Le contexte en traitement automatique des langues

Juan Luis Gastaldi, Richard Moot, Christian Retoré

Le contexte en traitement automatique des langues

J. L. Gastaldi [*], R. Moot, Ch. Retoré

October 2022

Résumé

Cet article essaie de définir ce qu'est un contexte en traitement automatique des langues, comment se calcule le contexte et comment le contexte aide au calcul des analyses. La notion même de contexte a été profondément modifiée par les approches récentes, quantitatives, du traitement automatique du langage naturel lesquelles sont principalement motivées par la réalisation d'applications efficaces. C'est pourquoi notre article se divise en deux partie, l'une consacrée à la notion de contexte dans le TAL symbolique, qui distingue différentes notions de contexte correspondant aux différents niveaux d'analyse linguistique ; l'autre, fondée sur des algorithmes quantitatifs (des méthodes probabilistes, puis d'apprentissage automatique) conduit à des notions numériques de contexte dont le lien avec tel ou tel niveau d'analyse linguistique, moins immédiat, nécessite un travail d'interprétation. Après ces deux parties, nous conclurons par une section sur le contexte dans le TAL hybride qui voit un système de traitement en une succession d'étapes de machine learning conduisant chacune à une structure numérique intermédiaire interprétable.

[*]Ce projet a reçu un financement du Programme de recherche et d'innovation Horizon 2020 de l'Union européenne en vertu de l'accord de subvention No 839730.

Table des matières

1 Introduction : le contexte langagier dans le TAL

Cet article sur le contexte présente les représentations formelles du contexte, les méthodes de calcul de ces représentations, ainsi que de l'utilisation de ces représentations pour l'analyse automatique des divers niveaux du langage naturel. Une notion de contexte, son mode de calcul et son utilisation seront traités dans la même partie de l'article.

Le contexte d'une expression, d'une phrase ou d'un discours peut bien-sûr être linguistique mais il peut aussi être extralinguistique, y compris dans une interaction avec une machine, comme en attestent les exemples suivants :

(1) Mets au premier plan la fenêtre du bas.

(2) Allons plutôt dans ce restaurant.

(3) Nous prendrons tous une pizza, sauf lui, qui prendra une entrecôte.

Afin de rester dans ce qui est maîtrisable le présent article ne parlera que du *contexte linguistique* d'autant que la machine peine à combiner des information de nature différente (image, geste, son,...).

Nous nous sommes aussi posé la question de distinguer ou non les différents niveaux d'analyse de la langue, sachant que les approches fondées sur l'apprentissage automatique ne font guère de différences entre ces niveaux. Nous avons décider de diviser l'article en deux parties, l'une consacrée aux méthodes symboliques, l'autre aux méthodes quantitative et plus particulièrement aux méthodes neuronales (qui ont intégré les approches statistiques).

Dans la partie consacrée aux méthodes symboliques, qui traitent différemment les différents niveau d'analyse de la langue, nous présenterons séparément les différentes notions de contexte associées aux différents niveaux d'analyse de la langue (morphologie, syntaxe, sémantique, pragmatique/énonciation).

Dans la partie consacré au méthodes quantitatives, qui traite simultanément tous les niveaux d'analyse de la langue, avec des notions de contextes similaires, nous ne distinguons pas les niveaux d'analyse de la langue ni les contextes associées.

L'utilisation de méthodes hybrides mêlant méthodes quantitatives et symbolique a nécessité l'ajout d'une troisième partie sur les notions de contextes associées aux dites méthodes hybrides.

Avant d'aborder le traitement du contexte dans le cadre du traitement automatique des langues, rappelons, comme cela est expliqué et débattu dans le reste de l'ouvrage, que la prise en compte du contexte est un aspect essentiel dans l'ana-

lyse et la compréhension humaines du langage naturel. Un locuteur humain a crucialement besoin du contexte pour comprendre ce qui est dit. Le contexte désambiguise souvent, c'est-à-dire qu'il permet de choisir un ou plusieurs structures linguistiques possibles et parfois même sans lui un énoncé est incompréhensible :

(4) a. Nicolas n'a pas pu soulever son fils, car il était trop faible.
    b. Nicolas n'a pas pu soulever son fils, car il était trop lourd.

(5) Je suis en retard, désolé, j'ai crevé.

(6) a. C'est une bonne classe.
    b. La classe a été repeinte pendant les vacances.

(7) Le jambon beurre veut un demi.

Dans l'exemple 4a, on comprend que le pronom "il" fait référence à Nicolas, mais en 4b à son fils. Pourtant, les deux phrases ont la même forme et c'est surtout notre connaissance du monde qui nous permet de faire cette distinction. On remarquera que la frontière entre contexte linguistique (le sens des mots) et contexte extra linguistique, n'est pas toujours claire : dans l'exemple, l'ambiguïté est lévé grâce à notre connaissance des liens sémantiques entre "soulever", "lourd", "faible". Dans l'exemple 5, c'est le fait que "crever" en français puisse porter sur la roue d'un véhicule (vélo, voiture,...), véhicule qui est identifié à son propriétaire, et que cela a occasionné un retard. Dans l'exemple 6, il s'agit d'une ambiguïté du sens du mot "classe". Parmi les différent sens possibles, exemple 6a fait référence à un ensemble d'élèves, et exemple 6b à une salle de cours. Dans le deux cas, l'ambiguïté est levée par le contexte lexical et grammatical. La dernière phrase nécessite un contexte très particulier, comme celui d'une brasserie.

La machine qui calcule assurément mieux que nous, mais qui a davantage de difficultés à combiner des informations de nature différentes a assurément encore plus besoin que nous des informations fournies par le contexte (cf. an invitation to cognitive sciences).

