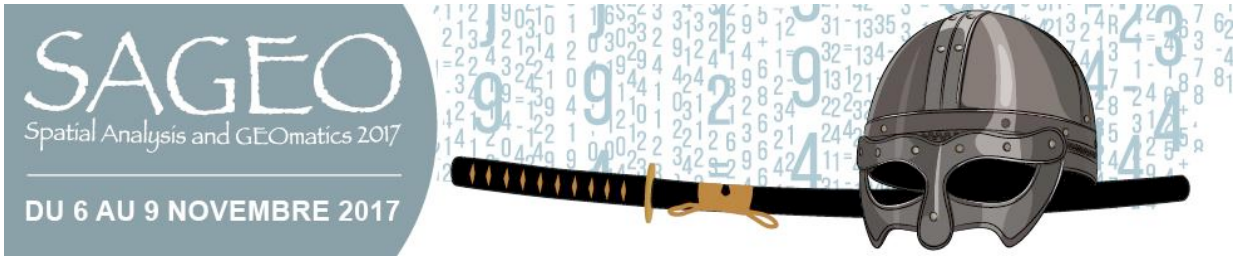

Actes de l'atelier

EXCES - **EXtraction de Connaissances à partir de donnéEs Spatialisées**

Eric Kergosien (GERiiCO, Université Lille 3)
Christian Sallaberry (LUIPPA, Université Pau, Pays de l'Adour)
Maguelonne Teisseire (IRSTEA, UMR TETIS, Montpellier)

https://sageo2017.sciencesconf.org/resource/page/id/20

Lundi 6 novembre 2017, Rouen

TABLE DES MATIÈRES

Session 1 : Analyse spatiale et médias sociaux

Conférencier invité
Davide Buscaldi (Université de Paris 13, LIPN)
Titre : « Information Géographique, textes et média sociaux»
Résumé : Au cours des années 2000, de nombreux chercheurs ont proposé diverses techniques pour améliorer la récupération de l'information géographique par l'analyse spatiale, le filtrage et le reclassement en fonction des toponymes identifiés dans les textes. Certaines estimations indiquent qu'au moins 70 % des informations contenues dans les textes en ligne contenaient des informations géographiques, souvent en forme de toponymes. Aujourd'hui, le déclin de la blogosphère et le succès des médias sociaux dans une société connectée rendent l'information géographique encore plus importante, au centre d'événements clés, tels que les phénomènes météorologiques, les catastrophes naturelles, les mouvements sociaux, les événements sportifs et plus encore. Extraire et analyser l'information géographique dans ces contextes présente des défis intéressants qui seront au centre de cet exposé.

« Gemedoc : Un outil pour annoter les correspondances entre les documents »
J. Fize, M. Teisseire, M. Roche

Session 2 : Exploitation automatisée de textes – toponymie, dynamiques spatiales

« Comment les hôtes et clients d'Airbnb parlent-ils des lieux ? Une analyse exploratoire à partir du cas parisien »
M. Guérois, M. Madelin

« Twitter comme corpus numérique d'analyse des représentations territoriales. Application au Parc national des Calanques de Marseille Cassis La Ciotat »
S. Fan, Ph. Deboudt, A. Fraisse, E. Kergosien

« Calcul de similarité entre événements sociaux »
A. Fotsoh, C. Sallaberry, A. Le Parc - Lacayrelle

Gemedoc : Un outil pour annoter les correspondances entre les documents

Jacques Fize[1], **Maguelonne Teisseire**[1], **Mathieu Roche**[1]

** UMR 9000 TETIS, Cirad, Irstea, CNRS, AgroparisTech, Univ. Montpellier Maison de la Télédétection, Montpellier, France*
{firstname}.{lastname}@teledetection.fr

RÉSUMÉ. *Nous présentons* Gemedoc, *une plateforme pour annoter la similarité inter-document pour un corpus sur différentes dimensions : thématique et spatiale. Pour évaluer la similarité, nous avons conçu un protocole d'annotation divisé en deux étapes : (1) l'identification de descripteurs pour chaque dimension; (2) l'annotation de la similarité sur une échelle de 4 degrés. À terme, les annotations récoltées doivent permettre de construire un corpus destiné à évaluer les méthodes et les représentations dans des applications de mise en correspondance de documents.*

ABSTRACT. *We present* Gemedoc *a platform for text similarity annotation for a corpus on different dimensions: spatial and thematic. In order to annotate the similarity between two documents, we designed an annotation protocol divided in two steps: (1) identification of dimension features; (2) similarity annotation on a 4-degree scale. Ultimately, gathered annotations will permit to build a corpus aimed to evaluate methods and representations in text matching applications.*

MOTS-CLÉS : *fouille de textes, mise en correspondance de textes, plateforme d'annotation*

KEYWORDS: *text mining, text matching, annotation platform*

Introduction

En Recherche d'Information, l'identification de correspondances ou l'alignement entre données textuelles est essentiel. De manière générale, cette recherche se concentre sur la mise en relation de deux objets, la requête et le(s) document(s) affilié(s). Ces deux objets comparés sont souvent de tailles différentes et pour surmonter cette différence, diverses méthodes sont mises en place pour extraire le plus d'informations possibles. Parmi celles-ci, nous pouvons mentionner les travaux d'extension de requêtes (Xu, Croft, 1996 ; Dalton *et al.*, 2014), qui ont pour objectif d'étendre l'ensemble des descripteurs associés à la requête.

