles plus sophistiqués, cette tendance à l'analyse de données linguistiques de plus en plus massives s'est vue fortement confirmée par une augmentation correspondante des performances, même si les bénéfices marginaux s'avèrent être décroissants. C'est, d'ailleurs, cette tendance qui a motivé l'appellation "Grands Modèles de Langage" (GML).

Enfin, le troisième trait caractérisant ces modèles depuis ses débuts est celui du manque de transparence des principes et procédures à la base de leur performance. Dans le cas des méthodes de TAL classiques, les principes théoriques d'intelligibilité souvent précédaient et guidaient la conception des implémentations formelles, y compris dans le cas de l'apprentissage, fût-ce au prix d'un manque d'adéquation entre ces principes et les données linguistiques brutes de la vie réelle. En revanche, comme nous l'avons dit, la performance gagnée par les modèles neuronaux basés sur des représentations vectorielles se paye par un besoin d'interprétabilité inédit en comparaison aux méthodes symboliques précédentes. Ce manque de transparence devient de plus en plus grave avec l'augmentation constante d'échelle à la fois des architectures (en nombre de paramètres) et des données d'entraînement, alors que le besoin d'intelligibilité devient de plus en plus urgent du fait du déploiement massif de ces modèles dans un nombre croissant d'aspects de nos sociétés.

1.3 Vecteurs, distributions et contextes

Au croisement de ces trois circonstances caractérisant les modèles neuronaux pour le TAL depuis leur renouveau, la notion de contexte occupe un rôle décisif. Dans la mesure où les vecteurs des mots représentent les éléments constitutifs de ces modèles, y compris dans leurs versions le plus sophistiquées, le rapport entre termes et contextes linguistiques qu'ils encodent reste le noyau minimal ultime dans le traitement automatique des corpus. Et de fait, ce rapport est souvent désigné comme principe explicatif privilégié pour l'efficacité de ces modèles. Pourtant, le lien entre les capacités linguistiques de ces modèles et une notion réfléchie de contexte douée de puissance explicative n'est pas évident et son traitement demeure insuffisant dans le cadre de ces travaux.

Une réponse est pourtant invariablement avancée lorsque la question devient incontournable, à savoir l'*hypothèse distributionnelle*. Attribué originalement à Harris (1960) et condensé dans la célèbre maxime de Firth : "On reconnaît un mot à ses fréquentations !" ("*You shall know a word by the company it keeps !*") (Firth, 1957), ce principe a trouvé de multiples formulations. L'article détaillé de Sahlgren (2008, p. 33-34) en propose un échantillon assez représentatif : "les mots ayant un sens similaire apparaissent dans des contextes similaires" (Rubenstein & Goodenough) ; "les mots dont le sens est similaire apparaîtront avec des voisins similaires si suffisamment de matériel textuel est disponible"

(Schütze & Pedersen) ; “une représentation qui capture une grande partie de la façon dont les mots sont utilisés dans un contexte naturel capturera une grande partie de ce que nous entendons par sens" (Landauer & Dumais) ; “les mots qui apparaissent dans les mêmes contextes ont tendance à avoir des sens similaires" (Pantel). Il apparaît clairement, alors, que la notion de contexte occupe un rôle central dans les tentatives explicatives associées aux modèles distributionnels d'apprentissage machine comme les RNP. En d'autres termes, la notion de contexte est censée ici informer les principes d'une sémantique distributionnelle.