Suivant le niveau d'analyse de la langue, le contexte est plus ou moins structuré : flux temporel de la parole en phonétique et en phonologie, paquets ou vecteurs de mots en analyse sémantique statistique, arbres et graphes en syntaxe, formules logiques (arbres enrichis de lien de quantification), structures discursives emboîtées. La prise en compte du contexte dans l'analyse automatique est donc différente suivant que sa structure est différente.

Bien évidemment les paquets de mots, les vecteurs de mots, et autres notions de contexte du TAL quantitatif, sont plus faciles à construire au fil du discours et

plus faciles à prendre en compte que les arbres syntaxiques, qui déterminent, par exemple, la portée d'une négation.

2 Tâches classiques du TAL

Les applications du traitement automatique des langues sont initialement apparues dans le contexte de la guerre froide, et des deux côtés du mur le principal objectif était la traduction automatique : de russe en anglais en Europe et en Amérique du Nord et très probablement de l'anglais vers le russe de l'autre côté du mur, même si nous avons moins d'informations.

Avant la fin des années 90 et le développement d'Internet, le grand public n'était pas vraiment concerné par les applications du TAL. Jusque là, le TAL était plutôt utilisé par les entreprises pour la gestion de documents internes : traduction automatique (notamment de contrats), classification et recherche de documents, l'aide à la gestion de la mémoire d'entreprise. Et certaines entreprises de service avait tout intérêt à développer le dialogue homme-machine surtout oral (serveurs vocaux accessibles par téléphone).

Aujourd'hui nombre d'applications utilisant du TAL sont devenues accessibles au grand public, notamment la traduction automatique (DeepL, Reverso context, Linguee), ou la correction grammaticale et orthographique (Word), ou tout simplement un moteur de recherche : aujourd'hui ces moteurs utilisent un peu de traitement automatique des langues, par exemple pour trouver des pages avec l'expression “production de minerai” ou "production laitière" à partir de “production minière” ou de “production laitière".

L'exemple emblématique d'une application du TAL, exemple central des origines à nos jour est la traduction automatique, apparue à la fin de la seconde guerre mondiale. Elle est devenue accessible au grand public avec Internet (avec Systran) dans les années 1990 mais la qualité des traducteurs automatiques généralistes n'était pas encore au rendez-vous. Le modèle à base de règles utilisé analysait la phrase de la langue source, sa syntaxe comme sa sémantique, pour construire une représentation du sens dans un langage pivot (abstrait), puis générait la phrase correspondante vers la langue cible. Ces systèmes nécessitaient des connaissances linguistiques assez complètes des deux langues, mais aussi une représentation de la connaissance du monde difficile à construire sans un domaine bien précis, avec ses ontologies et ses règles d'inférences. En fait, des logiciels de traduction ou d'aide à la traduction automatique de grande qualité existaient déjà, dans des domaines spécialisé, notamment un logiciel de traduction de contrat de

vente d'aéronefs de grande qualité développé par Xerox au milieu des années 90.

L'apparition ou plutôt la montée en puissance et en qualité des méthodes quantitatives (statistiques, apprentissage) par opposition aux méthodes symboliques à base de règles qui étaient les plus étudiées jusque là, a considérablement changé le domaine du TAL, ne serait ce parce que les méthodes quantitatives ne requièrent pas, ou très peu, de connaissances linguistiques, ce qui est une rupture épistémologique. Cette rupture explique que notre article soit écrit en deux parties.

Une première étape a été franchie vers le début des années 2000 avec la traduction statistique, qui en gros repère dans des corpus bilingue des expressions en vis-à-vis qui sont la traduction l'une de l'autre et qui lisse les phrases ainsi obtenues — les textes du parlement canadien et du parlement de l'union européenne fournissent une quantité impressionnante de corpus alignés.

Une seconde étape a été franchie avec l'utilisation des réseaux de neurones, vers 2010, et notamment des réseaux de neurones profonds, qui produisent des traductions encore plus impressionnantes.

Ces excellents résultats sont toutefois à relativiser, car ils reposent sur l'accessibilité de beaucoup de données pertinentes. Si les langues sont peu dotées (peu présente sur Internet), si les corpus accessibles ne correspondent pas à la langue utilisée (style, époque) la qualité du résultat se dégrade.

Avant l'arrivée des méthodes quantitatives mais encore aujourd'hui, le TAL conçoit des outils et des logiciels qui sont des composants intégrés à des applications réelles et visible, comme la recherche d'information, le dialogue homme-machine, la gestion automatique et notamment la classification de documents, les statistiques langagières à des fins linguistiques ou littéraires. Et bien sûr la traduction automatique et la correction grammaticale et orthographique, mais dans ce cas l'utilisateur a conscience qu'il y a du traitement automatique des langues là dessous !

Indépendamment des méthodes utilisées, les outils de traitement automatique des langues sont de deux sortes : la génération ou l'analyse d'énoncés en langage naturel. Dans cet article nous présentons surtout le TAL et la place du contexte dans l'analyse automatique du langage naturel écrit et nous ne parlerons ni du traitement de l'oral ni de génération.

Concernant l'oral, la transcription vers le texte utilise classiquement des modèles de Markov cachés pour la reconnaissance vocale et la prosodie pourtant porteuse d'informations utiles à la bonne compréhension est laissée de côté.

Concernant la génération, nous n'en parlons pas dans le contexte symbolique (d'autant que les grammaires formelles utilisées dans ce cadre ne font guère de différence entre analyse et production) mais nous parlerons des modèles neuro-

naux de génération tels que GPT et ses variants dans Section 4.

