Classification de relations pour l'intelligence économique et concurrentielle

Hadjer Khaldi, Amine Abdaoui, Farah Benamara, Grégoire Sigel, Nathalie Aussenac-Gilles

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Classification de relations pour l'intelligence économique et concurrentielle

Hadjer Khaldi[1, 2] Amine Abdaoui[1] Farah Benamara[2] Grégoire Sigel[1]
Nathalie Aussenac-Gilles[3]
(1) Geotrend, Toulouse, France
(2) IRIT, Université de Toulouse, France
(3) IRIT, CNRS, Toulouse, France
prenom.nom@irit.fr, prenom@geotrend.fr

RÉSUMÉ

L'extraction de relations reliant des entités par des liens sémantiques à partir de texte a fait l'objet de nombreux travaux visant à extraire des relations génériques comme l'hyperonymie ou spécifiques comme des relations entre gènes et protéines. Dans cet article, nous nous intéressons aux relations économiques entre deux entités nommées de type organisation à partir de textes issus du web. Ce type de relation, encore peu étudié dans la littérature, a pour but l'identification des liens entre les acteurs d'un secteur d'activité afin d'analyser leurs écosystèmes économiques. Nous présentons BIZREL, le premier corpus français annoté en relations économiques, ainsi qu'une approche supervisée à base de différentes architectures neuronales pour la classification de ces relations. L'évaluation de ces modèles montre des résultats très encourageants, ce qui est un premier pas vers l'intelligence économique et concurrentielle à partir de textes pour le français.

ABSTRACT

Relation Classification for Competitive and Economic Intelligence

Relation extraction aims at identifying semantic relations that may hold between entities in raw text. This task has been widely studied in the litterature focusing either on extracting generic relations like hyperonymy or domain-dependent relations like those linking genes and proteins. In this paper, we aim at extracting business relations between two organizations from web textual contents. In particular, we propose BIZREL, the first French annotated dataset for business relations as well as a supervised approach based on several neural architectures to classify these relations. Our results are encouraging and constitute a first step towards economic and competitive intelligence from French texts.

MOTS-CLÉS : Classification de relation, Relation économique, Ressource linguistique.

KEYWORDS: Relation classification, Business relation, Linguistic resources.

1 Motivations

L'extraction de relations sémantiques est une tâche d'extraction d'information qui permet de détecter les liens sémantiques reliant des entités à partir d'expressions en langage naturel, dans le but de produire de l'information structurée à partir de textes bruts (Aussenac-Gilles *et al.*, 2013). Les entités ainsi reliées peuvent être exprimées par des syntagmes nominaux désignant des classes (Hendrickx

Volume 2 : Traitement Automatique des Langues Naturelles, pages 27–39.
HAL : hal-02784753.

et al., 2010), comme *entreprise, produit*, ou des entités nommées (Mitchell *et al.*, 2005), comme *Google*, *Paris*.

Cette tâche, qui s'apparente à un problème de classification supervisée de fragments de phrases, a fait l'objet de nombreux travaux basés sur des approches à base de patrons (Aussenac-Gilles & Jacques, 2008), à base de traits ou de noyaux (Kambhatla, 2004; Culotta & Sorensen, 2004), ou neuronales (Wang *et al.*, 2016; Lee *et al.*, 2019). Nous distinguons ceux qui classifient des relations génériques telles que les relations *cause-effet*, *message-sujet* (Hendrickx *et al.*, 2010) ou encore des relations d'*hypernonymie* ou de *synonymie* (Hearst, 1992; Lee *et al.*, 2017), et ceux se focalisant sur des relations spécifiques telles que les relations biomédicales reliant des protéines ou des gènes (Zhou *et al.*, 2014; Zhang *et al.*, 2018; Fan *et al.*, 2018). Les approches s'appuyant sur de l'apprentissage supervisé utilisent des corpus manuellement annotés par des types de relations prédéfinis dont la plupart sont en anglais (tels que les corpus de SemEval-2010 Tâche 8 (Hendrickx *et al.*, 2010), TACRED (Zhang *et al.*, 2017) et BioNLP-OST 2019 (Bossy *et al.*, 2019)). Nous citons néanmoins quelques corpus dans d'autres langues comme ACE 2004 (Mitchell *et al.*, 2005) pour le chinois et l'arabe, ou ReRelEM (Freitas *et al.*, 2009) pour le portugais.

