Tentative d'approche multilingue en extraction d'information

Gaël Lejeune, Nadine Lucas, Antoine Doucet

▶ **To cite this version:**

Gaël Lejeune, Nadine Lucas, Antoine Doucet. Tentative d'approche multilingue en extraction d'information. JADT Journées internationales d'Analyse statistique des Données Textuelles, Jun 2010, rome, Italie. pp.1259-1267. hal-01067147

HAL Id: hal-01067147
https://hal.science/hal-01067147

Submitted on 23 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tentative d'approche multilingue en extraction d'information

Gaël Lejeune, Nadine Lucas, Antoine Doucet

GREYC, Université de Caen Bd du Maréchal Juin, 14000 Caen

Résumé

Nous présentons ici un système d'extraction d'information basé uniquement sur des critères positionnels et stylistiques. L'idée est de concevoir un modèle faisant appel à des ressources aussi limitées que possible de façon à limiter le travail nécessaire pour traiter plusieurs langues. Spécialisé dans le domaine de la veille épidémiologique sur la presse, le système que nous allons décrire s'appuie sur les lois du genre journalistique et les théories de la communication pour extraire des évènements épidémiologiques. Testée en premier lieu sur le français, cette approche s'est avérée efficace et très adaptée aux problématiques multilingues.

Abstract

Increasing amounts of information available on the web offer great possibilities for a domain like epidemic surveillance. The main issue is to give epidemiologists and health authorities the information they need. Information Extraction is intended to fulfill these needs but most of the systems cannot monitor more than one or two languages. Detecting epidemic events in some areas of the world might be therefore difficult or at least too slow for such strategic domains. The approach we will present here tries to deal with multilingual problems by using discourse rules and very few resources. The idea is to be able to monitor new languages very quickly. We will present here some good results which show we can do efficiently IE tasks with few resources.

Keywords: Information extraction, Natural Language processing, Epidemiological surveillance, multilingualism

1. Introduction

La question du traitement multilingue est au cœur de beaucoup de travaux en Extraction d'Information, dans un domaine tel que la veille épidémiologique il nous semble encore plus important de pouvoir analyser un grand nombre de langues efficacement.

1.1. La problématique multilingue

L'appropriation de certaines branches de la linguistique par les informaticiens a conduit ces derniers à accorder leur confiance à un certain type de modèles dans les différents domaines du Traitement Automatique des Langues Naturelles (TALN). Les évolutions récentes ont mis en lumière l'importance cruciale des ontologies et des motifs (patterns) comme mode de transmission à la machine de réalités langagières. La structure logique inhérente à ces représentations permet de modéliser de façon parfois très fine les modes d'expressions rencontrés dans les corpus.

Seulement ces modèles, qui ont pour objectif de décrire l'ensemble des actes langagiers, souffrent de deux problèmes. D'une part d'un problème interne qui se perçoit dans la confrontation entre des données figées (dictionnaires, ontologies...) et une langue vivante une course sans fin. Et

d'autre part d'un problème externe qui se manifeste quand on doit sortir d'une réalité unilingue pour tenter de traiter de nouvelles langues (Hull and Grefenstette, 1996).

Au premier problème, les systèmes d'apprentissage automatique apportent des réponses qui semblent de plus en plus probantes. L'extraction de motifs nouveaux à partir de motifs déjà existants et efficaces est ainsi très étudiée. A l'opposé, l'aspect multilingue est souvent simplement vu comme un processus pas à pas visant à traiter une langue A puis une langue B puis une langue C sans avoir possibilité de capitaliser massivement sur le travail déjà effectué. Cette nécessité de repartir de zéro, ou presque, pose le problème des délais de traitement des langues selon leur importance mais aussi plus généralement de la difficulté d'adapter des modèles conçus pour l'anglais à des réalités linguistiques bien différentes.

Le caractère pluridisciplinaire du domaine invite nécessairement à s'intéresser à d'autres modèles existant dans la littérature linguistique et à les tester de manière à définir lesquels offrent un meilleur compromis entre l'efficacité brute et l'adaptabilité à la machine. Nous tenterons de voir comment l'étude d'un grain différent, le texte plutôt que le mot ou la phrase, peut apporter des résultats intéressants et comment un modèle basé sur cette granularité peut se révéler efficace dans une perspective résolument multilingue.

