Analyse automatique en cadres sémantiques pour l'apprentissage de modèles de compréhension de texte

Gabriel Marzinotto, Delphine Charlet, Géraldine Damnati, Frédéric Béchet

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e 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

Analyse automatique en cadres sémantiques pour l'apprentissage de modèles de compréhension de texte

Gabriel Marzinotto[2] Delphine Charlet[2] Géraldine Damnati[2] Frédéric Béchet[1]
(1) Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
(2) Orange Labs, Lannion
(1) {prenom.nom}@lis-lab.fr
(2) {prenom.nom}@orange.com

RÉSUMÉ

Dans le cadre de la compréhension automatique de documents, cet article propose une évaluation intrinsèque et extrinsèque d'un modèle d'analyse automatique en cadres sémantiques (*Frames*). Le modèle proposé est un modèle état de l'art à base de GRU bi-directionnel, enrichi par l'utilisation d'embeddings contextuels. Nous montrons qu'un modèle de compréhension de documents appris sur un corpus de triplets générés à partir d'un corpus analysé automatiquement avec l'analyseur en cadre sémantique présente des performances inférieures de seulement 2.5% en relatif par rapport à un modèle appris sur un corpus de triplets générés à partir d'un corpus analysé manuellement.

ABSTRACT

Semantic Frame Parsing for training Machine Reading Comprehension models

In the framework of Machine Reading Comprehension this paper presents an intrinsic and extrinsic evaluation of a Semantic Frame parser. The proposed model is based on a state of the art bi-directionnal GRU enhanced by the use of transformer-based contextual embeddings. We show that a Machine Reading Comprehension model trained on a corpus of triplets generated from an automatically parsed corpus with our semantic frame parser only yields a 2.5% relative decrease in performance with respect to a model trained on triplets generated from a manually annotated corpus.

MOTS-CLÉS : Analyse en cadres sémantiques, Génération automatique de questions, Compréhension automatique de texte.

KEYWORDS: Semantic Frame Parsing, Question Generation, Machine Reading Comprehension.

1 Introduction

Les systèmes de *Question/Réponse* à partir de documents ont pour but de sélectionner un ou plusieurs passages d'un texte constituant la ou les réponses possibles à une question de compréhension sur le document. Cette tâche de *compréhension de texte* a été traitée avec deux types de modèles : d'une part des modèles basées sur des méthodes de *Recherche d'Information* utilisant un appariement entre requête et document reposant sur des représentations explicites du *sens* des requêtes et des *connaissances* contenues dans les textes avec éventuellement l'accès à des bases externes de connaissances (Kolomiyets & Moens, 2011; Shen & Lapata, 2007) ; d'autres part des méthodes d'appariement direct entre questions et passages de documents utilisant des apprentissages de type *end-to-end* rendu possibles grâce à la disponibilité de très grands corpus contenant à la fois documents, questions et réponses tells que SQuAD (Rajpurkar *et al.*, 2016).

Volume 2 : Traitement Automatique des Langues Naturelles, pages 288–295.
HAL : hal-02784778.

Récemment il a été proposé dans (Béchet *et al.*, 2019b) de combiner représentation explicite du sens et systèmes de question/réponse appris par appariement direct en utilisant un corpus annoté en cadres sémantiques pour générer le corpus d'apprentissage nécessaire à l'adaptation d'un système générique tel que BERT à un nouveau cas d'utilisation. Cette approche, évaluée dans (Béchet *et al.*, 2019c), a cependant une limitation importante : elle nécessite la disponibilité de corpus annotés manuellement en cadres sémantiques afin de générer le corpus d'adaptation. L'étude présentée dans cet article vise à relâcher cette contrainte en développant un analyseur sémantique de type *FrameNet* spécifiquement pour cette tâche de génération de corpus de type Question/Réponse sur documents. Nous étudions les performances de cet analyseur dans ce contexte particulier et proposons une analyse détaillée des résultats en fonction du type de questions posées. Nos contributions visent d'abord l'amélioration des modèles d'analyse en Frames pour le Français, puis l'étude sur la viabilité de la génération des corpus de Question/Réponse à partir des annotations automatiques en cadres sémantiques.

