

Gemedoc

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

by

Jacques Fize, Mathieu Roche, Maguelonne Teisseire

Introduction

Gemedoc

Appel à contribution

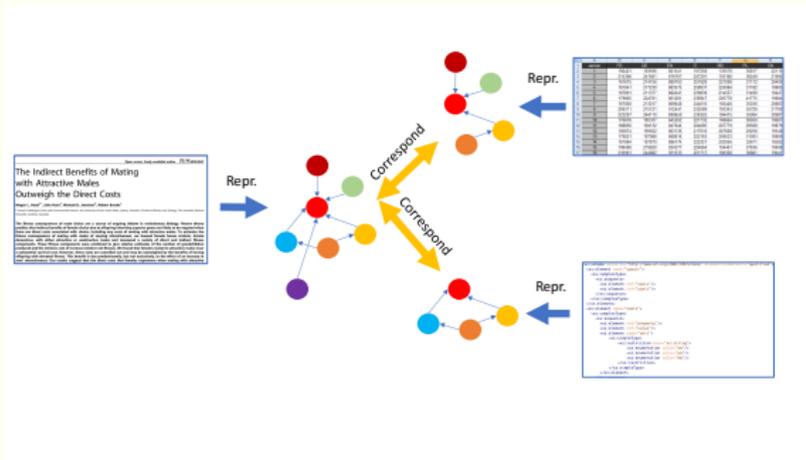
Conclusion

Introduction

Text Matching

Définition

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



- ❖ Généralement, **textes** comparés de **tailles différentes**
 - ❖ Moteur de recherche → 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)

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

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 → **Corpus annoté** selon la **similarité spatiale** entre ces documents

Généralement ...

- Similarité annotée selon sa dimension **thématique**
- ... peu d'informations **spatiales**
- Unités de textes comparés → phrases, paragraphes courts

Corpus annotés selon différentes dimensions

❖ **Corpus** dont la similarité inter-document est annotée selon :

- ❖ la thématique;
- ❖ la spatialité;
- ❖ et la temporalité.

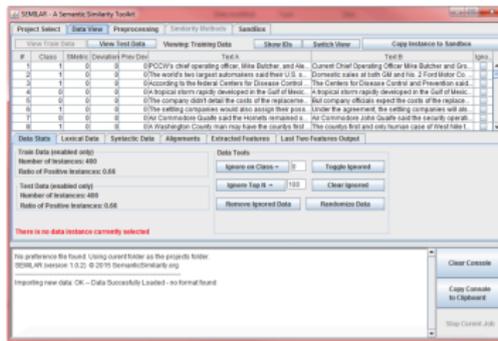
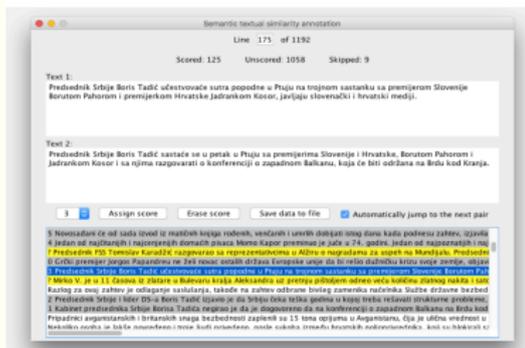
The diagram illustrates two dimensions of document similarity:

- Spatiale - Proche** (Spatial - Close): Represented by a blue arc connecting two documents that share a common geographical location (Syria).
- Thématique - Éloigné** (Thematic - Distant): Represented by a red arc connecting two documents that share a common theme (Syria) but differ in their primary focus (travel vs. news).

The left document is a travel agency page titled "Syria activities" with several tour options. The right document is a news article from "Proche-Orient" titled "Attaque chimique en Syrie : « Le risque est bien celui de l'impunité pour Bachar Al-Assad »".

Outils d'annotations disponibles

- SEMILAR¹ → Outil d'annotation de **similarité sémantique** entre **mots/phrases**
- STSANNO² → 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**

The screenshot displays the Gemedoc web application interface. At the top, there is a navigation bar with 'Gemedoc', 'About', 'Documentation', 'Annote', and 'Tools' menus, along with a 'Send Open' button and a user profile for 'Jacques Fle' with a 'Logout' link. The main content area is split into two columns: 'Documents' and 'Annotate'.