1.2. La veille épidémiologique

La veille épidémiologique consiste à rendre compte de la manière la plus efficace possible de l'irruption d'épidémies ou de cas isolés de maladies dans une perspective géographique globale. La partie informatique du domaine prend en charge l'analyse d'un certain nombre de sources pour en extraire des évènements concernant des maladies. Ces sources sont majoritairement des articles de presse généraliste et des articles spécialisés.

En conséquence si l'on considère que la veille épidémiologique a pour but premier de fournir des documents dûment sélectionnés et indexés, on ne peut pas se contenter de simples agrégateurs. De même si l'on cherche à mettre en avant la rapidité de la transmission de l'information à l'utilisateur final, ici l'épidémiologiste ou les autorités sanitaires, on ne peut se contenter de systèmes basés sur un filtre humain (Yangarber et al., 2008).

Pour l'aspect multilingue, on fera donc plus facilement référence à des systèmes qui sont véritablement automatiques comme par exemple Biocaster Health Monitor, ou Health Map (Freifeld et al., 2008). Globalement, ils se basent sur une série d'analyses « locales » qui prend le plus souvent le schéma suivant:

1. **Lemmatisation** : le document est déconstruit de façon à faire apparaître ses unités fonctionnelles autonomes, les lemmes.
2. **Analyse morphologique** : Les formes des lemmes extraits sont examinées. Aux éléments de l'énoncé lemmatisé on fait correspondre la forme canonique.
3. **Analyse syntaxique** : A l'aide des données morphologiques, on étudie les combinaisons formées par les unités significatives et leurs possibles relations.
4. **Interprétation/ analyse sémantique :** Ici on essaie de donner du sens aux informations préalablement extraites notamment en les comparant à des motifs préalablement validés.
5. **Analyse du discours, étude des coréférences :** De façon à ne pas prendre simplement les phrases isolément, quelques améliorations ont été apportées aux systèmes d'extraction d'information de façon à prendre en compte des éléments extra-phrastiques.

Chacune de ces étapes nécessite un certain nombre de ressources: étiqueteurs grammaticaux, base de motifs prédéfinis, dictionnaires, ontologies... Or, celles-ci ont un coût tant en terme financiers qu'en terme de temps. Et s'il y a aujourd'hui pour des langues telles que l'anglais

ou le français de nombreuses ressources utilisables, qu'en est-il de langues plus rares pour lesquelles ces ressources ne sont pas constituées ou constituables ?

Une tentation peut être d'utiliser la traduction automatique (Linge et al., 2009) pour réutiliser par exemple des motifs existants, mais cela pose le problème de s'appuyer sur un outil dont on sait d'avance qu'il génère lui même bon nombre d'erreurs. De plus l'efficacité de la traduction automatique est clairement dépendante du couple de langues. On pourra alors avoir de très bons résultats sur le français mais des résultats sans commune mesure sur le tchèque par exemple. L'intérêt et la fiabilité des données extraites pour l'épidémiologiste en souffriront immanquablement.

Dans cette optique nous proposons une approche différente basée sur l'utilisation d'un modèle communicationnel plutôt que d'un modèle syntaxico-sémantique afin de limiter au maximum la quantité de ressources nécessaires au traitement de nouvelles langues. Nous montrerons donc un système dont les différents éléments sont aussi « réutilisables » que possible. Notre objectif n'est pas tant d'avoir un niveau de détail très élevé que d'avoir une confiance importante dans les résultats et que ceux-ci puissent être comparables d'une langue à l'autre, chose capitale pour ce genre de données sensibles.

2. Notre proposition: une approche basée sur le grain texte

2.1. Fondements théoriques

Les branches de la linguistique exploitées à l'heure actuelle dans le cadre du Traitement Automatique des Langues Naturelles (TALN) se concentrent massivement sur le grain mot, ou sur le grain « Unité lexicale ». Nous cherchons ici au contraire à voir dans quelle mesure le grain « texte » peut être pertinent. En effet, dans la mesure où la langue n'est généralement plus vue aujourd'hui comme un simple sac de mots, nous proposons de ne plus analyser le texte comme un sac de phrases. Plutôt que d'utiliser la morphologie, la syntaxe et la sémantique nous choisissons de nous concentrer sur la pragmatique et la stylistique.

