DACCORD : un jeu de données pour la Détection Automatique d'énonCés COntRaDictoires en français

Maximos Skandalis, Richard Moot, Simon Robillard

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

DACCORD : un jeu de données pour la Détection Automatique d'énonCés COntRaDictoires en français

Maximos Skandalis[1,] Richard Moot[1,] Simon Robillard[1,]
(1) LIRMM, CNRS, Université de Montpellier, 34095 Montpellier, France
[prénom].[nom]@lirmm.fr

RÉSUMÉ

La tâche de détection automatique de contradictions logiques entre énoncés en TALN est une tâche de classification binaire, où chaque paire de phrases reçoit une étiquette selon que les deux phrases se contredisent ou non. Elle peut être utilisée afin de lutter contre la désinformation. Dans cet article, nous présentons DACCORD, un jeu de données dédié à la tâche de détection automatique de contradictions entre phrases en français. Le jeu de données élaboré est actuellement composé de 1034 paires de phrases. Il couvre les thématiques de l'invasion de la Russie en Ukraine en 2022, de la pandémie de Covid-19 et de la crise climatique. Pour mettre en avant les possibilités de notre jeu de données, nous évaluons les performances de certains modèles de transformeurs sur lui. Nous constatons qu'il constitue pour eux un défi plus élevé que les jeux de données existants pour le français, qui sont déjà peu nombreux.

ABSTRACT

DACCORD : a Dataset for Automated deteCtion of COntRaDictions between sentences in French

In NLP, the automatic detection of logical contradictions between statements is a binary classification task, in which a pair of sentences receives a label according to whether or not the two sentences contradict each other. This task has many potential applications, including combating disinformation. In this article, we present DACCORD, a new dataset dedicated to the task of automatically detecting contradictions between sentences in French. The dataset is currently composed of 1034 sentence pairs. It covers the themes of Russia's invasion of Ukraine in 2022, the Covid-19 pandemic, and the climate crisis. To highlight the possibilities of our dataset, we evaluate the performance of some recent Transformer models on it. We conclude that our dataset is considerably more challenging than the few existing datasets for French.

MOTS-CLÉS : détection automatique de contradictions, jeu de données, construction de corpus, tâche de paire de phrases, classification binaire, analyse sémantique de phrases, français.

KEYWORDS: automatic contradiction detection, dataset, corpus construction, sentence pair task, binary classification, semantic analysis of sentences, French.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 285–297.

1 Introduction

1.1 Contexte

Les désinformations (*fake news*) sur les réseaux sociaux sont, de nos jours, un enjeu sociétal majeur. Dans plusieurs pays, y compris la France, diverses entités (individus, états, entreprises) ont mené des campagnes de désinformation pour tenter de manipuler l'opinion publique, la législation, ou le résultat d'élections. Pendant la crise du Covid-19, puis la guerre en Ukraine, ce problème est devenu encore plus apparent. La création relativement récente d'agences comme East StratCom en Europe et Viginum en France montre l'intérêt politique de la lutte contre les fausses informations. De même, des journalistes élaborent des sites de vérification des faits (*fact-checking*) dans plusieurs langues, dont le français.

Pour détecter les *fake news*, plusieurs approches ont été proposées, parmi elles la vérification de rumeurs, la détection de techniques de propagande, l'analyse de sentiment et l'analyse de fiabilité des sources des phrases. Par contre, l'aspect sémantique, bien qu'essentiel pour parler de la véracité des propos, a été peu exploré dans ce cadre précis, surtout pour le français.

1.2 Définition de la tâche

La détection automatique de la relation logique de contradiction constitue une étape nécessaire pour identifier automatiquement des fausses informations. La plupart des fausses informations circulant contredisent un savoir scientifique ou une connaissance du monde établie ou largement acceptée. Par ailleurs, c'est dans la nature de la connaissance scientifique d'être réfutable.

La définition la plus courante des contradictions en langage naturel est que deux énoncés sont contradictoires s'il est impossible pour eux d'être tous les deux vrais dans la même situation (c.-à-d. dans le même modèle en logique). La détection automatique de contradictions entre énoncés consiste donc à indiquer si deux phrases données se contredisent ou non. Dans les données textuelles du monde réel néanmoins, il y a peu de chance de trouver, pour une phrase A, explicitement A et non-A, c'est pour cela que la tâche nécessite une analyse sémantique profonde des phrases confrontées, qui ne sont d'habitude pas identiques entre elles.

L'exemple 1.2.1 donne deux phrases qui sont contradictoires.