Pourtant, lorsque ces tentatives essaient de dépasser la simple invocation d'un principe et donner raison de l'action du contexte sur le sens des mots, les choses s'avèrent plus compliquées. Une idée couramment invoquée à ce propos est celle du sens comme “usage", généralement attribuée à Wittgenstein et soutenant que la signification linguistique est déterminée par la façon dont le langage est utilisé dans des circonstances déterminées (cf. (Manning and Schütze, 1999, p. 17), (Lenci, 2008, p. 1)). Dans sa version habituelle, ces “circonstances" d'utilisation sont associées aux contextes linguistiques qui déterminent les propriétés distributionnelles dont tirent parti les modèles neuronaux. L'invocation d'une théorie du sens comme usage pour rendre compte de l'efficacité du distributionalisme suggère alors que les locuteurs d'une langue emploient des mots dans des situations concrètes multiples et ont tendance à utiliser des mots ayant des sens similaires dans des situations similaires. Le lien entre contextes et sens est ainsi conçu comme effectué par l'intermédiaire d'un agent cognitif dont les facultés associatives dans des contextes pragmatiques concrets deviennent la source des co-occurrences statistiquement significatives. [5] Les co-occurrences présentes dans des corpus linguistiques, pour autant qu'elles reflètent ces usages, sont alors conçues comme des indicateurs (des *proxies*) pour des modèles distributionnels d'une similarité sémantique dont la source réside en dehors de ces corpus. [6]

Pourtant, cette interprétation cognitive de la théorie du sens comme usage ne semble pas faire entièrement justice au distributionalisme à l'œuvre dans les modèles neuronaux récents. Car, du point de vue cognitif, un contexte est conçu dans tous les cas comme un domaine ou une portée dans laquel des entités de même nature peuvent être présentées ensemble (*co-occur*) de sorte à être associées par un agent cognitif. Qu'il s'agisse de mots dans une portée linguistique spécifique, d'objets ou de faits dans une situation circonscrite ou de concepts dans un cadre inférentiel restreint, les contextes sont considérés comme cette région délimitée et restreinte sur fond de laquelle des agents individuels effectuent des opérations associatives. Si les contenus linguistiques sont liés aux propriétés distributionnelles des unités linguistiques, alors, d'une manière ou d'une autre, toutes ces versions fournissent une image dans laquelle ces dernières doivent être en quelque sorte corrélées avec cette faculté associative des agents individuels, et à travers elle, avec les conditions restreintes de son exercice que l'on peut alors appeler “contextes".

Pourtant, bien qu'appuyés sur l'analyse des contextes linguistiques, les modèles distributionnels effectuent un tout autre travail que celui de trouver des associations ou co-occurrences dans des contextes. Pour comprendre ce point dé-

5. Cf. Spence and Owens (1990) pour une étude classique rapportant co-occurrence et force associative.

6. Voir Lenci (2008) pour une analyse de l'approche cognitive du rapport entre distributionalisme et théorie du sens comme usage.

cisif, il est nécessaire de revenir sur ce que des modèles neuronaux de vecteurs de mots sont effectivement en train de faire. Car si les mécanismes de ces modèles sont directement opaques, des moyens indirects d'interprétabilité peuvent bel et bien être développés. C'est en particulier le cas des modèles comme word2vec. En effet, dans un article faisant suite à l'introduction de word2vec, Lévy et Goldberg (2014b) ont montré que le modèle Skip-gram pouvait être compris comme *la factorisation implicite d'une matrice terme-contexte*. Dans ces matrices, les lignes ainsi que les colonnes représentent tous les mots du vocabulaire et les valeurs à leur croisement exhibent une mesure de la capacité de l'une (typiquement celle de la colonne) à être (dans) le contexte de l'autre dans un corpus donné. [7] Plus précisément, les auteurs ont montré que les entrées de la matrice implicitement factorisée par Skip-gram correspondent a l'information mutuelle point à point (*pointwise mutual information* - PMI), décalée d'une constante globale.

Que des modèles neuronaux de vecteurs de mots soient en train de factoriser implicitement une matrice de ce type veut dire que les vecteurs de mots résultants peuvent être compris comme des lignes d'une matrice de basse dimension qui, lorsqu'elle est multipliée par une autre matrice, résulte dans une approximation de la matrice mot-contexte originale. La matrice de basse dimension résultant de la factorisation peut alors être utilisée à la place de la matrice originale, car elle encode l'information la plus essentielle de celle-ci. Qui plus est, a la suite de ces résultats, les auteurs ont montré que, en adoptant quelques-uns des hyperparamètres des modèles neuronaux concernant la définition des contextes pour un mot donné, une méthode explicite non neuronale de factorisation de matrice mot-contexte est capable d'attendre des performances comparables à celles de ces premiers modèles neuronaux (Levy et al. (2015)).