Dans cet article, nous nous intéressons à la classification de relations économiques prédéfinies entre deux entités nommées de type organisation (ORG) à partir de textes issus du web. Ces relations permettent de créer des réseaux commerciaux, valoriser des entreprises et suivre leurs activités. C'est donc un moyen crucial pour l'identification des liens entre les acteurs d'un secteur d'activité et pour l'analyse de leurs écosystèmes économiques. Contrairement aux relations génériques, peu de travaux se sont attaqués à la classification des relations économiques. Ces derniers se focalisent principalement sur deux types de relations, la *compétition* et la *coopération* (Zuo *et al.*, 2017; Lau & Zhang, 2011). Les techniques adoptées visent à adapter les travaux sur la classification de relations génériques pour prendre en compte la spécificité des relations économiques. Les approches varient de méthodes à base de patrons (Braun *et al.*, 2018) à des approches supervisées (Yan *et al.*, 2019) ou semi-supervisées (Lau & Zhang, 2011; Zuo *et al.*, 2017). Là encore, les données utilisées pour entraîner ces différents systèmes sont majoritairement en anglais (Lau & Zhang, 2011) avec quelques initiatives en allemand (Braun *et al.*, 2018) et en chinois (Yan *et al.*, 2019). À notre connaissance, il n'existe aucun corpus annoté pour la classification de relations économiques pour la langue française. Dans ce contexte, nous proposons :

1. *Une taxonomie* composée de cinq relations jugées pertinentes dans le cadre d'intelligence économique : *investissement, compétition, coopération, vente-achat* et *poursuite judiciaire*.
2. BIZREL, *le premier corpus français* de 10k instances de relations annotées manuellement selon cette taxonomie via une plateforme d'annotation collaborative. Ce corpus est mis à disposition de la communauté. [1]
3. *Une approche supervisée* pour la classification de relations économiques à base de différentes architectures neuronales. Nous évaluons en particulier le modèle de langue multilingue M-BERT (Devlin *et al.*, 2019) mais aussi ses variantes françaises CamemBERT (Martin *et al.*, 2019) et FlauBERT (Le *et al.*, 2019), les modèles de langues récemment développés pour le français. Ces deux derniers n'ont jamais été utilisés pour une tâche de classification de relations. Nous comparons les performances de ces modèles avec celles obtenues par d'autres modèles de classification de relations génériques à la Tâche 8 de SemEval-2010.
4. Une analyse d'erreurs montrant les limites des approches proposées.

1. BizRel_corpus.

Dans la suite de cet article, la section 2 présente un bref état de l'art des approches de classification de relations. La section 3 définit les relations économiques d'intérêt et rapporte la procédure suivie pour construire le corpus BIZREL. La section 4 décrit les modèles utilisés, ainsi que les résultats obtenus. Une analyse d'erreur est présentée en section 5. Nous présentons les perspectives envisagées des travaux futurs en section 6.

2 Classification de relations : état de l'art

2.1 Classification de relations génériques

Les approches classiques visent à identifier une combinaison de traits lexicaux, syntaxiques et sémantiques (Kambhatla, 2004; Zhou *et al.*, 2005; Nguyen *et al.*, 2007) à partir desquels un classifieur apprend à distinguer les différentes relations prédéfinies ou à exploiter les liens de dépendances entre les deux entités, en utilisant des représentations plus riches des instances de relations telles que des arbres syntaxiques ou de dépendances (Collins & Duffy, 2001; Culotta & Sorensen, 2004). Plus récemment, les approches neuronales, telles que des CNN, arrivent à extraire des traits aux niveaux du mot et de la phrase à l'aide de modèles de plongements de mots, tout en prenant en compte leurs positions relatives par rapport aux deux entités concernées par la relation (Zeng *et al.*, 2014; Nguyen & Grishman, 2015; dos Santos *et al.*, 2015). D'autres y rajoutent un mécanisme d'attention qui capture, pour chaque mot d'une instance de relation, les éléments de contexte qui lui sont pertinents, améliorant ainsi les performances (Wang *et al.*, 2016; Shen & Huang, 2016). Des modèles basés sur une architecture RNN Bi-LSTM [2] avec attention ont été proposés par (Lee *et al.*, 2019; Xiao & Liu, 2016) pour capter des traits à longue distance et à distance variable. Plus récemment, l'utilisation des modèles de langue à base de transformers a encore amélioré les performances. Par exemple, Wu & He (2019) ont utilisé BERT (Devlin *et al.*, 2019) en concaténant les représentations contextuelles des entités avec la représentation de la phrase produite par le transformer. Ce modèle a pu atteindre un score F1 de 89.25% sur le corpus SemEval-2010 Tâche 8, en dépassant les modèles neuronaux. Baldini Soares *et al.* (2019) a exploré l'efficacité de différentes représentations possibles d'une instance de relation où l'utilisation des représentations des marqueurs d'entités s'est avérée la plus efficace. Enfin, Tao *et al.* (2019) ont combiné la représentation fournie par BERT avec des traits syntaxiques de la phrase, améliorant ainsi les performances.