(1.2.1)
P : 58 caisses étaient en service dans le centre commercial Amstor le jour de l'attaque, enregistrant ce jour-là un chiffre d'affaires de 2,9 millions de hryvnia ukrainiennes, soit environ 97.000 euros. Des employés du centre commercial, blessés le 27 juin, ont témoigné auprès de l'AFP après l'attaque.
H : Amstor était fermé et vide au moment des frappes par les missiles russes.
R : Contradiction

Les caisses d'un magasin sont en service et produisent un chiffre d'affaires non-nul quand le magasin est ouvert et, sûrement, non-vide. La contradiction ici ne provient pas simplement de l'introduction d'une négation dans l'une des deux phrases. Il s'agit d'un cas difficile à détecter par une machine aujourd'hui, et il illustre l'importance de l'analyse sémantique des phrases pour cette tâche.

En elle-même, la tâche de détection d'énoncés contradictoires ne dit rien de la véracité des énoncés. En revanche, en comparant des textes suspects à une base de connaissances de confiance (par exemple,

réseaux sémantiques, archives d'agences de vérification croisée de faits, sites officiels d'organismes), elle peut être utilisée pour détecter des désinformations. Une comparaison, donc, à une base de connaissances pertinentes (avec la liste actualisée des membres de l'OMS, par exemple) devrait permettre d'établir quelle phrase, entre l'hypothèse et la prémisse, est celle qui est vraie, dans l'exemple 1.2.2.

(1.2.2) P : Madagascar reste toujours membre de l'OMS en mai 2022.
H : Madagascar a quitté, en mai 2022, l'OMS en raison d'un scandale lié au Covid-19.
R : Contradiction

Mais l'intérêt pour la détection automatique de contradictions ne se limite pas seulement à la détection de *fake news*. Pouvoir repérer automatiquement les contradictions dans des données textuelles relève aussi de la compréhension du langage naturel par les machines et du raisonnement automatique sur le langage naturel.

La détection automatique de contradictions textuelles pourrait aussi avoir une place déterminante dans des tâches combinant des données de différentes formes. L'exemple 1.2.3, qui pourrait être une description ou légende d'une vidéo, peut illustrer cette perspective pour une éventuelle tâche multi-modale. Par ailleurs, la diffusion de désinformations dans le monde numérique peut prendre plusieurs formes.

(1.2.3) P : Le camion-remorque de la vidéo transporte un long tube cylindrique, qui est une pièce destinée à une raffinerie de pétrole en Ouzbékistan.
H : Le camion-remorque de la vidéo transporte un missile nucléaire russe.
R : Contradiction

Dans le domaine du traitement automatique du langage naturel (TALN) écrit, les premiers travaux sur la typologie des contradictions (Harabagiu *et al.*, 2006) se concentraient plutôt sur l'usage de la négation et de paraphrases avec des antonymes.

de Marneffe *et al.* (2008) ont élaboré une classification plus détaillée, en analysant les jeux de données RTE 1, 2 et 3 (Dagan *et al.*, 2006) et en construisant leur propre jeu de données de 131 paires de phrases, toutes des contradictions. Ils ont conclu que deux catégories principales de contradictions, avec des sous-catégories, y apparaissent : (1) des contradictions qui se produisent via l'antonymie, la négation et l'incompatibilité date/nombre, qui sont relativement simples à détecter, et (2) des contradictions résultant de l'utilisation de mots factifs ou modaux, de contrastes structuraux et lexicaux subtils, ainsi que de la connaissance du monde.

Ritter *et al.* (2008) ont ajouté à la classification de de Marneffe *et al.* (2008) une sous-catégorie de contradictions sur la connaissance du monde, à savoir des contradictions qui découlent de relations fonctionnelles telles que la relation "x est né à y". Enfin, Tsytsarau *et al.* (2011; 2011) ont étudié les contradictions d'opinions et de sentiments, qu'ils ont divisées en contradictions synchrones et asynchrones (qui correspondent à un changement de sentiment).

1.3 Problématique et motivation

Le but du présent article est de présenter un nouveau jeu de données pour la tâche de détection automatique de contradictions logiques entre énoncés en français. La définition précédente de la contradiction

en langage naturel convient à ce but, car nous nous attachons ici à des cas de contradictions factuelles, qui satisfont cette définition.

La création de ce nouveau jeu de données est motivée par le fait qu'à l'époque où l'usage des modèles de transformeurs (Vaswani *et al.*, 2017) est devenu dominant au sein de la communauté de TALN, un grand besoin pour une quantité importante de jeux de données a émergé.

Un des jeux de données utilisable pour la détection de contradictions entre énoncés en langue française est FraCaS (Amblard *et al.*, 2020). Ce jeu, étant centré sur la tâche d'inférence textuelle, ne contient que 9% d'énoncés contradictoires. De surcroît, FraCaS est un jeu de données relativement petit (346 paires de phrases), par rapport à d'autres jeux de données (toujours) d'inférence textuelle disponibles en anglais. Le plus grand corpus pour la tâche d'inférence textuelle est XNLI (Conneau *et al.*, 2018), mais la plus grande partie (plus de 98%) de sa version française n'est issue que d'une traduction automatique de sa version originale en anglais.

