Gemedoc

Un outil pour annoter les correspondances entre les documents, **EXCES 2017**

by

Jacques Fize, Mathieu Roche, Maguelonne Teisseire

6 novembre 2017

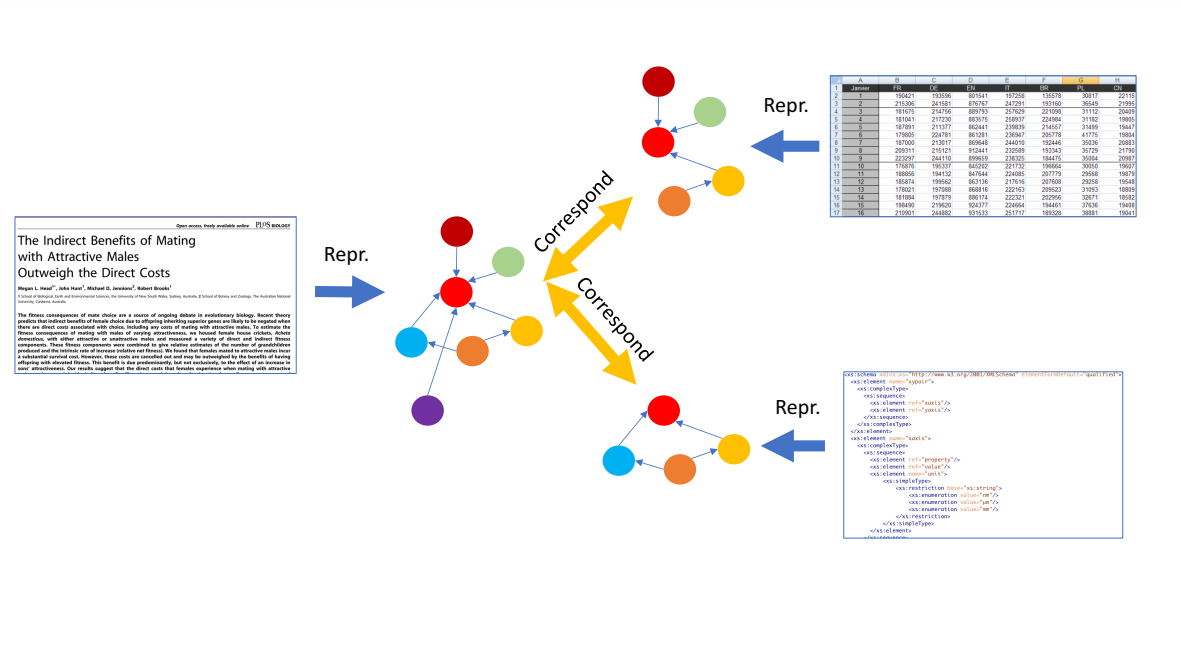
Sommaire

Introduction

Text Matching

Définition

- Méthodes et modèles dédiés à la recherche de similarité entre documents
  - Représentation de document
  - Mesure de similarité



Applications

- Généralement, **textes** comparés de **tailles différentes**
  - Moteur de recherche $\rightarrow$ requête $\stackrel{?}{=}$ document(s)
  - Surmonter cette différence : *Query Expansion* (Dalton et al., 2014)
- ... mais aussi de taille/format similaire
  - Détection de plagiat (Potthast et al., 2010)
  - Traduction utilisant l'alignement bilingue de documents (Zou et al., 2013)
  - Domaine de *Questions & Answering* (Voorhees et al., 1999 ; Dang et al., 2007)

Nos travaux

Approches multi-dimensionnelles

- Proposer une mise en correspondances sur 3 axes :
  - La thématique
  - La spatialité
  - La temporalité

Documents hétérogènes

- Structure des données textuelles : tableau, texte, etc.
- mais aussi la langue, le style d'écriture, etc.

Applications possibles

- Découvertes de connaissances, mise en relation des producteurs de données
- Cartographie de corpus
- Surveillance de phénomènes épidémiques, migratoires, etc.
- Synthèse automatique de textes

La spatialité dans les textes

Nous travaillons sur ...

