A proposal for a multilingual epidemic surveillance system

Gael Lejeune†, Mohamed Hatmi†, Antoine Doucet†, Nadine Lucas†, Roman Yangarber‡

†GREYC, University of Caen firstname.lastname@info.unicaen.fr
‡Doremi, University of Helsinki firstname.lastname@cs.helsinki.fi

Summary. In epidemic surveillance, monitoring numerous languages is a great issue. in this paper we will present a system designed to work on French, Spanish and English. The originality of our system is that we use only a few resources to do our information extraction tasks. Instead of using ontologies, we use structure patterns of newspapers articles. The results on these three languages are very good at this stage and we will present a few examples of interesting experiments in other languages.

1 Introduction

Recent epidemic events clearly showed that authorities have a great need of quick and precise information about disease spreading. It is also known that on-line news in this domain are spreading as well. Therefore selecting really relevant documents for specialists and health authorities is more and more important each day. Furthermore, as we are in a global world, it means that being able to monitor numerous languages with confidence should be a great improvement. As it is difficult to improve, even slightly, performance on monolingual systems, much effort is devoted to multilingual versions [2]. But the problem faced is that building an entire system for each new language is really time and money consuming. For instance we can guess that for major languages like English, Chinese, Spanish, Russian and Arabic one can have a valuable system. But they "only" represent 40% of world's population. For some parts of the world it will therefore take time for an epidemic to be detected by such systems.

Machine translation may be a way to fulfill these needs [10] although it has also its own limits. This paper will show an extension of an Information Extraction system for French to a Spanish and an English version, all three based on the same structural patterns. Its results will be compared to the state of the art to see how these results can be a road to an extension to other, and even rare, languages.

2 Related work

In epidemic surveillance, recent works use mostly two approaches: automatic processing with or without human post analysis. For instance in Health Map [1],

after potentially relevant documents have been selected automatically, experts have to check the relevance of the extracted events. On the opposite, Helsinki CS Department's PULS system [11] combines Information Retrieval with Information Extraction to perform a full automatized processing. These systems are mostly based on keywords and syntactic analysis.

To the contrary, our proposal is to use structure patterns [6] for extracting and selecting relevant content in newswires. The system worked at first for French and then experiments were made in small sets of English and Spanish documents to check if the position of relevant content can also be compared in these two languages. Here, the system is in a way considered as a non native speaker: as it does not know very well how the language function in details, it will only try to check if the terms searched are in some clearly identified positions. Details on this approach will be given in the next section.

3 Our approach

Our goal is to see how we could build an Information extraction system with very low resources and without machine translation. The idea was therefore to use a different grain in analysis, following the state of the art for press articles [3, 4]. Therefore the system uses the structure of the text to tell us where the relevant information may be. It is mostly based on the "5W rule" which, in press articles, tells that the answers to the main questions "What, Where, Who, When and Why" are to be easily found by the reader to help her check if the article is interesting for her. In order to find relevant content, the system divides the document in two parts:

- HEADER: title and first two sentences
- BODY: rest of the text

If a string is found in both "HEADER" and "BODY", using repeated string algorithm for [5], it is stored as a potentially relevant content. If one of these strings corresponds to a disease name then the document is considered potentially relevant.

In fact we assume that, according to relevance principle in human communication [7, 9], there is only one important event by article and as we want to control redundancy we consider that secondary events have been treated elsewhere as primary events. Therefore if numerous diseases are found in the "potentially relevant content", it means that the document is not an important alert (in which case only one disease is mentioned) and by the way less interesting for our purpose. Finally it is not necessary to have many different names for each disease because in newspaper articles only a few are really used. Finally to be able to monitor new languages easily, resources are limited: 200 diseases names , 400 toponyms and a few words for date matching in each of our languages. The slots we need to fill in the database were What (disease), Where (location), Who (cases, people affected by disease) and When (date).

To extract the location the following algorithm is applied: the relevant location correspond to a string in the "relevant content". If numerous locations matches, the system compares frequencies in the whole document: if one location is more than twice as frequent as others it is considered as the relevant one. Concerning cases, they are defined as the first numeric information found in the document and not related to money or date. Furthermore the extracted cases are considered more relevant if they appear twice in the document. Thus the system uses regular expressions to round up and compare them (see example 4).