1.4 Le sens du contexte

Il apparaît donc que le secret de word2vec et des modèles similaires de plongement de mots reposant sur des architectures neuronales réside dans la manière particulière dont les distributions d'unités linguistiques dans un corpus sont connectées les unes aux autres par une relation terme-contexte, qui peut être correctement saisie par la connexion entre les lignes et les colonnes d'une matrice. Ainsi, les composantes d'un vecteur de mot ne sont rien d'autre qu'un codage efficace de la distribution globale des contextes de ce mot dans un corpus. Si les mots peuvent être représentés de manière adéquate sous forme de vecteurs denses, et si des aspects significatifs de la structure linguistique peuvent être ainsi reflétés par l'espace que ces vecteurs définissent, la raison doit alors être cherchée dans la relation que ces matrices terme-contexte entretiennent avec le langage naturel. Mais alors la notion de contexte qui semble ici à l'œuvre se distingue d'une conception cognitive d'au moins trois façons décisives.

7. Les contextes linguistiques sont normalement définis, comme dans le cas de word2vec, par une fenêtre de mots autour du mot pris comme terme. Une telle définition comporte donc de multiples paramètres : taille de la fenêtre, symétrie gauche-droite, mesure d'association en fonction de la distance au terme, filtrage des mots fonctionnels, etc. D'autres modèles existent où les colonnes de la matrice représentent, non pas des mots, mais des passages entiers (eg. paragraphes, documents, etc.). Ces derniers sont surtout utilisés pour la modélisation thématique (*topic modelling*) ou la recherche d'information (*information retrieval*). Pour un aperçu des multiples manières de définir les contextes linguistiques, on pourra consulter Sahlgren (2006, §7).

D'abord, du fait de la manière dont la similarité est déterminée dans ce cadre, notamment en mesurant la distance entre des vecteurs lignes représentant des mots comme termes [8] il apparaît que la similarité est maximale non pas lorsque des mots apparaissent dans un même contexte, mais *lorsqu'ils sont susceptibles de s'y substituer l'un l'autre*. Bien que du point de vue de l'implémentation formelle cela ne fait que peu de différence, du point de vue de la philosophie qui sous-tend ces modèles, ainsi que des principes explicatifs qu'ils réclament et qu'ils méritent, la différence est radicale : deux mots n'ont pas un sens similaire lorsqu'ils apparaissent ensemble dans le même contexte (comme deux entités qu'un agent cognitif associerait du fait de les avoir devant soi), mais précisément lorsqu'ils n'apparaissent pas et ne peuvent pas apparaître dans le même contexte *en même temps* (puisque c'est bien cela ce que substitution veut dire). Si association il y a, elle doit alors être comprise moins sous le mode de mécanismes cognitifs enregistrant la co-occurrence statistique entre des stimuli (comme le suggèrent, par exemple, Miller and Charles (1991)) que sous celui des "rapports associatifs" *in absentia* mis en avant par Saussure pour désigner la "série mnémonique virtuelle" résultant de l'effet de la langue "chez chaque individu" (1980, p. 171).

