Filtering news for epidemic surveillance: towards processing more languages with fewer resources

Gaël Lejeune, Antoine Doucet, Roman Yangarber, Nadine Lucas

Filtering news for epidemic surveillance: towards processing more languages with fewer resources

Gaël Lejeune[1], **Antoine Doucet**[1], **Roman Yangarber**[2], **Nadine Lucas**[1]
[1]GREYC, University of Caen [2]CS department, University of Helsinki
first.last@info.unicaen.fr yangarbe@cs.helsinki.fi

Abstract

Processing content for security becomes more and more important since every local danger can have global consequences. Being able to collect and analyse information in different languages is a great issue. This paper addresses multilingual solutions for analysis of press articles for epidemiological surveillance. The system described here relies on pragmatics and stylistics, giving up "bag of sentences" approach in favour of discourse repetition patterns. It only needs light resources (compared to existing systems) in order to process new languages easily. In this paper we present here results in English, French and Chinese, three languages with quite different characteristics. These results show that simple rules allow selection of relevant documents in a specialized database improving the reliability of information extraction.

1 Multilingual techniques in information extraction

In natural language processing, information extraction is a task where, given raw text, a system is to give precise information fitting in a predefined semantic template.

1.1 Epidemic surveillance

Automated news surveillance is an important application of information extraction. The detection of terrorist events and economic surveillance were the first applications, in particular in the framework of the evaluation campaigns of the Message Understanding Conference (MUC) (MUC, 1992; MUC, 1993). In MUC-3 (1991) and MUC-4 (1992), about terrorism in Latin American countries, the task of participants was, given a collection of news feed data, to fill in a predetermined semantic template containing the name of the terrorist group that perpetrated a terrorist event, the name of the victim(s), the type of event, and the date and location where it occurred. In economic surveillance, one can for instance extract mergers or corporate management changes.

An application of information extraction that lately gained much importance is that of epidemiological surveillance, with a special emphasis on the detection of disease outbreaks. Given news data, the task is to detect epidemiological events, and extract the location where they occurred, the name of the disease, the number of victims, and the "case", that is, a text description of the event, that may be the "status" of victims (sick, injured, dead, hospitalised ...) or a written description of symptoms. Epidemiological surveillance has become a crucial tool with increasing world travel and the latest crises of SARS, avian flu, H1N1 ...

In this paper, we present an application to epidemic surveillance, but it may be equally applied to any subdomain of news surveillance.

1.2 Multilingual information extraction

As in many fields of NLP, most of the work in information extraction long focused on English data (Etzioni et al., 2008).Multilingual has often been understood as adding many

monolingual systems, except in pioneer multilingual parsing (Vergne, 2002). Whereas English is nowadays the *lingua franca* in many fields (in particular, business), we will see that for several applications, this is not sufficient. Most news agencies are translating part of their feed into English (e.g., AFP[1] and Xinhua[2] for which the source languages are respectively French and Chinese), but a good deal of the data is never translated, while for the part that is, the translation process naturally incurs a delay that is, by essence, problematic in a field where exhaustivity and early detection are crucial aspects.

Subsequently, the ability to simultaneously handle documents written in different languages is becoming a more and more important feature (Poibeau et al., 2008; Gey et al., 2009). Indeed, in the field of epidemiological surveillance, it is especially important to detect a new event the very first time it is mentioned, and this very first occurrence will almost always happen in the local language (except for countries like Iraq for instance). Therefore, it is not enough to be able to deal with several languages : It is necessary to handle many. For instance, the Medical Information System (Medisys) of the European Community gathers news data in 42 different languages (Atkinson and der Goot, 2009) (now 45[3]).

1.3 Current approaches

There are currently 2 main approaches to multilingual information extraction. The first approach relies on the prior translation of all the documents into one common language (usually English), for which a well-performing information extraction system has been developed (Linge et al., 2009). Whereas the simple design of this solution is attractive, the current state of the art in machine translation only allows for mediocre results. Most monolingual information extraction systems indeed rely on a combination of grammatical patterns and specialized lexicons (Grishman et al., 2002; Riloff, 1996).