D'autres travaux utilisent des méthodes d'alignement de documents, notamment, dans le domaine de questions-réponses (Voorhees *et al.*, 1999 ; Voorhees, 2001 ; Dang *et al.*, 2007), la détection de plagiat (Potthast *et al.*, 2010) ou encore la traduction utilisant l'alignement bilingue de documents (Zou *et al.*, 2013).

Dans nos recherches, nous nous intéressons à la mise en correspondance de documents hétérogènes. Il s'agit de développer des modèles de représentation et des méthodes dédiées à la recherche de similarité entre les documents selon différentes dimensions : la thématique, la spatialité et la temporalité. Les contributions sont nombreuses tels que : la découverte de connaissances, la mise en relation entre des producteurs de données, la cartographie de corpus, etc.

Dans cet article, nous présentons Gemedoc, un outil permettant d'annoter la similarité inter-document pour un corpus. Il est accompagné d'un protocole d'annotation fondé sur la similarité entre documents selon deux dimensions : la thématique et la spatialité. À termes, les résultats récoltés à travers diverses annotations doivent permettre de construire un corpus destiné à l'évaluation des méthodes de mise en correspondance de documents.

1. Évaluer la similarité entre deux documents

L'établissement d'un protocole d'annotation de la similarité entre deux documents selon une dimension, est difficile à définir. Il s'agit de trouver un équilibre entre deux extrêmes. L'un, qui est de définir strictement le processus d'annotation au risque de biaiser les résultats. L'autre, qui est de laisser l'affect de l'utilisateur inférer sur l'annotation et la rendre inutilisable.

Par conséquent, nous avons choisi de définir un protocole d'annotation simple, en laissant l'utilisateur libre sur la base de comparaison des documents, tout en lui donnant peu d'indices.

1.1. Modalités d'annotation

Pour annoter la similarité entre deux documents, nous avons choisi une échelle avec 4 degrés de similarités :

– **Ne sais pas.** L'annotateur ne sait pas évaluer la similarité entre les deux documents.

– **Différent.** L'annotateur indique que les documents n'ont (ou presque) rien en commun.

– **Similaire.** L'annotateur indique que les documents partagent quelques similarités.

– **Très similaire.** L'annotateur indique que les documents sont presque identiques.

1.2. Procédure d'annotation

Dans cette partie, nous illustrons la procédure d'annotation à l'aide des deux textes [1] ci-dessous. Ces deux documents traitent de la situation des migrants à Idomeni, en Grèce : l'un est un résumé, l'autre est un témoignage d'une infirmière présente sur les lieux.

Texte 1 The winding road across the wheat fields near the Greek village of Idomeni is full of people carrying large bags on their shoulders, babies in their arms and putting one step in front of the other. The stream of humanity continues day and night but not an average of 150 a day, (and only Syrians and the Iraqis who are lucky enough to have a passport or ID card from their home country) can continue the journey out of this place and across the border into the Former Yugoslav Republic of Macedonia (FYROM) and onwards to western and northern Europe. Few are leaving but more, many more keep coming, only to end up getting stranded in what is becoming unsustainable humanitarian situation. Today, in a transit camp that has the capacity to host 1,500 people, there are more than 11,000 crammed in trapped without information, in a mix of anxiety and delusion.

1. Source : Médecins Sans Frontières 2017

Texte 2 Daniela, an MSF nurse in Idomeni sums it up "there is confusion, stress. Lack of reliable information. There is a growing feeling of anger. Many refugees have been waiting here for over ten days. People are extremely exhausted." In the clinic that MSF operates in Idomeni, whole families, pregnant woman and kids arrive in a constant stream, as do many disabled people and elderly people suffering from chronic diseases. People, including babies and the elderly, are forced to sleep out in the cold, with just with a sleeping bag to keep them warm. The big tents made available by MSF have been full for days, and hundreds of small tents, are spread everywhere, even next to the train track. Omar, 24 years old, a Palestinian refugee from Homs camp in Syria is exhausted "This is making me very nervous, I don't know what is coming next. This waiting is killing me. We feel ignored here."

Comme énoncé précédemment, la similarité entre ces deux documents est évaluée selon deux dimensions : la thématique et la spatialité.

Similarité thématique
La similarité thématique est souvent perçu à travers le vocabulaire utilisé. Dans les deux exemples ci-dessus, les deux textes partagent de nombreuses thématiques communes telles que : l'aide humanitaire, la migration, les épreuves subites et la famille. Par conséquent, thématiquement, on considère que ces deux textes sont très similaires.