Ainsi, toutes les données structurelles que l'on peut extraire du document peuvent être utilisées: titre, chapeau, paragraphes... De cette façon nous pouvons nous appuyer sur un certains nombre de théories linguistiques de plus haut niveau (Lucas, 2004). On s'intéresse alors aux stratégies que l'auteur met en œuvre pour transmettre son message. En effet, on va traiter des textes qui sont faits pour être lus par des êtres humains. Derrière cette évidence se situe l'idée que l'on doit s'intéresser à la façon dont l'utilisateur humain décode le message.

Tout d'abord, quand il s'intéresse à un article de presse, le locuteur humain n'éprouve pas le besoin de décoder tout le texte pour savoir de quoi il traite. En effet, les lois du genre impliquent qu'après avoir lu le titre et les premières phrases, celui-ci sait si le document est digne d'intérêt pour lui. C'est ce que l'on nomme dans le genre journalistique la règle des « 5W ». Il faut que les réponses aux 5 questions « Why What When Where Why » soient fournies très tôt dans le document (Kando, 1996). Usuellement cela correspond à la partie d'un article de presse que l'on nomme chapeau ou au premier paragraphe du corps d'un article court. Cette propriété nous servira à éliminer les documents non pertinents.

Ensuite le principe d'efficacité (Sperber and Wilson, 1998) invite l'émetteur (ici le journaliste) à s'assurer que le récepteur (ici le lecteur) ait à faire « le minimum d'inférences logiques » pour traiter le message. C'est ce que l'on nomme aussi le principe du moindre coût cognitif ou du moindre effort (Reboul and Moeschler, 1998). Cela implique l'usage de la redondance et de

l'explicitation. On ne traitera pas deux sujets importants dans le même article, donc si l'on a deux sujets traités alors l'un est à coup sûr de moindre importance. De la même manière on a recours à des tournures explicites pour transmettre les éléments principaux de l'information. Contrairement au genre littéraire ou aux articles d'analyse, l'article d'information use donc peu ou pas de vocabulaire spécialisé et de tournures compliquées.

Enfin, le document est analysé dans sa dimension globale. En s'intéressant aux chaînes de caractères qui le composent nous essayons de nous adapter à des langues où le concept de mot et les modèles qui en découlent posent problème. Que ce soient des langues casuelles comme le russe, des langues agglutinantes comme le finnois ou des langues à idéogrammes comme le chinois.

2.2. Fonctionnement du système

Nous présentons ici le déroulement de l'algorithme en l'illustrant avec des exemples extraits des différentes langues sur lesquelles nous travaillons. Suivant les stratégies que met en œuvre le locuteur humain, le système sélectionne en premier lieu les articles potentiellement intéressants avant de les analyser en détail.

2.2.1. Première phase: écarter les documents non pertinents

Les ressources utilisées par le système présenté ici sont limitées: une liste de noms de maladies comprenant 200 termes, une liste de noms de pays et de leurs capitales ainsi qu'une dizaine de termes permettant de caractériser les articles de moindre actualité (campagne de vaccination ou de sensibilisation, programme de lutte...) que nous regroupons dans une « liste noire ».

Dans un premier temps pour déterminer si le document peut être intéressant, le système identifie les chaînes de caractères répétées dans le début (titre et premier paragraphe) et dans le corps du texte. Ces chaines forment ce que nous nommerons le « contenu pertinent » de l'article. S'il s'avère que dans ce contenu pertinent on détecte un terme correspondant à une maladie, le document est considéré comme potentiellement utile.

Si le contenu pertinent comprend aussi un des termes de la liste noire alors le document est considéré d'intérêt secondaire. Cela pourra être par exemple un article concernant une campagne de vaccination contre la grippe, il peut être proposé à l'utilisateur mais n'est pas analysé en profondeur.

