RAPPORT DE STAGE INTERMÉDIAIRE

Lucas Gonzalez—Leclercq
Master Androide UPMC

I. Descriptif du sujet

Le Question Answering est une tâche de NLU (Natural Language Understanding) qui consiste à trouver la réponse à une question donnée en langage naturel dans un corpus de connaissance plus ou moins structuré. Dans le cadre de ce stage, on se focalise sur le cas ou la connaissance est un document juridique (type PDF) dont la structure peut varier.

II. Objectif du stage

L'objectif du stage est d'étudier l'état de l'art en Question Answering et d'adapter une solution pour le domaine juridique. Il s'agit aussi d'essayer de rendre ce processus automatique afin qu'il soit facilement transposable à d'autres domaines.

Il existe 3 grands types d'approches en question answering, chacune présentant ses avantages et inconvénients:

- L'approche « Information retrieval » repose sur l'utilisation d'une base de connaissance explicite ou d'une collection de faits codés en dur. Dans le cas d'un document, on peut considérer chaque phrase comme un fait. Cette approche utilise des techniques traditionnelles pour mesurer la similarité entre la question et les phrases du document : TF-IDF (Term Frequency, Inverse Document Frequency).(préciser) Cette approche est efficace lorsque les données sont larges et souvent renouvelées dans des domaines variables.

- L'approche par ontologie nécessite un travail plus conséquent et pointilleux qui doit être réalisé par un expert du domaine. Elle requiert la construction (par itérations de tests et modifications) d'une représentation structurée du domaine. Ce processus est d'autant plus long et fastidieux que le domaine en question est complexe. Cette approche est la moins flexible mais présente un intérêt lorsque la structure du domaine joue un rôle important dans la tâche à résoudre (par ex: chatbot pour la réservation de billets de train).

- L'approche machine learning repose sur l'extraction de « patterns » dans le texte, et est relativement robuste aux subtilités des domaines complexes n'ayant pas été explicitement prises en compte dans un modèle, un algorithme ou une ontologie. Elle est donc facilement transposable d'un domaine à l'autre sans requérir le travail d'un expert. En revanche, elle repose sur la disponibilité de grandes quantités de données dans le domaine en question.

Le tableau suivant résume les avantages et les inconvénients de ces trois approches (1: avantage, 2: neutre, 3: inconvénient)

	Flexibilité et robustesse	Charge de travail humaine	Gestion de domaines dynamiques	Quantité de données nécessaire
« Information Retrieval »	2	2	1	2
Ontologie	3	3	3	1
Machine Learning	1	1	2	3

Pour construire un système de Q&A performant dans un domaine donné, il s'agit souvent de combiner ces trois approches.

Avant le machine learning, le Q&A nécessitait:

- Des données très structurées
- L'utilisation de techniques de comparaison sémantiques limitées ou une charge de travail importante de la part d'un expert

Le machine learning permet donc à la fois de réduire la charge de travail et d'augmenter la performance du système dans certains cas, **ce qui a motivé le choix d'une approche principalement basée sur le machine learning.**

III. Etat de l'art et positionnement du sujet par rapport à l'existant

Dataset SQuAD (Stanford): https://rajpurkar.github.io/SQuAD-explorer/

C'est un dataset constitué de triplets (contexte - question - réponse). Ce dataset sert de benchmark aux systèmes de Q&A ainsi que pour l'entraînement supervisé de modèles de MRC (Machine Reading Comprehension)

Outils de Q&A « clé en main » :

	Performance	**Zone d'action**	**Intelligence**	**Fonctionnement**
DrQA (Facebook)	10 points en dessous (SQuAD F1 score) du state-of-the art	Recherche la réponse dans une collection d'articles wikipedia	- Généralise bien à des articles généralistes - Capacité d'inférence limitée	- Cherche les N articles les plus probables de contenir la réponse - Prédit un extrait d'article avec un réseau de neurones récurrent prenant des word embeddings en entréee
DeepQA (IBM Watson)	Supérieur à l'humain au jeu tv Jeopardize	Framework à adapter à des problèmes spécifiques	? Certainement supérieure à drQA	Framework de QA complet
Dydu / Chat-Script	Couvre un maximum des questions les plus fréquentes	Listes pré-établie de questions / réponses Rayon limité à la liste des questions établie (pas de généralisation)	- Inférieure aux systèmes QA car pas de généralisation - Incapable de lire des documents - Intelligence conversationnelle	- Paramétrage par l'utilisateur des questions les plus fréquentes - Algorithme pour déterminer l'info demandée - Moteur de chatbot pour simuler des conversations réalistes
Amelia	?	?	Semble être un système 2 en 1 : compréhension de texte et intelligence conversationnelle	?