3 Le contexte dans le TAL symbolique

Pour les aspects historiques, qui importent dans cette collection nous suivrons les travaux de Jacqueline Léon, (28; 27; 26). En reprenant ses mots, nous pensons comme elle qu'il est difficile de définir "le traitement automatique des langues" (TAL) comme le montre la variété des dénominations : TAL, "traitement automatique du langage naturel" , "linguistique informatique" ou "linguistique computationnelle", "informatique linguistique", et en anglais "computational linguistics", "natural language processing", "natural language understanding" etc. S'agit-il d'un domaine scientifique, d'une technologie, ou d'une communauté de chercheurs et d'ingénieurs ?

3.1 La traduction automatique

Le début du TAL est apparu dès les débuts de l'informatique, juste après la seconde guerre mondiale, avec les premiers ordinateurs : le TAL, — à l'époque appelé "Machine Translation" /"Traduction Automatique" — est donc l'une des plus anciennes disciplines de l'informatique. Un premier rapport de prospective, pluôt optimiste, fut écrit par Bar-Hillel en 1951 (et parut en 1953) (3). En raison de la guerre froide, les langues considérées étaient l'anglais et le russe. Il s'agissait en particulier de pouvoir traduire les articles scientifiques, et tout particulièrement ceux de physique nucléaire. Se sont ajoutées à ses préoccupations militaires des intérêts économiques. De nos jours encore, une très large majorité des projets de TAL financé par la NSF aux USA concernent l'arabe et le chinois.

Dans ces années de guerre froide on ne s'étonnera pas que l'URSS se soit elle aussi engagée dans la traduction automatique, avec des différences, dues notamment au fait que les américains souhaitaient traduire du russe vers l'anglais, tandis que les russes souhaitaient traduire de l'anglais et du français vers le russe, mais aussi du russe vers les langues d'europe occidentale à des fins de propagande. Étant donné la diversité des langues, et les différents sens de la traduction, les russes se sont assez naturellement spécialisés dans la recherche d'une langue intermédiaire, voire d'une langue universelle. En raison du peu d'ordinateurs et de leur tradition scientifique, les recherches menées en URSS étaient plus théoriques, elle impliquaient aussi bien de la linguistique mathématique que de la linguistique comparée.

L'angleterre s'investit également à Londres et à Cambridge dans la traduction automatique à partir de 1955 et la France aussi à partir de 1959, la revue TAL (Traitement Automatique des Langues) s'appelait alors TA (Traduction Automatique) et au départ il s'agissait surtout de l'étude mécanisée du vocabulaire.

Qu'y a-t-il à retenir sur le *contexte* des premières recherches sur la traduction automatique ? Principalement les travaux du *Cambridge Language Research Unit (1955-1960)*. S'éloignant des grammaires formelles de Chomsky ou Bar-Hillel les travaux du CLRU mirent en avant : le thésaurus qui organise des sens élémentaires (le thésaurus était organisé en treillis, au sens mathématique du terme) et la grammaire vue comme un ensemble de morphèmes significatifs. Travaillant principalement sur le chinois et les pidgins ce centre a développé une vision de la langue où l'ingrédient principal est le sens lexical *en contexte*. Cela allait de pair avec ce qu'ils ont appelé "mechanical pidgin" sur un lexique anglais avec une grammaire minimale, plutôt portée par des morphèmes véhiculant une information sémantique (temps, aspect et mode font partie intégrante du thésaurus de CLRU). Plutôt que des mots, ils utilisent des "meaningfull chunks", qui peuvent être plus petits que le mots (désinences) mais aussi être constitué de plusieurs mots (expressions idiomatiques). Ces chercheurs étaient plutôt sceptiques à l'idée d'une langue pivot unique et pensaient plutôt que cette langue intermédiaire devaient être exprimée dans la langue cible.

Cette période faste de la traduction automatique se conclut par un rapport sévère de Bar-Hillel (4) sur l'impossibilité de ce qu'il nomme Fully Automatic High Quality Translation. Ce rapport contient du point de vue du présent article sur le contexte deux remarques fort intéressantes.

1. La première est qu'il donne un exemple de choses irréalisables par un traducteur totalement automatique en raison du contexte qu'on ne sait comment modéliser.

   the box was in the pen.

   ("pen" peut aussi designer un parc pour enfants)

2. La seconde concerne l'usage de méthodes statistiques (déjà présentes à l'époque) qui reposent intégralement sur le contexte et sa remarque est tout à fait pertinente, encore aujourd'hui pour les méthodes quantitatives : avec un traducteur automatique traduisant correctement 90% des phrases — une belle performance, surtout pour l'époque ! — il n'y a aucun moyen de savoir si le cas qui nous importe est dans ceux qui sont mal traduits, alors qu'avec des méthodes formelles, le problème est plutôt l'absence de proposition, et si jamais il y a erreur, il est plus facile de comprendre cette

erreur et d'en tirer partie pour améliorer le système. C'est une question encore plus pertinente avec la traduction statistique par alignement de texte ou avec les modèles neuronaux actuels, même s'ils sont très performants et que les traductions erronées sont bien plus rares. La question de Bar-Hillel reste d'actualité.