- STR, une représentation de la spatialité dans les textes
- Un ensemble de mesures de similarité associées

Évaluation des propositions

Pour évaluer cette structure $\rightarrow$ **Corpus annoté** selon la **similarité spatiale** entre ces documents

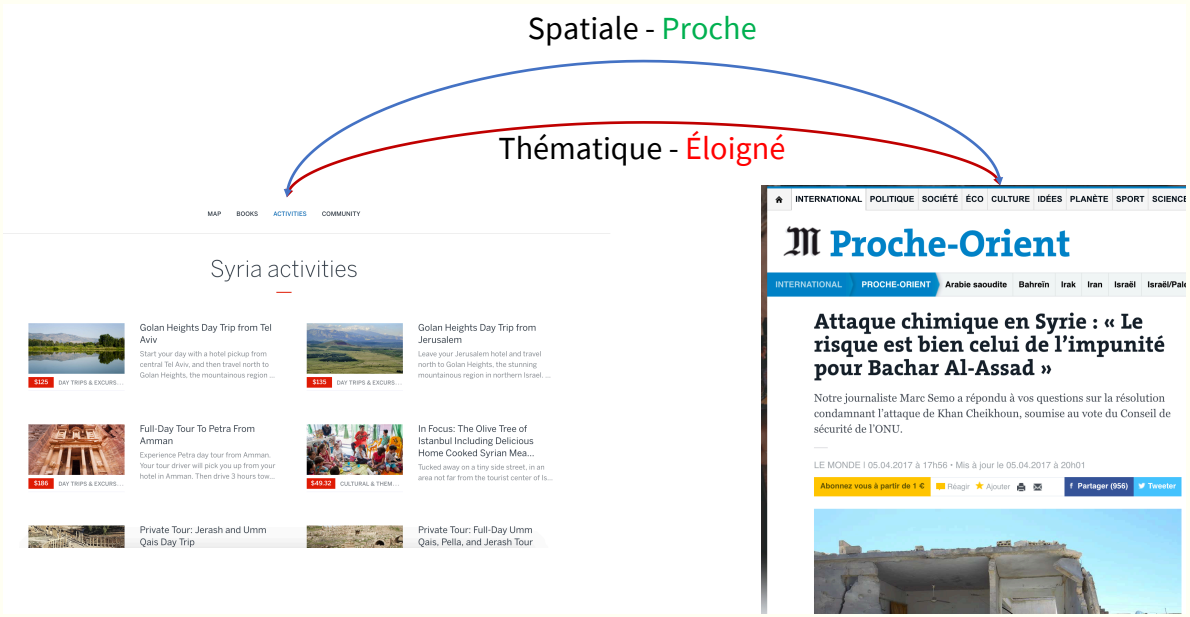
Corpus annotés disponibles

Généralement …

- Similarité annotée selon sa dimension **thématique**
- … peu d'informations **spatiales**
- Unités de textes comparés $\rightarrow$ phrases, paragraphes courts

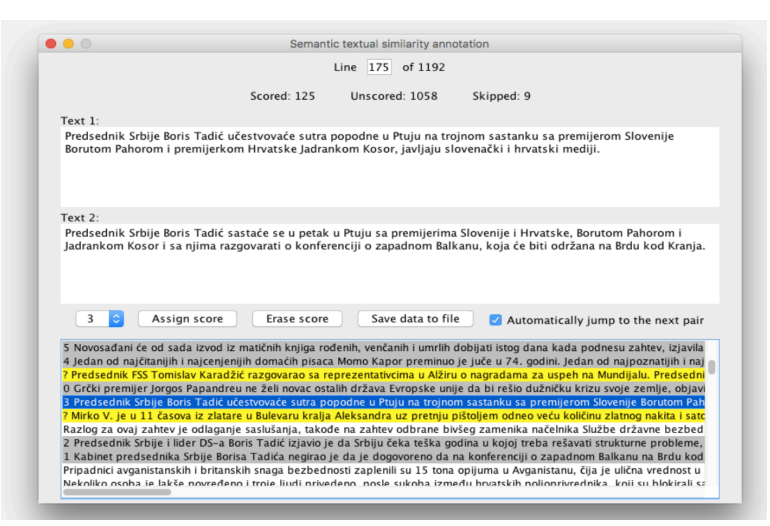
Corpus annotés selon différentes dimensions