La seconde différence concernant la notion de contexte a trait à la nature immédiate de leur identité. En effet, d'un point de vue cognitif, la possibilité de reconnaître que deux entités (mots ou autre) apparaissent dans un même contexte, même lorsqu'elles n'apparaissent pas en même temps, ne semble pas soulever de grandes interrogations. Pourtant, le fait que ce ne soit pas en même temps implique la possibilité que le contexte ne soit plus, à strictement parler, le même. Cette remarque, qui pourrait sembler purement spéculative, trouve pourtant une correspondance stricte dans l'analyse des données textuelles, car il y est question d'évaluer l'occurrence de deux mots différents dans le même contexte linguistique *à deux endroits différents* du corpus. Or, la solution triviale, consistant à identifier deux contextes linguistiques si et seulement s'ils sont constitués par la même séquence de mots, ne saurait suffire, du fait de la rareté de telles séquences pour peu qu'elles soient de longueur raisonnable. [9] Cette circonstance s'accorde d'ailleurs avec des formulations plus subtiles (et plus justes) de l'hypothèse distributionnelle, comme celles de Rubenstein & Goodenough rapportée plus haut, affirmant que des mots ayant un sens similaire apparaissent dans des contextes *similaires* (et non pas nécessaires *les mêmes*). Mais alors, comment déterminer si deux contextes sont en effet similaires ? L'intelligibilité gagnée au moyen des méthodes de factorisation de matrices offre une réponse à la fois évidente et profonde à cette question : puisque dans la manipulation des matrices le traitment des lignes (mots comme termes) et des colonnes (mots comme contextes) est tout à fait équivalent, la similarité des contextes est établie de façon parfaitement analogue à celle des mots, à savoir : deux contextes sont similaires si des termes similaires y apparaissent. On comprend alors que, dans des modèles distributionnels (qu'ils soient matriciels ou neuronaux) termes et contextes sont profondément co-déterminés suivant une symétrie qui est tout à fait étrangère à la conception cognitive des contextes et au rapport que ceux-ci

8. Typiquement au moyen de la distance cosinus ou parfois tout simplement du produit scalaire.

9. Ce qui faisait Chomsky soutenir que la notion de probabilité d'une phrase ne pouvait jouer aucun rôle dans l'analyse linguistique ((cf. Chomsky, 1969)).

sont censés entretenir avec les entités qui les habitent.[10] Étant donné que les unités contextuelles (les colonnes) représentent typiquement des entités de même nature que celle des termes, c'est-à-dire des mots, plus que comme domaine ou portée habilitant l'exercice de l'association entre des mots, ces modèles invitent à penser le contexte comme une dimension interne au mot lui-même, dans un rapport dual avec sa position de terme.

Enfin, la notion de contexte résultant des modèles matriciels se distingue de l'interprétation cognitive usuelle par leur *non localité*. Car, contrairement à ce qu'une conception empirique ou pragmatique des contextes laisserait penser, ces modèles sont, en principe, capables d'établir la similarité de deux termes dont la distribution est parfaitement disjointe, autrement dit qui ne partageraient aucun contexte. Il suffit pour cela que les deux ensembles disjoints de contextes correspondants soient vus comme similaires par le modèle grâce à l'action d'autres termes dans le corpus (mais non pas des deux termes par rapport auquel ils sont disjoints). Bien que d'une façon plus confuse, cet effet avait déjà était remarqué par Landauer à propos de l'Analyse Sémantique Latente ((Landauer et al., 2007, p. 16)). L'important pour nous est que non seulement l'identité des contextes n'est pas immédiatement donnée, mais la similarité dont elle dépend relève de la structure globale tant des contextes que des termes, telle qu'elle résulte des statistiques globales d'un corpus. À la différence de la localité assumée des contextes cognitifs, les contextes distributionnels comportent une dimension globale comme condition même de leur effectivité.

Il apparaît ainsi que la notion de contexte est centrale dans les modèles distributionnels, y compris les modèles neuronaux, mais que cette centralité réclame, lorsqu'elle est regardée de plus près, une conception originale qui soit capable de rendre compte de son efficacité au sein de ces modèles. Plus que cognitifs ou pragmatiques, les contextes dans les modèles distributionnels sont *formels*. Plus qu'une propriété du monde ou de l'esprit, ils constituent une dimension interne des unités linguistiques reflétant la structure globale de la langue. Il s'agit, notamment, de cette dimension par laquelle un mot accepte d'être regardé non seulement comme une unité actuelle (un terme) mais aussi comme une virtualité agissant réellement sur toutes les autres unités actuelles de la langue. C'est sans doute ce caractère formel des contextes qui fait que le même fonctionnement que l'on constate au niveau lexical puisse être constaté également à travers des niveaux sous- et supra-lexicaux, traversant de manière continue tous les niveaux de langue.

