Atelier de clôture du 3ème DÉfi Fouille de Texte

***

3 juillet 2007 – Grenoble (38)

***

Discussion finale

Quelle campagne pour DEFT 2008 ?

Minutes rédigées par: C. Grouin

MG: M. Généreux, PP: P. Paroubek, YT: Y. Toussaint, EC: E. Crestan, FI: F. Ibekwe, LO: L. Olivry, MP: M. Plantié, MHP: M. Hurault-Plantet.

MG Pour la prochaine campagne, nous pourrions inciter les participants à développer des méthodes générales en ne donnant pas aux participants de corpus d'apprentissage, seule serait alors définie la tâche à effectuer.

PP Avec l'expérience acquise dans GRACE et EASY, cela pose problème pour le calibrage des outils y compris pour les équipes expérimentées en évaluation. Il est impossible d'avoir une campagne d'évaluation sans phase de calage (phase d'essais). De plus si seule la tâche à effectuer est définie a priori sans corpus d'essai, nous allons autant évaluer la généricité des méthodes utilisées que la capacité des équipes à mettre en place rapidement une chaîne de traitements sur une problématique nouvelle (comme c'est le cas pour certaines campagnes d'évaluation proposées par le NIST). Ce serait d'autant plus dommageable pour DEFT que beaucoup d'équipes ont regretté le manque de temps qu'elles ont pu consacrer aux évaluations précédentes (beaucoup de soumissions sont faites 2h avant la limite, sans relecture).

YT N'est pas partisan des approches complètement génériques, car elles nécessitent toujours une adaptation des outils aux données et sont donc très proches au final des approches spécifiques. Il est plus enrichissant de partir d'un problème pour améliorer un système existant. Il serait plus intéressant de diversifier la nature des corpus pour tester des approches différentes. Si l'on part sur une approche trop générique on risque d'avoir des résultats moins fouillés.

EC Dans DEFT'07, les moins bons résultats portent sur les « relectures », un corpus rédigé par des personnes qui ne sont pas des professionnels des critiques tandis que les critiques de livres présentent un style littéraire plus formel. Pour information, *Yahoo! QR* dispose de 50% de questions de type opinion qui ne relèvent pas d'une syntaxe formelle mais plus du parlé quotidien.

YT Certaines questions politiques de *Yahoo! QR* ne portent que sur des éléments d'amélioration.

EC On pourrait envisager de prendre en compte la segmentation de l'opinion (par ex. 4 phrases dont 2 positives et 2 négatives). Il y a un fort intérêt pour l'extraction des phrases porteuses d'opinion.

FI Cela pose le problème de l'évaluation, comment construire les données de référence ? Quelles sont les phrases positives et négatives ?

PP N'oublions pas que jusqu'à présent DEFT n'a pas eu de financements, hormis des aides ponctuelles pour la publication des actes et l'organisation des ateliers de clôture. Cela pourrait changer si par exemple nous déposions une proposition de projet ANR.

MHP Pour la construction des données de référence, la première année à TREC dans la tâche question/réponse, chaque participants proposait un certain nombre de questions. Il serait possible que les participants définissent eux-mêmes les passages pertinents.

FI Cela pose le problème du biais induit par les participants dans les données de référence.

PP Initie une discussion pour une tâche à grain plus fin que le document.

FI DEFT'06 portait sur la segmentation thématique, si l'on compare les performances obtenues entre cette campagne et DEF'07, il reste une marge de progression importante, même si les F-scores sont supérieurs. Nous pourrions relancer une campagne d'évaluation sur la segmentation thématique afin de mesurer les progrès effectués dans l'intervalle de deux ans.

MHP Manuellement la tâche était difficile (problème de définir une segmentation de référence).

FI Mais c'est une tâche pour laquelle il existe beaucoup de travaux de recherche.

MP Concernant la possibilité de déposer un projet ANR, même les campagnes DEFT ne sont pas synchrones avec les appels ANR, nous pourrions envisager de déposer une demande en février 2008 pour l'édition 2009. Mais est-ce que du point de vue de l'ANR il s'agit d'un projet porteur ? Le fait d'être un groupement de laboratoires français de taille conséquente suffit-il ? Comme tâche d'apprentissage nous pourrions proposer : comment faire pour classer des textes quand on a très peu de données d'apprentissage ? Ce qui revient de manière plus générale à : comment produire des ensembles étiquetés automatiquement à partir de pas grand chose ?

YT Est-ce qu'il existe des méthodes particulières pour ce genre de tâche ? Tout dépend des étiquettes utilisées.

