Qu'est-ce qu'une donnée ?

Rôle des données dans un système d'IA

Au sens numérique général, les données correspondent aux informations qui sont utilisées, traitées et générées par les logiciels d'un système informatique.

Il n'y a pas d'IA sans données. Les données jouent un rôle central dans tous les processus d'apprentissage automatique, car elles sont utilisées à la fois pour la formation et pour les tests. Elles se présentent également sous la forme de paramètres utilisés pour gérer les processus de formation. Enfin, le système d'IA est une combinaison d'une certaine architecture logicielle avec tous les paramètres formés, ce que l'on appelle le modèle, qui est également une donnée.

Il est essentiel de comprendre le rôle des données dans les systèmes d'IA ainsi que la manière dont elles sont sélectionnées, documentées et diffusées pour pouvoir évaluer le comportement d'un système d'IA. C'est important en termes de reproductibilité ou pour comparer deux systèmes d'IA différents.

Dans le monde du traitement du langage naturel par exemple, la disponibilité de grandes quantités de données orales et écrites est centrale pour la performance des correcteurs orthographiques, la prédiction dans les moteurs de recherche, ou bien sûr la traduction automatique. Ces données sont utilisées pour construire ce que l'on appelle des modèles de langage, qui alimentent d'autres processus avec des représentations statistiques de combinaisons de mots ou de phrases.

La durabilité des systèmes d'intelligence artificielle dépend donc étroitement des méthodes de gestion des données utilisées lors de leur conception.

Données pour les systèmes d'IA supervisés et non supervisés

Comme nous l'avons déjà vu, les systèmes d'IA sont de deux types, selon la façon dont les données sont utilisées pour l'apprentissage. Les systèmes supervisés reposent sur la fourniture d'entrées et des sorties correspondantes. L'apprentissage consiste donc à apprendre au système à générer la sortie la plus probable à partir d'entrées inconnues. Il existe plusieurs façons d'obtenir de telles données. Par exemple, une base de données d'images où chaque image est associée à des mots-clés ou une collection de documents numérisés qui ont été transcrits par des annotateurs (cf. figure ci-dessous).

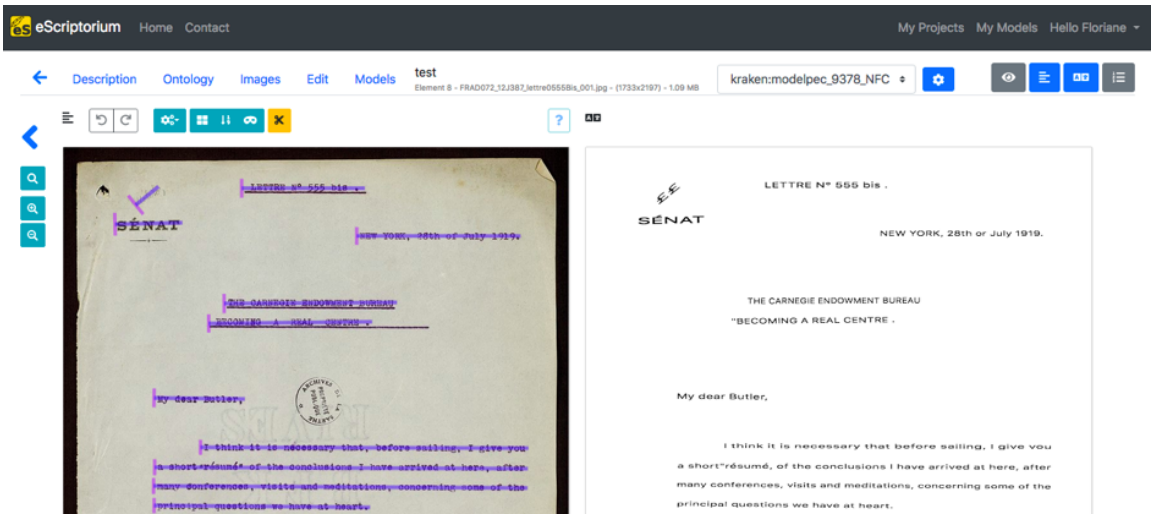

Figure : Transcription automatique d'une lettre de Paul D'Estournelles (avec l'aimable autorisation de F. Chiffoleau, Coll. Archives de la Sarthe).