2.2 Classification de relations économiques

Selon la nature de l'activité reliant deux organisations, Zhao *et al.* (2010) répartit les relations économiques en quatre types : coopération, investissement, vente et approvisionnement. D'autres travaux les classent en deux groupes uniquement (Zuo *et al.*, 2017; Yamamoto *et al.*, 2017) : co-opération et concurrence. Globalement, les relations économiques sont peu présentes dans les bases de connaissances telles que Freebase (Bollacker *et al.*, 2008) et DBpedia (Auer *et al.*, 2007) où figurent par exemple les relations *subsidiary* et *ownership_of* (Zuo *et al.*, 2017). Certaines relations économiques sont néanmoins annotées dans des corpus de relations génériques, comme par exemple *Employment/Membership/Subsidiary* dans le corpus ACE 2004 (Mitchell *et al.*, 2005), ou la relation *Component-Whole* dans le corpus SemEval-2010 Tâche 8 (Hendrickx *et al.*, 2010) mais avec des

2. Bidirectionnel à mémoire court terme étendue.

fréquences assez faibles (de l'ordre de 1k). Pour extraire ces relations, la plupart des travaux existants utilisent des approches semi-supervisées, soit à l'aide de patrons lexico-syntaxiques générés à partir d'arbres de dépendance (Braun *et al.*, 2018), soit grâce à une liste de mots clés représentatifs de chaque relation (Lau & Zhang, 2011). Les réseaux de neurones ont été utilisés dans (Yan *et al.*, 2019) qui propose un modèle entraîné sur un corpus chinois de 1k instances de relation économiques, afin d'aider les institutions financières à gérer le risque des crédits aux entreprises.

3 BIZREL : corpus français de relations économiques

BIZREL est le premier corpus annoté en relations économiques pour le français. Il décrit cinq types de relations binaires de nature économique, reliant deux entités nommées de type ORG, notés *EO*. D'abord, nous reprenons les quatre types de relations économiques proposés par Zhao *et al.* (2010) : *investissement*, *coopération*, *vente* et *approvisionnement*. Cependant, nous combinons les deux relations de *vente* et d'*approvisionnement* en une seule relation *vente-achat*, puisque nous ne nous intéressons pas pour le moment à l'orientation des relations entre les entités (i.e., $R(a, b) = R(b, a)$). Puis, nous rajoutons à la liste deux nouveaux types de relation, à savoir *compétition* et *poursuite judiciaire*.

La construction de ce corpus est passée par trois étapes principales : la collecte des phrases potentielles à annoter, l'annotation des entités nommées et enfin l'annotation des relations. Initialement, un ensemble de documents est collecté à partir du web en interrogeant les deux moteurs de recherche Google et Bing et ce en utilisant une liste de mots clés relatifs aux différents domaines d'activités d'entreprise comme *voitures autonomes, impression 3D, etc*. Les documents collectés passent par une étape de pré-traitement où les en-têtes, les pieds de page et les menus de navigation sont supprimés. Ensuite, le contenu de la page est segmenté en phrases et chaque entité nommée présente est identifiée automatiquement à l'aide des systèmes de reconnaissance d'entités nommées de SpaCy et StanfordNLP. Pour augmenter la précision, ne sont retenues que les entités organisations (*EO*) reconnues par les deux systèmes à la fois. Seules les phrases contenant au moins deux *EO* sont conservées. En revanche, celles dont les mots sont à 95 % des *EO* (comme c'est le cas dans des énumérations d'entreprises) sont rejetées.

Ces phrases sont annotées en relations en fonction du lien qu'entretiennent les *EO* annotées. Soulignons qu'entre deux mêmes *EO*, des relations de types différents peuvent exister dans des phrases différentes. De plus, une seule relation est annotée par phrase même si cette dernière peut exprimer plusieurs relations entre différents paires d'*EO* qui ne se chevauchent pas. Enfin, une relation est annotée si et seulement s'il existe dans la phrase un ou des indices explicites de la relation sans avoir recours à des connaissances externes, c'est l'un des principes d'annotation des corpus ACE. Ceci est illustré dans l'exemple suivant :

*"Présents dans la ville de Wuhan, Faurecian, **PSA**, **Renault** ou encore Valeo ont dû fermer leurs sites situés dans la zone de confinement en attendant le feu vert des autorités chinoises pour reprendre leurs activités."*.

Ici, aucune relation de type économique ne relie les deux *EO* ***PSA*** et ***Renault***, même si ces deux entités peuvent également être reliées par la relation ***Compétition*** car elles partagent le même marché de construction d'automobiles. Cependant, étant donnée que cette seconde relation n'est pas exprimée linguistiquement dans la phrase, elle ne sera pas annotée.

Les types de relations à annoter sont définis comme suit (tous les exemples sont extraits de notre corpus, les *EO* concernées par la relation sont en gras) :

— **Investissements** : une *EO* est filiale d'une autre *EO*, ou *EO* détient (toutes ou une partie) des actions d'une autre *EO*, i.e. : *Le missilier européen **MBDA** (filiale commune de Airbus, Leonardo et **BAE**) espère que l'accord signé à Helsinski lui donnera à terme accès à des financements pour développer de nouvelles versions de son missile antichar de moyenne portée (MMP).*

— **Compétition** : traduit une compétition/rivalité entre deux *EO* fournissant les mêmes biens ou services, ou voulant accéder à un même marché relativement restreint, par exemple : *Boeing et l'avionneur brésilien **Embraer**, rival de **Bombardier** sur les avions régionaux, ont annoncé discuter sur un éventuel rapprochement de leurs activités.*