Les modèles de MRC contenus dans ces outils et provenant de la recherche actuelle fonctionnent bien sur des articles wikipedia mais les différences entre articles wikipedia et documents juridiques sont trop importantes pour appliquer sur les seconds un modèle entraîné sur les premiers. Le sujet de ce stage est donc d'étudier et de développer un système de Q&A plus performant pour un domaine ciblé (le juridique).

IV. Thématiques du stage

Un système complet de question answering requiert une variété de technologies incluant :

- Classification et décomposition de question
- Acquisition automatique de données
- Extraction d'entités et de relations
- Représentation de la connaissance
- Raisonnement (inférence)
- Mesures d'évaluation

Chacun de ces axes de travail peut bénéficier des récentes avancées en machine learning. L'objectif du stage est de les traiter de manière à limiter le plus possible la charge de travail humaine à apporter pour adapter le système à de nouveaux domaines.

Acteurs de référence dans la recherche en Question Answering
Facebook, Google, Microsoft, Allen Institute for Artificial Intelligence, IBM Research

Vecteurs sémantiques (représentation de la connaissance)
Les word embeddings (vecteurs sémantiques) sont une technique de plongement des mots dans un espace de dimension très inférieure à la taille du vocabulaire, capable de capturer une représentation basée sur la sémantique. Dans un tel espace, des mots au sens proche auront des représentations vectorielles proches. Par exemple, on pourrait s'attendre à ce que les mots « chien » et « chat » soient représentés par des vecteurs relativement peu distants dans l'espace vectoriel obtenu. Ces vecteurs sont utilisés dans quasiment tous les travaux de recherche en NLP actuels et fournissent les entrées des réseaux de neurones

LSTM (inférence)
Architecture particulière de neurone permettant au réseau d'avoir une mémoire à court et à long terme (Long Short Term Memory), les réseaux de neurones récurrents les plus performants pour traiter les données séquentielles (comme le texte) utilisent ce genre de cellule.

Modèles de MRC (inférence)
BiDAF (Allen Institute for Artificial Intelligence, University of Washington)
: un des premiers modèles performants
R-Net (Microsoft Research) : un des plus performants

Transfer Learning (acquisition automatique de données et inférence)
Two-Stage Synthesis Networks for Transfer Learning in Machine Comprehension (Stanford University, Microsoft Research)

Avantage de cette approche : le Transfer Learning donne un aperçu rapide de la qualité des traitements effectués sans avoir à inventer et tester des questions à chaque fois: on juge la compréhension des docs sur la pertinence des couples (questions/réponses) inventés par Syn-Net. On fait ensuite confiance aux modèles MRC state-of-the-art comme R-Net.

Reconnaissance d'entités nommées (extraction d'entités et relations)

Permet d'identifier les réponses potentielles dans un texte avant de générer les questions associées avec Syn-Net. Cette étape, dépendante du domaine, est cruciale à la qualité du système final. La librairie SpaCy permet de détecter les types d'entités les plus courants (organisation, personne, lieu, date, etc…) et le projet Snorkel permet de définir et détecter d'autres types d'entités spécifiques au domaine.

Question / Answer analysis - LAT (Lexical Answer Type - classification et décomposition de question)
Building Watson: An Overview of the DeepQA Project (IBM Research)

Merging and ranking of answers
A framework for merging and ranking of answers in DeepQA (IBM Research)

Travaux de recherches en QA spécifiques au domaine juridique
A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker : basé sur l'ontologie LKIF (The Legal Knowledge Interchange Format) datant de 2008. Approche peu flexible, vieillissante, et demandant une certaine expertise du domaine juridique comparé au data-programming (Snorkel, prometteur mais pas encore testé).