2 Génération de corpus de questions

La tâche de compréhension automatique de texte rencontre un succès grandissant, préfigurant de nouvelles façons d'accéder à l'information contenue dans des documents. Les corpus disponibles sont composés de triplets *(document, question, passage de document constituant la réponse)*, ils sont très majoritairement en anglais (SQuAD (Rajpurkar *et al.*, 2016), MS MARCO (Nguyen *et al.*, 2016)) et leur construction est coûteuse. Dans (Béchet *et al.*, 2019c), les auteurs proposent une approche alternative pour la constitution de corpus d'apprentissage, utilisant un corpus annoté en cadres sémantiques pour générer automatiquement ces triplets à l'aide de patrons. Dans ce protocole, lorsqu'une phrase est annotée en cadre sémantique pour la *Frame* F, pour chaque *Frame Element* E, on peut générer une question dont la réponse est E, à partir de patrons appliqués sur F et les autres *Frames Elements* présents dans la phrase.

Un exemple de phrase annotée en cadre sémantique et les questions générées à partir des annotations est donné ci-dessous :

M. Wildon a laissé plus clairement entendre que si l'Allemagne exécutait sa menace contre le commerce neutre, [*l'Amérique*]$_{\texttt{Speaker}}$ [*lui*]$_{\texttt{Addressee}}$ [*déclarerait*]$_{\texttt{Statement}}$ [*la guerre*]$_{\texttt{Message}}$ *et* [*il*]$_{\texttt{Speaker}}$ *a* [*demandé*]$_{\texttt{Request}}$ [*aux neutres*]$_{\texttt{Addressee}}$ [*de se joindre à lui dans son action*]$_{\texttt{Message}}$.

Questions générées

— *Qui est-ce qui a demandé de se joindre à lui dans son action ?*
— *À qui est-ce que l'Amérique a déclaré la guerre ?*

Dans cet exemple la première question porte sur l'élément `Speaker` de la Frame `Request` ; la deuxième question porte sur l'élément `Addressee` de la Frame `Statement`.

Dans ce protocole, le seul travail manuel requis pour produire les corpus de questions est de définir les patrons générateurs de questions pour chaque *Frame*. Le corpus `CALOR-QUEST` (Béchet *et al.*, 2019a) ainsi généré à partir d'un corpus annoté manuellement en cadres sémantiques a été utilisé pour entraîner un modèle de compréhension de lecture et a donné des résultats très encourageants. Nous proposons dans cet article de développer un analyseur automatique en cadres sémantiques spécifiquement mis au point pour permettre de générer automatiquement, à partir de textes sans annotation manuelle, un corpus d'apprentissage pour les modèles de compréhension de lecture.

3 Analyseur sémantique pour la génération de questions

Dans cette étude nous avons développé un analyseur en cadre sémantique se basant sur un étiqueteur de séquence tel que proposé dans (Marzinotto *et al.*, 2018b). C'est un modèle `biGRU` avec 2 couches de `GRU` bidirectionnelles dans lequel les cadres sémantiques sont codés à l'aide de structures plates reprenant le codage *Begin, Inside, Outside* (`BIO`). Les couches de `biGRU` ont 150 neurones dans chaque direction, et un dropout de 30% entre chaque couche. Nous utilisons Adam comme optimiseur, avec un taux d'apprentissage de $lr = 0.00005$ et des mini-batchs de taille 32. Les séquences d'apprentissage ont une longueur maximale de 120 tokens/word-pieces. Cette taille est suffisante pour tenir compte de plus de 99% des exemples annotées dans CALOR.