The second main approach consists in leaving documents in their original language but to translate the lexicons and extraction patterns into that language (Efimenko et al., 2004; Linge et al., 2009). However, the same problems occur as in the first approach because the patterns are strongly language-related. Yet, to "translate the system" seems more realistic than to translate the documents, as it can be done manually, and offline (once and for all, and not as documents arrive). The bottleneck is then that the amount of work for each language is enormous: it naturally requires the complete translation of the lexicon (for all trigger words), but the more challenging issue is the translation of patterns, whose language-dependence might well mean that the amount of work needed to translate them comes close to that required for writing them from scratch. In addition, this task must necessarily be achieved by a domain expert, with excellent skills in the languages at hand. One could want to tackle this problem by using machine learning but she will need training data in many languages. In practice, this will often mean that only a few major languages will be dealt with, whilst all the others (amongst which all low-resource languages), will again be totally discarded. One can then only wish that epidemics will chose to occur in locations handled by surveillance systems...

Both approaches additionally require a number of linguistic processing tools, in a number comparable to the number of languages to be dealt with: tokenizer, stemmer, syntactic analyzer, ... One might therefore conclude that such techniques are not properly multilingual but rather monolingual methods that may be adapted to other languages individually.

In this paper, we explore a third approach to multilingual information extraction. We restrain ourselves to the sole use of truly mul-

[1] http://www.afp.com/afpcom/en
[2] http://www.xinhuanet.com/english2010/
[3] http://medusa.jrc.it/medisys/aboutMediSys.html

tilingual elements, facts that are equally true for any language. The approach hence relies on universals, relying, e.g., on stylistics and rhetorics.

2 Rationale of the experiment

The objective of the system is to monitor news in a variety of languages to detect disease outbreaks which is an important issue for an alert system in epidemic surveillance. For this task a simple and clear framework is needed in order to limit the amount of work for new languages while keeping good reliability. The main idea of our work is using text granularity and discourse properties to write rules that may be language independent, fast and reliable (Vergne, 2002). For this study, regularities at text level are exploited. These phenomena can be related to stylistics and pragmatics. It has already been shown that news discourse has its own constraints reflected in press articles of different languages (Van Dijk, 1988; Lucas, 2004).

2.1 Stylistic rules

Journalists all over the world know how to hook their potential readers. These methods are described in journalism schools (Itule and Anderson, 2006). One very important rule for journalists seems to be the"5W rule" which emphasise on the fact that answering to the questions"What","Where","When", "Why" and "Who" is a priority at the start of a paper. Only after that can journalists develop and give secondary information. This phenomenon is genre dependent and is exploited for processing texts by searching for repetitions.

Example 1 shows a piece of news where the disease name is found in the beginning of the news article and developed later on. No local pattern is needed to detect what the article is about, repetition phenomena is sufficient.

Example 2 is a counter example, where a disease name is found but not repeated. This French document reports on a pop music band being the "coqueluche" of Hip-Hop, which can mean "pertussis", but here means "fashion" in a figurative sense (underlining the fast spread of the band's popularity). Usually, figurative meanings are not used twice in the same article (Itule and Anderson, 2006) and hence the repetition criteria allows one to rightfully ignore this article.

2.2 Pragmatic rules

As press articles are made for humans, strong effort is exerted to ensure that readers will understand the main information with as few inferences as possible (Sperber and Wilson, 1998). In fact, the more inferences the reader has to make, the more errors he is likely to make and the more probability he will get confused and not read the full article. Repetitions are there to relieve the memory effort. A point that journalists pay much attention to is leaving as few ambiguities on main facts as possible. It means that potentially unknown or complicated terms will be used quite rarely. Only one main story will be developed in an article, other facts that are important will be developed elsewhere as main stories.