YT Pour l'ANR, est-ce que les participants peuvent être financés ?

PP Oui à condition d'avoir identifié les participants comme partenaires dans la proposition. Mais il paraît difficile de soumettre un projet sur une tâche que l'on n'aurait pas déjà explorée. De plus se pose la question du financement global du projet, il ne faut pas espérer que chaque participant récupère un financement abondant (50 k€ par labo est impossible, les financements

actuels pour des campagnes d'évaluation sont plus de l'ordre de quelques milliers d'euros par participant). Pour DEFT, l'intérêt porte pour les marchés de veille, d'étude et d'analyse d'image. Nous pourrions peut-être envisager la soumission d'un projet qui combine campagne d'évaluation et développement de pré-prototype industriel. Cela augmenterait les chances que le projet soit accepté.

YT Aucune idée de qui a participé aux éditions de DEFT ? Qui a participé plusieurs fois ? Faire ce vers quoi les gens ont envie d'aller. La tâche d'annotation binaire positif/négatif est un peu réductrice. Elle a l'avantage d'être simple à évaluer. Mais nous pourrions aller vers des tâches nécessitant plus une production de contenus (résumé de contenus, analyse de l'innovant, tendances etc.). Il faut s'orienter vers des compétences propres à un domaine particulier. Attention, la segmentation n'est pas la fouille de texte.

PP Avec des évaluation plus orientées sur le contenu on retrouve le problème de la définition et de la construction du des données de référence.

FI Propose de travailler sur la classification. Les classes qui peuvent être découvertes automatiquement ne regroupant des documents ne correspondent en général pas à des taxonomies existantes.Un moyen indirect d'évaluation pour une annotation en « concepts » est d'évaluer la classification des documents ou extraits de document où sont représentés les concepts.

EC Nous pourrions trouver des données de référence déjà annotées dans sur le web, par ex. avec les blogs, où les documents sont étiquetés par les utilisateurs (avec des mots clés).

PP Un peu comme les folksonomies de *Wikipédia* ?

EC Pas exactement, il s'agit de pointeurs étiquetés vers des sites préférés (par ex. *Daily Motion*). A quel point est-on capable de générer les classes de pointeurs ?

FI Arrivent-t-ils à utiliser les mêmes classes pour plusieurs sites ? Comment sont effectués les regroupement de plusieurs traits comparables ?

EC Les documents sont étiquetés par plusieurs utilisateurs distincts. Dans 90% des cas, les mots clés se retrouvent dans les sites cibles des pointeurs.

PP Est-ce qu'on a suffisamment de données ?

EC C'est un système comparable à celui des publications scientifiques avec leurs mots-clés.

MP Pour les prochaines campagnes, il faudrait augmenter le nombre de valeurs de jugements (très positif, positif, neutre...). Trier des documents dans un domaine, puis dans des sous-thématiques (ceux associés aux acteurs, aux scénarios)...

EC Propose d'utiliser les résumés des livres dans *Amazon*, étiquetés par catégorie de genre.

MHP Rappelle que pour DEFT07, sur le nombre de classes d'opinion, des changements d'échelles ont été effectués car dès que l'échelle s'élargit, les humains ont du mal à s'accorder sur

la valeur à attribuer à un item. Augmenter le nombre de valeurs d'opinion est très difficile.

MP Mais lorsque l'échelle est trop réduite, la note exprime quelque chose de synthétique qui ne correspond pas toujours à l'opinion exprimée.

MHP Les personnes qui mettent les étoiles (marques de qualité) dans « *À voir à lire* » ne sont pas les mêmes que celles qui ont rédigé les critiques. Par rapport à la classification de textes d'opinion, est-ce qu'il y aurait d'autres tâches ? Faut-il passer à autre chose ?

FI Nous pourrions étendre le protocole actuel en typant les traits qui ont conduit aux classifications, pour faire émerger en surface les éléments qui ont permis la classification, mais cela pose encore une fois le problème de l'évaluation (comment comparer les indices).

PP Les données de référence peuvent parfois être disponibles, par exemple nous pourrions travailler sur l'historique des pages de *Wikipédia* et sur les tentatives de vandalisme, en proposant d'identifier automatiquement les tentatives de vandalisme à partir des différentes versions d'un page qui sont préservées dans la base *Wikipédia*.

MP C'est une tâche qui rejoint la détection de spam.

PP Pas complètement, car la détection de vandalisme ne porte que sur une partie du document (celle qui a été réécrite) et sur plusieurs versions successives (évolution dans le temps).

