*e-Health – For Continuity of Care*
*C. Lovis et al. (Eds.)*


*doi:10.3233/978-1-61499-432-9-995*

# Electronic Processing of Informed Consents in a Global Pharmaceutical Company Environment

Dina VISHNYAKOVA[a,c,1] , Julien GOBEILL[b,c], Fatma OEZDEMIR-ZAECH[d], Olivier KREIM[d], Therese VACHON[d], Thierry CLADE[d], Xavier HAENNING[d], Dmitri MIKHAILOV[e] and Patrick RUCH[b,c]
[a] *Division of Medical Information Sciences, University Hospitals and University of Geneva, Switzerland*
[b] *BiTeM, HEG/University of Applied Sciences of Western Switzerland, Geneva Switzerland*
[c] *SIBTex, SIB Swiss Institute of Bioinformatics, Geneva Switzerland*
[d] *Novartis Institute for BioMedical Research, Basel, Switzerland*

**Abstract.** We present an electronic capture tool to process informed consents, which are mandatory recorded when running a clinical trial. This tool aims at the extraction of information expressing the duration of the consent given by the patient to authorize the exploitation of biomarker-related information collected during clinical trials. The system integrates a language detection module (LDM) to route a document into the appropriate information extraction module (IEM). The IEM is based on language-specific sets of linguistic rules for the identification of relevant textual facts. The achieved accuracy of both the LDM and IEM is 99%. The architecture of the system is described in detail.

**Keywords.** Informed consent, information retrieval, language detection

## Introduction

Almost all potential volunteers or simple patients enrolled in some biomedical study have been handed a stack of papers containing informed consent (IC) [1] to sign. Today, most of collected questionaries' forms are paper-based, as it is the most convenient instrument to persistently store such legal & medical information in the long run. The exponential growth of paper-based documents pledges the integration process of paper-based collections with digital documents for further processing. Furthermore, if the collected documents are issued in the framework of an international research it adds an additional difficulty of information sharing, due to the multilingual formats and the needed compliance with local or government regulations in each of the respective locations with data are issued.

We believe that the application of natural language processing (NLP) technologies is the best solution to provide an easy access to the information in digitalized document [1], [2]. NLP approaches to tackle information extraction tasks can be split between

[1] Corresponding Author.

data-driven and manually crafted models. In any case, a predefined domain-oriented text corpus is required to tune and/or assess these models. Therefore to perform our experiments, we rely on simple and robust pattern matching. Such methods are cost effective when no data is available to train a machine-learning model [3] .

An additional source of complexity is introduced by the Optical Character Recognition (OCR) needed to transform the source-digitalized document into machine-readable contents. Thus the collection must be regarded as a relatively noisy content. Again micro-grammars have been shown to perform well on noisy documents such as clinical records [4], [5] and OCRized corpora [6]. Such a dimension is key in our experiments, as it has been shown that the quality of information extraction is directly related to the quality of an OCRized document [7]. Similarly, retrieval effectiveness of search engines is also directly dependent on the ability to handle corruption and to restore quality in the indexed collection [8]. Finally, in a case where the collected document set contains documents in different languages, a language detection pre-processing step is mandatory.

To meet the above-mentioned challenges such as multilingual environment, variations of quality of the OCRized documents, we developed a robust system based on a language detection module to detect document's language and on name entity recognition (NER) method.

## 1. Data and Methods

In this section, we describe the data we needed to process as well the original pipeline we developed to perform the language-specific information extraction tasks.

### *1.1. Data overview*

A collection of documents' sets (146 documents), see Table 1, which contain occurrences of information expressing the duration of a consent in 10 languages, was used in our experiments. The collection was acquired using a random sample strategy. The resulting set is however very heterogeneous. For example informed consents in Chinese varied depending on the country, thus the Taiwanese documents were in traditional Chinese while documents generated to recruit patients from China were in modern Chinese. Within a given language, various types of forms were provided. Thus, only for the English language: different forms are used in the UK or the US. Further, within the body of the forms, we find legal statements likely to supersede other passages (e.g. specific regulations in particular States in the United States).

The quality check of the supplied corpora revealed some semantic interoperability inconsistencies. For instance, some documents had no information about the time duration, see Table 1. Difficulties were also introduced by the OCR module, which resulted sometimes in error in metadata needed to describe the encoding of the document (e.g. a Russian document marked up with a Latin encoding format).

### *1.2. Architecture*

The overall architecture of the tool with the following components is shown in Figure 1:

- FilesConverter is the module which converts files either of .doc or .rtf extension to a plain text;
- LanguageDetector is the module responsible for the detection of language in the given text. This module is based on a Language Detection (LD) API [9]. The achieved accuracy[2] is the main advantage of LD over its competitors such as Apache Tika and Compact Language Detector. Furthermore, LD has a relatively wide linguistic coverage with more than 50 languages;
- SentenceSplitter splits the text into sentences. By operating at the sentence level, entity mention extraction tasks will be faster and more focused. It should be noticed that traditional Chinese sentences have a specific punctuation where dot is represented by the mark "。", see Figure 2;
- RulesModule is responsible for providing list of rules for the detected language and applying them to text;
- TextTagger tags detected patterns in sentences, see Figure 2.

