Les 10 ans du défi fouille de texte DEFT

CYRIL GROUIN[1]

[1]LIMSI–CNRS, UPR 3251, Orsay, France
cyril.grouin@limsi.fr

1er juillet 2014

Créé en 2005, le défi fouille de texte [1] (DEFT) est une campagne d'évaluation annuelle francophone qui vise à confronter, sur un même jeu de données, les systèmes produits par plusieurs équipes issues de laboratoires de recherche publique ou privée. Les campagnes DEFT se veulent exploratoires et proposent aux participants de travailler sur des thématiques régulièrement renouvellées. L'édition 2014 du défi est la dixième de la série. Cet anniversaire constitue le prétexte pour un premier bilan de la décennie écoulée.

Parce que l'organisation de ces campagnes se fait sans financement, deux problématiques apparaissent chaque année : (*i*) l'obtention de corpus librement accessibles et distribuables, et (*ii*) la constitution des données de références sur chaque tâche de la manière la plus automatique possible. Le premier point est complexe à traiter car, sauf exceptions notables, [2] il n'est pas souhaitable d'enfermer une campagne d'évaluation dans le cadre d'un seul et même jeu de données réutilisé chaque année, au risque de ne plus évaluer que des systèmes capables de ne traiter qu'un seul type de corpus (typiquement des articles de presse). La raison du deuxième point est plus pragmatique dans la mesure où il n'est pas envisageable, faute de financement, d'organiser des campagnes d'annotation humaine de corpus (forcément longues et coûteuses) pour constituer le *gold standard*.

Si les premières éditions reposaient sur les mesures d'évaluation classiquement utilisées en recherche d'information, dans le cadre des dernières éditions, nous avons creusé la problématique de l'évaluation plus en détail, tant pour accorder du sens aux résultats calculés que pour illustrer les possibilités offertes en matière d'évaluation, sortant ainsi du tryptique habituel « Rappel, Précision, F-mesure ».

Cette évolution a émergé lors de l'édition 2011 consacrée à la prédiction de l'année de publication d'un article de presse sur une période de 144 ans. Plutôt que de considérer la tâche comme une tâche de classification de documents parmi cent quarante quatre classes, nous avons réalisé une évaluation reposant sur la similarité entre l'année de référence et l'année prédite en nous fondant sur une fonction gaussienne.

En matière de distance entre prédiction et référence, l'édition 2013 attendait des participants qu'ils classent des recettes de cusisine selon quatre niveaux de difficulté. De manière à pénaliser un écart élevé entre niveau de difficulté prédit et niveau de difficulté indiqué dans la

1. `http://deft.limsi.fr/`

2. Les exceptions concernent, soit le besoin de mesurer l'évolution des performances des systèmes sur des tâches complexes (traduction automatique), soit le besoin de disposer d'outils adaptés à un domaine spécifique (les documents cliniques dans le challenge i2b2, les données en biologie dans le challenge BioNLP, etc.).

référence (e.g., *très facile* vs. *très difficile*), nous avons eu recourt à l'exactitude en distance relative à la solution moyenne (EDRM).

Au-delà du choix des mesures se pose également la question de décider si l'évaluation doit être stricte (la prédiction doit correspondre parfaitement à la référence) ou lâche (on autorise une variation possible entre la prédiction et la référence). Sur l'édition 2012 consistant à produire une correspondance entre articles scientifiques et mots-clés indexant ces articles, nous avons introduit dans la mesure d'évaluation une étape de normalisation de la casse et de lemmatisation des mots-clés prédits. Ce choix revient à opérer une évaluation plus lâche, évitant de pénaliser inutilement un système qui aurait prédit un mot-clé du même champ sémantique que celui renseigné dans la référence.

Si une campagne d'évaluation est perçue, côté participant, comme le moyen de développer ou d'adapter des outils à de nouvelles données et de nouveaux enjeux, du côté des organisateurs, la campagne est le moyen de réfléchir aux mesures d'évaluation les plus pertinentes pour les tâches considérées. La question de l'évaluation et du sens accordé aux valeurs obtenues par les mesures retenues constitue à elel seule un thème de recherche porteur et complexe.