— **Coopération** : ce type de relation apparaît lorsqu'il existe une coopération contractuelle entre deux *EO*, que deux *EO* travaillent ensemble pour le même projet, par exemple : *Depuis le 25 novembre 2017, 32 associations et startups , 400.000 citoyen.nes, la Fondation **Kering**, **Facebook** et la Région Île-de-France ont travaillé ensemble avec Make.org pour élaborer le premier plan de actions de la société civile contre les violences faites aux femmes.*

— **Poursuite judiciaire** : c'est lorsqu'une *EO* lance une poursuite judiciaire contre une autre *EO*, comme dans : *Grégoire Triet a représenté **Shionogi** dans une action en contrefaçon de brevet portant sur un médicament contre le VIH, qui l'a opposé à **Merck** et ses filiales.*

— **Vente-achat** : une *EO* est cliente de l'autre, ou elle lui fournit des biens ou services, par exemple : *Le capot d'un réacteur d'un **Airbus** A320 de la compagnie **Frontier Airlines** s'est rompu en plein décollage.*

Une classe **Autres** est rajoutée, qui regroupe tous les autres types de relations possibles qui peuvent exister entre deux *EO* et qui ne sont pas des relations économiques.

Un corpus de 10k instances a été construit et annoté par six annotateurs francophones non-experts via la plateforme d'annotation collaborative *Isahit*[3]. L'annotation a été faite par lots contenant chacun 2k instances de relations. Pour chaque lot, 10 % des données annotées sont ré-annotées par des experts. Ceci permet d'évaluer la qualité des annotations et la qualité du guide d'annotation fourni aux annotateurs, et de les améliorer à travers une démarche rétroactive itérative. Sur un total de 1k instances de relations, la moyenne des accords Kappa de Cohen (1960), calculés entre les annotateurs et les experts, est de 0.685. Nous considérons cette valeur comme un bon accord étant donné la diversité des relations économiques à annoter. Le tableau 1 présente le nombre total des relations annotées ainsi que la répartition des instances en corpus de développement, d'entraînement et de test.

	Invest.	**Compét.**	**Coopérat.**	**Poursuit.**	**Vente.**	**Autres.**	**Total**
Entrain.	220	1 229	598	41	188	4 747	**7 023**
Dev.	48	263	128	9	40	1 017	**1 505**
Test.	47	263	129	8	40	1 018	**1 505**
Total	**315**	**1825**	**855**	**58**	**268**	**6 782**	**10 033**

TABLE 1 – Répartition des données annotées par type de relation et type de corpus.

3. https://isahit.com/en/

Le tableau 2 présente quelques statistiques sur notre corpus, avec notamment le nombre total d'*EO* uniques dans le corpus, le nombre de paires d'*EO* ainsi que le nombre maximum, minimum et la moyenne des *EO* par instance de relation. Nous observons que le nombre moyen de *EO* par phrase est de 5 *EO*. Ceci dit, un maximun de 10 relations, en moyenne, pourraient exister dans une seule phrase entre différentes paires de *EO*. De plus, des instances de relations ayant 41 mots en moyenne reflètent la complexité des contextes dans lesquels ces relations économiques sont exprimées dans le web.

	Entrain.	**Dev.**	**Test.**	**Total**
Nb. total EO uniques	1 096	650	688	**1 182**
Nb. paires EO uniques	4 315	1 302	1 305	**5 280**
EO par instance	max = 47 min = 2 moy. = 5	max=40 min = 2 moy.= 5	max = 32 min = 2 moy. = 5	**max = 47** **min = 2** **moy. = 5**
Mots par instance	max = 203 min = 5 moy. = 41	max = 253 min = 7 moy. = 42	max = 189 min = 7 moy. = 42	**max = 253** **min = 5** **moy. = 41**

TABLE 2 – Statistiques sur les EOs dans les relations annotées par type de corpus.

4 Modèles proposés et résultats

Le corpus BIZREL, composé de phrases préalablement annotées en *EO* (deux *EO* par phrase), a été utilisé pour entraîner différents modèles neuronaux afin qu'ils identifient le type de relation économique qui relie les deux *EO*. Notre objectif est de répondre aux deux questions suivantes : (a) Les modèles de la littérature proposés pour la classification de relations génériques sont-ils adaptés pour la classification de relations économiques ? (b) Comment adapter ces modèles pour le français ?

Pour répondre à ces questions, nous proposons cinq modèles. Les trois premiers (M_1, M_2 et M_3) ont obtenu les meilleures scores à la tâche 8 de SemEval 2010 de classification de relations génériques, en anglais. Ces modèles étant indépendants de ressources lexicales et d'outils externes, nous souhaitions tester leurs performances pour la classification de relations économiques. Les deux derniers modèles (M_4 et M_5) sont des adaptations de plongements de mots contextuels issus de modèle de langue basé sur les transformers bidirectionnels pour le français, qui sont utilisés ici pour la première fois pour une tâche de classification de relations. Nous utilisons les plongements de mot français pré-entraînés sur Wikipedia et Common Crawl *FastText* (Joulin *et al.*, 2016, 2017) ayant une dimension de 300 pour les modèles (M_1) et (M_2). Les hyper-paramètres des modèles ont été ajustés sur le corpus de développement. Nous préseontons dans ce qui suit les modèles, puis nous détaillons les résultats obtenus par les expérimentations.

