Analyse automatique de documentation technique – application sur des retours d'essais en développement

C. Berthou - Safran Aircraft Engines - celine.berthou@safrangroup.com

Résumé

Des essais sont réalisés lors de la phase de développement du moteur et un rapport d'essai est émis lorsqu'un évènement se produit. L'opérateur réalisant l'essai décrit l'évènement en langage naturel. Cette source d'information, sous format textuel, n'est pas ou peu capitalisée à ce jour.

La finalité des travaux entrepris est la mise en place d'un modèle d'apprentissage automatique de la cause de l'évènement, à partir de ces rapports d'évènements, dans un cadre non supervisé faute de label systématiquement renseigné et viable. L'objectif est de capitaliser sur les évènements produits en développement afin d'améliorer l'aide au diagnostic pour les éventuels futurs évènements en service.

Cet article présente les modèles testés pour répondre à la problématique. Nous avons testé un premier modèle de topic modeling LDA (Latent Dirichlet Allocation) puis un modèle de langage neuronal BERT (Bidirectional Encoder Representations from Transformers). L'application de ces deux modèles nous a permis de réaliser une classification automatique des descriptifs d'évènements par cause d'évènement selon deux approches distinctes.

Mots-clés

NLP, apprentissage non supervisé, topic modeling, LDA, transformer, BERT.

Abstract

Tests are carried out during the engine development phase and a test report is issued when an event occurs. The operator performing the test describes the event in natural language. This source of information, in textual format, is not or little capitalized until now.

The purpose of this work is to set up an automatic learning model of the cause of the event, based on these event reports, in an unsupervised learning for lack of a systematically informed and viable label. The objective is to capitalize on the events produced in development in order to improve diagnostic assistance for any future events in service.

This article presents the models tested to solve the problem. We tested a first LDA (Latent Dirichlet Allocation) topic modeling model and then a BERT (Bidirectional Encoder Representations from Transformers) neural language model. The application of these two models allowed us to carry out an automatic classification of the descriptions of events by cause of event according to two distinct approaches.

Keywords

NLP, unsupervised learning, topic modeling, LDA, transformer, BERT.

1 Introduction

De manière générale, nous disposons d'une grande quantité de données texte dans différents domaines d'application : description d'essais, description d'évènements, descriptif de maintenance etc… L'ensemble de ces données est peu exploité alors qu'elles détiennent des informations essentielles à une meilleure connaissance du moteur, son développement, son bon fonctionnement et sa maintenance.

Un besoin est donc identifié d'exploiter cette importante source d'information et de développer des méthodes et des outils de *text-mining* permettant de traiter ces données texte de manière automatique. Plus particulièrement, nous disposons de rapports d'essais d'évènements en développement détaillant, selon le cas, l'essai réalisé, le contexte de l'évènement et sa description.

Nous souhaitons labelliser de manière automatique l'ensemble des rapports d'évènements en développement selon le type d'évènement produit, en utilisant des techniques de *text-mining* (exploitation de données non structurées telles que du texte écrit en langage naturel).

Cet article présente la classification automatique obtenue des rapports d'évènements par typologie d'évènement. Les sections §2 et §3 présentent les deux approches utilisées pour aboutir à ce résultat : en premier lieu, le modèle de topic modeling LDA (*Latent Dirichlet Allocation*) et ensuite, le modèle de réseaux de neurones transformer BERT (*Bidirectional Encoder Representations from Transformers*). La discussion en §4 résume les acquis et les constats afin de proposer des pistes de travaux futurs.

2 Modèle de topic modeling LDA

2.1 Chaine de traitement

Le topic modeling (identification de topics ou sujets) est une méthode de classification non supervisée de documents, équivalente au clustering pour des données numériques [1].

Pour mettre en œuvre ce type de modèle, nous allons appliquer la chaine de traitement présentée en Figure 1.

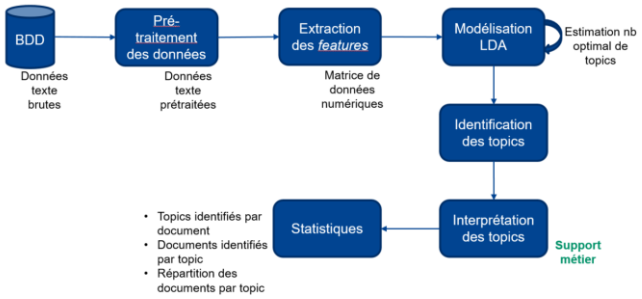

Figure 1. Chaine de traitement pour l'application du modèle LDA.

