Structure patterns in Information Extraction: a multilingual solution?

Gael Lejeune

Department of Computer Science, University of Helsinki
GREYC, University of caen

Gustaf Hallstromin katu 2B, Helsinki

E-mail: lejeune@cs.helsinki.fi

Abstract

IE systems nowadays work very well, but they are mostly monolingual and difficult to convert to other languages. We maybe have then to stop thinking only with traditional pattern-based approaches. Our project, PULS, makes epidemic surveillance through analysis of On-Line News in collaboration with MedISys, developed at the European Commission's Joint Research Centre (EC-JRC). PULS had only an English pattern-based system and we worked on a pilot study on French to prepare a multilingual extension. We will present here why we chose to ignore classical approaches and how we can use it with a mainly language-independent based only on discourse properties of press articlestructure. Our results show a precision of 87% and a recall of 93%. And we have good reasons to think that this approach will also be efficient for other languages.

1 Introduction

In a domain like epidemic surveillance, having an IE system limited to only one language is insufficient. For instance, for countries like France or Togo it will be difficult to find press articles about diseases written in English. Therefore some crucial information might take time to be shown

by the system. If we wait for a better Machine Translation (as a recent article on Euro- surveillance proposed [1] it might take a long time as well. Much effort as been devoted to multilingual analysis. How should one extend a system to another language? As we will have different steps to follow to extract information, our goal will be to have as few language-dependent steps as possible. The purpose of this article is to show a method whose aim is modest but may also be simple and efficient.

2 Classical works in NLP area

The concept of text as a "bag of words" seems to be declining, its obviously because from now on linguistic approaches in NLP are quite powerful. These approaches are mainly based on academical subdivisions of linguistics:

- Lexical normalization (aiming to morphemes)
- Morphological analysis (identifying "words")
- Syntactic analysis (combining words)
- Semantic analysis (meaning representation, sentence level)
- Discourse analysis (combination of sentences or documents)

Improvements have been made to some of these tasks, using predicate/argument constructions and ontology-driven analysis. Also, Part of Speech (POS) taggers are quite efficient now. However, in the current view, a multilingual parser is a sum of language-dependent parsers. Building these resources is a long and hard job and even automatic learning always needs human fixing. Furthermore as it requires many steps and many tools, many different errors can occur during the process. In this "standard vision" the granularity used is the word (or the lexical item) and almost nothing is said about two other fields of linguistics:

- Stylistics
- Pragmatics

These parts are maybe supposed to be somewhat useless for our purpose or to have a lack of models. But there are some useful works that might help Computer Scientists.

For instance, pragmatics tell us that language is governed by effectiveness rules[2] or said differently by pertinence rules [3]. Human beings use speech acts principles to limit the cognitive cost of the exchange. For instance, as a journalist has to prove to his reader that his article is worthy of reading he will be very careful about his title and first sentences. He will also try to focus on an important and easy understandable fact. It means that in such a special type of text we have possibilities to guess what the main information is just by refering to those rules. There is also an important probability that there is only one interesting piece of information in each article, if a piece of information is worthy it should need its own article.

Studies with text as different granularity levels exist for different text types [4]. For press articles which are the main part of our corpus, many models have been used. The structure model that our approach is based on was elaborated by Nadine Lucas [5] It is conveyed by the "5W rule" saying that answers to Who,What, Where, When Why have to be given in the very beginning of documents. It works in both French and English and we can guess that it would be a good candidate for a multilingual rule. This rule says that the main information is to be found in the top of the document. As in our epidemic surveillance system PUlS the corpus only contains press articles, we have tried to apply this rules to our task. As an experiment our simplified goal was to identify:

- What: disease
- Where: country
- Who: cases (people affected by diseases)
- When: date (in this version we extract only the documents date)

3 PULS French System

PULS's aim is to monitor as many language as possible to help epidemiologists in their task. The French system which we present here is therefore

Type of event	Explanations	To be extracted
Highly relevant	new information	Yes
Quite relevant	important update	Yes
Less relevant	review article	Yes
Very low relevance	historical, not current	No
Not relevant	non-specific event	No
Not relevant	wrong event	No

Table 1: Relevance scoring

intended to be a real seedwork for monitoring new languages with as few human labor as possible:

First the documents is divided into two parts,

- HEAD: title and two first sentences
- BODY: rest of the document

3.1 Disease

If a disease from the database (150 items) is found in both parts then the document is considered possibly relevant. Then a small blacklist is used to filter out less relevant documents (cf table 1). It is very important to control redundancy and to give to the user really relevant documents. If more than one disease are matched, the document is also considered less relevant referring to the pertinence rule.