1.5 Vecteurs contextuels

Comme il a été déjà dit, les modèles neuronaux de vecteurs de mots dont on a parlé dans la section précédente ne constituent que des modèles trop simples, dont les performances, bien que surprenantes au moment de leur apparition, restent modestes comparées aux modèles qui définissent l'état de l'art en 2022. De manière significative, ce qui caractérise ces nouveaux modèles, c'est encore un certain rapport aux contextes, car leur motivation a été avant tout celle de produire des representations vectorielles *contextuelles*.

10. La méthode la plus répondue de factorisation, à savoir SVD, produit une matrice correspondant aux termes (lignes) et une autre aux contextes (colonnes). Word2vec produit également deux matrices, mais, comme Lévy & Goldeberg (2014a) le remarquent, le modèle ne garde que celle des termes, en abandonnant celle des contextes.

En effet, dans les modèles que nous avons vus jusqu'ici, l'ensemble des contextes d'un mot à travers un corpus est mobilisé pour produire une et une seule représentation vectorielle. Pourtant, malgré le fait que l'ensemble de contextes soit tout ce qui détermine le contenu d'un mot, rien n'est dit sur le problème plus classique soulevé par le contexte linguistique, à savoir que le sens d'un même mot peut varier en fonction du contexte. L'idée est alors apparue de produire une représentation vectorielle pour chaque occurrence d'un mot en fonction de son contexte. Plutôt que de produire une représentation vectorielle statique comme dans le cas de embeddings classiques, de nouveaux modèles se sont orientés vers l'objectif d'entraîner un modèle neuronal capable de les encoder au moment de l'exécution et de les utiliser (décoder) pour des tâches spécifiques si besoin.

Quelques travaux ont été développés dans ce sens dans les années qui ont suivi l'essor des premiers modèles vectoriels ((cf. Liu et al., 2020, pour un aperçu)). Pourtant, la nature séquentielle des architectures neuronales employées à cette époque (massivement convergeant autour de modèles LSTMs) imposait des limites sévères au passage à l'échelle de la puissance de calcul, alors que la "contextualisation" des représentations vectorielles réclamait, de par sa nature, un changement d'échelle dans les données traitées pour l'entraînement des modèles.

Une véritable révolution est alors survenue avec l'introduction de l'architecture neuronale appelée "Transformeur", par Vaswani et al. (2017). La clé de cette architecture réside dans le glissement de l'accent des mécanismes de "mémoire" (propres aux architectures récurrentes, comme les LSTMs) vers des mécanismes d'"*attention*". Ces derniers étaient déjà employés dans quelques-unes des architectures de l'époque, mais toujours comme un élément auxiliaire. L'originalité des Transformeurs a été d'organiser l'architecture neuronale autour de ce mécanisme, en abandonnant la récurrence et libérant ainsi le calcul des contraintes séquentielles permettant une parallélisation s'ouvrant sur un passage à l'échelle inédit, à la fois en nombre de paramètres et en quantité de données traités à l'entraînement.

Un traitement détaillé de la complexe architecture des Transformeurs dépasse le cadre de ces pages.[11] On se contentera ici, donc, de donner l'idée générale du mécanisme d'attention. Le Transformeur reçoit en entrée une suite de mots, chacun représenté par un vecteur de mot classique (i.e., word2vec ou similaire). La clé du mécanisme d'attention consiste à transformer chacun de ces vecteurs de mots en trois vecteurs différents, appelés respectivement "requête" (*query*), "clé" (*key*) et "valeur" (*value*). On les obtient en multipliant le vecteur en entrée par trois matrices initialisées aléatoirement et dont les valeurs seront ajustées pendant l'entraînement du modèle. De manière significative, les deux premiers de ces nouveaux vecteurs, requête et clé, peuvent être interprétés comme les deux faces du mot que nous avons vu se dégager dans la section précédente, à savoir : terme et contexte. Ainsi, pour produire une représentation contextuelle d'un mot dans le contexte de la suite donnée, le vecteur de requête du mot pris comme terme est multiplié par le vecteur clé de chacun des autres mots dans le contexte (y compris celui du mot en question), produisant ainsi autant de valeurs que de mots dans le contexte. Après normalisation, ces valeurs