3 Our system

The system is based on the comparison of repetitions in the article to find documents relevant for epidemic surveillance and extract where the disease occurs and how many people are concerned.

3.1 String repetitions: relevant content

A system is not a human reader, so objective discourse marks are used by the system. Repetitions are known since the ancient times as reflecting discourse structure. A press article is divided into two parts, roughly the head and the rest of the news. The title and the first two sentences form the head or thematic part and the rest of the text is considered to be a development in an expository discourse.

Measles outbreak spreads north in B.C.
Number of cases hits **44** provincewide B.C.'s **measles** outbreak appears to have spread to northeastern areas of the province, after doctors confirmed two new cases of the disease in the Fort St. John and Fort Nelson areas on Thursday.

The new cases bring the total number of confirmed cases in the province to **44**, not including suspected but unconfirmed cases, said the B.C. Centre for Disease Control. Northern Health spokeswoman Eryn Collins said the virus had not been detected in the north in more than six years and the two new cases involve people who weren't immunized. [...] "It is suspected that at least two out-of-country visitors brought **measles** into Vancouver sometime in February or early March, as two separate strains of the virus have been identified," said a statement from the B.C. Centre for Disease Control earlier this week. So far, 17 cases of the **measles** have been detected in the Fraser Valley, 17 in the Vancouver area, seven in the southern Interior, two in northern B.C. and one on Vancouver island.

Figure 1: Example in English: repetition of disease name and cases

Cameroun/Musique : X-Maleya nouvelle **coqueluche** du Hip-Hop camerounais !
Le trio Hip-Hop Cameounais X-Maleya, a le vent en poupe. Le groupe qui s'illustre dans la tendance Hip-Hop, est aujourd'hui l'une des valeurs sres musicales grâce son second opus Yelele.
Derrière ces trois prénoms : Roger, Auguste et Haïs, se cachent un trio camerounais qui s'illustre dans le monde du Hip-Hop. [etc.] C'est donc, une nouvelle valeur sûre qu'incarnent eux trois Roger, Auguste et Haïs. Le groupe rencontre en effet, une ascension fulgurante. Les trois faiseurs de Hip-Hop, ont une seule idée en tête, continuer de se produire pour ceux qui les apprécient, toujours composer de belles mélodies et, ne pas oublier d'où ils viennent.

Figure 2: Example in French: no repetition

Strings that are present in both parts will be referred to as"relevant content". They are found in the beginning of the news and repeated in the development. To process as many languages as possible, repeated character strings will be searched (not words because Chinese for instance does not use graphic words).

3.2 Defining epidemic event

Epidemic events are captured through these information slots:

- Disease (What)
- Location (Where)
- Case, i.e.,People concerned (Who)

3.3 Selecting potentially relevant documents

This discourse related heuristic rule limits resources needed by the system. Many character strings that are repeated in the text reflect important terms. However, repetition alone does not allow to fill IE templates with detailed information as required. Accordingly, a lexical filter is applied on the repeated strings. 200 common disease names are used to filter information and find disease names. The idea behind the restricted list is that a journalist will use a common name to help his readers understand the message. Similarly, for locations, a list of country names and capitals provided by UN is

WHO checks smallpox reports in **Uganda**
LONDON, Thursday
The World Health Organisation said today it was investigating reports of suspected cases of the previously eradicated disease smallpox in eastern **Uganda**.
Smallpox is an acute contagious disease and was one of the worlds most feared sicknesses until it was officially declared eradicated worldwide in 1979.
"WHO takes any report of smallpox seriously, Gregory Hartl, a spokesman for the Geneva-based United Nations health agency, told Reuters via email.
"WHO is aware of the reports coming out of **Uganda** and is taking all the necessary measures to investigate and verify."[etc.]

Figure 3: Example in English: repetition and location

used (about 500 items). Finally, in order to comply with a specific demand of partners, blacklist terms were used to detect less relevant articles (vaccination campaign for instance).