D'autres jeux de données sur l'inférence textuelle ont parfois une vision qui n'est pas purement logique de ce qui constitue une contradiction. Ces jeux peuvent voir leur tâche comme un calcul de score de similarité sémantique entre deux phrases. Cette approche nous semble inadaptée pour la détection de contradictions, car elle peut aboutir à des situations où deux phrases qui ne parlent pas du même sujet sont considérées comme contradictoires.

D'autres jeux de données parus récemment sur la désinformation (Li *et al.*, 2020; Kochkina *et al.*, 2018) sont composés de messages issus de médias sociaux tels que Twitter. Les messages de Twitter, néanmoins, sont en général d'une qualité variable, avec des phrases typiquement courtes et beaucoup d'erreurs et de bruit. De plus, la détection de contradictions sur des messages de Twitter s'effectue très souvent par analyse de sentiment. Par contre, les articles de presse, qui nous intéressent davantage, comprennent des structures linguistiques plus riches et plus complexes. D'autre part, nous travaillons sur des contradictions logiques, donc l'analyse de sentiment n'est pas la voie à prendre.

L'article suit le plan suivant : après cette définition, dans l'introduction, des contradictions et de la tâche de détection automatique de contradictions, la Section 2 présente les jeux de données uni-lingues (anglais ou français) ou bien multi-lingues, qui sont disponibles, au moment de la publication de l'article, pour la tâche en question. Elle aborde aussi les différences entre la tâche de détection de contradictions et la tâche d'inférence textuelle. La Section 3 décrit la démarche suivie pour établir notre nouveau jeu de données sur la Détection Automatique d'énonCés COntRaDictoires en français. Dans la Section 4, nous évaluons certains récents modèles d'apprentissage profond sur le jeu de données construit.

2 État de l'art

2.1 Jeux de données disponibles en anglais

Il existe en anglais plusieurs jeux de données sur la tâche d'inférence textuelle, à savoir RTE (Dagan *et al.*, 2006; Dzikovska *et al.*, 2013) [1], SICK (Marelli *et al.*, 2014), SNLI (Bowman *et al.*, 2015),

1. Nous ne parlons pas ici de la version de RTE intégrée dans GLUE (Wang *et al.*, 2018), où RTE est divisé en deux catégories (inférence, pas d'inférence) et non en trois (inférence, contradiction, ni l'un ni l'autre).

MultiNLI (Williams *et al.*, 2018), XNLI (Conneau *et al.*, 2018) qui est multi-lingue, FraCaS (Cooper *et al.*, 1996).

Ce dernier, FraCaS, contient 346 problèmes d'inférence à traiter. SICK, quant à lui, contient 9840 exemples, dont 14% sont des contradictions. Leur annotation a été faite par *crowdsourcing* et il s'agit de phrases simplifiées en anglais, similaires à celles de FraCaS, comme le notent ses auteurs par ailleurs.

Ensuite, il existe huit corpora RTE, au même nombre que les compétitions homonymes organisées de 2006 à 2013. RTE-1 contient au total 1367 paires de phrases (800 dans le sous-ensemble de test, 287 dans le sous-ensemble dev_1 et 280 dans le sous-ensemble dev_2), dont 252 sont contradictoires (18,43% du corpus), 682 sont des inférences et 433 des cas neutres. RTE-2 est composé de 1600 paires (800 pour le sous-ensemble de test et 800 pour le sous-ensemble dev), dont 215 des contradictions (13,44% des cas), 800 des cas d'inférence et le reste (585) des cas neutres. Enfin, RTE-3 compte également 1600 paires de phrases (800-800), dont 152 des contradictions (9,5%), 822 des cas d'inférence et 626 des cas neutres.

Nie *et al.* (2020) ont transformé le jeu de données anglais FEVER, initialement construit par Thorne *et al.* (2018) avec des phrases de Wikipedia modifiées par des annotateurs et comparées à la section d'introduction d'une liste de pages Wikipedia pour une tâche de vérification automatique de faits, à un jeu de données pour l'inférence textuelle (donc, avec une étiquette "inférence", "contradiction" ou "neutre" pour chaque paire de prémisse-hypothèse fixée pour cette version).

D'autres jeux de données en anglais, qui essaient de traiter en particulier le problème des désinformations, existent, souvent basés sur des messages de Twitter, par exemple PHEME (Kochkina *et al.*, 2018). Pour PHEME, des messages de Twitter ont été classifiés par un journaliste comme "rumeurs" ou "non-rumeurs", puis ceux, qui sont des rumeurs, en "vrais", "faux" ou "non-vérifiées". Additionnellement, des réponses à ces *tweets* ont été annotées comme interrogatives, soutenant, niant ou commentant la rumeur.