Ce phénomène de répétition d'informations importantes à des positions données peut prendre différentes formes. Nous l'avons testé sur un corpus de 1200 documents en français dont 210 avaient été identifiés manuellement comme pertinents. Seul un de ces 210 documents avait été écarté d'emblée par le système, ce qui semble garantir la fiabilité de la détection des articles potentiellement intéressants.

Grippe A : un nouveau décès en France

En France, le décès d'une femme à Creil vient alourdir, à 38, le bilan des morts de malades porteurs du virus **H1N1**. Au niveau mondial les vaccins vont manquer selon l'organisation mondiale de la santé (OMS).

Le bilan continue de s'alourdir en France et des inquiétudes se font sentir au niveau mondial sur l'approvisionnement de vaccins, selon l'OMS (Organisation mondiale de la santé). En France, une femme de 38 ans est décédée à Creil (Oise) samedi dernier et porte à 38 le nombre de décès dû à la grippe **H1N1**.

[...]

D'autres responsables sanitaires américains ont déploré le retard de la production vaccinale, expliqué notamment par le développement lent du virus **H1N1** dans les œufs où le virus est incubé pour la fabrication du vaccin. Ils la comparent à la production plus rapide du vaccin contre la grippe saisonnière effectué sur des cellules souches.

Quant à l'OMS, elle s'est inquiétée de la protection des pays pauvres car elle estime que des « milliards de doses » vont manquer pour protéger toute la population.

En Europe, selon la commissaire européenne à la Santé, Androulla Vassiliou, jusqu'à près de 30% de la population pourrait être touchée dans les prochains mois. Elle appelle les Européens à se vacciner afin de limiter le développement de la pandémie.

Exemple 1 : (français): document pertinent identifié

One was an arrogant bully, the other a nervous wreck... so what is the truth about Van Gogh's hear?

Gauguin, a journalist's son who had formerly made his living as a banker, had abandoned his wife and five children to paint. He knew Van Gogh from painting sessions in Brittany, and finally accepted his invitation to the South after Van Gogh bombarded him with illustrations of the seductive life there.

He sketched the trees, the fields, the house, even the furniture to persuade his friend to join him. He even painted the famous Sunflowers to decorate Gauguin's room. Gauguin arrived in October and the two started to work together, producing several now famous canvasses. At night they repaired to the local bars and brothels. [...] Most telling of all, they say, is Van Gogh's description in a letter to Theo about his friend's aggressive character. He writes it was lucky Gauguin didn't have a machine gun or other firearms. Gauguin took himself off to Tahiti where he entertained under-age mistresses, consumed vast quantities of absinthe and morphine and died of **syphilis** in 1903. [...]

Exemple 2 : (anglais): document non-pertinent écarté

Comme on peut le voir (exemples 1 et 2), le « contenu pertinent » permet à la fois d'identifier aisément des documents potentiellement intéressants et d'écarter ceux dans lesquels une maladie n'est évoquée que comme fait secondaire. La chaîne de caractère répétée « **H1N1** » dans le premier document caractérise de façon très sûre un document potentiellement pertinent. Au contraire dans le second document, la chaîne « syphilis » ne figure pas dans le contenu pertinent extrait par le système, ce qui cadre avec le fait que cette maladie n'est qu'un événement secondaire dans l'article en question.

2.2.2. Deuxième phase: extraire les événements

A ce stade la date extraite (*When*) est celle du document, des essais sont en cours pour affiner l'extraction de cette donnée. L'extraction des causes (*Why*) ne fait a priori pas partie des besoins prioritaires dans le domaine que nous étudions.

Notre tâche première est alors d'identifier la paire suivante:

Maladie+Lieu

Qui correspond dans la règle des 5W à :

What+Where

La maladie extraite par le système correspond aux termes découverts dans le contenu pertinent. Si plusieurs maladies y figurent alors le document est considéré de pertinence moindre puisqu'il traite de plusieurs maladies différentes.

Pour extraire le lieu nous appliquons le même algorithme à ceci près que si plusieurs lieux sont identifiés nous tentons de désambigüiser de la manière suivante.

Le lieu pertinent est celui qui a une fréquence dans le document global deux fois supérieure aux autres lieux « candidats ». S'il y a toujours plus d'un lieu candidat alors le document est considéré de moindre importance.