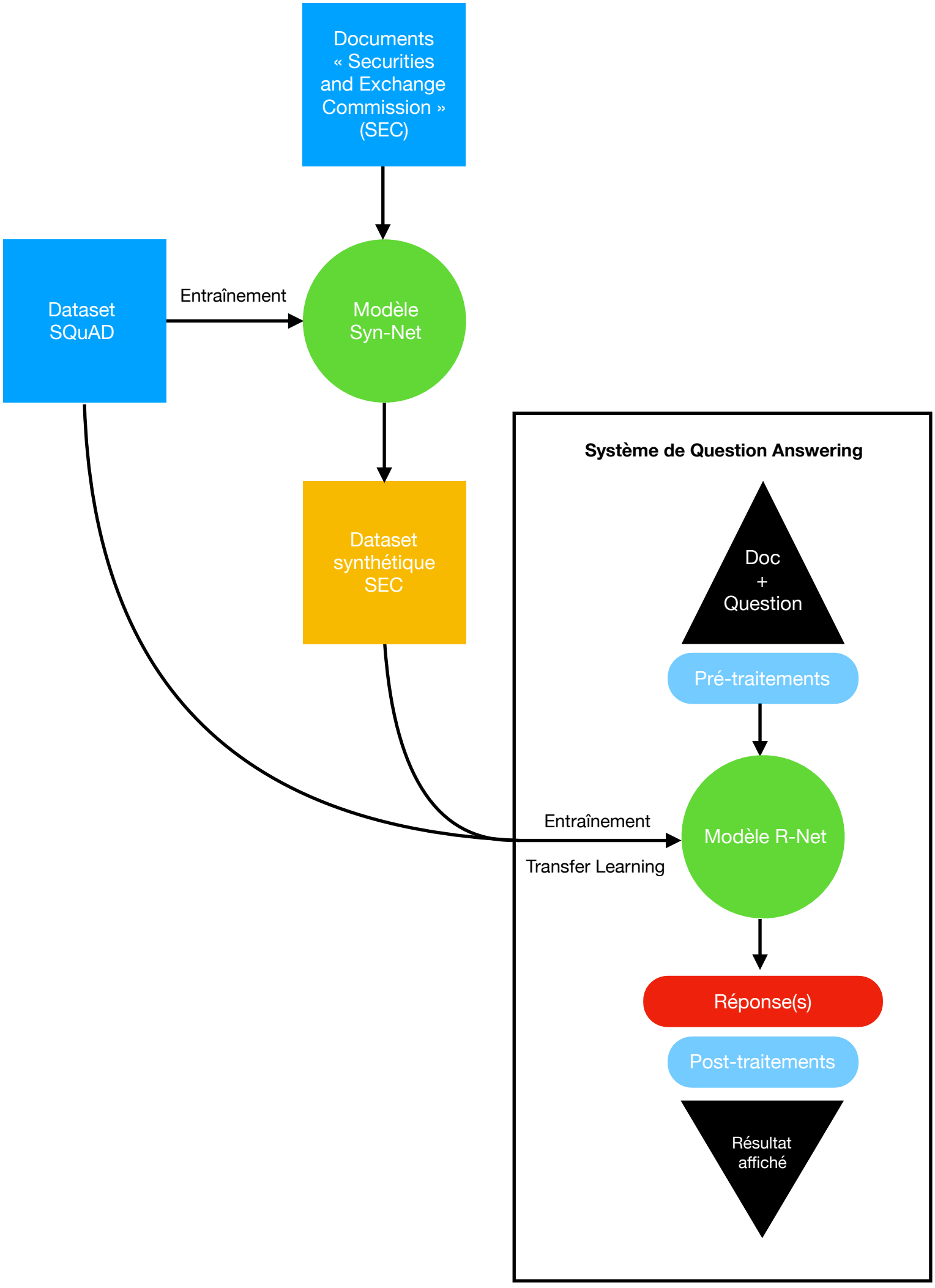
V. Travaux réalisés ou en cours dans le cadre du stage

Utiliser l'état actuel de la recherche pour adapter une solution de Q&A à des documents du domaine juridique.

Particularités des docs juridiques

- Semi-structurés
- Entités nommées spécifiques au domaine
- Longueur variable et parfois très importante (gourmand en mémoire des réseaux neuronaux)
- Vocabulaire propre au domaine

La solution envisagée pour adapter les modèles de MRC au domaine du juridique repose sur une technique de Transfer Learning illustrée sur le schéma suivant.


Documents « Securities and Exchange Commission » (SEC)
Dataset SQuAD
Entraînement
Modèle Syn-Net
Dataset synthétique SEC
Système de Question Answering
Doc + Question
Pré-traitements
Entraînement
Transfer Learning
Modèle R-Net
Réponse(s)
Post-traitements
Résultat affiché

Travaux en cours ou effectués:

- Constitution d'un corpus de documents juridiques (téléchargés sur le site de la Securities and Exchange Commission, onglet Final rules) : conception d'un script pour parser le contenu et la structure des documents pdf. (terminé)

- Étude des modèles de MRC (Machine Reading Comprehension) : R-Net (Microsoft), le plus performant sur SQuAD (implémentation existante et open-source)

- Transfer learning: passage du domaine général à un domaine spécifique en générant automatiquement des données annotées pour l'entraînement supervisé. Implémentation du modèle Syn-Net de Microsoft (modèle implémenté, entraînement en cours)

- Système de QA : assemblage des autres constituants autour du modèle MRC

Possibles améliorations du système à prévoir:

- Reconnaissance d'entités nommées spécifiques (projet Snorkel: data programming) pour étendre les capacités du système

- Traduction automatique du dataset obtenu par Syn-Net pour faire du Question Answering en français

- Question decomposition