Notre système utilise en entrée soit des plongements de mots de type `word2vec` soit des plongements contextuels issus de `BERT` (Devlin *et al.*, 2019). Ces plongements contextuels apparus récemment ont apporté des gains considérables sur des tâches similaires à l'analyse en cadres sémantiques, comme l'analyse en rôles sémantiques de type PropBank (Peters *et al.*, 2018). Cependant, l'impact de ces représentations des mots dans la tâche d'analyse FrameNet a été étudiée uniquement au niveau de la sélection du cadre sémantique (Tan & Na, 2019), et l'étude se limite au corpus FrameNet en anglais. Notre étude étend l'utilisation de BERT à toute la chaîne d'analyse sémantique et l'applique à des corpus en français. Dans tous nos modèles, nous incorporons également des traits linguistiques comme les dépendances syntaxiques, POS, morphologie, capitalisation, préfixes et suffixes des mots de la phrase. Les analyses syntaxiques et morphologiques ont été faits avec un modèle UDPipe (Straka & Straková, 2017) appris sur la FTB d'Universal Dependencies 2.0. Lors de l'apprentissage, nous ajustons les plongements des mots `BERT` ou `word2vec`.

L'originalité de notre approche est d'appliquer au moment du décodage n fois notre analyseur sur chaque phrase sur les n occurrences de déclencheurs potentiels de Frame au sein de la phrase. Les paires { phrase, déclencheur } sont traitées séparément par le réseau, qui prend en entrée une feature indicateur du mot déclencheur. C'est ainsi que chaque paire génère une probabilité de distribution sur les Frames et Frame Elements pour chaque mot de la phase. A partir de l'ensemble des hypothèses produites une dernière phase de décodage implémentant une stratégie de décodage A* similaire à celle proposée par (He *et al.*, 2017) est appliquée. Cette dernière étape permet de garantir la cohérence à la fois des étiquettes `BIO`, mais aussi des relations sémantiques entre Frame et Frame Elements.

L'avantage de cette approche est de permettre facilement l'optimisation de notre modèle par rapport à un point de fonctionnement particulier en terme de précision et rappel pour la détection des Frame et Frame Elements, il suffit pour cela de filtrer parmi toutes les hypothèses produites par l'analyseur de séquence. Nous présentons dans le paragraphe suivant une évaluation intrinsèque de cet analyseur se focalisant sur le type de *questions* qui peuvent être générées par les analyses produites et proposant plusieurs points de fonctionnement qui seront évalués dans l'évaluation extrinsèque sur la tâche de compréhension de documents.

3.1 Évaluation intrinsèque

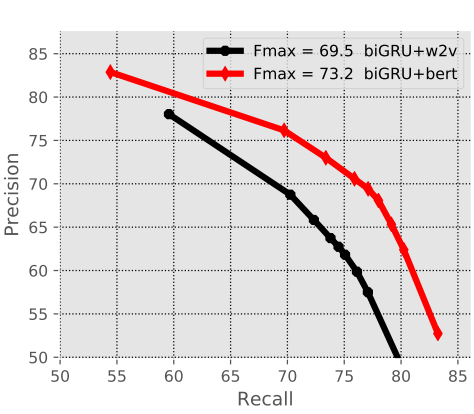
Le modèle est appris et évalué sur le corpus CALOR (Marzinotto *et al.*, 2018a) [1], un corpus de textes encyclopédiques en français annotés en cadres sémantiques (Frames) selon le formalisme FrameNet (Fillmore *et al.*, 2004). Ce corpus est annoté sur un ensemble de 53 Frames différentes pour un total de 31440 occurrences de déclencheurs. Pour cette première série d'expériences le corpus est séparé