11. Le lecteur intéressé pourra consulter directement l'article de Vaswani et al (2017), ainsi que les multiples présentations didactiques sur Internet (cf. The Annotated Transformer ; Jay Alammar, The Illustrated Transformer [références]).

peuvent être utilisés comme mesure de l'importance de chaque mot de la suite pour le contenu du terme en question. En utilisant cette mesure, il s'agit, enfin, d'effectuer une somme pondérée des vecteurs de valeur (le troisième des vecteurs introduits) correspondant à chaque mot de la suite. De cette manière, la couche d'attention dans cette architecture produit un vecteur pour chaque mot de la suite, qui est à chaque fois le résultat d'une somme des vecteurs de valeurs de tous les mots de la suite, pondérée par des coefficients censées capturer l'importance de chaque mot pour le mot en question. Le vecteur résultant pour un mot est ainsi capable d'enregistrer de manière sélective (selon la valeur résultant du rapport requête-clé) l'information (i.e., vecteur valeur) de tous les mots dans le contexte spécifique donnée. De façon évidente, lorsqu'un mot apparaît dans deux contextes différents, le modèle calculera un vecteur différent dans chaque cas.

Les Transformeurs ne se réduisent pas à ce mécanisme élémentaire d'attention. Celui-ci est inscrit au milieu d'une architecture très sophistiquée. D'abord, une couche d'attention ne compute pas un seul vecteur d'attention, mais plusieurs (huit "têtes", dans la version originale), combinés de manière non triviale (au moyen d'une matrice, elle aussi entraînée). De plus, les vecteurs en entrée de cette couche sont préalablement enrichis d'une information positionnelle relativement à la suite, car autrement le calcul en parallèle n'aurait pas accès à cette information. Le vecteur en sortie de cette couche est encore additionné au vecteur d'input et normalisé. Ensuite, ce dernier est donné en entrée à une couche neuronale entièrement connectée, et le résultat encore additionné au vecteur d'entrée de cette dernière couche et normalisé. Enfin, l'ensemble de ces transformations constitue seulement une couche complexe de l'architecture, qui est censée se composer avec plusieurs autres couches de même nature, la sortie de l'une étant l'entrée de la suivante (six dans la version originale). Et cela uniquement pour l'encodeur, car dans sa forme première, le Transformeur comporte également un décodeur de structure et complexité similaire.

Le modèle original introduit par Vaswani et al. (2017) a été capable de montrer des améliorations significatives dans la traduction automatique des textes. Mais plus remarquablement, il a été capable de le faire après un temps d'entraînement représentant une petite fraction du temps requis par les meilleurs modèles de l'époque. Pourtant, la vraie puissance des Transformeurs s'est révélée par la suite, lorsque des implémentations spécifiques du modèle original ont été proposées, exhibant des résultats inattendus. Deux modèles méritent d'être mentionnés à ce sujet : BERT et GPT-3.

BERT (de l'anglais *Bidirectional Encoder Representations from Transformers*) a été introduit par Devlin et al. (2018). Cette implémentation a proposé une stratégie d'entraînement des Transformeurs consistant à pré-entrainer le réseau sur deux tâches génériques pour obtenir un modèle capable d'être affiné ensuite sur une multiplicité des tâches spécifiques avec un coût minimal. L'essentiel de l'apprentissage a donc lieu pendant le pré-entraînement, guidé par l'exigence de prédire des unités aléatoirement masquées au milieu d'une suite de mots, ainsi que de prédire si deux phrases données proviennent de deux phrases consécutives dans un corpus ou non. En affinant sur des tâches spécifiques ce modèle ainsi pré-entrainé, les auteurs ont été capables d'exhiber des améliorations parfois très substantielles à travers non moins de 11 tâches linguistiques différentes. Depuis, BERT a été utilisé pour le traitement des tâches les plus variées, allant de la médecine jusqu'aux mathématiques. Il est important de remarquer

que ces applications, dont les résultats sont parfois étonnants, manquent généralement de solidité à la fois théorique, technique, voire éthique, et devraient être pris plus comme des preuves de concept que comme de résultats établis. Dans tous les cas, BERT constitue de nos jours le modèle neuronal pour le TAL le plus populaire, tant dans le domaine de la recherche que des applications.