Les systèmes d'IA basés sur l'apprentissage non supervisé ne seront pas conçus pour un comportement spécifique mais conçus pour supporter la propriété statistique des données d'apprentissage. C'est le cas par exemple des modèles de langage tels que BERT qui tendent à associer des positions similaires dans un espace mathématique à des mots ayant le même comportement syntaxique ou sémantique, observé à partir de la fourniture d'un grand nombre de phrases échantillons pour chaque mot. De tels modèles sont par exemple très efficaces pour prédire les synonymes ou les mots suivants d'une séquence donnée.

Sources - sélection, documentation, préparation, annotation

La conception d'un système d'IA repose essentiellement sur la conception adéquate de l'ensemble de données qui est utilisé pour l'entraîner. Parmi les différents facteurs qui entrent en jeu, on peut citer la pertinence des données pour la tâche à accomplir, la taille des données qui doit correspondre à la complexité de l'architecture du logiciel d'IA --- plus il y a de paramètres mathématiques à entraîner, plus il faut de données --- et la variété des échantillons qui doit refléter la complexité de la tâche.

Selon les sources de données, les données doivent être sélectionnées et souvent nettoyées avant d'être introduites dans le processus de formation. Si l'on prend l'exemple de l'apprentissage d'un modèle de langage sur la base d'un contenu Web, les différents échantillons doivent être triés en fonction de la langue réelle, débarrassés du code Web qui les accompagne (HTML, Javascript, etc.) et éventuellement mélangés pour éviter toute violation des droits d'auteur. Un bon exemple d'une telle préparation des données est la conception du corpus OSCAR [1].

La conception de données annotées pour les systèmes d'IA supervisée est plus complexe car elle implique la conception d'un schéma d'annotation, l'organisation de campagnes

d'annotation et le contrôle de la qualité des données annotées, par exemple en évaluant l'accord entre les annotateurs sur les mêmes données.

Dans l'ensemble, il est essentiel que le processus de conception soit bien documenté pour pouvoir remonter à la source d'éventuels comportements défectueux dans le système entraîné qui en résulte.

Systèmes d'IA biaisés

Comme nous venons de le mentionner, le comportement d'un système d'IA reflète étroitement la nature des données qui ont été utilisées pour le former. Cela peut à son tour induire toute une série de biais possibles qui peuvent résulter de la sélection sous-jacente des ensembles de données. Par exemple, un modèle de langage formé uniquement à partir d'articles de journaux couvrira des types d'expressions et de sujets complètement différents de ceux pour lesquels on a choisi la littérature ou le contenu des réseaux sociaux. De la même manière, les systèmes de génération d'images refléteront la taille et la variété des bases de données d'images sources (par exemple, des œuvres artistiques) qui ont été considérées.

Dans le cas des systèmes supervisés, un biais spécifique peut provenir de la manière dont les étiquettes d'annotation sont conçues ainsi que de la manière dont les annotateurs, avec leur propre bagage culturel, interpréteront les données. Si, par exemple, vous voulez identifier les discours de haine sur les réseaux sociaux, la façon dont les sentiments sont interprétés par les annotateurs peut varier en fonction de l'âge, de la culture et des sentiments personnels des annotateurs vis-à-vis du matériel à annoter.

Dans l'ensemble, il faut toujours garder à l'esprit que les systèmes d'IA sont, par construction, très conservateurs en ce qui concerne leurs données d'entraînement et donc les observables existants. On ne peut pas s'attendre à une réelle créativité de la part d'un système d'IA.

Hébergement, mise en commun, distribution des données

En raison de la taille et de la complexité possible des données d'entraînement des systèmes d'IA ainsi que des modèles qui en résultent, diverses initiatives ont été mises en place pour permettre leur hébergement et leur distribution.

Les ensembles de données et les modèles ouverts peuvent être hébergés dans des dépôts spécialisés (par exemple, la ressource de données d'image[2]) ou dans des dépôts génériques nationaux ou internationaux (par exemple, Zenodo[3]). Ces dépôts fournissent généralement l'infrastructure nécessaire à la gestion des auteurs, des licences, des versions et de l'archivage de leur contenu.

Dans le cas de tâches complexes, où plusieurs équipes travaillent en parallèle sur l'annotation d'une variété d'échantillons de données, certaines initiatives font office de catalogues pour les sources de données correspondantes. C'est le cas par exemple de l'initiative HTR United[4] qui

rassemble les métadonnées des documents annotés pour la reconnaissance de textes (manuscrits).

---

1. Site en anglais Corpus OSCAR : https://oscar-corpus.com/ ↩
2. Site en anglais de ressources de données d'images : https://idr.openmicroscopy.org/ ↩
3. Site en anglais Zenodo : https://zenodo.org/ ↩
4. Site en anglais HTR United : https://htr-united.github.io ↩