Similarité spatiale
Contrairement à la similarité thématique, la similarité spatiale peut dépendre de plusieurs facteurs qui varient selon ce qu'on cherche. Une première approche serait de comparer le contexte générale des deux textes, ici Idomeni. Une deuxième approche, consisterait à comparer les entités spatiales identifiées (Syrie, Idomenie, Macédoine, etc.). Enfin, dans l'étude de textes migratoires, une troisième approche comparerait les deux documents en se focalisant sur la similarité des parcours des individus.

Si l'on fonde notre raisonnement selon la première approche, les deux textes sont très similaires. Tandis que la deuxième approche fait ressortir quelques différences. Par conséquent, dans leur spatialité, ces deux textes sont partiellement similaires.

Gemedoc permet de capturer ces différences de similarités entre deux documents, selon la dimension étudiée.

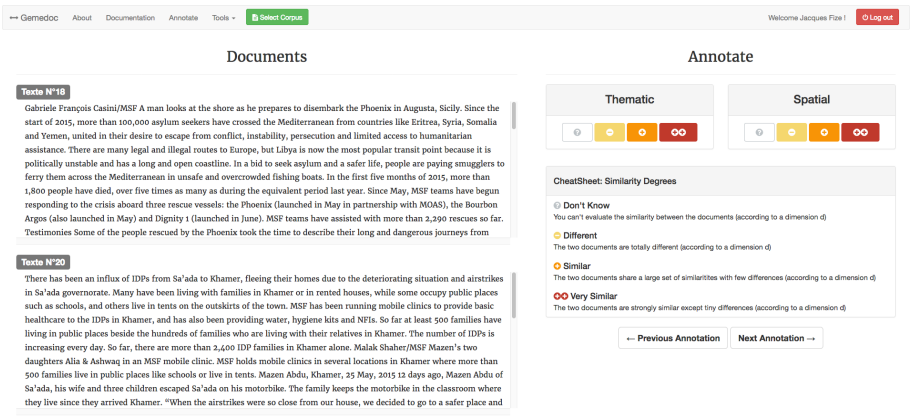

FIGURE *1. Aperçu de l'interface de* GEMEDOC

1.3. Corpus utilisés

Notre objectif étant de réaliser une mise en correspondance selon différentes dimensions (thématique, spatialité), nous récoltons divers corpus où ces deux dimensions sont exploitées. Pour cause, si la thématique est intrinsèque à tous documents, ce n'est pas forcément le cas de la spatialité.

2. Un outil dédié : Gemedoc

Afin d'annoter la similarité entre les documents d'un même corpus, nous avons décidé de développer un outil dédié : GEMEDOC. GEMEDOC est une application Web fonctionnant à l'aide d'un programme Python utilisant le module Flask [2]. Nous avons choisi ce format pour faciliter la mise en place de l'outil que cela soit :

– dans sa conception : interface en HTML5/CSS ;

– dans sa publication : hébergé donc aucune installation nécessaire pour l'annotateur.

La Figure 1 montre l'interface principale de GEMEDOC et ses différentes composantes.

2. http://flask.pocoo.org/

3. Conclusion

À travers l'atelier EXCES, nous souhaitons effectuer une première campagne d'annotation, et ainsi profiter de la présence d'un public expert dans le traitement de la spatialité. Nous souhaitons mettre à disposition différents corpus de textes à plusieurs groupes. À l'issue de l'atelier, nous collecterons les résultats de chaque groupe, à partir desquelles nous identifierons les convergences sur la similarité entre les documents. Puis, une fois les résultats obtenus, nous évaluerons la pertinence de notre modèle d'évaluation et les possibles améliorations.

Une fois les différents corpus annotés et le modèle d'annotation fixé, nous envisageons de fusionner les résultats au sein d'un unique corpus. Une fois le corpus généré, ce dernier nous permettra d'évaluer les représentations et les mesures de similarité destinées à la mise en correspondance de documents hétérogènes.

Bibliographie

Dalton J., Dietz L., Allan J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval*, p. 365–374.

Dang H. T., Kelly D., Lin J. J. (2007). Overview of the trec 2007 question answering track. In *Trec*, vol. 7, p. 63.

Potthast M., Stein B., Barrón-Cedeño A., Rosso P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, p. 997–1005.

Voorhees E. M. (2001). The trec question answering track. *Natural Language Engineering*, vol. 7, n° 4, p. 361–378.

Voorhees E. M. *et al.* (1999). The trec-8 question answering track report. In *Trec*, vol. 99, p. 77–82.

Xu J., Croft W. B. (1996). Quary expansion using local and global document analysis. In *Acm sigir forum*, vol. 51, p. 168–175.

Zou W. Y., Socher R., Cer D., Manning C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, p. 1393–1398.