Le second modèle est la troisième version d'un Transformeur génératif pré-entrainé, appelé GPT de son nom anglais *Generative Pre-trained Transformer*, introduit par Brown et al. Brown et al. (2020). GPT-3 est un modèle de langage autoregressif, ce qui veut dire qu'il est entraîné à générer un texte, mot après mot, en prenant comme entrée une séquence des mots initialement donnée, récursivement augmentée du mot produit par le propre modèle. Cette architecture comporte plusieurs spécificités par rapport aux modèles précédents, qui ne méritent pas forcément qu'on s'y attarde ici (cf. Jay Alammar [reference]). Pourtant, deux caractéristiques singularisent GPT-3 et se trouvent à la base de sa célébrité. D'abord un étonnant passage à l'échelle. Alors que BERT comportait un maximum de 340 millons de paramètres dans sa version originale et qu'au moment de la sortie de GPT-3 le modèle le plus large en comportait 17 millard, GPT-3 était composé de 10 fois plus que ce dernier, plus précisément : 175 millard de paramètres. Le passage à l'échelle concerne également les données traitées : environ 3300 millions de mots (*tokens*) pour BERT contre presque 500 millards pour GPT-3. Cette échelle sans précédent pour un modèle de langage a déclenché une course vers des modèles de plus en plus massifs qui est devenue prohibitive en dehors des plus grandes compagnies privées du numérique (cela exclue, en particulier la recherche universitaire). D'autres problèmes et dangers ont été également associés à la taille croissante de ces modèles (cf. Bender et al. (2021)), génériquement identifiés sous le nom de Grands Modèles de Langage (LLMs, de son sigle en anglais).

Toujours est-il que ce changement d'échelle a entraîné un saut qualitatif dans le traitement des langues. Ceci est particulièrement évident lorsque l'on considère la seconde des caractéristiques singulères de GPT-3, à savoir l'*apprentissage en contexte* (*in-context learning*). En effet, GPT-3 suit une stratégie d'entraînement différente des modèles comme BERT. Au lieu de pré-entrainer un modèle de manière générique pour l'affiner par la suite, GPT-3 propose que l'entraînement s'arrête à la partie générique d'un modèle géneratif (devenue pourtant substantiellement plus massive) et que la tâche soit spécifiée, soit sous la forme d'une poignée d'exemples, soit au moyen d'une courte description, comme partie du texte proposée en entrée au modèle entraîné. Autrement dit, on peut inclure la tâche à apprendre dans ce qui, pour ces modèles, constitue le contexte linguistique à partir duquel ils établissent des propriétés linguistiques (ici, la continuation de ce contexte, sous la forme d'un modèle génératif auto-régressif). Ainsi, non seulement GPT-3 est capable de continuer une histoire de façon étonnement cohérente à partir d'un fragment de texte donnée, mais aussi, il est capable d'accomplir des tâches qui seraient soit exemplifiées soit décrites dans un texte proposé en entrée. Cette capacité est exhibée à travers une série assez diverse de tâches, telles la réponse aux questions, traduction, résolution d'anaphores, compréhension de textes, arithmétique, etc. ((cf. Brown et al., 2020)).

L'ensemble de ces résultats indique que la contextualisation des unités linguistiques, opérée sur une détermination déjà purement contextuelle de l'identité de ces unités, est capable de capturer des leviers essentiels des mécanismes des langues naturelles, au point d'aboutir à une sémantisation telle du contexte lin-

guistique que celui-ci peut être utilisé pour la spécification de tâches en langue naturelle.