YT Propose une tâche de suivit de thème, càd identifier dans des documents des passages qui parlent de la même chose.

MP Cela revient à ce qui nous avions fait dans DEFT avec la tâche Mitterrand/Chirac.

PP Une semaine avant la chute du mur de Berlin, la police secrète de l'ex-Allemagne de l'Est (STASI) a détruit les archives de ses espions. L'unique broyeuse de documents est tombée en panne. Les documents ont été déchirés à la main en attendant que des camions viennent les prendre pour les brûler. Les camions ne sont jamais venus. Une cellule a ensuite été mise en place pour reconstituer les documents. Un laboratoire allemand travaille sur la reconnaissance optique des morceaux de documents pour les réassocier. Grâce aux outils informatiques, les archives seront reconstituées en 5 ans au lieu de 50 ou 100 ans si l'on devait faire le travail à la main. Une tâche pour DEFT 2008, pourrait consister à reconstituer des documents à partir de fragments artificiellement créés. C'est une tâche qui s'apparente fortement au suivi de thème. Problème de la constitution du corpus initial surtout si pris sur le Web, qui pourrait être trop facilement accessible aux participants.

MG Quelle application correspond à cette tâche?

PP La STASI utilise uniquement des paramètres visuels (couleurs des documents, formes de découpe, critère de texture). Rien sur l'analyse sémantique. Apparemment, pas d'OCR.

YT Application sur la reconstitution de chèque. Cette application étant tellement d'actualité, nous pourrions avoir facilement des financements pour les campagnes.

PP Application possible : réalignement de corpus textuels. Actuellement, le réalignement dynamique pose problème si les modifications apportées sur le document sont trop importantes.

YT Reconstitution d'un texte par rapport à son origine est artificielle. Intérêt pour la manière dont chaque article parle de item particulier (ontologie). Comment chaque article parle d'un concept donné ? Il s'agit d'un tâche complexe.

PP Quelle référence ? Dictionnaire électronique.

YT Est-ce qu'on sait reconstituer un texte ?

PP Oui, il existe des algorithmes et méthodes que l'on pourrait combiner. Ouvre un large panel de domaines (critères formels, sémantiques, thématiques etc.). Cf. par ex. algorithmes développés pour l'analyse du génome. Par contre, un des grands avantages de la tâche DEFT'07 est que le domaine applicatif est clairement identifié.

LO Pourquoi ne pas faire une tâche sur la classification de documents ?

MHP Les problèmes pour cette tâche ont déjà été évoqués, elle pose le problème des corpus accessibles sur le web.

YT Travaille avec des astronomes à Strasbourg sur une base de données d'objets célestes classés dans des catégories (thésaurus spécifique, ex. étoile, étoile-double, astéroïde). Est-ce que la classification utilisée correspond à celle utilisée dans les textes ? Nous pourrions faire automatiquement des classes d'objets célestes à comparer avec la base de données des astronomes (référentiel). Dans une table, nous disposons des objets et de leur classe. Dans une seconde table nous avons les caractéristiques issues de textes (environ 3 000 résumés). La tâche consisterait à partir des articles d'astronomes soit à identifier un objet particulier à partir des caractéristiques mentionnées dans les contextes où apparaît l'objet (par ex. « *F358 a un période de rotation de 28h, émet une lumière bleue* », etc.), soit proposer une classification des objets, à partir des résumés, qui soit cohérente avec la nomenclature de référence.

FI Est-ce que tous les objets ont des noms « barbares » comme F358 ou bien existe t-il des mots conventionnels ?

YT Ce ne sont pas des noms, juste des identificateurs normalisés (ex. NCG456). Mais les noms de leurs caractéristiques ne sont pas standardisés dans les résumés où ils sont mentionnés. D'ailleurs, Strasbourg a normalisé les colonnes de caractéristiques dans la table de référence.

FI Intérêt de la tâche réside dans le classement des objets à partir de leurs caractéristiques, avec les problèmes d'ambiguïté usuelle, par ex. plusieurs objets différents peuvent être mentionnés dans un même résumé. Mais si les noms sont trop exotiques, est-ce que ce n'est pas plutôt une tâche de reconnaissance d'entités nommées ?

YT Le problème n'est pas le nom de l'objet, mais celui de trouver les caractéristiques de l'objet.

EC Cela va poser le problème de la distance entre deux objets.

PP Propose de continuer la discussion par courrier électronique sur une liste de discussion que mettra en place le LIMSI pour l'organisation de DEFT 2008. En particulier la pister proposée par YT sera examinée en détails au moyen d'échantillons de corpus.