Il est vrai que les traductions automatiques sont restées de piètre qualité tant qu'elles ont utilisées des méthodes formelles et des notions linguistiques. L'approche avec un langage pivot, le sommet du triangle de Vauquois (53), vers lequel on analyse le texte en langage source, puis à partir duquel on en gendre le texte dans la semble cible, bien qu'intellectuellement séduisante, n'a pas donné les résultats escomptés. La traduction statistique, nourrie de corpus alignés apparue dans les années 90 a eu de bien meilleurs résultats. Mais c'est surtout avec l'apprentissage neuronal profond depuis 2010 que la Fully Automated High Quality Translation avec par exemple DeepL de Google a fait de grand progrès, comme on le verra ci-après. Reste néanmoins la deuxième remarque de Bar-Hillel : comment savoir que la traduction erronée, et si l'utilisateur s'en rend compte, et comment savoir d'où proviennent les erreurs ? Y compris à l'heure du TAL neuronal ces questions restent d'actualité.

3.2 Le contexte en syntaxe

Après cette période portée par l'espoir d'une traduction automatique totalement automatisée et fiable, les années 60 sont l'avènement d'une "computational linguistics" dans laquelle la syntaxe joue un rôle central.

Initialement portée par des automates, puis par des grammaires formelles, l'analyse syntaxique devient une fin en soi, et d'un point de vue du TAL l'analyse syntaxique (parsing) devient une activité à part entière d'autant que les grammaires analysant les programmes dans langages de programmation utilisent le même style de grammaires.

Les aspects algorithmiques et applicatifs de ces recherches prennent le nom de "natural language processing" incorporant aussi des techniques probabilistes.

La syntaxe est reine car les linguistes considèrent que c'est à ce niveau que s'articulent les différents domaines de la linguistique, phonologie, syntaxe, prosodie, ainsi que les nouvelles venues que sont sémantique et pragmatique jusque là relevant plutôt de la philosophie du langage.

La formalisation informatique de la syntaxe s'appuie surtout sur les grammaires classiques ou formelles qui soint des grammaires de réécriture (phrase

structure grammars) simples plutôt que sur les grammaires transformationnelles de Chomsky, dont les déplacements de constituants et les éléments vides ne conduisent pas à des algorithmes d'analyse efficaces. C'est pourquoi on peut leur préférer des grammaires formelles standards, plus simples, comme les grammaires hors contexte, éventuellement enrichies par des traits d'unification, ou alors les grammaires d'arbres adjoints, et même les grammaires catégorielles, qui après des études florissantes dans les années 60 par Bar-Hillel et al. (5), ont été remise au goût du jour lorsque les travaux de Montague (44) sur la sémantique formelle se sont diffusés.

Venues de l'autre du rideau de fer, la formalisation pour le TAL les grammaires de dépendance est une alternative fort intéressante qui ne présuppose pas un ordre des mots par défaut. Les relations de dépendance peuvent être orthogonales à l'ordre des mots (grammaires de dépendance non projectives). Il est bien plus difficile de rendre compte de la flexibilité de l'ordre des mots dans une grammaire générative usuelle, d'où la nécessité d'introduire des transformations qui compliquent l'analyse automatique.

Nous venons de distinguer deux familles de théories syntaxiques, l'une appelée grammaire générative, ou grammaire transformationnelle, développée dans en Amérique du Nord et en Europe suivant et modifiant les idées de Noam Chomsky (11) (voir par exemple (46) pour un aperçu des versions successives jusqu'au programme minimaliste) l'autre issue de la vision de Lucien Tesnière (51) et développée à Prague, en URSS (puis au Canada) (38). La grammaire générative lie fortement l'ordre des mots à la structure hiérarchique, tandis que les grammaires de dépendance les distinguent : les langues slaves ayant des cas et par conséquent un ordre des mots relativement libre, tandis que l'anglais ainsi que le français ont un ordre des mots assez strict dont se déduit la structure grammaticale. Il faudrait distinguer une troisième voie issue de la logique, aussi très liée à l'ordre des mots, les grammaires catégorielles, qui permettent de traiter au niveau syntaxique la structure sémantique (logique) de la phrase.

Pour en revenir au sujet du présent article, quelles notions de contextes sont utilisées en syntaxe ? Afin de tenter de répondre à cette question, demandons-nous quelle est la structure d'une analyse syntaxique. Le plus souvent la structure syntaxique (parse structure) est un arbre, notamment dans les grammaires génératives, mais cela peut aussi être un graphe, par exemple un graphe exprimant les dépendances.

Pour de telles structures, qu'est ce qu'un contexte, et de quoi est-il le contexte ? Dans des structures comme des termes, des arbres ou même des graphes un contexte est une structure avec un "trou" (place holder) ce "trou" pouvant être com-

blée par une autre structure du même style, arbre ou graphe. La structure la plus simple qui puisse venir combler ce "trou" est un seul sommet, qui correspond généralement à un mot. Mais on peut aussi y placer une expression, voire y insérer un constituant discontinu si le "trou" est connecté à plusieurs endroits.

Cette notion de contexte syntaxique qui décrit la position structurelle (et/ou sa position dans la suite des mots) d'une expression intervient directement dans la définition des certaines grammaires dites transformationnelles, avec une notion de déplacement, comme les grammaires génératives de Chomsky : grammaires transformationnelles, government and binding, grammaires minimalistes. Dans ce style de grammaires, le mouvement est déclenché par la position respective de deux constituants, pour expliquer le déplacement de l'un qui laisse alors une trace, un élément vide.

(8) J'ai lu [combien de livres que Chomsky a écrits].
(en forme profonde)

(9) [Combien de livres que Chomsky a écrits], ai-je lus [trace] ?
(en forme de surface)

Outre leur position respective, il est également possible de modéliser ces déplacements par l'annulation de traits de polarités opposées.

Les traits portent en eux-mêmes une notion de contexte, car ils vont avec une notion de portée. Un trait porté par la tête d'un sous arbre, porte sur tout le sous arbre dont c'est la tête. Mais les traits définissent aussi un ensemble de mots, souvent un constituant, dont le spécifieur et la tête portent effectivement ce trait.

