satw it's all about technology

# Le Traitement Automatique du Langage Naturel

## Recherches conduites par le Laboratoire de Systèmes d'Information Répartis de l'EPFL

Le Traitement Automatique du Langage Naturel (TALN) est une discipline qui vise à donner aux ordinateurs la capacité de comprendre le langage humain, dit «naturel», qu'il soit écrit ou oral. Le TALN permet ainsi à la machine de comprendre les requêtes soumises par les humains, de les interpréter et d'y apporter les réponses adéquates.

Le TALN est un domaine multidisciplinaire qui combine les connaissances de l'informatique, de la linguistique et de l'intelligence artificielle. Depuis une décennie, le TALN a connu un essor sans précédent rendu possible par les progrès continuels de l'intelligence artificielle – apprentissage automatique (*machine learning*) et apprentissage profond (*deep learning*) – ainsi que la disponibilité de vastes volumes de données en libre accès. Le TALN fait déjà partie de nos vies: moteurs de recherche, assistants personnels vocaux, assistants virtuels (*chatbots*) ou filtrage des mails figurent parmi ses applications les plus largement répandues.

À l'EPFL, le Laboratoire de Systèmes d'Information Répartis (LSIR) étudie actuellement comment adapter le TALN à des domaines spécialisés et à d'autres langues que l'anglais. La mission de ce groupe est de produire des informations fiables à partir de l'énorme quantité de données disponibles sur Internet – un défi majeur dans la société de l'information d'aujourd'hui. Ses chercheurs développent des méthodes et des systèmes destinés à transformer des données non structurées, hétérogènes et non fiables en informations significatives, fiables et compréhensibles.

**Des réseaux de neurones récurrents aux *Transformers***

Pour apprendre un langage aux machines, il est nécessaire de constituer un modèle mathématique du langage naturel. L'approche est probabiliste, basée sur des algorithmes. Les premiers modèles reproduisaient la façon dont nous lisons: un traitement séquentiel de l'information, un mot après l'autre. Dans cette approche, dite «récurrente» et basée sur les réseaux de neurones récurrents (ou RNN pour *Recurrent Neural Networks*), l'apprentissage se construit sous la forme d'un vecteur d'état qui contient une représentation des données rencontrées jusque-là dans la phrase. À chaque nouveau mot rencontré, le vecteur d'état est modifié pour inclure celui-ci et créer un nouvel état. Cette approche permet une compréhension basique et l'exécution de tâches simples. La figure 1 montre une représentation graphique d'un RNN. →

Les RNN classiques sont exposés au problème de disparition de gradient, une dégradation de l'information à mesure qu'ils grossissent et se complexifient. Depuis une dizaine d'années, les progrès de l'apprentissage profond et l'augmentation des capacités ont permis l'apparition des modèles *Transformers* qui ont constitué une véritable révolution. Contrairement aux RNN, le *Transformer* ne traite pas nécessairement les données dans l'ordre, le début de la phrase avant la fin. À la place, il identifie le contexte qui confère un sens à chaque mot de la phrase en s'appuyant sur un mécanisme d'attention pour établir des dépendances globales. Ces modèles ont fondamentalement changé la façon de travailler avec les données textuelles et ont permis l'entraînement des algorithmes sur des ensembles de données plus volumineux qu'auparavant. Cela a conduit au développement de systèmes pré-entraînés tels que BERT, acronyme anglais de *Bidirectional Encoder Representations from Transformers,* développé par Google et publié en 2018. BERT est disponible dans le domaine public et peut être utilisé et adapté par les utilisateurs.

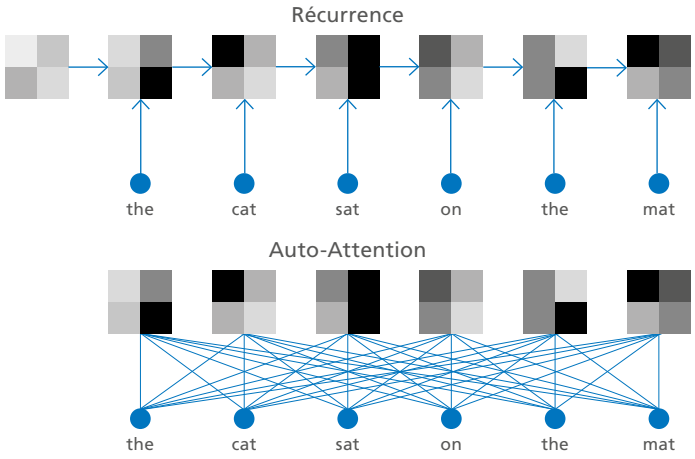


**Figure 1**
Comparaison d'un réseau de neurones récurrent et d'un modèle basé sur un mécanisme d'attention. Dans un RNN, chaque mot est conditionné en fonction des mots précédents. Les dépendances à long terme entre mots sont par conséquent difficiles à modéliser. Avec un modèle basé sur un mécanisme d'attention, chaque mot est conditionné en fonction de tous les autres mots appartenant à la même séquence. Des dépendances entre mots distants sont ainsi modélisables.