2.2 Jeux de données disponibles en français

Comme déjà mentionné (1.3), en français, nous disposons d'une traduction française récente de FraCaS (Amblard *et al.*, 2020). FraCaS classe ses exemples dans les catégories suivantes : Quantificateurs Généralisés, Pluriels, Anaphore (nominale), Ellipse, Adjectifs, Comparatifs, Référence temporelle, Verbes, et Attitudes. Le jeu de données est disponible sous forme de question-réponse et sous forme de prémisse(s)-hypothèse.

XNLI (Conneau *et al.*, 2018) contient des traductions manuelles de l'anglais au français des sous-ensembles de validation et de test de MNLI initial. En particulier, Son sous-ensemble de validation est composé de 2490 paires de phrases (830 contradictions, 830 cas neutres, 830 inférences), alors que le sous-ensemble de test contient 5010 paires (1670 contradictions, 1670 cas neutres, 1670 inférences). Par contre, le sous-ensemble d'entraînement de XNLI en français n'est qu'une traduction automatique (par Conneau *et al.* (2018) et par Hu *et al.* (2020)) de celui de MNLI.

Enfin, MM-COVID (Li *et al.*, 2020) est un jeu de données multi-modal et multi-lingue, qui inclut aussi des exemples en français, mais qui sont tous des messages issus de médias sociaux. Pour ce qui est de son contenu en français, il inclut 2821 tweets dits "faux" (et 4459 réponses), contre 166 tweets dits "vrais" (et 5095 réponses).

2.3 Inférence textuelle et détection des contradictions

La détection d'énoncés contradictoires, avec comme finalité la détection de désinformations, et la tâche d'inférence textuelle, bien qu'elles partagent certains points communs, ne sont pas les mêmes.

D'un autre côté, il est important de faire la distinction entre la détection d'énoncés contradictoires et la détection de désinformations en tant que tâche de détection de rumeurs (Gorrell *et al.*, 2019) ou de techniques de propagande (Da San Martino *et al.*, 2020). De même, nous ne jugeons pas, dans le cadre de notre jeu de données, les sources des phrases ni ne caractérisons leur fiabilité en les séparant en sources fiables et sources non-fiables, comme le font Guibon *et al.* (2019). Nous nous limitons aux événements décrits par les phrases elles-mêmes, car le but est de détecter des contradictions logiques. Ceci dit, le corpus contient des phrases du monde réel, ce qui signifie aussi que, lors de la détection de contradictions logiques, il peut y avoir (et il y en a) des prémisses cachées.

Notre jeu de données sur la détection de contradictions se rapproche plutôt de jeux de données construits pour la tâche d'inférence textuelle (*textual entailment*). Cependant, il existe une différence entre la détection de contradictions et l'inférence textuelle, à savoir que la première est une tâche de classification binaire (les phrases sont contradictoires ou non), alors que la dernière est souvent vue comme une tâche de classification multi-classe ("oui/inférence", "non/contradiction" ou "inconnu/neutre" [2]). Dans notre cas, nous avons choisi de nommer nos deux classes "contradiction" et "compatibles".

Un autre point important est que, par comparaison à la tâche de détection de contradictions, les contradictions sont sous-représentées dans les jeux de données construits pour la tâche d'inférence textuelle. Pour la tâche d'inférence textuelle à trois étiquettes, les contradictions ne constituent qu'un tiers (ou même moins) des exemples dans les jeux de données dédiés. Dans FraCaS par exemple, les contradictions représentent 9% de l'ensemble du corpus, 52% des énoncés sont des inférences, 27% sont des cas neutres, et 12% des exemples nécessitent une réponse plus détaillée.

La raison pour laquelle nous avons choisi d'établir ce corpus sur la relation de contradiction et non pas sur l'inférence textuelle plus généralement est que la contradiction est symétrique. Il est important de voir si des modèles neuronaux sont capables de percevoir et prédire cette symétrie. Ainsi, nous avons intégré dans le corpus quelques exemples qui sont similaires mais inversés, c'est-à-dire que la prémisse prend la place de l'hypothèse et l'hypothèse celle de la prémisse, par rapport à l'exemple précédent. L'inférence textuelle n'est quant à elle pas symétrique, d'où aussi une différence entre les tâches de classification contradiction/pas de contradiction et inférence/pas d'inférence.

Toutefois, ces différences n'empêchent pas que nos résultats puissent s'intégrer avec des approches plus étendues visant des tâches telles que l'inférence textuelle (en divisant notre catégorie "compatibles" en "inférence" et "neutre" par exemple, ou en fusionnant ces deux catégories des autres jeux de données, même si cela aboutirait à des catégories qui ne sont pas équilibrés, ou bien en combinant notre jeu de données avec d'autres jeux).