Nous reparlerons ci-après des notions de contexte à la frontière entre syntaxe et sémantique notamment les opérations logiques qui ont une portée, comme les quantificateurs ou la négation (qui régit l'utilisation des items de polarité négative comme "moindre" ou "grand-chose" en français ou "any" ou "ever" en anglais)

D'un point de vue pratique, le contexte en syntaxe permet de trouver la catégorie syntaxique d'un mot, de lever des ambiguïtés afin d'analyser correctement la phrase. Par exemple, la phrase 10a permet deux interprétations. Dans la première interprétation (10b), le verbe est "briser" et "la glace" est analysé comme un déterminant suivi d'un nom commun. Dans la deuxième interprétation (10c), "la glace" est interprété comme un pronom suivi d'un verbe.

(10) a. La petite brise la glace.
b. [$_{Suj}$ La petite] [$_{V}$brise] [$_{Obj}$la glace].
c. [$_{Suj}$La petite brise] [$_{Obj}$la] [$_{V}$glace].

Les techniques standard pour réaliser cet étiquetage grammatical s'appuient toutes sur la notion de contexte : les règles de Brill (8) qui attribue au mot des catégories possibles et au "trou" du contexte, règles qui sont corrigées en prenant en compte le contexte local ; les modèles de Markov cachés qui attribuent la probabilité de voir un mot avec un catégorie donnée, une probabilité qu'une catégorie soit suivie d'une catégorie (il s'agit d'un modèle probabiliste auto-apprenant par renforcement par ré-estimation des probabilités au vu des exemples) (35; 36) ; les réseaux de neurones qui sont très performants pour cette tâche, y compris lorsque les catégories sont des termes complexes, cf. ci-après.

3.3 Le contexte en sémantique et pragmatique

Dans cette sous section nous parlons pour commencer de structure sémantique et discursive (formal semantics) tandis que la sémantique lexicale qui a des notions de contexte fort différentes, sera discutée dans la prochaine sous section.

Dans ces interactions syntaxe sémantique le liage (binding, notion introduite par Chomsky, voir par exemple (46)) qui régit le calcul des antécédents possible d'un pronom réfléchi ou non, dépend des positions *syntaxique* respectives du pronom et du possible référent et donc du contexte syntaxique de ces deux mots.

(11) a. Paul n'aime pas qu'il le regarde.
b. il≠Paul et (le=Paul ou le ≠Paul les deux sont possibles)

(12) a. Paul n'aime pas qu'il se regarde.
b. il≠Paul et se=il

La sémantique de la phrase comporte aussi une notion de contexte intraphrastique (qui peut grâce au pronoms sortir de la phrase). Ce contexte résulte des opérateurs logique, de négation, de quantification qui ont une portée déterminée par la structure syntaxique. Dans les grammaires catégorielles, on peut même dire, et c'est ce qui fait leur intérêt, que la structure syntaxique calculée par une grammaire catégorielle, un lambda-terme coïncide avec sa structure sémantique : il suffit d'insérer à leur place les lambda-termes lexicaux. Sous la portée d'une négation il faut utiliser les items de polarité négative, et en dehors, c'est impossible. La référence des pronoms est affectée par la quantification, ainsi "il" ne permet pas de référer à la variable liée par "tout étudiant"' : [1]

1. Notons que le pluriel "ils" passe mieux dans l'exemple 14b, car on réfère à l'ensemble des entités sur lesquelles porte la quantification.

(13) a. Tout étudiant à un badge pour entrer dans les bâtiments de l'université, mais pas dans les salles.

b. * Ainsi il peut accéder aux couloirs s'il pleut, mais ne peut entrer dans une salle en l'absence de professeur.

(14) a. Aucun coureur qui s'est entraîné n'a abandonné.

b. Aucun coureur n'a abandonné. * Il s'est entraîné.

Dans le second exemple, c'est encore plus frappant : "il" ne peut référer à "aucun étudiant" s'il est en dehors de la proposition

La structure sémantique la plus commune est une formule logique, obtenue par composition des éléments de sens suivant la structure syntaxique de la phrase. Cette vision est une adaptation par Montague du principe de compositionalité de Frege. Formellement, l'analyse syntaxique est vue comme un lambda-terme (qui est plus aisément obtenu s'il s'agit d'une analyse faite dans une grammaire catégorielle), et le sens logique des mots aussi (par exemple "elle-même" s'applique à une relation entre deux entités $P(x,y)$ et construit la propriété d'une entité $P(x,x)$. La réduction de ce lambda-terme donne une formule logique qui correspond à la phrase. Hormis la portée des opérations logiques il n'y a pas de notion spécifique de contexte lié à ce niveau d'analyse.

(15) a. si (un paysan possède un âne) alors (il le bat)

b. pour tout paysan pour tout âne si le paysan possède l'âne alors ce paysan bat cet âne

C'est en remarquant les limites de la compositionnalité avec des phrases comme l'exemple classique ci-dessus qu'est apparue la nécessité de prendre en compte le contexte pour identifier les référents possibles des pronoms qui peuvent être dans une autre sous structure logique que le pronom.