4 Some results

In the experiments of Spanish and English (see example 1 and example 2), it has been found that the algorithm used for French still works. Documents for the experiments were selected at random and manually tagged. The combination between position and frequency seems as reliable in these two languages as we had found it to be in French. Some problems may occur concerning cases, see for instance in example 2 where the system will detect "50" and consider it relevant because of repetition. It is impossible at this stage for the system to detect which kind of case is concerned by the number (human or animals for instance). In previous experiments, results (Table 1) were slightly better than in Table 2 and Table 3. More precisely the disease extracted is relevant in 90% of documents while location seem more dificult to detect correctly: around 80% are good.

HONG KONG , Sept. 21 (Xinhua) – **Hong Kong** 's Department of Health said Monday that it had advised eight primary schools and three secondary schools to suspend classes for seven days starting Tuesday to stop the possible spread of ***A/H1N1 flu*** in the schools. The advice was made following the outbreaks of the special flu in the schools involving **824 students** aged between six and 17, a spokesman for the department said.[...] **Hong Kong** reported 446 newly confirmed cases of ***influenza A/H1N1*** in the 24 hours up to 2:30 p.m. Monday, bringing the number of the city's cumulative cases to 22,500. Meanwhile, the city reported another fatal case of ***Influenza A/H1N1*** , bringing to 16 the total number of fatal cases of the special ***flu*** in the city.

Example 1: English ***disease*** **country** **cases**

Un cocinero de 50 años es la cuarta víctima mortal de la ***gripe A*** en Cataluña Unos familiares lo encontraron muerto en su casa **Un cocinero de 50 años** residente en L'Estartit, en Torroella de Montgrí (Girona), es la última víctima de la ***gripe A*** en **España**, según ha confirmado una portavoz de Salud. [...] El número de fallecidos en **España** por la enfermedad es ya de 33, a la espera de que mañana el Ministerio de Sanidad actualice las cifras.

Example 2: Spanish ***disease*** **country** **cases**

Manually tagged	Extracted	Rejected	Results
Relevant documents	196	14	Recall 93%
Non relevant documents	28	962	Precision 87.5%

Table 1. Results on French

Manually tagged	Extracted	Rejected	F-measure 84%
Relevant documents	44	6	Recall 88%
Non relevant documents	11	39	Precision 80%

Table 2. Results on English

Manually tagged	Extracted	Rejected	F-measure 85%
Relevant documents	61	6	Recall 91%
Non relevant documents	15	25	Precision 80%

Table 3. Results on Spanish

Work is in progress on Russian, Finnish and Turkish from documents of similar dates. At this stage, algorithms seem to keep good reliability (Examples 3, 4 and 5). String repetitions permit to extract interesting informations even in languages with declension. Using position also allow the system to prune the last sentence of the example in Russian which concerns Spanish flu. Then comparing different collections might quicken the acquisition of new terms by quasi alignments techniques [8].

« ***Свин***ой ***грипп***» шествует по **мир**у: уже **4379** заболевших в 29 странах Опубликована: 10 мая 2009 19:53:11 По данным Всемирной организации здравоохранения количество заболевания **грипп**ом A/H1N1 увеличилось до **4379** в 29 странах **мир**а.
Еще в субботу ВОЗ сообщал, что количество заболевших 3440 человек. НА сегодняшний момент 45 человек уже умерло от « ***свин***ого ***грипп***а» в Мексике, 2 – в США, 1 – в Канаде, 1 – в Коста-Рике: итого – 49 человек. [...] Ранее ученые неоднократно заявляли, что нынешняя эпидемия ***грипп***а вряд ли повторит "испанку", которая в 1918-1920 годах унесла более 20 миллионов жизней, поскольку теперь медики и эпидемиологи намного больше знают о возбудителе ***грипп***а и механизмах распространения болезни, сообщает РИА Новости.

Example 3: Russian ***disease*** **country** **cases**

The disease identified here is "swine flu" (repeated string "Свин грипп") worldwide (repeated string "мир"). The system extracts "4379" as the number of cases with confidence because it appears twice.