Documents Column:

- Texte N°18:** Subtitled 'Subtitled Francais Castel/MSP A man looks at the shore as he prepares to disembark the Phoenix in Augusta, Sicily. Since the start of 2015, more than 100,000 asylum seekers have crossed the Mediterranean from countries like Eritrea, Syria, Somalia and Yemen, united in their desire to escape from conflict, instability, persecution and limited access to humanitarian assistance. There are many legal and illegal routes to Europe, but Libya is now the most popular transit point because it is politically unstable and has a long and open coastline. In a bid to seek asylum and a safer life, people are paying smugglers to ferry them across the Mediterranean in unsafe and overcrowded fishing boats. In the first five months of 2015, more than 1,800 people have died, over five times as many as during the equivalent period last year. Since May, MSP teams have begun responding to the crisis aboard three rescue vessels: the Phoenix (launched in May in partnership with NGOs), the Douarbes Argos (also launched in May) and Dignity (launched in June). MSP teams have assisted with more than 3,200 rescues so far. Testimonies Some of the people rescued by the Phoenix took the time to describe their long and dangerous journeys from...'
- Texte N°20:** 'There has been an influx of IDPs from Sa'ada to Khamer, fleeing their homes due to the deteriorating situation and airstrikes in Sa'ada governorate. Many have been living with families in Khamer or in rented houses, while some occupy public places such as schools, and others live in tents on the outskirts of the town. MSP has been running mobile clinics to provide basic healthcare to the IDPs in Khamer, and has also been providing water, hygiene kits and NTFs. So far at least 500 families have living in public places beside the hundreds of families who are living with their relatives in Khamer. The number of IDPs is increasing every day. So far, there are more than 3,400 IDP families in Khamer alone. Malik Shahr/MSP Maoun's two daughters Aha & Fahwa in an MSP mobile clinic. MSP mobile clinics in several locations in Khamer where more than 500 families live in public places like schools or live in tents. Maoun Abda, Khamer, 25 May, 2015 12 days ago, Maoun Abda of Sa'ada, his wife and three children escaped Sa'ada on his motorbike. The family keeps the motorbike in the classroom where they live since they arrived Khamer. "When the airstrikes were so close from our house, we decided to go to a safer place and

Annotate Column:

- Two tabs: 'Thematic' and 'Spatial', each with a set of five colored buttons (grey, yellow, orange, red, dark red).
- Checkbox: Similarity Degree**
- Don't Know**
The two documents are totally different (according to a dimension d)
- Different**
The two documents are totally different (according to a dimension d)
- Similar**
The two documents share a large set of similarities with few differences (according to a dimension d)
- Very Similar**
The two documents are strongly similar except for differences (according to a dimension d)

Navigation buttons: 'Previous Annotation' and 'Next Annotation'.

Protocole d'annotation

Deux étapes

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

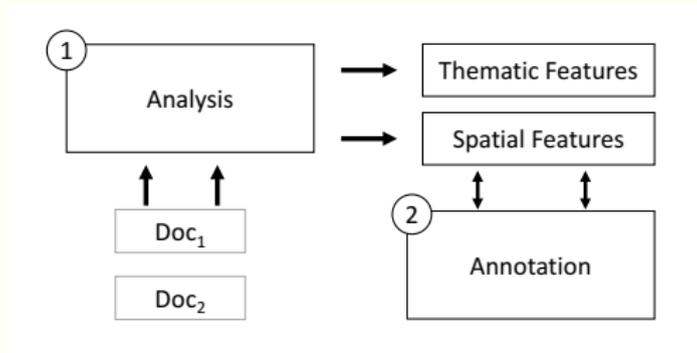
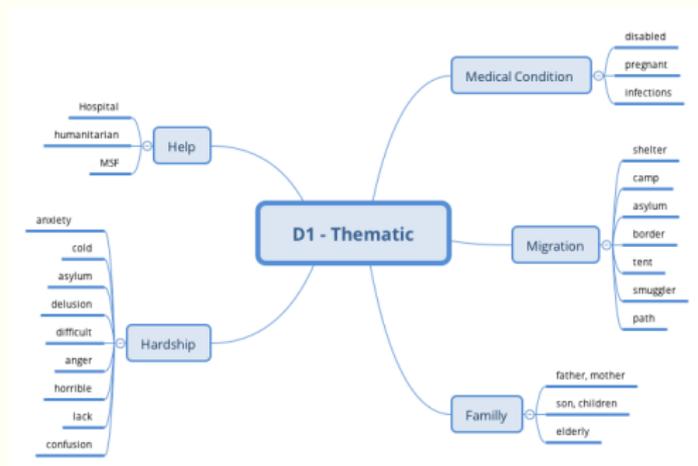


FIGURE 1 – Protocole d'annotation pour deux documents

Indicateurs thématiques

La thématique

- Se distingue par son vocabulaire



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 texte³ (spatial entities, thematic indicators)

3. Source : MSF

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* → *Macedonia*

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*.

- ❖ Dans un premier temps → 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

- ❖ 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

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 → Explore Corpus), lisez et analysez chaque document.
4. Enfin, annotez chaque combinaison de document dans la partie **Annotate**

Retour

À l'issue de la récolte des différentes annotations :

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

Instant démo !

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

- ❖ 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

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 → 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.