La structure la plus communément utilisée par les linguistes pour formaliser les référents de discours et la sémantique logique d'une ou plusieurs phrases est la DRT (Discourse Representation Theory) de Kamp and Reyle (23) qui inclut une notion de contexte dans les structures utilisées pour représenter les analyses. Celles-ci, appelées Discourse Représentation structures, sont des boîtes possiblement imbriquées. Une boites est constituées de deux parties. Une partie est dévolue aux référents de discours possibles : entités, événements et instants. L'autre partie de la boîte contient des opérations logiques entre boîtes (négation, disjonction, implication) et des formules logiques atomiques (c'est-à-dire des prédicats tel que $paysan(x)$ et $bat(x,y)$). A partir de la structure des boîtes, une relation appelée accessibilité permet de savoir quel est le contexte à l'intérieur d'une boite,

c'est-à-dire quels sont les référents de discours (entités, événements, instants) accessibles à cet endroit là. Pour simplifier, disons que de l'extérieur d'une boite, on ne peut accéder aux référents qui sont à l'intérieur, tandis que de l'intérieur d'une boite on peut accéder aux référents des boites qui le contiennent.

En DRT, l'interprétation d'une phrase correspond à un mise-à-jour du contexte. Un article indéfini comme "un" rajoute un nouveau référent au contexte, tandis qu'un négation supprime des référents. Les pronoms tels que "il" et "elle" sélectionnent un référent du contexte. Ceci explique le contraste entre 14a, où un référent correspondant à "un coureur" est introduit qui est accessible pour pour le pronom "il" et 14b où "aucun" introduit un référent correspondant à un coureur et ce référent n'est accessible que localement, mais pas à l'extérieur de la phrase. C'est la même chose avec "tout étudiant" et "il" dans les exemples 13a et 13b, même si la phrase incorrecte est davantage acceptable dans ce cas là.

Un modèle formel appelé S-DRT (Segmented Discourse representation Theory) (2) étend la DRT à la structure du discours, en structurant les relations qu'entretiennent entre-elles les propositions (segments) du discours : narration, elaboration concession,... C'est un arbre (et donc une structure hiérachique récursive) dont les noeuds sont des DRS (de la DRT) qui peut en plus des relations de narration avoir des relation de narration (succession temporelle dans un récit) de la gauche vers la droite entre noeuds de même niveau. Des contraintes de cohérence impose par exemple de faire référence (pronom, sens particulier d'un mot, phrase entière) à des expressions situées à la frontière droite d'une sous-structure. Il y la une notion de contexte fort intéréssante correspondant à ce à quoi il est possible de se référer à tel ou tel moment du discours. Cependant, on ne peut pas dire qu'il y ait d'analyse automatique du discours avec la S-DRT, pas plus qu'avec le formalisme similaire, plus ancien et moins riche appelé RST (Rhetorical Structure Theory). (33)

3.4 Le contexte en sémantique lexicale

La sémantique lexicale concerne le sens des mots. La linguistique l'associe plutôt aux autres propriétés du mot (et notamment à la morphologie dérivationnelle, et aux propriétés grammaticales du mot) qu'à la sémantique de la phrase.

Bien sûr, c'est le contexte aussi bien grammatical que sémantique (global ou local) qui choisit le sens d'un mot polysémique.

(16) L'avocat a raté sa plaidoirie.

(17) La classe a été repeinte pendant les vacances.

(18) Le hautbois donne le la.

La thématique générale permet d'exclure l'avocat-fruit dans le premier exemple. Dans le deuxième exemple, même dans un contexte scolaire, le sens de "la classe" n'est pas déterminé : il est obtenu en prenant en compte le contexte local et "repeindre" présuppose que sont objet soit un objet physique inanimé. Le troisème et dernier exemple est purement grammatical, un modèle de Markov caché le traite aisément.

Les techniques d'apprentissage sont particulièrement bien adaptées à trouver le sens des mots en contexte, comme nous le verrons ci-après.

En revanche les méthodes d'apprentissage ne permettent pas ou du moins pas encore d'apprendre les données nécessaires au calcul du sens logique de la phrase, et encore moins lorsque celui-ci intègre le sens lexical, comme le font (39; 1). Il faudrait pour cela savoir apprendre les lambda termes typés qui permettent de calculer le sens logique et les coercions lexicales.

4 Le contexte dans le TAL quantitatif et en particulier neuronal

Les méthodes statistiques et l'apprentissage machine ont joué une rôle important dans le développement de la linguistique informatique depuis la naissance de ce domaine. Pendant des périodes, c'était l'approche dominante en traitement automatique des langues. Aujourd'hui encore, ça reste l'approche le plus utilisé pour les applications dans le domaine.

Une approche très simple, mais très efficace de la notion de contexte utilisait ce qu'on appelle *term frequency-inverse document frequency*. Cet approche consiste à compter le nombre de fois un mot apparaît dans un document et de multiplier ce résultat par un chiffre représentant le nombre de fois ce même mot apparaît dans tout les documents, correspondant à un type de pénalité pour des mots qui sont présents dans beaucoup de documents. Cette mesure est une indication de l'importance des mots pour un document : si un mot est fréquent dans tout les documents, comme c'est le cas pour des mots tel que "le" ou "est" le score *tf-idf* sera proche de zéro, indiquant que ces mots sont pas très informatif parce qu'ils sont trouvés dans beaucoup de contextes. Par contre, d'autres mots, tel que "Covid-19", pourraient avoir de valeurs *tf-idf* élevés, indiquant qu'ils portent beaucoup d'information. Malgré la simplicité de la notion de contexte qu'il utilise, des variants

de la mésure *tf-idf* ont trouvé beaucoup d'applications, notamment en recherche d'information, grâce à leur efficacité et à la simplicité des calculs nécessaires.

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 (34), 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, –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.

4.1 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) [2] 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, 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

2. 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 (17) ou (10). Pour un livre de référence sur le TAL neuronal, cf. (16), et pour les vecteurs de mots, on pourra consulter (45).