Kolera tappanut jo yli **3000** **Zimbabwe**ssa
julkaistu 28.01. klo 12:37, päivitetty 28.01. klo 14:10
Eteläafrikkalaisessa **Zimbabwe**ssa ***kolera***an kuolleiden määrä on noussut jo yli kolmen tuhannen. Maailman terveysjärjestön WHO:n mukaan kuolleita on **3 028**. Lisäksi yli 57 000 ihmistä on sairastunut. Tiistain jälkeen on rekisteröity 57 uutta kuolemantapausta ja yli 1 500 uutta tartuntaa.
Elokuussa puhjennut epidemia on Afrikan pahin 14 vuoteen.
Zimbabwen presidentti Robert Mugabe on väittänyt, ettei **Zimbabwe**ssa enää ole ***kolera***a. [...]
Kolera tarttuu bakteerin saastuttamasta ruoasta tai juomavedestä. Tauti aiheuttaa ripulia, oksentelua ja niistä johtuvan nestehukan. Pahimmassa tapauksessa kolera johtaa kuolemaan.

Example 4: Finnish ***disease*** **country** **cases**

This document tells us that there were 3028 cases of cholera identified in Zimbabwe, 3000 cases were mentioned in the title but it was identified as a round up of "3028". Repetition analysis permits the system to identify "Zimbabwe"(repeated string "Zimbabve") as the location of the main event of the document.

Zimbabve'de ***kolera*** salgınında ölenlerin sayısı, **3868'e** çıktı
25 Şubat 2009 Çarşamba 04:30
CENEVRE -AA- **Zimbabve**'de ***kolera*** salgınında ölenlerin sayısı **3868'e** yükseldi. Dünya Sağlık Örgütü yetkilileri, ağustostan bu yana görülen vak'a sayısının 83 bini aştığını bildirdi. Yetkililer, ***kolera***dan ölüm oranının yüzde 4,7 olduğuna dikkat çekti. Dünya ortalaması ise yüzde 1'in altında. Hastalığın yayılmasında kirli sular baş rolü oynuyor.

Example 5: Turkish ***disease*** **country** **cases**

This document shows that the system does not extract Geneva (string "CENEVRE") as a relevant location because there is no repetition for the concerned string.

5 Conclusion

The results for English and Spanish seem to show comparable reliability to those for French. It means that this high-grain method can be useful. One can also see that first experiments on such different languages like Russian, Finnish and Turkish are interesting because they are from different families. The agglutinative aspect of Finnish would be for instance a great problem for a word-based approach. However, work is still needed to improve the case extraction in order to make a truthful comparison to the PULS English system. Our first goal will be to evaluate our system in a multi-event perspective. We shall then ponder on the trade-off between the subsequent loss and the possibility of monitoring new languages quickly and efficiently.

References

1. Freifeld, Mandl, Reis, Brownstein *HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports* in *Med Inform Assoc.* vol 15 pp 150-157 (2008)
2. Hull and Grefenstette *Querying across languages: a dictionary-based approach to multilingual information retrieval.* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (1996)
3. Itule and Anderson *News writing and reporting for today's media.* Mcgraw-Hill College (1991)
4. Kando *Text Structure Analysis as a Tool to Make Retrieved Documents Usable.* 4th International Workshop on Information Retrieval with Asian Languages, Taipei Nov. 11-12, pp. 126-132. (1999)
5. Kärkkäinen and Sanders *Simple linear work suffix array construction.* in Proc. 30th International Conference on Automata, languages and Programming volume 2719 of LNCS, pages 943-955. Springer (2003)
6. Lucas *The enunciative structure of news dispatches: A contrastive rhetorical approach* in Ilie, ed., *Language, culture, rhetoric: Cultural and rhetorical perspectives on communication* ASLA, Stockholm, pp. 154-164 (2004)
7. Reboul and Moeschler *La pragmatique aujourd'hui. Une nouvelle science de la communication.* Paris: Le Seuil (1998)
8. Riloff, Schafer, Yarowski *Inducing information extraction systems for new languages via cross language projection.* 19th international conference on Computational linguistics - Volume 1, Taipei, Taiwan, Association for Computational Linguistics. pp. 1-7. (2002)
9. Sperber and Wilson *Relevance: Communication and cognition.* Blackwell Press (1998)
10. Linge, Steinberger, Weber, Yangarber, Van der Goot, Al Khudhairy, Stilianakis *Internet surveillance systems for early alerting of health threats* in *Eurosurveillance* 14 (13), (2009)
11. Steinberger, Fuart, Van der Goot, Best, Von Etter, Yangarber *Text mining from the Web for medical intelligence in Mining massive data sest for security.* Amsterdam, the Netherlands OIS Press (2008)