2. L'hypothèse n'est pas conséquence de la prémisse, mais elle ne la contredit pas non plus.

3 DACCORD, un nouveau jeu de données pour la détection de contradictions

3.1 Méthode de construction du jeu de données

Notre jeu de données a été construit à partir d'articles sur le site factuel.afp.com. AFP Factuel est un site français de vérification de faits par Agence France-Presse. Les paires de phrases ont été manuellement sélectionnées en lisant les articles du site susmentionné et en gardant les phrases qui nous paraissaient d'intérêt pour la tâche étudiée.

Le jeu de données couvre actuellement trois thématiques : l'invasion russe en Ukraine, la pandémie de Covid-19 et la crise climatique. À notre connaissance, c'est en général le tout premier jeu de données de TALN couvrant le conflit entre Russie et Ukraine.

Le jeu de données est composé de 1034 paires de phrases, dont 515 (49,81%) forment des contradictions. Parmi elles, 472 paires de phrases (215 contradictions) ont été recueillies à partir de 106 articles sur la guerre russo-ukrainienne, datant du 24 février 2022 au 3 novembre 2022 (inclus). 450 paires (251 contradictions) ont été retenues à partir de 164 articles publiés du 20 octobre 2021 au 23 novembre 2022 sur la pandémie de Covid-19. Enfin, les 112 paires de phrases (49 contradictions) sur le réchauffement climatique sont issues de 33 articles datant du 16 juillet 2021 au 24 octobre 2022.

3.2 Propriétés du jeu de données

Nous avons choisi AFP Factuel comme source pour recueillir des phrases pour le jeu de données, cependant nous ne nous positionnons pas par rapport à la vérité des énoncés choisis, les thématiques traitées étant délicates et la notion de vérité n'étant formellement pas triviale. Il est indiqué dans le corpus quand une paire de phrases forme une contradiction ou quand les deux phrases d'un exemple sont inter-compatibles, mais il n'est pas indiqué lequel parmi les énoncés est vrai et lequel ne l'est pas.

De plus, chaque paire de phrases est un monde clos. Tout le contexte (par exemple, dates et/ou lieu) se trouve dans ces mêmes phrases. Par contre, il est clair, en regardant les différentes paires, qu'elles peuvent contenir beaucoup de prémisses cachées, qui devront être prises en compte pour appliquer des approches purement formelles/logiques sur le jeu de données.

Les exemples 3.2.1 et 3.2.2, extraits de la partie du corpus sur la pandémie de Covid-19, seraient des exemples de contradiction structurelle d'après la typologie de de Marneffe *et al.* (2008).

(3.2.1) P : De nombreux vaccins utilisés aujourd'hui n'induisent, comme ceux contre le Covid-19, qu'une immunité effective. Le vaccin contre la variole est, quant à lui, un exemple d'immunité "stérilisante".
H : Les vaccins contre le Covid-19 sont un exemple d'immunité stérilisante.
R : Contradiction

(3.2.2) P : Interrogée par l'AFP, l'Autorité régionale de santé (ARS) de Guadeloupe déplore une fausse information circulant et précise que ce n'est jamais elle qui passe les commandes de médicaments.
H : C'est une fausse information que ce n'est pas l'Autorité régionale de santé (ARS) de Guadeloupe qui passe les commandes des médicaments.

R : Contradiction

La contradiction dans 3.2.1 résulte des expressions “ne... que” et “quant à lui” de la prémisse, mais aussi d’une confusion concernant le groupe nominal qui sert de sujet du groupe verbal “être un exemple d’immunité stérilisante”. 3.2.2 pourrait être considéré comme un exemple méta-référentiel de fausse information.

La paire de phrases 3.2.3, qui provient du sous-ensemble sur le conflit ukrainien-russe, est un cas difficile de contradiction numérique.

(3.2.3) P : Cent-trente neuf pays sur les 193 membres de l’Assemblée générale des Nations Unies ont voté contre une résolution demandant à la Russie d’arrêter l’opération d’invasion de l’Ukraine et de retirer l’armée du territoire.
H : Toutes les résolutions présentées devant l’Assemblée générale au sujet de l’Ukraine ont été approuvées par au moins deux tiers des membres présents et votants conformément à la Charte des Nations Unies, et ont, donc, été adoptées par l’Assemblée générale.
R : Contradiction

Il s’agit d’un cas difficile car il n’est pas direct. Il faudrait que la machine arrive à faire le raisonnement que l’expression “cent-trente neuf pays sur les 193” correspond à 72%, et donc que, selon la prémisse, 28% des pays ont voté pour la résolution, alors que l’hypothèse parle d’“au moins 66%” qui votaient pour. Nous pourrions aussi voir cet exemple comme un cas d’antonymie contradictoire, entre les verbes “voté contre” et “approuvées” dans la prémisse et l’hypothèse, respectivement.