Les rapports d'évènements d'essais en développement sont extraits de la base de données sous format textuel brut. Les données texte sont ensuite nettoyées et les caractéristiques (*features*) sont extraites. On passe ainsi de données texte à une matrice de données numériques sur laquelle on peut appliquer le modèle LDA, après avoir estimé le nombre optimal de topics d'évènements à extraire de l'ensemble des rapports d'évènements considérés. Une fois les topics identifiés par le modèle LDA, on cherche à les interpréter avec l'aide du métier, et on en déduit un ensemble de statistiques.

2.2 Résultats

Le modèle LDA retenu permet d'obtenir 65 topics. Chaque topic est caractérisé par ses douze termes les plus représentatifs qui sont classés par ordre décroissant d'importance. On peut, à partir de ces termes représentatifs, inférer le topic d'évènement avec le support des métiers. On représente graphiquement ces 65 topics d'évènements en deux dimensions via une ACP (Analyse en Composantes Principales), à partir de la librairie pyLDAvis de Python [2].

Cette approche va nous permettre, après l'interprétation des topics d'évènements identifiés, de constituer une liste de classes d'évènements la plus exhaustive possible, que l'on va utiliser dans l'approche neuronale que l'on a souhaitée aussi tester.

3 Modèle transformer BERT

3.1 Chaine de traitement

BERT est l'acronyme de Bidirectionnal Encoder Representations from Transformers. C'est un modèle de langage pré-entraîné, développé en 2018 par Google, qui repose sur des réseaux neuronaux et est très utilisé en classification de textes. Pour la mise en place de ce modèle, nous allons suivre la chaine de traitement présentée en Figure 2 :

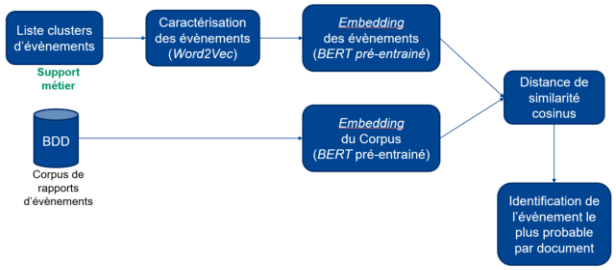

Figure 2. Chaine de traitement pour l'application du modèle BERT.

Après avoir consolidé avec les métiers la liste exhaustive des évènements, nous les caractérisons avec le modèle de word embedding Word2Vec. Puis nous réalisons l'embedding d'une part des évènements précédemment caractérisés, et d'autre part, du corpus de rapports d'évènements d'essais en développement. Nous allons procéder ensuite par distance de similarité cosinus entre un rapport d'évènement donné et chaque cluster d'évènement, pour identifier à chaque rapport d'évènement la combinaison d'évènements associée et ainsi le cluster d'évènement le plus probable.

3.2 Résultats

Maintenant que l'embedding de chaque rapport d'évènement et de chaque cluster d'évènement est réalisé, il reste à mesurer la distance entre chaque rapport d'évènement et chaque cluster d'évènement. On utilise pour cela la mesure de similarité cosinus entre les deux vecteurs correspondants. Pour un rapport d'évènement donné, le cluster d'évènement le plus probable est celui dont la probabilité de similarité est la plus élevée.

4 Discussion et perspectives

Nous disposons d'une base de données importante de rapports d'essais en développement où lorsqu'un évènement se produit en essai, l'évènement est décrit en langage naturel par l'opérateur. L'objectif fixé est de labelliser de manière automatique l'ensemble de ces rapports selon une typologie d'évènement.

Pour cela, plusieurs méthodes ont été testées. Comme les données considérées ne sont labellisées que dans 8% des cas et que la fiabilité de cette labellisation n'est pas assurée, nous avons testé des méthodes d'apprentissage non supervisé.

Nous avons utilisé le modèle LDA de topic modeling. Ce modèle permet d'estimer, par score de cohérence, le nombre optimal de topics d'évènements, définis par les termes les plus représentatifs identifiés par le modèle. Nous avons testé ensuite le modèle neuronal de type transformer BERT. On fournit en entrée du modèle une liste exhaustive d'évènements, et le modèle prédit à partir de la description de l'évènement les clusters d'évènements les plus probables par pourcentages de similarité. On constate, dans tous les cas, la nécessité d'être guidé par la connaissance métier.

Ce projet a permis d'appliquer différentes méthodes de NLP (*Natural Language Processing*) pour des données réelles d'essais, et de développer des outils exploitables et réutilisables pour d'autres types de données texte, symboliques, liées aux connaissances de l'ingénierie : les données de maintenance, les données des questions posées par les clients à Safran avec les réponses associées. Ce premier travail pourra donc être adapté à ces nouvelles données et complété en termes de méthodologies et d'outils.

Références

[1] D. M. Blei, *Probabilistic topic models*, 2012.
[2] Carson Sievert, Kenneth E. Shirley, *LDAvis: A method for visualizing and interpreting topics*, 2014.