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 (22; 42; 41). Word2vec comprend en réalité deux modèles, Skip-gram and CBOW, [3] 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" [4] 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). [5] 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 précé-

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

4. 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.

5. 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.

demment 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 ont rapidement surpassé de manière significative les performances des modèles existants à travers des tâches linguistiques de diverse nature ((cf. 6)). 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 (42) ou des différences de degré associées à des aspects sémantiques de groupes de mots (18).

4.2 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 (41; 40). 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.

4.3 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 !*") (15), ce principe a trouvé de multiples formulations. L'article détaillé de Sahlgren (2008) 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. (35, p. 17), (25, 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. [6] 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. [7]

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".

6. Cf. (50) pour une étude classique rapportant co-occurrence et force associative.

7. Voir (25) pour une analyse de l'approche cognitive du rapport entre distributionalisme et théorie du sens comme usage.

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 décisif, 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 (2014a) 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. [8] 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 ((31)).

8. 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).

4.4 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.

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 [9] 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, (43)) 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).

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

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

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. [10] 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 sont censés entretenir avec les entités qui les habitent. [11] 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

10. Ce qui faisait Chomsky soutenir que la notion de probabilité d'une phrase ne pouvait jouer aucun rôle dans l'analyse linguistique ((cf. 12)).

11. 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 (2014b) le remarquent, le modèle ne garde que celle des termes, en abandonnant celle des contextes.

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 ((24, 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.

4.5 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*.

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. 32, 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. [12] 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 nou-

12. 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]).

veaux 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 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'entraî-

nement 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 (13). 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. (9). 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 précé-

dent 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. (7)), 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. 9)).

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 linguistique 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 (54) et de (14). 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.

5 Hybridation et contexte

Même si les approches neuronales obtiennent des excellents résultats pour beaucoup de tâches en linguistique informatique, ils est souvent difficile d'obtenir les mêmes performances sur de tâches qui ne sont que légèrement différentes. Ceci est une probléme classique en apprentissage automatique appelée *overfitting* (surapprentissage). Un exemple standard de surapprentissage serait un système qui mémorise simplement les données utilisés pour son entraînement et qui échoue après sur des données jamais vues.

Par exemple, pour la tâche de détection d'inférence, on donne à l'application les phrases 19a et 19b, et le système doit conclure que phrase 19a entraîne la phrase 19b, c'est-à-dire que si la première phrase est vraie, le deuxième doit l'être aussi. Par contre, quand on donne les phrases 19a et 19c au logiciel, il doit conclure que les deux phrases se contredisent, et qu'elles ne peuvent pas être vraies dans la même situation.

(19) a. Je serais heureux de répondre à toutes questions que les membres du sous-comité pourraient avoir.
b. J'adorerais répondre aux questions.
c. Je ne réponds pas aux questions.

Pour cette tâche, les approches neuronales obtiennent de très bons résultats : pour l'anglais, le meilleur système pour cette tâche donne la réponse correcte dans 96% des cas. On pourrait alors penser que ce problème est plus ou moins résolu. Mais de nombreux auteurs ont montré que la plus grande prudence est de mise avec

l'interprétation des ces résultats, car il y a des stratégies pour induire en erreur ces systèmes neuronaux.

Par exemple, Gururangan et al. (19) ont montré qu'il est possible d'obtenir des bons résultats sur cette tâche en présentant la phrase 19b ou 19c sans la phrase 19a. Ceci veut dire qu'au moins une grande partie des performances est due à des régularités dans les données : la présence de la négations "ne ... pas" est une indice assez fiable qu'il s'agit d'une contradiction, mais cet indicateur n'est fiable que pour les données d'entraînement, et pas en général.

Un autre type d'exemples d'implication sont les phrases suivantes.

(20) a. La fille de la directrice a dansé.
b. La directrice a dansé.

A partir de la phrase 20a, on ne devrait pas pouvoir conclure que la phrase 20b est nécessairement vraie, c'est possible mais certainement pas obligatoire. Par contre, McCoy et al. (37) montrent que les systèmes d'implicature ont beaucoup de mal avec des exemples tel que 20. Ce genre difficultés sont l'une des principales motivations pour combiner les méthodes neuronaux avec des méthodes symboliques. Le but de ces approches neuro-symboliques est d'utiliser les points forts des systèmes symboliques pour combler les lacunes des systèmes neuronaux et vice-versa. Les approches hybrides utilisent les deux notions de contexte discuté des les sections précédentes. La structure syntaxique et sémantique peut permettre de déterminer que ~19c tout seul n'est pas une contradiction et qu'il n'y a pas d'implication entre 20a et 20b. Dans l'autre sens, les vecteurs de mots peuvent être utilisés pour des tâches pour lesquelles il n'existe pas de solutions purement symboliques avec des performances équivalentes, comme la desambiguísation du sens des mots.

D'un coté, il y a les approches neuronales. Elles sont très performantes pour beaucoup de tâches mais elles ne permettent aucune réflexion sur les résultats : si un modèle de question-réponse répond "oui" à une question, c'est parce que des millions (voire des milliards) d'opérations algébriques ont donnée une certaine valeur et pas une autre. En outre, une telle approche a besoin de beaucoup de données.

De l'autre coté, il y a les approches symboliques, qui sont souvent très loin d'avoir les performances des approches neuronales, mais qui permettent d'analyser les résultats et de corriger des erreurs.