Au contraire, si aucun lieu ne figure dans le contenu pertinent, on considère que le pays concerné est celui de la source. En effet, quand dans un journal français on parle de la France, il n'est pas nécessaire de le préciser pour que le lecteur le comprenne immédiatement, inconsciemment celui-ci situera l'évènement.

El cólera se extiende en Zimbabue mientras prosigue la lucha de poder

Más de 6.000 infectados y casi 300 muertos en un brote comenzado el martes y agravado por los hospitales cerrados, fruto de la pésima situación económica del país.

Son ya más de 6.000 infectados y 294 las muertes confirmadas por el último brote de cólera que se declaró el martes pasado en el sur de Zimbabwe. La epidemia se ve agravada por la pésima situación económica del país, cuya moneda acumula ya una inflación del 100.000%. Los servicios públicos están bajo mínimos y los hospitales cerrados por falta de pago. [...]

Por otra parte, **Sudáfrica** ha anunciado que retendrá 23 millones de euros que tenía previsto enviar a modo de ayuda humanitaria hasta que no haya un Gobierno claramente definido en la que parece la primera medida punitiva adoptada contra Mugaba por el sabotaje de las negociaciones.

Exemple 3 : (espagnol): identification du lieu pertinent

Dans notre exemple en espagnol (exemple 3) on voit que l'utilisation du concept de contenu pertinent permet d'identifier le pays concerné par l'évènement principal comme étant « Zimbabue » (Zimbabwe), qui est répété, et non « Sudáfrica » (Afrique du Sud).

Pour ce qui est des cas (*Who* dans la règle des 5W), le système extrait la première chaîne correspondant à une donnée numérique et satisfaisant aux conditions suivantes:

- être située dans le premier tiers du texte ou les 2 premiers paragraphes.
- ne pas caractériser une date, une distance ou une somme d'argent.

Le cas extrait est affecté d'un degré de confiance plus grand si la chaîne de caractère où il figure est répétée, en tenant compte d'une éventuelle troncature. Si par contre aucune information n'est identifiée de cette façon, le système extrait la première phrase contenant la chaîne de caractères correspondant à la maladie identifiée. En effet on suppose que le nombre de cas est alors écrit en toutes lettres.

Red eye plagues Quang Ngai

Quang Ngai Province General Hospital on the central coast has treated more than **1.200 people** with the form of conjunctivitis known as red eye in the past week.

Most of them have been children under six years old, a hospital spokesman said. Private clinics are also overloaded, with each treating between 100 and 120 cases of red eye since one week ago.

In the provincial capital, also called Quang Ngai, many kindergarten classes are two thirds empty as infected children are forced to stay home until their eyes heal.

Quang Ngai was one of the provinces worst affected by Typhoon Ketsana, which killed 163 people last month, and a succession of tropical lows since then has caused more flooding.[...]

Unlike in Quang Ngai and Can Tho, red eye seems to be tapering off in Ho Chi Minh City after running amuck for two weeks, HCMC Ophthalmology Hospital reported on Sunday without giving related statistics.

At one stage, the city's major eye hospital was treating nearly 500 inpatients and outpatients for the viral disease every day, almost double the normal number.

Exemple 4 : (anglais): identification des cas

Over 200 swine flu cases detected in Pakistan: minister

ISLAMABAD, Jan. 12 (Xinhua) -- Pakistani Health Ministry informed the parliament on Tuesday that **219** cases of swine flu have so far been detected in the country. Health Minister Makhdoom Shahabuddin told the National Assembly that 14 casualties have been reported due to the global disease.

He said the federal government has decided to install scanners at all international airports and ground entry points to control swine flu. He said the government has also provided necessary medicines to the provincial governments and Pakistan Army for treatment of swine flu. The deaths from H1N1 influenza since May 2009 have led to rising public fear and concern, and experts are calling on the government to introduce more preventative measures. [...]

Shahabuddin said that 828 patients have been tested and **219** of them have been tested positive. He said that the World Health Organization had promised to provide vaccine by Jan. 10 but it has now sought a week more time to supply vaccine. Officials say about two million vaccines would be available this month, with vulnerable people, including pregnant women and health workers, to be inoculated first.