When a disease name is found in the relevant content, the article is selected as potentially relevant and the system tries to extract location and cases.

3.4 Extracting location and cases

To extract the location, the following heuristic is applied: the relevant location corresponds to a string in the"relevant content". For instance, Example 3 shows that it allows for the system to find that the main event concerns Uganda but not London.

If numerous locations match, 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. If no location is found, the location of the source is selected by default. In fact according to pragmatic rules when one reads an article in the Washington Post, she will be sure that it is about the United States even if it is not explicitly mentioned. To the contrary if the article is about Argentina it will be clearly mentioned so the reader has less chances of misunderstanding.

Concerning the cases, they are related to the first numeric information found in the document, provided the figures are not related to money or date (this is checked by a blacklist and simple regular expressions).

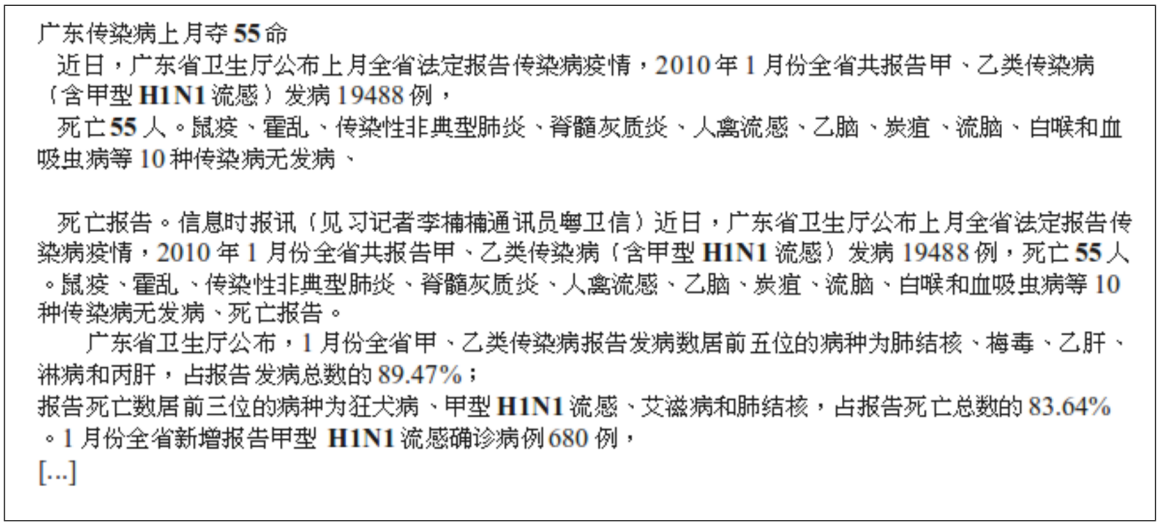
Furthermore the extracted cases are considered more relevant if they appear twice in the document, the system uses regular expressions to round up and compare them. See Example 4 where the number of dead people "55" is the first numeric information in the beginning and is repeated in the development (we chose an example where it is easy even for a non Chinese speaker to see the repetition). One can also note that the second repeated figure is "19488" which is the number of infected people.

4 Evaluation

It is important to insist on the fact that our system extracts the main event from one article, considering that secondary events have been or will be mentioned in another article. Often, the more topics are presented in one article, the less important each one is. In the case of epidemic surveillance, review articles or retrospectives are not first-hand, fresh and valuable information.