## Données et apprentissage

Un grand nombre de données est nécessaire à l'entraînement des algorithmes de TALN. Internet et, en particulier, Wikipédia ainsi que les bibliothèques de livres en ligne sont des sources très utiles de grandes quantités de textes généraux.

Ces textes ne peuvent pas être utilisés tels quels. Ils doivent être transformés en données utilisables par des techniques de traitement comme, par exemple, la *tokenization*, qui consiste en la segmentation du langage en petites unités lexicales (*tokens*).

Le processus d'apprentissage consiste à alimenter l'algorithme avec ces données et à lui assigner des tâches spécifiques qui permettent d'ajuster sa compréhension. BERT a été pré-entraîné avec des tâches de modélisation du langage masqué et de prédiction de la phrase suivante. Le processus d'apprentissage comporte deux étapes. Pendant l'entraînement initial, le modèle est entraîné sur des données non étiquetées. Pour le réglage fin, tous les paramètres sont ajustés en utilisant des données étiquetées.

## Étude de cas I: adaptation du modèle BERT au langage juridique

BERT a été pré-entraîné sur un corpus de langage général. Dans de nombreux cas, il est possible d'utiliser BERT tel quel pour exécuter des tâches linguistiques. Mais des études ont montré qu'il est toujours utile d'adapter un modèle pré-entraîné au domaine spécialisé d'une tâche cible pour améliorer ses performances.

La figure 2 présente un exemple de distribution et de recouvrement des données d'un domaine spécialisé avec le domaine général ainsi que les données spécifiques à la tâche (en gris). Le recouvrement est possible, mais pas nécessaire, ce qui illustre l'importance d'un entraînement spécifique au domaine spécialisé.

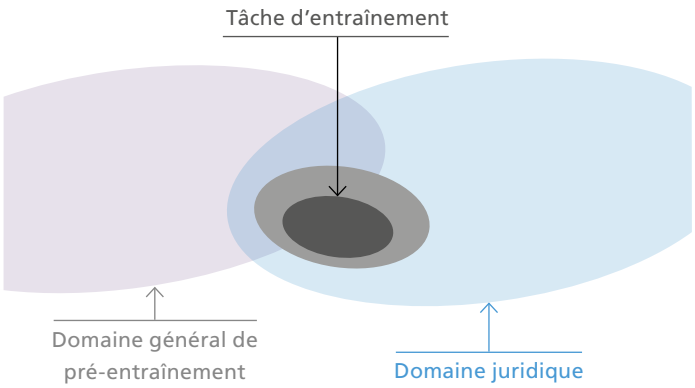


**Figure 2**
Une illustration des distributions de données. Les données de la tâche sont constituées d'une distribution observable de la tâche, généralement échantillonnée de manière non aléatoire à partir d'une distribution plus large (ellipse gris clair) au sein d'un domaine cible encore plus vaste, qui n'est pas nécessairement l'un des domaines inclus dans le domaine de pré-entraînement initial - bien qu'un chevauchement soit possible (adapté de Gururangan et al., ACL 2020).

Le traitement du langage juridique présente plusieurs particularités importantes qui constituent autant de défis. Le langage juridique est complexe et formel par rapport au langage «ordinaire». Le volume de données textuelles disponibles dans le domaine public est limité car un grand nombre de documents sont confidentiels. Enfin, l'utilisation du TALN dans les domaines juridique, réglementaire ou politique peut avoir des conséquences majeures, ce qui n'autorise pas le droit à l'erreur.

L'objectif de l'étude de cas menée par le LSIR de l'EPFL était d'évaluer dans quelle mesure le pré-entraînement adaptatif au domaine juridique d'un modèle dérivé de BERT améliorait sa capacité à réaliser une tâche de classification spécifique.