La paire de phrases 3.2.4 est donnée à titre d’exemple non-contradictoire.

(3.2.4) P : Le calcul du total des factures d’énergie n’est pas uniquement fondé sur les prix de gros des marchés, et diffère pour les consommateurs britanniques et français, la majorité de ces derniers bénéficiant des tarifs réglementés, dont la hausse a été plafonnée par le gouvernement jusqu’à fin 2022.
H : Il faut bien faire la différence entre les prix de marché et les tarifs réglementés, qui sont fixés par les autorités et qui concernent la plupart des particuliers consommateurs d’électricité en France.
R : Compatibles

Selon le tokéniseur de NLTK (Bird *et al.*, 2009), l’ensemble du jeu de données contient 63326 *tokens* au total (y incluse la ponctuation). La prémisse la plus courte du jeu contient 9 *tokens* (la ponctuation toujours incluse) et se trouve dans le sous-ensemble sur la pandémie de Covid-19, alors que la prémisse la plus longue est de 156 *tokens*, dans la partie sur la guerre entre Russie et Ukraine. Concernant les hypothèses, la plus courte fait partie du sous-ensemble sur le Covid-19 et contient 6 *tokens*, tandis que l’hypothèse la plus longue est composée de 111 *tokens* et porte de nouveau sur la guerre russo-ukrainienne.

Le Tableau 1 donne des détails sur le nombre de *tokens* des phrases par thématique pour DACCORD, ainsi que pour XNLI et FraCaS, à titre comparatif.

Jeux de données		Prémisse plus courte	Prémisse plus longue	Moyenne par prémisse	Somme de *tokens* des prémisses	Hypothèse plus courte	Hypothèse plus longue	Moyenne par hypothèse	Somme de *tokens* des hypothèses
DACCORD	Climat	20	112	50,13	5.614	9	111	43,61	4.884
	Covid-19	9	100	30,29	13.631	6	76	22,25	10.014
	Guerre Rus-Ukr	10	156	34,86	16.455	5	147	26,98	12.734
XNLI (test et val)		2	59	22,55	169.092	3	46	11,66	87.485
FraCaS		2	28	9,13	4.805	4	41	9,49	3.245

TABLE 1 – Nombre de *tokens* dans DACCORD, XNLI et FraCaS

4 Expériences

4.1 Protocole expérimental

Afin d'évaluer la performance de l'état de l'art sur notre nouveau jeu de données, nous avons choisi d'utiliser des modèles d'apprentissage profond basés sur l'architecture de transformeur (Vaswani *et al.*, 2017). Les modèles retenus pour l'évaluation sur le jeu de données DACCORD sont DistilmBERT (Sanh *et al.*, 2019), XLM-R (Conneau *et al.*, 2020), mDeBERTa-v3 (He *et al.*, 2021), et CamemBERT (Martin *et al.*, 2020). Ils sont tous entraînés en partie (DistilmBERT, XLM-R et mDeBERTa) ou entièrement (CamemBERT) sur des données françaises. Pour leur évaluation, nous avons utilisé des versions ajustées à XNLI, disponibles sur huggingface.co.

4.2 Résultats

Le Tableau 2 présente les résultats des expériences menées sur DACCORD. À titre de comparaison, des résultats calculés sur XNLI y sont aussi indiqués. Les modèles mentionnés et accessibles par hyperliens dans le Tableau 2, même quand évalués sur DACCORD, sont des modèles pour l'instant entraînés sur le sous-ensemble d'entraînement de XNLI. Ces données sont issues d'une traduction automatique en français qui limite leur qualité, mais la quantité de données fournie est nécessaire pour permettre l'entraînement des modèles actuels.

L'étude portant sur la détection de contradictions, nous avons calculé l'*accuracy* et le score F1 [3] sur la probabilité prédite par les modèles que l'étiquette "contradiction" soit vraie.

En regardant les résultats, nous constatons, d'abord, une évolution progressive et constante des performances des modèles multi-lingues même sur les tâches uni-lingues (par exemple, mDeBERTa par opposition à DistilmBERT).

De plus, les modèles entraînés sur XNLI présentent, sans exception, des performances inférieures sur DACCORD que sur XNLI. Cela ne constitue pas une surprise, puisque DACCORD est construit de sorte à éprouver les capacités des modèles existants.