Exemple 5 : (anglais): utilisation de la troncature pour valider les cas détectés

On voit ici (exemple 4) que l'évènement principal décrit par l'auteur de cet article concerne « 1.200 people » (1.200 personnes). Les autres données ont trait à des faits plus anciens, ou annexes comme le nombre de cas traités en clinique: « between 100 and 120 cases ». Le second exemple (exemple 5) illustre l'enrichissement de la détection de la répétition par le rapprochement de données numériques tronquées.

3. Résultats

Les premiers tests pour valider le modèle ont été effectués sur le français. Puis nous avons ajouté une base de données de lieux et de maladies en espagnol et en anglais en utilisant le même principe de ressources limitées. Les tableaux suivants présentent la précision, nombre de documents pertinents par rapport au nombre extrait, et le rappel, nombre de documents pertinents extraits par rapport aux documents pertinents attendus.

Étiquetage manuel	*Extraits*	*Ignorés*	F-mesure: 90%
Documents pertinents	196	14	Rappel: 93%
Documents non-pertinents	28	962	Précision: 87,5%

Tableau 1: Résultats en français

Les résultats globaux en Français sont très intéressants. Les documents ignorés sont pour leur plus grande partie ceux auxquels le système a donné un moindre indice de confiance (plusieurs maladies traitées, plusieurs pays concernés) alors que parfois l'évaluateur humain peut identifier un événement pertinent. Au niveau des éléments identifiés, ce sont les maladies qui sont le mieux repérées (93% d'identification correcte) alors que les lieux extraits sont corrects à 82% et les cas corrects à 83% par rapport à ce que l'étiquetage manuel avait identifié.

Étiquetage manuel	*Extraits*	*Ignorés*	F-mesure: 85%
Documents pertinents	61	6	Rappel: 91%
Documents non-pertinents	15	25	Précision: 80%

Tableau 2: Résultats en espagnol

Étiquetage manuel	*Extraits*	*Ignorés*	F-mesure: 84%
Documents pertinents	44	6	Rappel: 88%
Documents non-pertinents	11	39	Précision: 80%

Tableau 3: Résultats en anglais

Les résultats en espagnol et en anglais (tableaux 2 et 3) sont bien sûr à pondérer étant donné que les corpus utilisés sont de taille plus limité que pour le français. Néanmoins on voit que ces résultats restent du même ordre dans les trois langues. Cela semble accréditer la thèse que les règles utilisées se prêtent bien à un traitement multilingue. Les maladies extraites sont correctes dans plus de 90% des cas dans les deux langues. De la même façon qu'en français, la précision des lieux et des cas détectés est moindre, entre 76 et 85%. Malgré tout nous cela offre des perspectives intéressantes pour pouvoir comparer des évènements issus de documents en différentes langues.

В Москве от свиного гриппа умерла 53-летняя женщина

В столице зарегистрирован первый летальный исход от гриппа A /**H1N1**

53-летняя женщина скончалась сегодня в 9.00 в инфекционной больнице №1, сообщает со ссылкой на информацию Департамента здравоохранения Москвы.

Ранее сообщалось о том, что в Забайкальском крае зарегистрировано три летальных исхода от гриппа A/H1N1. В частности, в Чите скончались две женщины 29-ти и 50-ти лет.

На Сахалине зарегистрировано 100 подтвержденных случаев свиного гриппа. В Красноярском крае число госпитализированных людей с диагнозом грипп A/**H1N1** за сутки выросло с 236 до 285 человек.

В Москве число случаев свиного гриппа за последнюю неделю выросло до 444. По последним данным, число заболевших свиным гриппом в России составляет 1349 человек.

Exemple 5 : (russe): Maladie, lieu, cas

Le russe (Exemple 5) est une des langues sur lesquelles nous travaillons actuellement, dans cet exemple l'identification dans le contenu pertinent des chaînes de caractères « H1N1 », « Москве » et « 53-летняя женщина » permet au système d'identifier qu'il s'agit d'un cas de Grippe H1N1 à Moscou concernant une femme de 53 ans.