Dans un premier temps, les chercheurs ont collecté plusieurs dizaines de milliers de documents de droit public européen en libre accès en anglais, français, allemand et italien. La taille du corpus de textes en chaque langue, entre 5 et 6 GB, était relativement petite comparé à la taille des corpus utilisés dans d'autres domaines (table 1).

| Domaine | Nom du corpus | Taille (en GB) | Taille (en GB) |
|---|---|---|---|
| Biomédical | S2ORC | 7,55 | 47 |
| Informatique | S2ORC | 8,10 | 48 |
| Informations | REALNEWS | 6,66 | 39 |
| Avis de consommateurs | AMAZON REVIEWS | 2,11 | 11 |
| Juridique | LEGAL-ROBERTA | 1,01 | 4,9 |

**Table 1**
Caractéristiques de différents corpus de données dans plusieurs domaines. Ces données non étiquetées permettent de réaliser un apprentissage non supervisé pour l'adaptation de modèles aux domaines spécifiques.

Ces données ont été utilisées pour le pré-entraînement adaptatif au domaine juridique du modèle RoBERTa – une version de BERT pré-entraînée sur un corpus dix fois plus grand de données. Une tâche de modélisation du langage masqué a été utilisée (figure 3). Ce nouveau modèle a été dénommé LegalRoBERTa.

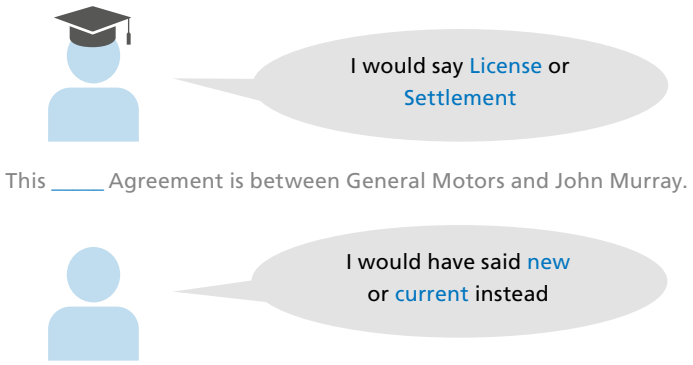


**Figure 3**
Exemple de différence de prédiction de mots masqués entre le modèle LegalRoBERTa (en haut) adapté au domaine juridique, et le modèle BERT (en bas) entrainé sur un corpus de langage général.

L'apprentissage du domaine légal a été affiné par un apprentissage supervisé, soit un apprentissage sur des données annotées de manière adaptée à la tâche en aval, avec des tâches de classification de documents et de recherche d'information sur un ensemble de données de près de 4000 cas juridiques annotés.

LegalRoBERTa a ensuite été évalué sur une sélection de 600 documents précédemment classifiés par des expert·e·s en deux catégories: document juridiquement contraignant et/ou document juridiquement non contraignant.

Les résultats des tâches de classification et de recherche d'information de textes ont été comparés avec ceux produits par BERT et RoBERTa. Un résultat typique de la tâche de recherche d'information assignée à ces trois modèles est montré dans la table 2. Cette tâche fonctionne comme un moteur de recherche: à partir d'une courte phrase (la requête de recherche), l'objectif est de retrouver les documents juridiques les plus pertinents dans un ensemble de plusieurs milliers de documents.

| Modèle | Recall@10 | Rang médian |
|---|---|---|
| BERT | 43,7% | 14 |
| RoBERTa | 46,6% | 13 |
| LegalRoBERTa | **48,3%** | **12** |

**Table 2**
Résultats de la tâche de recherche d'information. À partir d'une expression juridique résumant un long document, le Recall@10 mesure la probabilité de retrouver un document pertinent dans les 10 premiers résultats.

LegalRoBERTa a surpassé les deux autres modèles sur la plupart des métriques et ce, malgré la petite taille du corpus de documents utilisé pour son entraînement spécifique au domaine.

### Étude de cas II: adaptation multilingue

La plupart des modèles de TALN sont pré-entraînés sur des corpus de données en langue anglaise. Il existe une version multilingue de BERT, M-BERT, pré-entraînée sur 104 langues disponibles de Wikipédia, mais cette version n'est pas adaptée au langage juridique.

Afin d'entraîner M-BERT au domaine juridique dans les trois principales langues nationales suisses, deux possibilités sont envisageables: traduire les documents en anglais à l'aide de traducteurs standards, par exemple Google Translate, ou utiliser des adaptateurs de langue.

L'adaptateur de langue est entraîné par apprentissage non supervisé sur l'ensemble des documents en anglais, français, allemand et italien. Ensuite, un adaptateur de tâche est entraîné avec une méthode d'apprentissage à partir de zéro (*zero shot*) qui consiste à demander au modèle de classer des objets sans informations préalables ou à l'aide de représentations intermédiaires. Dans ce cas, il s'agissait de classer les documents français et allemands sur la base des documents anglais.