4.1 Description des modèles

Nos modèles sont comme suit :

(M_1) : CNN (Zeng *et al.*, 2014). L'architecture de ce modèle repose sur 3 types de couches :

— Une couche d'entrée qui utilise les plongements de mots pré-entraînés *FastTest* pour associer un vecteur à chaque mot de l'instance. Des plongements de position de dimension 5 sont également calculés par mot pour encoder sa position relative par rapport aux deux EO.

— Trois couches cachées de convolution constituées de 100 filtres chacune, de tailles multiples (des fenêtres de 3,4,5 vecteurs), qui permettent d'extraire des traits au niveau de la phrase à partir de plongements de mot et de position. Au niveau du mot, les plongements des deux *EO* ainsi que de leurs contextes sont utilisés comme traits.

— Une couche de sortie entièrement connectée *(fully connected)* dotée d'un *softmax* qui exploite la concaténation des traits aux niveaux du mot et de la phrase pour calculer la distribution de probabilités des différents types de relation et d'en déduire la relation candidate en prenant le type ayant la probabilité maximale.

L'apprentissage des paramètres du modèle est fait sur 200 *epochs*[4] sur des lots de données *(batch)* de 100 instances, en utilisant la méthode d'optimisation *Adam* (Kingma & Ba, 2014) avec un taux d'apprentissage de 10^{-3}. Une régularisation de type *dropout* (Srivastava *et al.*, 2014) à un taux de 50% est appliquée entre les différentes couches afin de prévenir le sur-apprentissage lors de l'entraînement.

(M_2) : Bi-LSTM avec mécanisme d'attention (Zhou *et al.*, 2016). L'architecture de ce modèle est composée comme suit :

— Une couche d'entrée qui utilise les plongements de mots pré-entraînés *FastTest* pour associer un vecteur à chaque mot de l'instance, avec un taux de dropout de 70% lors de l'entraînement.

— Deux couches cachées composées de cellules LSTM (Hochreiter & Schmidhuber, 1997) bidirectionnelles, de 100 unités par direction avec un taux de dropout de 70% lors de l'entraînement. Ces cellules récurrentes dites "à mémoire court et long terme", permettent de calculer la sortie d'une cellule en prenant en considération l'élément courant de la séquence, mais aussi l'historique passé des cellules précédentes. De plus, la bidirectionnalité permet de considérer le contexte passé et futur pour tout élément de la séquence.

— Une couche d'attention qui fusionne les traits extraits au niveau du mot par le BiLSTM en un vecteur de niveau phrase en les multipliant par un vecteur de poids calculé à partir des sorties des couches LSTM. Ceci permet de capturer les informations sémantiques les plus importantes dans une phrase.

— Une couche de sortie entièrement connectée dotée d'un *softmax* qui exploite les traits extraits par le BiLSTM pour calculer la distribution de probabilités des différents types de relation et d'en déduire la relation candidate en prenant le type ayant la probabilité maximale.

L'apprentissage est fait sur 100 *epochs* sur des lots de données *(batch)* de 10 instances, en utilisant la méthode d'optimisation *Adam* (Kingma & Ba, 2014) avec un taux d'apprentissage de 1, , un taux de décroissance du taux d'apprentissage de 0.9 et un *facteur de régularisation* L_2 de 10^{-5}. Pour prévenir le sur-apprentissage, un *dropout* global à un taux de 50% est appliqué entre les couches cachées et la couche de sortie.

(M_3) : R-mBERT (Wu & He, 2019). Ce modèle est une réadaptation du modèle de langue pré-entraîné BERT (Devlin *et al.*, 2019) pour la tâche de classification de relation en prenant en compte, en plus de la représentation de la phrase, les représentations des deux *EO*.

Un marqueur [CLS] désignant la sortie de classification est rajouté au début de chaque instance de relation, et un marqueur [SEP] est utilisé pour séparer les phrases dans une même instance. Comme

4. Une passe complète sur les données d'entraînement.

réadaptation du modèle, des marqueurs spéciaux sont utilisés pour identifier le début et la fin des *EO* afin de capturer des traits de position les concernant : [E_{11}],[E_{12}] pour *EO1* et [E_{21}],[E_{22}] pour *EO2*. Les instances ainsi modifiées sont introduites au modèle de langue pré-entraîné pour du *fine-tuning* où tous les paramètres pré-entraînés du modèle sont réajustés. À la sortie du modèle de langue, seul l'état caché final est exploité. Les moyennes des vecteurs d'états cachés finaux pour *EO1* et *EO2* sont concaténées avec le vecteur de l'état caché final du marqueur [CLS]. La représentation obtenue passe par une couche entièrement connectée doté d'un classifieur *softmax* qui identifie le type de la relation.