Dans d'autres documents sur lesquels nous avons testé nos algorithmes, nous avons aussi pu contourner les problèmes posés par des variations morphologiques typiques des langues à déclinaisons. Voici un exemple illustrant ce principe sur le Finnois (Exemple 6) qui présente en plus la particularité d'être une langue plus rare où peu de ressources sont disponibles. Le système identifie dans cet exemple un événement concernant le choléra (« Kolera ») au Zimbabwe (« Zimbabwe ») et comprenant a priori 60.000 cas.

Zimbabwessa koleraan sairastuneita jo yli 60.000

Kolerasta kärsivää lasta hoidetaan sairaalassa Hararessa.

Geneve. Jo yli 60.000 ihmistä on sairastunut koleraan **Zimbabwe**ssa, Maailman terveysjärjestö WHO kertoi perjantaina. Epidemian alkuvaiheessa WHO arvioi, että kolera voi pahimmassa tapauksessa levitä jopa 60.000 ihmiseen. Nyt raja on ylitetty, WHO:sta kerrottiin.

Tautiin on elokuusta lähtien kuollut yli 3.160 ihmistä. "**Zimbabwe**n humanitaarinen tilanne on akuutti ja pahenee edelleen", sanoi YK:n humanitaarisen avun järjestön OCHA:n edustaja Elisabeth Byrs.

"Apua tarvitaan enemmän kuin koskaan. Elämme kriittisen tärkeitä hetkiä".

YK on vedonnut jäsenmaihinsa, jotta ne antaisivat liki 570 miljoonaa dollaria hätäapua **Zimbabwe**en. Yhtään lahjoituksia ei ole kuitenkaan tullut, Byrs sanoi. **Kolera**epidemian pelätään pahenevan keväällä alkavalla sadekaudella. **Kolera** leviää likaisen veden ja ruuan välityksellä.

*Exemple 6 : (finnois): **Maladie**, lieu, **cas***

4. Conclusion

Les tests sur les trois premières langues de notre étude que sont le français, l'espagnol et l'anglais montrent une réelle convergence dans la qualité des résultats. Ceux-ci sont proches des attentes dans le domaine malgré un coût de construction des ressources nettement moins élevé. Les éléments principaux de la méthode décrite ici sont d'ores et déjà utilisés, en collaboration avec le projet PULS de l'université d'Helsinki, sur la plateforme Medisys de l'Union européenne [1]. Ainsi des évènements extraits à partir de sources en langue anglaise est combinée avec ceux extraits de documents en français, le tout en temps réel.

Les travaux que nous menons actuellement sur d'autres langues, le russe et le finnois mais aussi le chinois et le turc, semblent démontrer que les grandes lignes de notre modèle peuvent être conservées dans des langues aussi différentes. Les perspectives sont importantes car ce sont précisément des langues où les modèles basés sur le grain mot éprouvent le plus de difficulté.

Références

Freifeld C., Mandl K., Reis B. and Brownstein J. (2008). HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Med Inform Assoc*, vol. 15 : 150-157.

Hull D. and Grefenstette G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49-57.

Kando N. (1996). Text structure analysis based on human recognition: cases of newspaper articles and English newspaper articles. *Bulletin of the National Center for Science Information Systems*, 8 : 107-129.

Lejeune G., Hatmi M., Doucet A., Lucas N. and Huttunen S. (2009). A proposal for a multilingual system for epidemic surveillance. In *Proceedings of the 1st International ICST Conference on UCMedia, workshop on Mining User-Generated Content for Security Workshop*, Venezia, Italy, December 9-11.

Linge J.P., Steinberger R., Weber T.P., Yangarber R., Van der Goot E., Al khudhairy D.H. and Stilianakis N. (2009). Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14, Issue 13.

Lucas N. (2004). The enunciative structure of news dispatches: A contrastive rhetorical approach. In *ASLA*, Stockholm, pp. 154-164.

Reboul A. and Moeschler J. (1998). *La pragmatique aujourd'hui. Une nouvelle science de la communication.* Paris: Le Seuil.

Sperber D. et Wilson D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Yangarber R., Steinberger R., Fuart F., Van der Goot E., Best C. and Von Etter P. (2008). Text mining from the Web for medical intelligence. *Mining massive data sets for security*. Amsterdam: OIS Press.

[1] http://medusa.jrc.it/medisys/helsinkiedition/fr/home.html.