4.1 Corpus and Languages

For each language we randomly extracted documents from the Medisys website. Medisys documents are gathered using keywords: medical terms (including scientific disease names), but also weaker keywords such as casualties, hospital... This implies that some news document not related


广东传染病上月夺 **55** 命
近日，广东省卫生厅公布上月全省法定报告传染病疫情，2010 年 1 月份全省共报告甲、乙类传染病（含甲型 **H1N1** 流感）发病 19488 例，
死亡 **55** 人。鼠疫、霍乱、传染性非典型肺炎、脊髓灰质炎、人禽流感、乙脑、炭疽、流脑、白喉和血吸虫病等 10 种传染病无发病、

死亡报告。信息时报讯（见习记者李楠楠通讯员粤卫信）近日，广东省卫生厅公布上月全省法定报告传染病疫情，2010 年 1 月份全省共报告甲、乙类传染病（含甲型 **H1N1** 流感）发病 19488 例，死亡 **55** 人。鼠疫、霍乱、传染性非典型肺炎、脊髓灰质炎、人禽流感、乙脑、炭疽、流脑、白喉和血吸虫病等 10 种传染病无发病、死亡报告。
广东省卫生厅公布，1 月份全省甲、乙类传染病报告发病数居前五位的病种为肺结核、梅毒、乙肝、淋病和丙肝，占报告发病总数的 89.47%；
报告死亡数居前三位的病种为狂犬病、甲型 **H1N1** 流感、艾滋病和肺结核，占报告死亡总数的 83.64%。1 月份全省新增报告甲型 **H1N1** 流感确诊病例 680 例，
[...]

Figure 4: Example in Chinese: 55 deaths from H1N1

to epidemic surveillance, but to accident reports for instance, are liable to be found in the database.

We must underline that in this framework, recall can only be estimated, notably because the news documents are keyword-filtered beforehand. However, our aim is not to provide an independent system, but to provide quick sorting of irrelevant news, prior to detailed analysis, which is the key issue of a surveillance and alert system. 200 documents were extracted for each language and manually tagged by native speakers with the following instructions:

- Is this article about an epidemic?
- If it is, please give when possible:
  Disease(s)
  Country (or Worldwide)
  Number of cases

100 of these annotated documents were used for fine-tuning the system, 100 others for evaluating. We chose for this study 3 fairly different languages for checking the genericity of the approach

- French, with its rather rich morphology,
- English,a rather isolating language with poor morphology,
- Chinese, a strict isolating language with poor morphology.

4.2 Results

These results were computed from a set of 100 annotated documents, as described in section 4. Table 1 shows recall, precision and F-measure for document selection(more examples are available online [4]) Table 2 compares automatically extracted slots and human annotated slots, therefore if an event is not detected by the system it will count as an error for each slot.

Table 1 shows that selection of documents is quite satisfactory and that recall is better than precision. This is mostly due to the fact that the system still extracts documents with low relevance. We found it impossible to predict if this is a general bias and whether it can be improved. The result analysis showed that many false negatives are due to cases when the piece of news is quite small, see for instance Example 5 where “Swine flu” is only found in the first two sentences, which implies the repetition criteria does not apply (and the system misses the document).

Table 2 shows the accuracy of the information entered into semantic slots, respec-

[4] http://sites.google.com/site/iesystemcoling2010

China has 100 cases of swine flu: state media
China has 100 confirmed cases of swine flu, state media said Tuesday, as data from the World Health Organization showed the disease had spread to 73 countries.
"The health ministry has reported that so far, China has 100 confirmed cases of A(H1N1) flu," said a news report on state television CCTV. The report said the 100 cases were in mainland China, which does not include Hong Kong or Macau.

Figure 5: Example in English: Disease name not in "relevant content"

Language	Recall	Precision	F-measure
French	93%	88%	90%
English	88%	84%	86%
Chinese	92%	85%	88%

Table 1: Selecting documents

Language	Diseases	Locations	Cases
French	88%	87%	81%
English	81%	81%	78%
Chinese	82%	79%	77%

Table 2: Accuracy in filling slots

tively name of disease, location and number of cases. It is important to say that the descriptors extracted are really reliable in spite of the fact that the annotated set used for evaluation is fairly small: 100 documents per language, 30 to 40 of which were marked as relevant. The extraction of cases performs a bit worse than that of locations but the location is the most important to our end-users.