From a functional point of view the application we developed receives a bulk of documents in .doc or .rtf format and output a .csv file containing the filename, the sentence containing the time duration, the language assigned to the document and a status/feedback flag to record a possible error during the annotation process.

### *1.3. Language Detection*

The language detector is especially important because some documents may contain texts written in more than one language. Thus, some non-English documents, sometimes, contained section in two languages: English and non-English. In these cases, we experimentally defined a threshold: when a candidate language at rank 1 receives a score very similar to a candidate language at rank 2, and if language at rank 1 is English, then we force the system to assign the language proposed in rank 2.

### *1.4. Information Extraction Rules*

A total of 198 regular expressions, to detect time duration, were written, see Table 1.

Some languages such as French or Polish required the duplication of regular expressions due to misspellings introduced by earlier steps (e.g. at Document Authoring or OCRization). Some specific characters, such as accents and cedillas were substituted with the same character without the diacritics (e.g. *ç->c, é->e*).

Additionally, due to some overlapping between Traditional and Modern Chinese rules, the Chinese set includes both of them. The construction of rules is based on the key-words observation and as well as on local culture of using numbers as words or digits in particular languages. The rules combine a set of trigger words (e.g. up to, during, years) and specific micro-grammars such as the recognition of digit sequences (e.g. 12, 15) as proposed to identify named entities [10], [11].

[2] http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html .

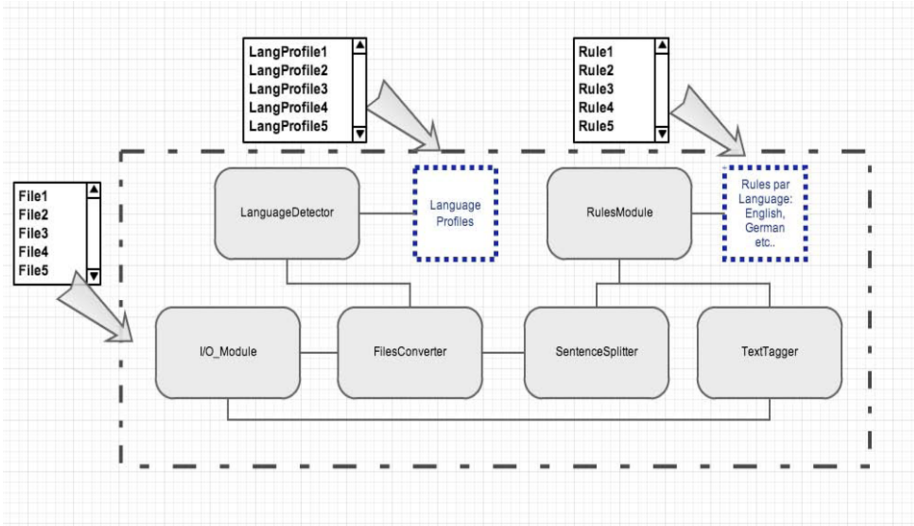
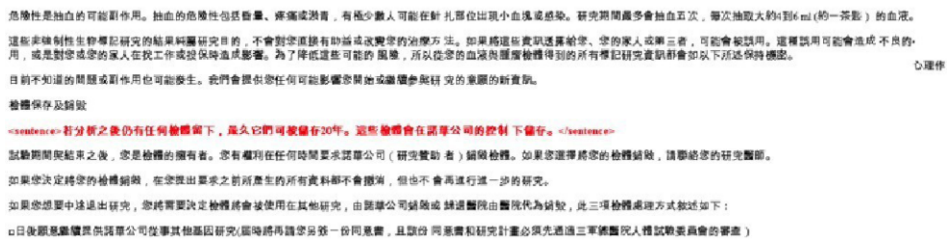



**Figure 1.** Components of the electronic capture tool of informed consents service. The dotted blue line represents dynamic modules, which can be modified without recompilation of the whole service. This adds an appropriate flexibility to maintain the application.