1. Corpus disponible : `https://gitlab.lis-lab.fr/alexis.nasr/calor-public/`

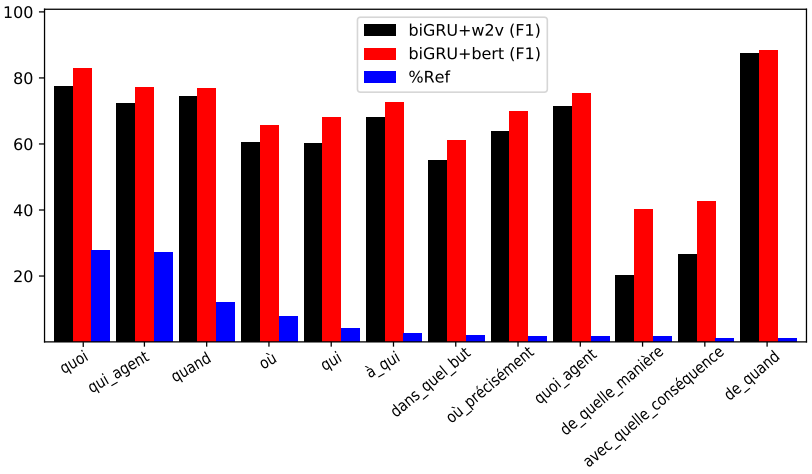
selon une partition de 70% pour l'apprentissage, 10% pour la validation et 20% pour le test. Pour évaluer notre modèle nous utilisons la *F-mesure* sur la tâche de détection et classification des Frame Elements. Nous considérons une détection comme correcte si le recouvrement entre la référence et l'hypothèse contient au moins un mot en commun. Dans l'évaluation, nous propageons les erreurs faits dans l'étape de sélection de la Frame. C'est-à-dire, si un Frame est mal sélectionne, ces Frame Elements seront également faux.

Compromis entre précision et rappel Nous présentons les courbes précision/rappel (P/R) en utilisant différents seuils d'acceptation sur les hypothèses de cadres et de rôles sémantiques produites par nos modèles. Pour dessiner ces courbes, nous utilisons un paramètre $\delta \in (-1; 1)$ qui est soustrait à la probabilité de sortie de l'étiquette *nulle* (ou *Outside*) $P(y_t = O)$ de chaque mot. Par défaut, avec $\delta = 0$, l'hypothèse non nulle la plus probable est sélectionnée si sa probabilité est supérieure à $P(y_t = O)$. Faire varier $\delta < 0$ (ou $\delta > 0$) équivaut à être plus strict (ou moins strict) sur l'hypothèse non nulle la plus élevée. Nous pouvons ainsi étudier le compromis P/R de nos modèles.

Deux variantes du modèle `biGRU` sur le corpus CALOR ont été apprises et évaluées : l'une avec `word2vec` et l'autre avec `BERT`. Le modèle `word2vec` est un modèle cbow de dimension 300 appris sur Wikipedia en français. Le modèle `BERT` utilisé est le modèle multilingue multi_cased_L-12_H-768_A-12 (Devlin *et al.*, 2019). Dans la figure 1a nous observons les courbes précision et rappel pour ces deux variantes. Le modèle `BERT` est supérieur au modèle `word2vec` classique, il atteint une performance de $Fmax = 73.2\%$, avec un $\delta = 0$, soit trois points d'amélioration absolue sur la $F1$ par rapport au précédent modèle à base de `word2vec`. Ce point de fonctionnement est proche du point d'égale erreur ($P \approx R \approx Fmax$). La figure 1a montre aussi plusieurs points de fonctionnement possibles. Les deux points extrêmes pour le modèle basé sur `BERT` présentent respectivement une précision maximale de $P = 83\%$ pour rappel minimal de $R = 55\%$ ($\delta = -0.9$) et un rappel maximal de $R = 84\%$ pour une précision minimale de $P = 53\%$ ($\delta = +0.9$).