Un modèle pré-entainé de BERT multilingue `bert-base-multilingual-cased`[5], disponible dans la librairie python *HuggingFace*, est utilisé. Le modèle, constitué d'un vocabulaire de $110k$ sous-mots, couvre 104 langues les plus utilisées sur Wikipedia, dont le français. Il compte 12 couches de 768 unités chacune et ayant $110M$ paramètres. Le modèle a été pré-entraîné de façon non-supervisée sur les deux tâches qui sont : (1) *Modèle de langue masqué* (MLM) qui apprend à prédire les sous-mots masqués de façon aléatoire ; et (2) *Prédiction de phrase suivante* (NSP) dans laquelle le modèle apprend à prédire si B est la phrase suivante réelle qui suit A, étant donné une paire de phrases en entrée A, B. Pour l'ajustement des paramètres de ce modèle sur la tâche de classification de relation, la méthode d'optimisation *Adam* est utilisée. La table 3 présente les principaux hyper-paramètres utilisés pour cela.

Nb. epochs	5
Taille de lots par gpu	16
Longueur max. phrase	260
Taux d'apprentissage Adam	2^{-5}
L2 régul. lambda	5^{-3}

TABLE 3 – Réglages des hyper-paramètres.

(M_4) **: R-CamemBERT**. L'architecture du modèle de langue CamemBERT (Martin *et al.*, 2019) est basée sur celle de RoBERTa (Liu *et al.*, 2019), une version optimisée de BERT. Ce modèle est monolingue pré-entraîné sur du textes français issus du corpus multilingue OSCAR (Suárez *et al.*, 2019). R-CamemBERT est adapté à la tâche de classification de relation selon l'architecture proposée par Wu & He (2019), décrite dans M_3. La seule différence du modèle R-mBERT réside dans les marqueurs de classification et de séparation des phrases d'une instance qui sont $< s >$ et $< /s >$, respectivement, pour le modèle R-CamemBERT.

Un modèle pré-entraîné de CamemBERT `camembert-base`, disponible dans la librairie python *HuggingFace*, est utilisé. Ce modèle, constitué d'un vocabulaire de $32k$ sous-mots pour le français, compte 12 couches de de 768 unités chacune et ayant $110M$ paramètres. Il a été pré-entraîné sur la tâche de *Modèle de langue masqué* (MLM) qui consiste ici à prédire des mots masqués aléatoirement au lieu de sous-mots. La méthode d'optimisation Adam est utilisée pour ajuster les paramètres de CamemBERT sur notre tâche. Les mêmes valeurs d'hyper-paramètres qui sont présentées dans la table 3 sont utilisées pour cela.

(M_5) **: R-FlauBERT**. L'architecture du modèle de langue FlauBERT (Le *et al.*, 2019) est basée sur celle de BERT (Devlin *et al.*, 2019). Ce modèle est monolingue pré-entraîné sur du textes français provenant de différentes sources, couvrant différents sujets et des styles d'écriture allant du texte formel (par exemple Wikipédia et livres)[6] au texte aléatoire extrait d'Internet (par exemple Common

5. https://github.com/google-research/bert/blob/master/multilingual.md
6. http://www.gutenberg.org

Crawl [7]). R-FlauBERT reprend également la même architecture proposée par Wu & He (2019), décrite dans M_3, pour prendre en considération les représentations des deux *EO* lors de la classification des relations. Cependant, R-FlauBERT utilise un seul marqueur qui est $< /s >$ pour marquer la classification et la séparation des phrases.

Un modèle pré-entraîné de FlauBERT `flaubert-base-cased`, disponible dans la librairie python *HuggingFace*, est utilisé. Ce modèle, constitué d'un vocabulaire de $50k$ sous-mots pour le français, compte 12 couches de 768 unités chacune et a $138M$ paramètres. Il a été pré-entraîné sur les deux tâche de *Modèle de langue masqué* (MLM) et de *Prédiction de phrase suivante*(NSP). La méthode d'optimisation Adam est utilisée pour ajuster les paramètres de ce modèle sur notre tâche. Les mêmes valeurs d'hyper-paramètres qui sont présentées dans la table 3 sont utilisées pour cela.

4.2 Résultats

Modéle	**Exactitude**	**Précision**	**Rappel**	**F1-score**
M_1 : CNN (Zeng *et al.*, 2014)	76.2	66.8	51.3	57.0
M_2 : Att-Bi-LSTM (Zhou *et al.*, 2016)	74.2	56.6	55.0	53.3
M_3 : R-mBERT (Wu & He, 2019)	80.3	71.2	**64.1**	**67.1**
M_4 : R-CamemBERT	**81.2**	74.6	53.8	59.5
M_5 : R-FlauBERT	80.7	**77.2**	59.7	66.3

TABLE 4 – Résultats de classification de relations économiques sur le corpus de test.

