DAnIEL : Veille épidémiologique multilingue parcimonieuse

Gaël Lejeune, Romain Brixtel, Charlotte Lecluze, Antoine Doucet, Nadine Lucas
Normandie Université; UNICAEN, GREYC, CNRS UMR 6072, F-14032 Caen
prénom.nom@unicaen.fr

RÉSUMÉ

DAnIEL est un système multilingue de veille épidémiologique. DAnIEL permet de traiter un grand nombre de langues à faible coût grâce à une approche parcimonieuse en ressources.

ABSTRACT

DAnIEL, parsimonious yet high-coverage multilingual epidemic surveillance

DAnIEL is a multilingual epidemic surveillance system. DAnIEL relies on a parsimonious scheme making it possible to process new languages at small cost.

MOTS-CLÉS : extraction d'information, recherche d'information, veille, multilinguisme, genre journalistique, grain caractère.

KEYWORDS: information extraction, information retrieval, news surveillance, multilingualism, news genre, character-level analysis.

DAnIEL (*Data Analysis for Information Extraction in any Language*) est un système multilingue de veille épidémiologique développé au GREYC. Les systèmes de veille peinent à couvrir un grand nombre de langues du fait d'un coût élevé en ressources (Steinberger, 2011) : lemmatiseur, analyseur syntaxique ou encore ontologie du domaine. DAnIEL est au contraire conçu pour pouvoir traiter de nouvelles langues avec un **coût marginal minimal** (Lejeune *et al.*, 2012). Ainsi, il est possible de détecter un évènement dès le premier article publié, indépendamment de la langue dans laquelle celui-ci est rédigé.

DAnIEL se base sur les propriétés du genre journalistique d'une part et sur une analyse au grain caractère d'autre part. Un document décrit un évènement épidémiologique si des **chaînes de caractères** particulières sont répétées à des **positions clefs**. Ces chaînes de caractères sont choisies via un algorithme de détection de chaîne de caractères répétées maximales conjointement à un lexique minimal. Cela permet à DAnIEL d'être indépendant de toute description grammaticale locale. DAnIEL s'affranchit ainsi de l'usage de grammaires et facilite le traitement des langues à morphologie riche ; langues pour lesquelles les ressources sont rares (finnois, grec, polonais, tchèque...).

Le traitement d'une nouvelle langue par DAnIEL nécessite une quantité limitée de lexique de manière à faciliter l'extension du système, que ce soit de manière automatique (aspiré sur *Wikipedia*) ou par le biais d'un utilisateur (épidémiologiste). Ces ressources sont aisément modifiables, ce sont environ 50 mots-clés par langue.

La Figure 1 présente un exemple d'extraction d'évènement en grec. DAnIEL a été évalué sur 17 langues pour mesurer la plus-value offerte vis-à-vis du système manuel de référence *ProMED-mail* (Lejeune *et al.*, 2013). Cette expérience a montré que DAnIEL comble les lacunes de

couverture et accélère considérablement le délai de détection des événements épidémiologiques dans des régions du globe mal couvertes : Afrique et Asie du Sud-Est mais aussi Europe centrale. Le coût marginal de traitement d'une nouvelle langue par le système est de deux heures-homme (contre plusieurs mois ordinairement). Toutefois, quelques minutes suffisent pour obtenir des premiers résultats fiables. Les résultats extraits par DAnIEL sont disponibles en ligne [1].

Durant cette démonstration nous aurons l'occasion d'utiliser DAnIEL sur les cas suivants :
– traitement de langues morphologiquement riches ;
– test du système sur des documents proposés par des utilisateurs ;
– détection d'événements sur des fils de presse multilingues.

Nous souhaitons promouvoir l'utilisation de méthodes simples et reproductibles, adaptées au traitement de données multilingues. La combinaison d'un modèle de document, dépendant du genre de texte et non de la langue, et d'une analyse au grain caractère, permet d'envisager d'autres applications.

FIGURE 1 – Extraction de l'évènement grippe , Chine (γρίπη , Κίνα) dans un article en grec

Références

LEJEUNE, G., BRIXTEL, R., DOUCET, A. et LUCAS, N. (2012). DAnIEL : Language Independent Character-Based News Surveillance. *In Advances in Natural Language Processing, Springer LNAI 7614*, pages 64–75.

LEJEUNE, G., BRIXTEL, R., LECLUZE, C., DOUCET, A. et LUCAS, N. (2013). Added-value of automatic multilingual text analysis for epidemic surveillance. *14th Conf. Artificial Intelligence in Medicine AIME, Murcia, May*.

STEINBERGER, R. (2011). A survey of methods to ease the development of highly multilingual text mining applications. *Language Resources and Evaluation*, pages 1–22.

1. https://daniel.greyc.fr