(a) Courbe P/R du modèle `biGRU` avec W2V ou `BERT`



(b) Performace du `biGRU` pour chaque type de question

FIGURE 1

Analyse des résultats Afin de mieux percevoir les performances du modèles, nous avons établi un regroupement des *Frame Elements* (FE) selon une question prototypique à laquelle ils pourraient potentiellement répondre. Par exemple, pour la Frame `Hiding`, l'élément Hiding_place est associé à la question prototypique `où`, `Hiding_object` est associé à `quoi` et `Agent` à `qui_agent`.

Pour "qui" et "quoi", nous traitons séparément le cas où le FE est agent de la Frame, auquel cas la question prototypique est `qui_agent` ou `quoi_agent`. Cette distinction permet d'évaluer les cas comme "qui attaque qui". Cette association a été établie pour les 53 Frames du corpus, conduisant à un regroupement des 495 Frame Elements en 64 catégories selon la question prototypique à laquelle ils peuvent répondre.

Les graphiques ne reprennent que les catégories les plus représentées. Ainsi, la figure 1b montre à la fois la performance de l'analyse automatique en Frame par type de question, ainsi que la distribution des types de questions associées aux FEs dans le corpus CALOR. Nous observons que les types les plus fréquents sont *quoi*, *quoi_agent*, *quand*, *où* et *qui*. Les FEs répondant à des questions plus abstraites comme *dans_quel_but*, *de_quelle_manière* et *avec_quelle_conséquence*, sont beaucoup plus rares mais aussi plus difficiles à repérer. Nous observons aussi que `BERT` donne systématiquement de meilleurs résultats que ceux obtenus avec `word2vec` dans tous les types de FEs et que les apports les plus significatifs sont pour les catégories les plus difficiles.

3.2 Evaluation extrinsèque

Pour répondre à la tâche de détection de la réponse à une question donnée dans un texte, nous utilisons une version adaptée sur cette tâche d'un modèle de langue contextuel, ici le modèle BERT multilingue (multi_cased_L-12_H-768_A-12 (Devlin *et al.*, 2019)), avec les hyperparamètres utilisés par ce auteurs pour l'entraînement sur le corpus SQUAD. Afin de se placer dans des conditions équivalentes à SQUAD, les documents de CALOR sont découpés en paragraphes d'une longueur proche de la longueur moyenne des paragraphes de SQUAD (environ 120 tokens).

L'évaluation standard proposée dans SQUAD consiste à comparer, en supprimant les articles, l'ensemble des mots présents dans la réponse détectée à l'ensemble des mots présents dans la réponse de référence. Cette comparaison est faite de façon stricte ("exact-match") pour donner une valeur binaire, ou par le calcul d'une F-mesure issue de la précision et du rappel sur la comparaison des ensemble de mots. Nous reprenons ici cette évaluation standard, en adaptant la liste d'articles au français, ainsi que le protocole expérimental proposé sur CALOR-QUEST par (Béchet *et al.*, 2019a).

Pour entraîner le modèle de compréhension de lecture, nous considérons les corpus suivants : le corpus des questions générées par patrons à partir de l'annotation manuelle en cadres sémantiques (Gold), et les corpus des questions générées par patrons à partir de l'annotation automatique en cadres sémantiques par le meilleur modèle obtenu précédemment (BiGRU-BERT), à différents points de fonctionnement de l'analyseur automatique (de façon à pouvoir traiter l'ensemble du corpus CALOR, les analyses automatiques sont produites par un mécanisme de k-Fold avec k=9).

Le corpus d'évaluation est constitué de 2069 triplets (paragraphe,question,réponse) produits par des annotateurs humains. Ces annotateurs observaient une `Frame`, un `FE Reponse` et des `FEs Contexte`), ensuite ils produisent une question qui a comme réponse le FE indiqué. La phrase originale n'était pas affichée pour laisser plus de liberté aux annotateurs dans les choix lexicaux effectués pour rédiger les questions. Les annotateurs étaient libres de choisir les FE du contexte qu'ils allaient inclure dans leurs questions. Même si les questions sont limités aux Frames et Frame Elements de CALOR, il faut clarifier que ces Frames ont été sélectionnées pour être les plus représentatifs des documents (Marzinotto *et al.*, 2018a). Par ailleurs, le grand nombre de Frame Elements et le dégrée de détail des annotations FrameNet induit une variabilité importante dans les questions produites. Afin de réduire au maximum le biais dû au fait que les questions sont restreints aux Frames du corpus CALOR. Nous générons les exemples d'apprentissage à partir des documents distincts à ceux du qui