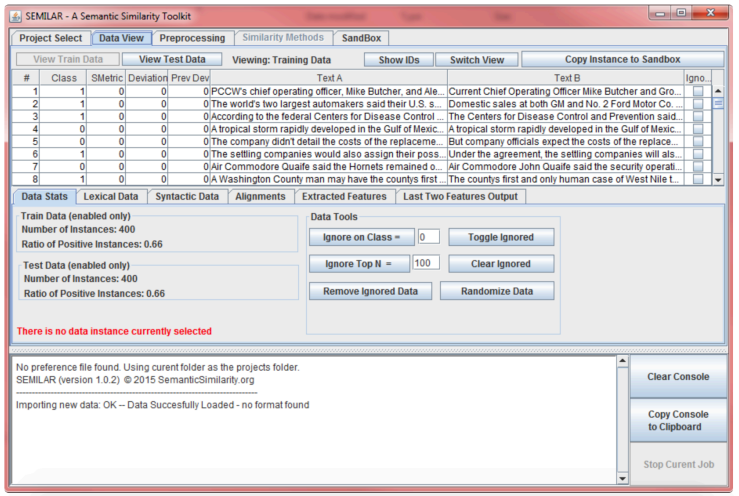
- **Corpus** dont la similarité inter-document est annotée selon :
  - la thématique ;
  - la spatialité ;
  - et la temporalité.



Outils d'annotations disponibles

- SEMILAR [1] → Outil d'annotation de **similarité sémantique** entre **mots/phrases**
- STSAnno [2] → Outil d'annotation de **similarité sémantique** entre deux unités (**phrases**) de textes



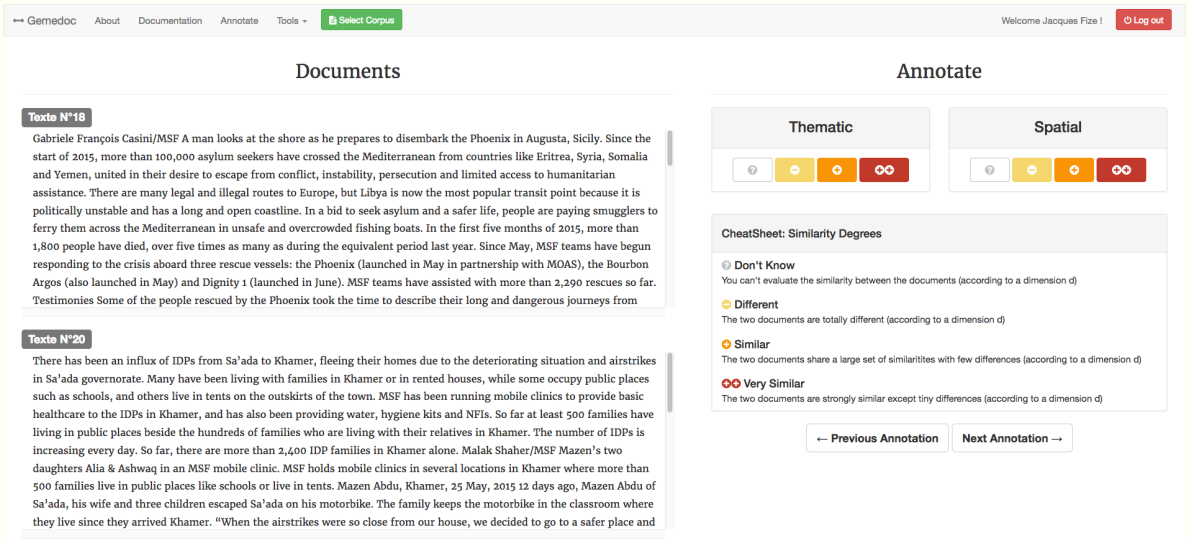


1. http://www.semanticsimilarity.org/
2. https://vukbatanovic.github.io/STSAnno/

Gemedoc

Qu'est que c'est ?