La table 4 présente les résultats obtenus. Nous observons que les modèles basés sur les transformers dépassent ceux basés sur des architectures neuronales de type *CNN* ou *RNN*. Nous constatons également que le modèle (M_3), qui donne aujourd'hui les meilleurs résultats sur le corpus SemEval-2010 Tâche 8 pour la classification de relations génériques, est aussi bon pour la classification de relations économiques, offrant un bon compromis entre *Précision* et *Rappel* (*F1-score*). D'un autre côté, l'adaptation des transformers français à notre tâche (modèles (M_4) et (M_5)) donne des résultats assez bons avec R-FlauBERT qui dépasse légèrement R-CamemBERT. Ceci est semblable aux résultats rapportés par Le *et al.* (2019) sur une tâche de classification de texte du benchmark FLUE.

L'analyse des matrices de confusion telles que obtenues par (M_3), (M_4) et (M_5) montre que ces modèles arrivent à bien distinguer les relations économiques entre elles. Les principales sources d'erreurs proviennent de la relation *Autres*, ce qui montre la difficulté de distinguer les relations économiques des non économiques. Par exemple, pour les relations *Compétition* et *Coopération*, respectivement 163 et 85 instances sont bien classées, 97 et 36 sont classées comme *Autres* et 3 et 5 sont classées comme *autres relations économiques* par le modèle M_3.

5 Analyse d'erreurs

Une analyse manuelle plus poussée des erreurs de classification montre deux autres sources d'erreur. La première concerne les phrases contenant plusieurs relations mais seule celle qui relie les

7. http://data.statmt.org/ngrams/deduped2017

deux *EO* marquées est à identifier, comme dans l'exemple (1). Ici la relation prédite par (M_3) est *Coopération*(EO_2, EO_3) alors que l'annotation de référence indique *Autre*(EO_2, EO_3). En revanche, il existe bien une relation de coopération dans cette phrase mais entre EO_1 et EO_2.

(1) [Sikorsky]$_1$ et ses partenaires [**Hensoldt**]$_2$, Liebherr-Aerospace , MTU , [**Rheinmetall**]$_3$ et ZF lancent le CH-53K qu'il construit pour l'US Marine Corps et qui est le digne successeur de l'actuel CH-53 .

La seconde source d'erreur provient des relations exprimées implicitement, comme dans l'exemple (2) où c'est l'expression métaphorique *se ranger derrière* qui déclenche la relation de compétition entre EO_1 et EO_2. Dans ce cas, le modèle *(M3)* prédit souvent la relation *Autre*.

(2) Plusieurs entreprises de la Silicon Valley dont Google Facebook, [**Dell**]$_1$ et [**HP**]$_2$ se sont rangées derrière Samsung.

6 Conclusions et perspectives

Cet article présente le premier corpus annoté en relations économiques entre organisations pour le français ainsi qu'une approche supervisée pour la classification de ces relations dans des textes issus du web. L'évaluation des modèles de l'état de l'art proposés pour la classification de relations génériques sur notre corpus a donné des résultats prometteurs, ce qui est un premier pas vers l'intelligence économique et concurrentielle à partir de textes pour le français. La prochaine étape est l'adaptation de ces modèles pour prendre en compte les spécificités des relations économiques.

Références

AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R. & IVES Z. G. (2007). Dbpedia : A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 de *LNAI*, p. 722–735 : Springer. DOI : 10.1007/978-3-540-76298-0_52.

AUSSENAC-GILLES N. & JACQUES M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with caméléon. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, **14**(1), 45–73.

AUSSENAC-GILLES N., KAMEL M., BUSCALDI D. & COMPAROT C. (2013). Construction d'ontologies à partir de pages web structurées. In *Journées Francophones d'Ingénierie des Connaissances (IC 2013), Lille*, p. 1–17 : AFIA.

BALDINI SOARES L., FITZGERALD N., LING J. & KWIATKOWSKI T. (2019). Matching the blanks : Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2895–2905, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1279.

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase : A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, p. 1247–1250, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1376616.1376746.

BOSSY R., DELÉGER L., CHAIX E., BA M. & NÉDELLEC C. (2019). Bacteria biotope at bionlp open shared tasks 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, p. 121–131.

BRAUN D., FABER A., HERNANDEZ-MENDEZ A. & MATTHES F. (2018). Automatic relation extraction for building smart city ecosystems using dependency parsing. In P. BASILE, V. BASILE, D. CROCE, F. DELL'ORLETTA & M. GUERINI, Éds., *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 22nd to 23rd, 2018*, volume 2244 de *CEUR Workshop Proceedings*, p. 29–39 : CEUR-WS.org.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

COLLINS M. & DUFFY N. (2001). Convolution kernels for natural language. In *Proceedings of the 14th International Conference on Neural Information Processing Systems : Natural and Synthetic*, NIPS'01, p. 625–632, Cambridge, MA, USA : MIT Press.

CULOTTA A. & SORENSEN J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 423–429, Barcelona, Spain. DOI : 10.3115/1218955.1219009.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DOS SANTOS C., XIANG B. & ZHOU B. (2015). Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 626–634, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1061.