Il reste que ces mécanismes demeurent profondément méconnus. Car par l'orientation que le domaine a pris, massivement gouverné par des applications et des résultats plus que par la recherche d'une intelligibilité, il a privilégié la complexité et l'augmentation de ressources à la parcimonie aux explications théoriques. De multiples tentatives d'élucidation des ressorts ultime de ces modèles sont aujourd'hui à l'œuvre, définissant un domaine très actif de recherche. Parmi les plus stimulants de ces travaux, on peut mentionner ceux de Weiss et al. (2021) et de Elhage et al. (2021). Malgré ces tentatives, la question de l'interprétabilité des modèles de vecteurs contextuels actuels reste largement ouverte.

La notion de contexte a été centrale dans l'évolution de cette ligne de recherche dans le TAL, motivant ce qui a toutes les caractéristiques d'une véritable révolution scientifique définissant, dans tous les cas, l'état de l'art dans le domaine à l'heure où l'on écrit ces pages. Pourtant, le prix à payer pour une telle efficacité a été l'obscurcissement radical de la notion même de contexte, devenue un nom générique pour parler d'une structure de la langue dont one ignore peut-être même plus qu'avant. Il reste à espérer que, comme dans toutes les révolutions scientifiques, le jour arrivera où des principes stables, même si temporaires, permettront de rendre évident la façon dont ces deux notions - contexte et structure - s'éclairent l'une l'autre.

Références

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 238–247. Association for Computational Linguistics, 2014. doi : 10.3115/v1/P14-1023.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi : 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering : Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2 edition, 2022. doi : 10.1017/9781009089517.

Noam Chomsky. *Quine's Empirical Assumptions*, pages 53–68. Springer Netherlands, Dordrecht, 1969. ISBN 978-94-010-1709-1. doi : 10.1007/978-94-010-1709-1_5. URL https://doi.org/10.1007/978-94-010-1709-1_5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https ://transformer-circuits.pub/2021/framework/index.html.

John Rupert Firth. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford, 1957.

Yoav Goldberg. *Neural Network Methods for Natural Language Processing*, volume 10. 2017. doi : 10.2200/S00762ED1V01Y201703HLT037. URL https://doi.org/10.2200/S00762ED1V01Y201703HLT037.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge MA, London UK, 2016.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. 2022. ISSN 2397-3374. doi : 10.1038/s41562-022-01316-8. URL https://doi.org/10.1038/s41562-022-01316-8.

Zellig Harris. *Structural linguistics*. University of Chicago Press, Chicago, 1960. ISBN 0226317714 0226217714.

Google inc. word2vec, https ://code.google.com/archive/p/word2vec/, 2013.

Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA, 2007.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *From context to meaning : distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics*, 1(20) :1–31, 2008.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA, 2014a. MIT Press.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180, 2014b.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3 :211–225, 2015.

Qi Liu, Matt J. Kusner, and Phil Blunsom. A survey on contextual embeddings, 2020. URL https://arxiv.org/abs/2003.07278.

Christopher D. Manning. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4) :701–707, 12 2015. ISSN 0891-2017. doi : 10.1162/COLI_a_00239. URL https://doi.org/10.1162/COLI_a_00239.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013b.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the ACL : Human Language Technologies*, pages 746–751. ACL, 2013c.

George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1) :1–28, 1991. doi : 10.1080/01690969108406936.

Mohammad Taher Pilehvar and Jose Camacho-Collados. Embeddings in natural language processing : Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4) : 1–175, November 2020. doi : 10.2200/s01057ed1v01y202009hlt047. URL https://doi.org/10.2200/s01057ed1v01y202009hlt047.

Magnus Sahlgren. *The Word-Space Model : Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. PhD thesis, Stockholm University, Stockholm, Sweden, 2006.

Magnus Sahlgren. The distributional hypothesis. *Special issue of the Italian Journal of Linguistics*, 1(20) :33–53, 2008.

Saussure. *Cours de linguistique générale*. Payot, Paris, 1980.

Donald P. Spence and Kimberly C. Owens. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19(5) :317–330, Sep 1990. ISSN 1573-6555. doi : 10.1007/BF01074363.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. *CoRR*, abs/2106.06981, 2021. URL https://arxiv.org/abs/2106.06981.