- Une plateforme web dédiée à l'annotation de similarités inter-documents
- Similarités annotées entre deux documents → **spatialité** et **thématique**



Protocole d'annotation

Deux étapes

1. Analyse du corpus
2. Annotation des similarités inter-documents

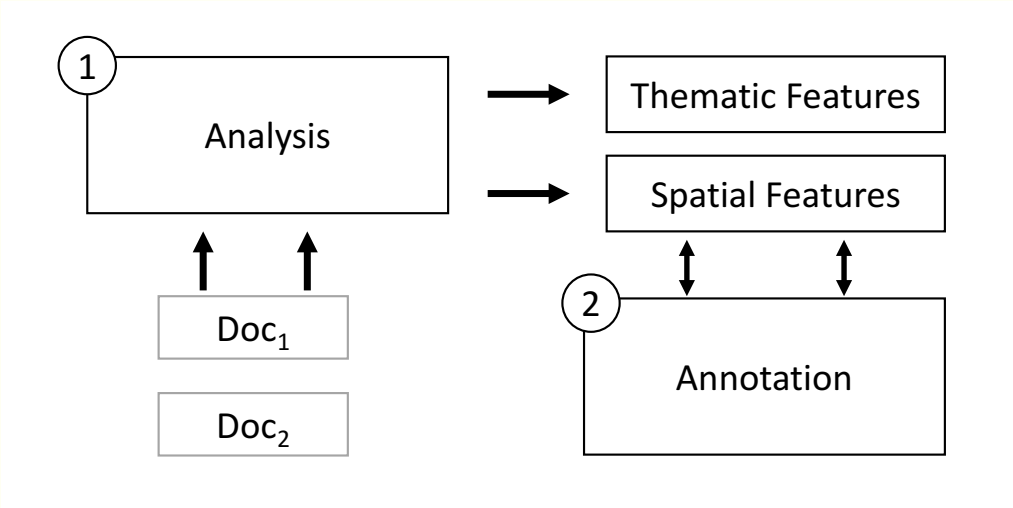
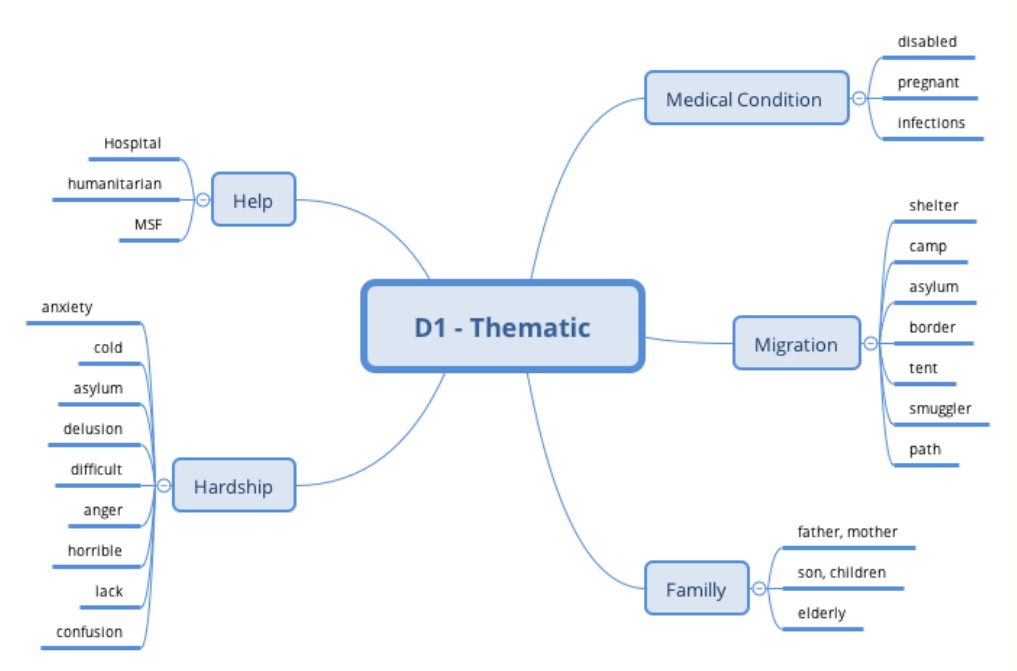