3.2 Location

Any location from the database (400 items) found in the "Head" part is considered possibly the good one. If there is no location in the head, we search locations that appears at least twice in the body. If there is still no location matched we consider, according to the pertinence rule, that the event is happening in the default country of the news source which we find in our source database (30 items).

If more than one country is matched, this algorithm is applied:

- The relevant location appears more than 2 times in the head and twice as many as any other location
- If its undecided, the same algorithm is applied to the whole text
- Finally if its still undecided, the document is marked as less relevant assuming that the pertinence rule suggests that if there is an important fact in a document you must talk mostly about this fact and therefore about only one location.

3.3 Cases (descriptor)

To find cases (descriptors) we apply this rule: Cases in a relevant document are specified as the first numeric information in the first half of the text that is not related to money, distance or time. We use a stoplist which includes names of months, currency names and date nouns (20 items). When no descriptor is matched its mostly because the number of cases is in letters therefore we extract the first phrase containing the disease name.

4 Results

Indian **swine flu** death toll hits **100**
MUMBAI A total of **100** **people** have died from **swine flu** in **India** since the first fatality was recorded one month ago, the government said.
The health ministry announced in a statement late on Monday that the most recent victims were four people in western Maharashtra state, which has seen 55 deaths from the **A(H1N1)** virus.
Indian 's first confirmed case of **swine flu** was in May. The first death was on August 3. Fears about contracting **swine flu** led to huge queues forming outside government hospitals while the rising death toll led to the temporary closure of schools and cinemas in and around Mumbai and Pune.

Example on English: **disease** **country** **cases**

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

Table 2: Results

The corpus is provided by Medical Information System, MedISys, which gather reports concerning Public Health. We worked with a sample of approximately 1200 files (from December 3 to 14 2008), from which 210 documents were manually tagged as relevant according to our scale (table1 :score 1 to 3). it is important to say here that we are very careful about which documents we consider relevant for our purpose.

Our results (Table 2) are difficult to compare to other systems since we have only language-dependent systems to compare with. However we are already very close to the English version of PULS [6] which is pattern-based.

5 Conclusion

The promising scores we got from the above experiment has convinced us that there are still improvements to get from the existing models. Our next step will be to test our system on other Romance languages (for instance Italian and Spanish) then to other Indo-European ones. If we can keep the idea and the simplicity of it in a number of language families we would be able to say that we can monitor an important part of the epidemic data in the world.

Acknowledgements

AFFRST, Content factory, Algodan and Europeans Commission JRC for their support. Nadine Lucas for basic algorithm and useful ideas. Roman Yangarber and Antoine Doucet for advices and reviews. Charlotte Lecluze and Calliopi Sachtouri for previous work on news reports.

References

[1] JP.Linge, R.Steinberger, T.P Weber, R.Yangarber, E van der Goot, D H Al Khudhairy, N I Stilianakis *Eurosurveillance Vol 14.* Issue 13, 02 April 2009

[2] Reboul and Moeschler *La pragmatique aujourdhui. Une nouvelle science de la com- munication.* Paris: Le Seuil 1998(Points)

[3] Sperber and Wilson *Relevance: Communication and cognition.* Blackwell Press 1998.

[4] N. Kando *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 (1996), pp. 107-129.

[5] N.Lucas (2004) *The enunciative structure of news dispatches: A contrastive rhetorical approach A contrastive rhetorical approach, in C. Ilie, ed., Language, culture, rhetoric: Cultural and rhetorical perspectives on communication* ASLA, Stockholm, 2004, pp. 154-164.

[6] R.Steinberger, F.Fuart, E.Van der Goot, C.Best, P. Von Etter, R.Yangarber *Text mining from the Web for medical intelligence in Mining massive data sest for security.* Amsterdam, the Netherlands OIS Press 2008.