On peut, toutefois, observer une cohérence entre les performances sur XNLI des modèles étudiés et

3. Pour rappel, la mesure F1 est la moyenne harmonique de la précision et du rappel.

Modèles	DACCORD		XNLI	
	Accuracy	Score F1	*Accuracy*	Score F1
$\text{DistilmBERT}_{\text{Base-cased}}$	63,73	52,59	79,98	68,01
$\text{XLM-R}_{\text{Base}}$	71,57	67,62	87,17	81,14
$\text{CamemBERT}_{\text{Base, 3-class}}$	77,76	76,19	89,64	85,09
$\text{mDeBERTa-v3}_{\text{Base, XNLI}}$	80,75	78,30	90,98	86,39
$\text{mDeBERTa-v3}_{\text{Base, NLI-2mil7}}$	80,95	78,47	90,76	85,89
$\text{XLM-R}_{\text{Large}}$	82,01	80,00	**96,49**	**94,74**
$\text{CamemBERT}_{\text{Large, 3-class}}$	83,27	81,01	92,30	88,12
$\text{CamemBERT}_{\text{Large, 2-class}}$	**84,24**	**82,49**	91,70	87,66

TABLE 2 – Résultats de détection de contradictions par les transformeurs sur DACCORD et XNLI

leurs performances sur DACCORD, les deux meilleurs modèles pour XNLI par exemple (XLM-R et CamemBERT) étant aussi les meilleurs modèles pour DACCORD, même si les résultats sont inférieurs à ceux obtenus sur XNLI.

Pour information, tous les modèles évalués dans l'article ont échoué à détecter la contradiction dans l'exemple 3.2.2, et seulement $\text{mDeBERTa-v3}_{\text{Base, NLI-2mil7}}$ a réussi à trouver la bonne étiquette pour l'exemple 3.2.1. Quant à l'exemple 3.2.3 donné, tous les modèles ont correctement prédit son étiquette, sauf pour $\text{XLM-R}_{\text{Base}}$. Enfin, l'étiquette pour l'exemple 1.2.1 n'a pas été correctement prédit par DistilmBERT et $\text{CamemBERT}_{\text{Large, 3-class}}$.

Nous avons déjà fait remarquer dans 2.3 que les contradictions sont naturellement symétriques. Afin de tester le comportement des modèles d'apprentissage profond vis-à-vis de cette symétrie, nous avons effectué une dernière expérience, en échangeant la place de toutes les prémisses avec celle des hypothèses. Ses résultats sont consultables dans le Tableau 3.

Modèles	DACCORD		XNLI	
	Accuracy	Score F1	*Accuracy*	Score F1
$\text{XLM-R}_{\text{Large}}$	77,47	73,79	**83,77**	**73,12**
$\text{CamemBERT}_{\text{Large, 2-class}}$	**82,50**	**80,22**	81,64	69,72

TABLE 3 – Détection des contradictions avec la place des prémisses et des hypothèses inversée

Dans ce nouveau test, $\text{CamemBERT}_{\text{Large, 2-class}}$ obtient maintenant un score d'*accuracy* de 82,5% et un score F1 de 80,22% sur DACCORD, au lieu de 84,24% et 82,49%, respectivement. De même, il obtient un score d'*accuracy* de 81,64% et un score F1 de 69,72% sur XNLI, au lieu de 91,7% et 87,66%, respectivement. Les résultats de $\text{XLM-R}_{\text{Large}}$ sont aussi en baisse dans ce scénario avec les prémisses et les hypothèses échangées : sur XNLI, son score d'*accuracy* devient 83,77% et son score F1 73,12%, par opposition à 96,49% et 94,74%, respectivement, avant. Enfin, sur DACCORD, son *accuracy* est en baisse à 77,47% et son score F1 à 73,79%. Ces derniers résultats pourraient suggérer que cet aspect de la relation de contradiction n'est pas suffisamment pris en compte par les jeux de données existants utilisés largement pour l'entraînement des modèles.

5 Conclusion et perspectives

Dans cet article, nous avons présenté DACCORD, un nouveau jeu de données pour la tâche de détection automatique d'énoncés contradictoires en français. Il est composé de 1034 paires de phrases, toutes récupérées manuellement, dont 515 contradictions. À notre connaissance, c'est le premier corpus en français exclusivement dédié à la tâche de détection de contradictions et couvrant les thématiques du Covid-19 et de la guerre entre Russie et Ukraine. De plus, c'est le jeu de données le plus compliqué pour cette tâche, étant données la longueur des prémisses et des hypothèses incluses et la nature des sources utilisées (articles de presse et non messages de médias sociaux). Lors de l'évaluation des modèles examinés, DACCORD s'avère être plus difficile pour eux que le jeu de données sur l'inférence textuelle XNLI.

Nous comptons par la suite expérimenter avec des méthodes neuro-symboliques sur le corpus construit. Nous souhaiterions, enfin, enrichir le corpus avec davantage de phrases et de thématiques de l'actualité mais peu incorporées dans les jeux de données disponibles.

Remerciements

La recherche présentée a été réalisée avec le soutien financier du Ministère des Armées – Agence de l'innovation de défense (AID), que nous en remercions. Ce travail a également bénéficié du soutien de l'ICO, Institut Cybersécurité d'Occitanie, financé par la Région Occitanie, France, auquel nous exprimons aussi notre gratitude.