FIGURE 1 – Protocole d'annotation pour deux documents

Indicateurs thématiques

La thématique

- Se distingue par son vocabulaire



Indicateurs spatiaux

La spatialité

- Se distingue de plusieurs manières :
  - Position du narrateur
  - Liste d'entités spatiales absolues (Paris, Coutances, Madagascar, etc.) présentes
  - Parcours, itinéraires détectés

Exemple

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.

FIGURE 2 – Échantillon de texte [3] (spatial entities, thematic indicators)

3. Source : MSF

Exemple - Indicateurs extraits

Indicateurs thématiques

- L'aide humanitaire
- La misère, la souffrance
- La famille
- L'immigration

Indicateurs spatiaux

- Le narrateur se trouve à *Idemoni*
- Entités spatiales mentionnées : *Idemoni*, *Europe*, *Macedonia*, etc.
- Un parcours : *Greece* $\rightarrow$ *Macedonia*

Annotation

4 degrés de similarité

- **Ne sais pas.** L'annotateur *ne sait pas évaluer* la similarité entre les deux documents.
- **Différent.** L'annotateur indique que les documents n'ont *(quasiment) rien en commun*.
- **Similaire.** L'annotateur indique que les *documents partagent quelques similarités*.
- **Très similaire.** L'annotateur indique que les documents *sont presque identiques*.

Corpus utilisés

- Dans un premier temps $\rightarrow$ Corpus de petite taille (10 documents = 40 combinaisons à annoter)
- Chaque document, comprend des indicateurs spatiaux
- Corpus en anglais de documents extraits de :
  - News provenant du site de *Médecins sans Frontières*
  - News provenant de journaux concernant des phénomènes épidémiologiques

Appel à contribution

Objectifs

- Avoir un retour sur l'utilisation de l'outil Gemedoc
  - Lisibilité de l'interface
  - Fonctionnement
- Éprouver notre protocole d'annotation
  - Accord inter-annotateurs
  - Y a-t-il d'autres indicateurs pertinents ?
- Débuter la construction du corpus

Organisation I

Comment participer ?

1. Inscrivez-vous sur Gemedoc : `https://gemedoc.jacquesfize.com/signup`
2. Sélectionnez un corpus (MSF10 ou EPI10)
3. À l'aide de l'outil de lecture (Tools $\rightarrow$ Explore Corpus), lisez et analysez chaque document.
4. Enfin, annotez chaque combinaison de document dans la partie **Annotate**

Organisation II

Retour

À l'issue de la récolte des différentes annotations :

1. Analyse des résultats
2. Envoie d'un document à tous les participants $\rightarrow$ Résumé des observations faites sur les annotations récoltées.

Instant démo !

Conclusion

Conclusion

Nous proposons ...

- Gemedoc, un outil d'annotation de similarité inter-documents
- Un protocole d'annotation de similarité en deux étapes

Dans l'objectif …

- Évaluer nos modèles et méthodes de Text Matching
- Proposer un corpus d'alignement de texte selon différentes dimensions :
  - La spatialité
  - La thématique

Perspectives

- À l'issue de cette première campagne
  - Évaluer le fonctionnement de l'outil, du protocole
  - Lancement d'une campagne d'annotation sur des corpus plus larges

Questions ?

Merci Beaucoup ! Questions ?

Rappel

1. Inscrivez-vous sur Gemedoc : `https://gemedoc.jacquesfize.com/signup`
2. Sélectionnez un corpus (MSF10 ou EPI10)
3. À l'aide de l'outil de lecture (Tools $\rightarrow$ Explore Corpus), lisez et analysez chaque document.
4. Enfin, annotez chaque combinaison de document dans la partie **Annotate**

Références

Dalton, J., Dietz, L., and 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*, pages 365–374. ACM.

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

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

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

Zou, W. Y., Socher, R., Cer, D., and 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*, pages 1393–1398.