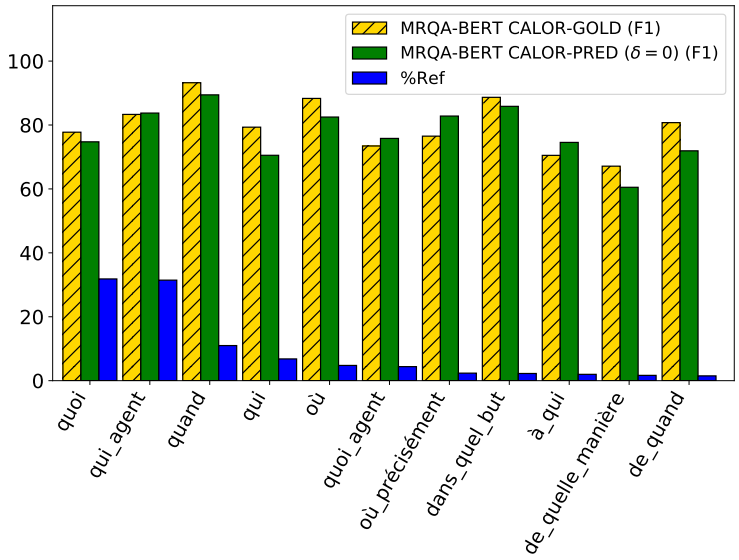


FIGURE 2: Performance du MRQA pour chaque type de question

ont servi pour générer les question manuelles.

Le tableau 1 présente les performances obtenues selon les corpus d'apprentissage utilisés. Pour le corpus d'apprentissage issu de l'analyse automatique, nous considérons les 3 points de fonctionnement suivants : maxP, maxF et maxR. maxP est le point de fonctionnement favorisant la précision de l'annotation en cadre sémantique (80%), maxF celui favorisant la Fmesure (73%) et maxR celui favorisant le rappel (80%). Pour maxP, le corpus est relativement restreint mais avec peu d'erreurs, tandis que pour maxR, le corpus est plus gros mais avec plus d'erreurs. On constate qu'il vaut mieux ne pas favoriser la précision de l'analyse en cadres sémantiques pour la génération du corpus, et qu'il est préférable d'avoir plus de questions, même entachées d'erreurs. En effet, les points de fonctionnement maxF et maxR donnent des performances sensiblement équivalentes. Finalement, les performances obtenues avec un corpus d'apprentissage issu de l'annotation automatique ne sont que faiblement dégradées par rapport aux performances obtenues par le corpus d'apprentissage Gold. Les résultats sur la configuration Gold montrent que même si le corpus de Frames induit un biais sur l'annotation du corpus de test, le corpus obtenu est loin d'être trivial car les modèles de MRQA BERT ont encore une marge d'amélioration considérable.

point de fonctionnement analyseur	#nbtrain	Exact-Match	F-mesure
maxP	8779	62.6	75.6
maxF	12254	67.4	78.6
maxR	13692	67.0	78.7
Gold	17423	69.9	80.6

TABLE 1: Performance de réponse aux questions selon la qualité du corpus d'apprentissage

La figure 2 montre le détail de ces performances de compréhension de lecture selon le type de questions. Le type de question est obtenu par le même protocole d'association que celui présenté à la section 3.1. Il s'agit en effet de la catégorie associée au FE répondant à la question. Les types de questions sont triés par fréquence décroissante et seuls ceux ayant plus de 30 occurrences dans le corpus de test sont présentés ici. La figure permet de comparer, pour chaque type de question, les performances obtenues par apprentissage sur annotations Gold et par apprentissage sur annotations

automatiques. On peut constater que les performances pour chaque type de questions sont assez proches, que le corpus d'apprentissage soit manuel ou automatique.