FAN Z., SOLDAINI L., COHAN A. & GOHARIAN N. (2018). Relation extraction for protein-protein interactions affected by mutations. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, p. 506–507, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3233547.3233617.

FREITAS C., SANTOS D., MOTA C., GONÇALO OLIVEIRA H. & CARVALHO P. (2009). Relation detection between named entities : report of a shared task. In *Proceedings of the Workshop on Semantic Evaluations : Recent Achievements and Future Directions (SEW-2009)*, p. 129–137, Boulder, Colorado : Association for Computational Linguistics.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2 : The 15th International Conference on Computational Linguistics*, p. 539–545 : Association for Computational Linguistics.

HENDRICKX I., KIM S. N., KOZAREVA Z., NAKOV P., SÉAGHDHA D. O., PADÓ S., PENNACCHIOTTI M., ROMANO L. & SZPAKOWICZ S. (2010). Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, p. 33–38, USA : Association for Computational Linguistics.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JÉGOU H. & MIKOLOV T. (2016). Fasttext.zip : Compressing text classification models. arXiv preprint : 1612.03651.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431, Valencia, Spain : Association for Computational Linguistics. .

KAMBHATLA N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, p. 22–es, USA : Association for Computational Linguistics. DOI : 10.3115/1219044.1219066.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. arXiv preprint : 1412.6980.

LAU R. & ZHANG W. (2011). Semi-supervised statistical inference for business entities extraction and business relations discovery. In *SIGIR 2011 workshop, Beijing, China*, p. 41–46.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. In *Proceedings of 12th Language Resources and Evaluation Conference (LREC 2020)*. arXiv : 1912.05372.

LEE J., SEO S. & CHOI Y. (2019). Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry*, **11**(6). DOI : 10.3390/sym11060785.

LEE J. Y., DERNONCOURT F. & SZOLOVITS P. (2017). MIT at SemEval-2017 task 10 : Relation extraction with convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 978–984, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/S17-2171.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. arXiv preprint : 1907.11692.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. arXiv preprint : 1911.03894.

MITCHELL A., STRASSEL S., HUANG S. & ZAKHARY R. (2005). Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*.

NGUYEN D. P. T., MATSUO Y. & ISHIZUKA M. (2007). Relation extraction from wikipedia using subtree mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, p. 1414–1420 : AAAI Press.

NGUYEN T. H. & GRISHMAN R. (2015). Relation extraction : Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, p. 39–48, Denver, Colorado : Association for Computational Linguistics. DOI : 10.3115/v1/W15-1506.

SHEN Y. & HUANG X. (2016). Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2526–2536, Osaka, Japan : The COLING 2016 Organizing Committee.

SRIVASTAVA N., HINTON G., KRIZHEVSKY A., SUTSKEVER I. & SALAKHUTDINOV R. (2014). Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15**(1), 1929–1958.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p.9̃.

TAO Q., LUO X., WANG H. & XU R. (2019). Enhancing relation extraction using syntactic indicators and sentential contexts. In *IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 1574–1580. DOI : 10.1109/ICTAI.2019.00227.

WANG L., CAO Z., DE MELO G. & LIU Z. (2016). Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1298–1307, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1123.

WU S. & HE Y. (2019). Enriching pre-trained language model with entity information for relation classification. arXiv preprint : 1905.08284.

XIAO M. & LIU C. (2016). Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1254–1263, Osaka, Japan : The COLING 2016 Organizing Committee.

YAMAMOTO A., MIYAMURA Y., NAKATA K. & OKAMOTO M. (2017). Company relation extraction from web news articles for analyzing industry structure. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, p. 89–92 : IEEE.

YAN C., FU X., WU W., LU S. & WU J. (2019). Neural network based relation extraction of enterprises in credit risk management. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, p. 1–6 : IEEE.

ZENG D., LIU K., LAI S., ZHOU G. & ZHAO J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 2335–2344, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

ZHANG Y., LIN H., YANG Z., WANG J., ZHANG S., SUN Y. & YANG L. (2018). A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, **81**, 83–92.

ZHANG Y., ZHONG V., CHEN D., ANGELI G. & MANNING C. D. (2017). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 35–45, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1004.

ZHAO J., JIN P. & LIU Y. (2010). Business relations in the web : Semantics and a case study. *Journal of Software*, **5**(8), 826–833.

ZHOU D., ZHONG D. & HE Y. (2014). Biomedical relation extraction : from binary to complex. *Computational and mathematical methods in medicine*, **2014**.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 427–434, Ann Arbor, Michigan : Association for Computational Linguistics. DOI : 10.3115/1219840.1219893.

ZHOU P., SHI W., TIAN J., QI Z., LI B., HAO H. & XU B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 207–212, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-2034.

ZUO Z., LOSTER M., KRESTEL R. & NAUMANN F. (2017). Uncovering business relationships : Context-sensitive relationship extraction for difficult relationship types. In M. LEYER, Éd., *Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Rostock, Germany, September 11-13, 2017*, volume 1917 de *CEUR Workshop Proceedings*, p. 271 : CEUR-WS.org.