**Table 1.** Distribution of rules per language

| Language | # of rules | # of provided documents / # documents with no consent |
|---|---|---|
| English | 3 | 25/2 |
| German | 39 | 18 |
| French | 59 | 23 |
| Russian | 28 | 8/2 |
| Chinese (Traditional+Modern) | 29 | 6/2 |
| Spanish | 5 | 9 |
| Portuguese | 8 | 11/10 |
| Polish | 13 | 15 |
| Greek | 15 | 10 |
| Swedish | 2 | 8 |


危險性是抽血的可能副作用。抽血的危險性包括血量、疼痛或淤青，有極少數人可能在針孔部位出現小血塊或感染。研究期間最多會抽血五次，每次抽取大約4到6 ml (約一茶匙) 的血液。

這樣來做制性生物標記研究的結果純屬研究目的，不會對您直接有助益或改變您的治療方法。如果將這些資訊選擇給您、您的家人或第三者，可能會被誤用。這種誤用可能會造成不良的影響。為了降低這些可能的風險，所以從您的血液與腫瘤檢體得到的所有標記研究資訊都會如以下所述保時機密。

目前不知道的問題或副作用也可能發生。我們會提供您任何可能影響您開始或繼續參與研究的意願的新資訊。

檢體保存及銷毀

<sentence>若分析之後仍有任何檢體留下，最久它們可被儲存20年。這些檢體會在諾華公司的控制下儲存。</sentence>

試驗期間與結束之後，您是檢體的擁有者。您有權利在任何時間要求諾華公司（研究贊助者）銷毀檢體。如果您選擇將您的檢體銷毀，請聯絡您的研究醫師。

如果您決定將您的檢體銷毀，在您提出要求之前所產生的所有資料都不會銷毀，但也不會再進行進一步的研究。

如果您想要中途退出研究，您將需要決定檢體將會被使用在其他研究，由諾華公司銷毀或歸還醫院由醫院代為銷毀，此三項檢體處理方式敘述如下：

□日後願意繼續提供諾華公司從事其他基因研究(屆時將再請您另簽一份同意書，且該份同意書和研究計畫必須先通過三軍總醫院人體試驗委員會的審查)

**Figure 2.** An output fragment of the web-based interactive version of the service (Chinese).

## 2. Results and Discussions

The selected language-detection achieved an average precision of 99% for the language classification task. The high accuracy of 98% achieved on the full dataset by information extraction module showed that the authored rules set have excellent effectiveness. Table 1 shows that the most modest and standardized language in documents is English. The relatively high number of rules needed to process French documents compare to others can be explained by high frequency of diacritics along with numbers written in words (*un, deux, trois* etc.). At the same time Russian and German had more variations of key words compared to English. Chinese rules required deeper linguistic investigations due to the analytical nature [12] of the language.

The processing time for a document depends on several parameters such as: text encoding, number of languages in the text and, finally, number of rules defined for the given language. On average, a document is fully processed, from OCRization to language recognition and information extraction, in less than 5 seconds.

As a conclusion, the proposed system successfully demonstrated its power. The system is indeed able to process noisy textual contents in order to identify consents' durations in a highly multilingual environment, which go beyond traditional work with European languages such as French or German, and expand to Russian as well as Asian language with the very same architecture. Today, the system has been successfully deployed at Novartis Institute for Biomedical Research as a bulk-processing tool to support database curation of clinical trial contents.

## References

[1] E.J Emanuel, D. Wendler, C. Grady, What makes clinical research ethical?, *JAMA: the journal of the American Medical Association* ***283*** (2000), 2701-11.

[2] P. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, S. Huff, editors, A natural language understanding system combining syntactic and semantic techniques, *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1994).

[3] A.-L. Veuthey, A. Bridge, J. Gobeill, P. Ruch, J.R. McEntyre, L. Bougueleret, et al., Application of text-mining for updating protein post-translational modification annotation in UniProtKB, *BMC bioinformatics* ***14*** (2013), 104.

[4] P. Ruch, R. Baud, A. Geissbühler, Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records, *International Journal of Medical Informatics* ***67*** (2002), 75-83.

[5] P. Ruch, R. Baud, A. Geissbühler, Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, *Artificial intelligence in medicine* **29** (2003), 169-84.

[6] R. Ananthanarayanan, V. Chenthamarakshan, P. M. Deshpande, R. Krishnapuram, editors, Rule based synonyms for entity extraction from noisy text, *Proceedings of the second workshop on Analytics for noisy unstructured text data* (2008).

[7] R. Pereda, Information Extraction in an Optical Character Recognition Context, Diss. University of Nevada, Las Vegas 2011.

[8] P. Ruch, J. Gobeill, C. Lovis, A. Geissbühler, Automatic medical encoding with SNOMED categories. *BMC medical informatics and decision making*, **8**(Suppl 1), (2008), S6.

[9] C. Flanagan, S. N. Freund, editors, Type-based race detection for Java. *ACM SIGPLAN Notices* (2000).

[10] P. Ruch, J. Wagner, P. Bouillon, R. H. Baud, A.-M. Rassinoux, J.-R. Scherrer, editors, MEDTAG: tag-like semantics for medical document indexing, *Proceedings of the AMIA Symposium*; (1999).

[11] P. Ruch, R. H. Baud, A.-M. Rassinoux, P. Bouillon, Robert G, editors, Medical document anonymization with a semantic lexicon, *Proceedings of the AMIA Symposium* (2000).

[12] C. N. Li, S. A. Thompson, *Mandarin Chinese: A functional reference grammar,* University of California Pr., 1989.