5 Discussion and Conclusion

Most research in Information Extraction (IE) focuses on building independent systems for each language, which is time and resource consuming. To the contrary, using common features of news discourse saves time. The system is not quite independent, but it allows filtering news feeds and it provides reasonable information even when no resources at all are available. Our results on English are worse than some existing systems (about 93% precision for Global health Monitor for instance) but these systems need strong resources and are not multilingual. We then really need a multilingual baseline to compare both approaches.

Recall is important for an alert system, but is very difficult to assess in the case of epidemiological surveillance. This measure is always problematic for web based documents, due to the fact that any randomly checked sample would only by sheer luck contain all the positive documents. The assumption here is that no important news has been missed by Medisys, and that no important news filtered from Medisys has been rejected.

One explanation for missed articles lies in the definition of the article header: it is too rigid. While this is fine for standard size news, it is inappropriate for short news, hence meaningful repetitions are missed in the short news. This is a flaw, because first alerts are often short news. In the future, we may wish to define a discourse wise detection rule to improve the location slot filling. The extraction of locations is currently plagued by a very long list of countries and capitals, most of which is not useful. Locations are actually mentioned in data according to states, provinces, prefectures, etc. The country list might be abandoned, since we do not favour external resources.

The methods that are presented here maintain good reliability in different languages, and the assumption that genre laws are useful has not been challenged yet. Light resources, about 750 items (to be compared to tens of thousands in classical IE systems), make it possible to strongly divide the amount of work needed for processing new languages. It might be attempted to refine the simple hypotheses underlying the pro-

gram and build a better system for filtering relevant news. This approach is best suited when combined with elaborate pattern-based IE modules when available. Repetition can be checked for selecting documents prior to resource intensive semantic processing. It can also provide a few, easily fixable and efficient preliminary results where language resources are scarce or not available at all.

References

Atkinson, Martin and Erik Van der Goot. 2009. Near real time information mining in multilingual news. In *18th International World Wide Web Conference (WWW2009)*.

Efimenko, Irina, Vladimir Khoroshevsky, and Victor Klintsov. 2004. Ontosminer family: Multilingual ie systems. In *SPECOM 2004: 9th Conference Speech and Computer*.

Etzioni, Oren, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Gey, Fredric, Jussi Karlgren, and Noriko Kando. 2009. Information access in a multilingual world: transitioning from research to real-world applications. *SIGIR Forum*, 43(2):24–28.

Grishman, Ralph, Silja Huttunen, and Roman Yangarber. 2002. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246.

Itule, Bruce and Douglas Anderson. 2006. *News Writing and Reporting for Today's Media*. McGraw-Hill Humanities.

Linge, JP, R Steinberger, T P Weber, R Yangarber, E van der Goot, D H Al Khudhairy, and N I Stilianakis. 2009. Internet surveillance systems for early alerting of threats. *Eurosurveillance*, 14.

Lucas, Nadine. 2004. The enunciative structure of news dispatches, a contrastive rhetorical approach. *Language, culture, rhetoric*, pages 154–164.

MUC. 1992. *Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992*.

MUC. 1993. *Proceedings of the 5th Conference on Message Understanding, MUC 1993, Baltimore, Maryland, USA, August 25-27, 1993*.

Poibeau, Thierry, Horacio Saggion, and Roman Yangarber, editors. 2008. *MMIES '08: Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, Morristown, NJ, USA. Association for Computational Linguistics.

Riloff, Ellen. 1996. Automatically generating extraction patterns from untagged text. In *AAAI/IAAI, Vol. 2*, pages 1044–1049.

Sperber, Dan and Deirdre Wilson. 1998. *Relevance: Communication and cognition*. Blackwell press, Oxford U.K.

Van Dijk, T.A. 1988. *News as discourse*. Lawrence Erlbaum Associates, Hillsdale N.J.

Vergne, Jacques. 2002. Une méthode pour l'analyse descendante et calculatoire de corpus multilingues: application au calcul des relations sujet-verbe. In *TALN 2002*, pages 63–74.