Beaucoup de chercheurs tentent, parfois avec succès, de combiner les approches neuronales avec des approches symboliques, (Hamilton et al.). L'espoir

des méthodes hybrides est de garder les aspects positifs des deux approches, l'interprétabilité des approches symboliques et la performance des approches neuronales. Il y a aujourd'hui un continuum entre approches purement symboliques, approches symboliques augmenté par des approches neuronales, approches neuronales augmenté par des approches symboliques et approches purement neuronales. Par exemple, certains analyseurs syntaxiques confie à un modèle neuronal le choix de l'une des différentes procédures que l'analyseur peut appliquer pour produire une analyse.

A l'inverse, un modèle neuronal peut directement produire plusieurs représentations sémantiques pour un texte, avec une module symbolique pour éliminer les interprétations qui sont incohérentes avec une base de connaissances donnée.

Les méthodes hybrides sont particulièrement utiles lorsque les données sont rares, lorsque les résultats doivent être interprétés et justifiés, mais aussi quand l'absence de réponse est préférable à une réponse fausse. Par exemple, dans le contexte d'un système de questions-réponses pour des conseils médicaux, des réponses fausses peuvent avoir des conséquences très graves, et même pour des réponses correctes, il est important de savoir *pourquoi* elles sont correctes.

Références

[1] Asher, N. (2011). *Lexical Meaning in context – a web of words*. Cambridge University press.

[2] Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

[3] Bar-Hillel, Y. (1953). The present status of automatic translation of languages. *American Documentation*, 2(4) :229–237.

[4] Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, 1 :91–163.

[5] Bar-Hillel, Y., Gaifman, C., and Shamir, E. (1964). On categorial and phrase structure grammars. In Bar-Hillel, Y., editor, *Language and Information. Selected Essays on their Theory and Application*, pages 99–115. Addison-Wesley, New York.

[6] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vec-

tors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 238–247. Association for Computational Linguistics.

[7] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). 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. Association for Computing Machinery.

[8] Brill, E. (1992). A simple rule-based part of speech tagger. In *Speech and Natural Language : Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*. Morgan Kaufmann.

[9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[10] Brunton, S. L. and Kutz, J. N. (2022). *Data-Driven Science and Engineering : Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2 edition.

[11] Chomsky, N. (1957). *Syntactic structures*. Janua linguarum. Mouton, The Hague.

[12] Chomsky, N. (1969). *Quine's Empirical Assumptions*, pages 53–68. Springer Netherlands, Dordrecht.

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[14] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https ://transformer-circuits.pub/2021/framework/index.html.

[15] Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford.

[16] Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*, volume 10.

[17] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge MA, London UK.

[18] Grand, G., Blank, I. A., Pereira, F., and Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings.

[19] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

[Hamilton et al.] Hamilton, K., Nayak, A., Božić, B., and Longo, L. Is neuro-symbolic AI meeting its promises in natural language processing ? a structured review. *Semantic Web*, (Preprint) :1–42.

[21] Harris, Z. (1960). *Structural linguistics*. University of Chicago Press, Chicago.

[22] inc., G. (2013). word2vec, https ://code.google.com/archive/p/word2vec/.

[23] Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

[24] Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., editors (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

[25] Lenci, A. (2008). 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.

[26] Léon, J. (1997). La langue intermédiaire dans la traduction automatique en URSS (1954-1960). *Histoire Epistémologie Langage*, 19(2) :105–132.

[27] Léon, J. (2000). Traduction automatique et formalisation du langage : les tentatives du cmabriudge language research unit (1955-1960). In Desmet, P., Jooken, L., and Schmitter, P., editors, *The history of linguistics and grammatical praxis*, pages 369–394. Peeters.

[28] Léon, J. and Cori, M. (2002). La constitution du TAL — Étude historique des dénominations et des concepts. *Traitement Automatique des Langues*, 43(3) :21–55.

[29] Levy, O. and Goldberg, Y. (2014a). 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.

[30] Levy, O. and Goldberg, Y. (2014b). 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. MIT Press.

[31] Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3 :211–225.

[32] Liu, Q., Kusner, M. J., and Blunsom, P. (2020). A survey on contextual embeddings.

[33] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text & Talk*, 8 :243 – 281.

[34] Manning, C. D. (2015). Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4) :701–707.

[35] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

[36] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

[37] McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons : Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

[38] Mel'čuk, I. (1997). Vers une linguistique sens-texte. leçon inaugurale, collège de france. http ://www.fas.umontreal.ca/ling/olst/melcuk/.

[39] Mery, B., Bassac, C., and Retoré, C. (2007). A Montagovian generative lexicon. In *Formal Grammar*. CSLI.

[40] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[41] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[42] Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). 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.

[43] Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1) :1–28.

[44] Montague, R. (1974). The proper treatment of quantification in ordinary English. In Thomason, R., editor, *Formal Philosophy. Selected Papers of Richard Montague*. Yale University Press, New Haven.

[45] Pilehvar, M. T. and Camacho-Collados, J. (2020). Embeddings in natural language processing : Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4) :1–175.

[46] Pollock, J.-Y. (1997). *Langage et cognition : le programme minimaliste de la grammaire générative*. Presses Universitaires de France, Paris.

[47] Sahlgren, M. (2006). *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.

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

[49] Saussure (1980). *Cours de linguistique générale*. Payot, Paris.

[50] Spence, D. P. and Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, 19(5) :317–330.

[51] Tesnière, L. (1959). *Éléments de syntaxe structurale*. Éditions Klincksieck. 5^{e} édition : 1988.

[52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

[53] Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In Morrel, A. J. H., editor, *Information Processing, Proceedings of IFIP Congress 1968, Edinburgh, UK, 5-10 August 1968, Volume 2 - Hardware, Applications*, pages 1114–1122.

[54] Weiss, G., Goldberg, Y., and Yahav, E. (2021). Thinking like transformers. *CoRR*, abs/2106.06981.