Références

AMBLARD M., BEYSSON C., DE GROOTE P., GUILLAUME B. & POGODALLA S. (2020). A French version of the FraCaS test suite. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5887–5895, Marseille, France : European Language Resources Association.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

BOWMAN S. R., ANGELI G., POTTS C. & MANNING C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 632–642, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1075.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S. R., SCHWENK H. & STOYANOV V. (2018). Xnli : Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

COOPER R., CROUCH R., VAN EIJCK J., FOX C., VAN GENABITH J., JASPARS J., KAMP H., PINKAL M., MILWARD D., POESIO M., PULMAN S., BRISCOE T., MAIER H. & KONRAD K. (1996). *Using the Framework*. Rapport interne, FraCaS : A Framework for Computational Semantics. FraCaS deliverable D16, 136 pages, also available by anonymous ftp from ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz.

DA SAN MARTINO G., BARRÓN-CEDEÑO A., WACHSMUTH H., PETROV R. & NAKOV P. (2020). SemEval-2020 task 11 : Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, p. 1377–1414, Barcelona (online) : International Committee for Computational Linguistics. DOI : 10.18653/v1/2020.semeval-1.186.

DAGAN I., GLICKMAN O. & MAGNINI B. (2006). The pascal recognising textual entailment challenge. In J. QUIÑONERO-CANDELA, I. DAGAN, B. MAGNINI & F. D'ALCHÉ BUC, Éds., *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, p. 177–190, Berlin, Heidelberg : Springer Berlin Heidelberg.

DE MARNEFFE M.-C., RAFFERTY A. N. & MANNING C. D. (2008). Finding contradictions in text. In *Proceedings of ACL-08 : HLT*, p. 1039–1047, Columbus, Ohio : Association for Computational Linguistics.

DZIKOVSKA M., NIELSEN R., BREW C., LEACOCK C., GIAMPICCOLO D., BENTIVOGLI L., CLARK P., DAGAN I. & DANG H. T. (2013). SemEval-2013 task 7 : The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 263–274, Atlanta, Georgia, USA : Association for Computational Linguistics.

GORRELL G., KOCHKINA E., LIAKATA M., AKER A., ZUBIAGA A., BONTCHEVA K. & DERCZYNSKI L. (2019). SemEval-2019 task 7 : RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, p. 845–854, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/S19-2147.

GUIBON G., ERMAKOVA L., SEFFIH H., FIRSOV A. & LE NOÉ-BIENVENU G. (2019). Multilingual Fake News Detection with Satire. In *CICLing : International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France. HAL : halshs-02391141.

HARABAGIU S., HICKL A. & LACATUSU F. (2006). Negation, contrast and contradiction in text processing. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, p. 755–762 : AAAI Press.

HE P., GAO J. & CHEN W. (2021). Debertav3 : Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. DOI : 10.48550/ARXIV.2111.09543.

HU J., RUDER S., SIDDHANT A., NEUBIG G., FIRAT O. & JOHNSON M. (2020). XTREME : A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In H. D. III & A. SINGH, Éds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 de *Proceedings of Machine Learning Research*, p. 4411–4421 : PMLR.

KOCHKINA E., LIAKATA M. & ZUBIAGA A. (2018). All-in-one : Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 3402–3413, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

LI Y., JIANG B., SHU K. & LIU H. (2020). Mm-covid : A multilingual and multimodal data repository for combating covid-19 disinformation.

MARELLI M., MENINI S., BARONI M., BENTIVOGLI L., BERNARDI R. & ZAMPARELLI R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 216–223, Reykjavik, Iceland : European Language Resources Association (ELRA).

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

NIE Y., WILLIAMS A., DINAN E., BANSAL M., WESTON J. & KIELA D. (2020). Adversarial NLI : A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4885–4901, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.441.

RITTER A., SODERLAND S., DOWNEY D. & ETZIONI O. (2008). It's a contradiction – no, it's not : A case study using functional relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 11–20, Honolulu, Hawaii : Association for Computational Linguistics.

SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2019). Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. DOI : 10.48550/ARXIV.1910.01108.

THORNE J., VLACHOS A., CHRISTODOULOPOULOS C. & MITTAL A. (2018). FEVER : a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 809–819, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1074.

TSYTSARAU M. & PALPANAS T. (2011). Towards a framework for detecting and managing opinion contradictions. In M. SPILIOPOULOU, H. WANG, D. J. COOK, J. PEI, W. WANG, O. R. ZAÏANE & X. WU, Éds., *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, p. 1219–1222 : IEEE Computer Society. DOI : 10.1109/ICDMW.2011.167.

TSYTSARAU M., PALPANAS T. & DENECKE K. (2011). Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, **1**, 9–16.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WILLIAMS A., NANGIA N. & BOWMAN S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1112–1122 : Association for Computational Linguistics.