Une analyse plus détaillée des erreurs produites par le modèle d'analyse en Frames peut permettre d'expliquer pourquoi les performances ne sont que peu dégradées avec le corpus produit automatiquement. Pour le point de fonctionnement `maxF` par exemple, les erreurs commises sur l'identification des FE sont pour 54.2% d'entre elles des insertions et pour 32.0% d'entre elles des omissions. Seules 13.8% des erreurs sont des substitutions dues principalement à des confusions entre deux cadres sémantiques, or les substitutions sont les erreurs les plus enclines à générer des questions erronées. Les erreurs d'omission sont présentes pour tous les types de FE, mais sont plus fréquentes pour les arguments sémantiques abstraits et difficiles. Les omissions n'ayant d'autre impact sur le processus d'apprentissage que de réduire le nombre d'exemples d'apprentissage, les conséquences sur ces types de questions sont moindres.

4 Conclusion

Nous avons présenté un nouveau modèle d'analyse en cadres sémantiques basé sur un modèle bi-GRU associé à des embeddings contextuels de type BERT. Une évaluation intrinsèque détaillée originale sous l'angle de questions prototypiques a permis de révéler une typologie de rôles sémantiques plus ou moins difficiles à détecter et identifier. De façon complémentaire, une évaluation extrinsèque est proposée ou le corpus analysé automatiquement est utilisé pour générer un corpus d'apprentissage pour une tâche de compréhension de lecture. Les expériences montrent qu'une analyse automatique en Frames peut permettre efficacement de générer un corpus d'apprentissage conduisant à des modèles fauiblement dégradés par rapport à l'utilisation d'une annotation manuelle. Ces résultats encourageants pourront être confortés par la suite par des expériences complémentaires sur des données issues de domaines applicatifs différents, au delà des textes encyclopédiques.

Références

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019a). CALOR-QUEST : generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations. In *MRQA : Machine Reading for Question Answering - Workshop at EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. HAL : hal-02317018.

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019b). CALOR-QUEST : un corpus d'entraînement et d'évaluation pour la compréhension automatique de textes. In *TALN 2019*, Toulouse, France. HAL : hal-02377119.

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019c). Calor-quest : generating a training corpus for machine reading comprehension models from shallow semantic annotations. In *MRQA2019, second workshop on machine reading comprehension, satellite workshop EMNLP2019*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint : 1810.04805.

FILLMORE C. J., BAKER C. F. & SATO H. (2004). FrameNet as a "net". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA). Anthologie ACL L04-1221.

HE L., LEE K., LEWIS M. & ZETTLEMOYER L. (2017). Deep semantic role labeling : What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

KOLOMIYETS O. & MOENS M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, **181**(24), 5412–5434.

MARZINOTTO G., AUGUSTE J., BECHET F., DAMNATI G. & NASR A. (2018a). Semantic Frame Parsing for Information Extraction : the CALOR corpus. In *LREC2018*, Miyazaki, Japan. HAL : hal-01959187.

MARZINOTTO G., BÉCHET F., DAMNATI G. & NASR A. (2018b). Sources of Complexity in Semantic Frame Parsing for Information Extraction. In *International FrameNet Workshop 2018*, Miyazaki, Japan. HAL : hal-01731385.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). Ms marco : A human generated machine reading comprehension dataset. arXiv preprint : 1611.09268.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

SHEN D. & LAPATA M. (2007). Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, p. 12–21.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

TAN S.-S. & NA J.-C. (2019). Positional attention-based frame identification with bert : A deep learning approach to target disambiguation and semantic frame selection.