Dans son étude de cas, le LSIR de l'EPFL a comparé ces deux approches pour des documents en français et en allemand. Les résultats de l'expérience ont montré que l'utilisation des adaptateurs de langue peuvent fournir de meilleurs résultats que l'utilisation de données traduites.

**L'Académie suisse des sciences techniques (SATW) soutient des projets dans le domaine de l'intelligence artificielle**

De nombreuses organisations cherchent à savoir comment l'analyse automatisée des textes peut les aider à mieux utiliser leurs vastes ressources documentaires. Les géants de la technologie accélèrent le développement grâce à l'intelligence artificielle (IA). Mais les conditions cadres sont très différentes. Par exemple, ces géants de la technologie disposent d'énormes ensembles de données et de nombreux feedbacks d'utilisateurs·trices qui peuvent être utilisés pour former des algorithmes d'analyse de texte.

Toutefois, dans de nombreuses applications industrielles et administratives – aussi dans l'administration publique – les ensembles de données sont limités et le feedback provient tout au plus d'experts. Pour ces organisations, la question se pose donc de savoir dans quelle mesure les progrès techniques mondiaux en matière d'analyse de documents peuvent également être utilisés avec profit dans des domaines d'application spécialisés, afin de rendre les informations disponibles plus exploitables et de décharger les utilisateurs·trices finaux des tâches de routine.

La SATW a soutenu une équipe de chercheurs de l'EPFL dirigée par le professeur Karl Aberer, expert SATW, dans la planification, la mise en œuvre et la diffusion du projet pilote présenté dans cette fiche. Les études de cas ont été réalisées pour le compte du Département fédéral des affaires étrangères (DFAE) et représentent des situations comparables dans l'industrie et l'administration publique. Un intérêt particulier a été porté à l'analyse textuelle automatisée des accords internationaux et des obligations juridiques afin d'en faciliter le monitoring et d'en améliorer la qualité.

Dans le domaine de l'intelligence artificielle (IA), la SATW dispose d'une plateforme thématique et d'un programme prioritaire, organise régulièrement des discussions et édite des publications - toujours en collaboration avec des experts et des organisations partenaires. La SATW est également très active dans le domaine des données, car l'IA ne fonctionne pas sans données. Au printemps 2021, la SATW a fondé, avec la Direction du droit international public du DFAE, l'Office fédéral de la communication et la Swiss Data Alliance, le «Réseau d'autodétermination numérique» national pour permettre aux citoyen·ne·s, aux entreprises et aux institutions publiques d'utiliser l'économie des données sur la base des valeurs fondamentales de la démocratie libérale.

**Impressum**

**Auteur*es:** Anne May (Radar RP) et Rémi Lebret (EPFL) | **Chef de projet:** Karl Aberer (EPFL) | **Équipe de projet:** Rémi Lebret (EPFL), Saibo Geng (EPFL), Xiangcheng Cao (EPFL), Stephan Michel (EDA), Claude Schenker (EDA), Manuel Kugler (SATW), Adriana Cantaluppi (SATW) | **Traduction:** Life Science Communication AG | **Rédaction:** Beatrice Huber, Esther Lombardini et Alexandre Luyet

**Publication complète sur le projet**

– Geng, Saibo et al. «Legal Transformer Models May Not Always Help». arXiv preprint arXiv:2109.06862 (2021).

**Autres publications sur le sujet**

– Chalkidis, Ilias, et al. «LEGAL-BERT: The muppets straight out of law school». arXiv preprint arXiv:2010.02559 (2020).
– Devlin, Jacob, et al. «Bert: Pre-training of deep bidirectional transformers for language understanding». arXiv preprint arXiv:1810.04805 (2018).
– Gururangan, Suchin, et al. «Don't stop pretraining: adapt language models to domains and tasks». arXiv preprint arXiv:2004.10964 (2020).
– Liu, Yinhan, et al. «Roberta: A robustly optimized bert pretraining approach.» arXiv preprint arXiv:1907.11692 (2019).
– Pfeiffer, Jonas, et al. «Mad-x: An adapter-based framework for multi-task cross-lingual transfer». arXiv preprint arXiv:2005.00052 (2020).

satw it's all about technology

Swiss Academy of Engineering Sciences SATW
St. Annagasse 18 | 8001 Zürich | 044 226 50 11 | info@satw.ch | www.satw.ch

Photo: Adobe Stock