*Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*
*A. Ugon et al. (Eds.)*


*doi:10.3233/978-1-61499-852-5-715*

# Automatic Processing of Anatomic Pathology Reports in the Italian Language to Enhance the Reuse of Clinical Data

Natalia VIANI[a,1], Lorenzo CHIUDINELLI[a,b], Cristina TASCA[b], Alberto ZAMBELLI[b], Mauro BUCALO[c], Arianna GHIRARDI[b], Nicola BARBARINI[c], Eleonora SFREDDO[b], Lucia SACCHI[a], Carlo TONDINI[b] and Riccardo BELLAZZI[a]
[a] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy*
[b] *ASST Papa Giovanni XXIII Hospital, Bergamo, Italy*
[c] *BIOMERIS, Pavia, Italy*

**Abstract.** Medical reports often contain a lot of relevant information in the form of free text. To reuse these unstructured texts for biomedical research, it is important to extract structured data from them. In this work, we adapted a previously developed information extraction system to the oncology domain, to process a set of anatomic pathology reports in the Italian language. The information extraction system relies on a domain ontology, which was adapted and refined in an iterative way. The final output was evaluated by a domain expert, with promising results.

**Keywords.** information extraction, text mining

## 1. Introduction

Textual medical reports include a great amount of valuable information that can be exploited for research purposes. To enable the reuse of such information, it is important to convert the available unstructured texts into structured data to be queried and examined in an automatic way. In the Papa Giovanni XXIII hospital in Bergamo (Italy), there are ongoing efforts to implement an i2b2 platform [1] for research in the oncology field. The objective is to create a repository that integrates all the data available for more than 23,000 cancer patients, and make them available for researches to answer a variety of questions. Since a lot relevant data are currently stored in the form of free text, developing automatic information extraction (IE) techniques is fundamental. Although many systems have been developed to process clinical narratives written in English, the related research for other languages, such as Italian, is still limited [2].

In previous work, we developed a pipeline that processes medical reports in the Italian language to identify mentions of events (e.g., diagnostic procedures) and their related attributes (e.g., test results) [3]. The pipeline, which was designed on a set of molecular cardiology reports, exploits a domain ontology to define the events and the attributes to be extracted from the text. This feature facilitates the extension of the IE

[1] Natalia Viani, Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100 Pavia, Italy; E-mail: natalia.viani01@universitadipavia.it.

system to a variety of different domains, since only an ontology modification would be needed to adapt the pipeline. In this work, we showed how the developed pipeline could be tailored to the oncology domain, with a particular focus on the field of breast cancer.

## 2. Materials and Methods

### *2.1. Dataset and Reports Structure*

The corpus considered in this paper was provided by the Papa Giovanni XXIII hospital, and consists of 221 anatomic pathology reports belonging to patients with breast cancer. Pathologists generate the reports using an electronic form, which includes a set of predefined sections to be filled in. Figure 1 shows an example of a complete report, with four sections: "clinical information", including the references to previous tests, "sent specimen", which lists all the analyzed specimens, "specimen description", containing details about these specimens, and "diagnosis", which reports the diagnostic conclusions.

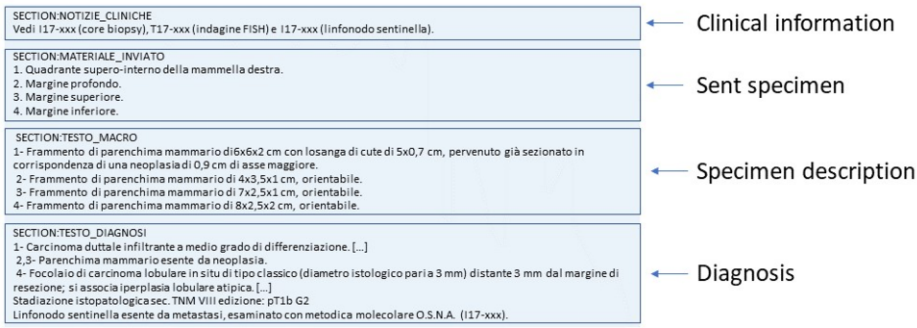


**Figure 1.** Example of an anatomic pathology report.

It is important to point out that, whenever multiple items are mentioned in the "Sent specimen" section, each of them is identified with a different number (*specimen number*). These numbers are used in the other sections to keep track of the specific item that is being referred to. As another important remark, each report might include different diagnoses, each related to one specific specimen. For example, in the report shown in Figure 1, an invasive ductal carcinoma was found in the first analyzed specimen (a breast quadrant), while the second specimen (a margin) did not show any sign of neoplasia.

### *2.2. Information Extraction Task*

In this work, an ontology-driven IE pipeline for the extraction of events and their attributes was exploited. In this IE pipeline, the events of interest are extracted by performing a search on external dictionaries. Then, for each identified event, the related attributes are searched for by exploiting the relations defined in the ontology, which is manually created. This ontology is structured in Event and Attribute classes, each related to a regular expression which allows searching for concept mentions inside the text.

To adapt the ontology to the oncology domain, it was first necessary to formalize the IE problem, defining the information to be extracted from the texts. To this end, a set of 20 reports was randomly selected to be manually reviewed and discussed with

physicians (*set for ontology design*), thus selecting the relevant concepts to be included in the ontology. Moreover, to facilitate the identification of concepts' variants, the n-grams (i.e., sequences of n words) that are most frequent in the considered dataset were extracted. As the result of these analyses, the following relevant entities were identified:

- **Specimen**. Anatomic pathology reports describe pathologic findings on one or more specimens, such as core biopsies or organ portions.
- **Diagnosis**. Each document contains relevant diagnostic conclusions (even in a negated form).
- **Histopathological stage**. In the case a breast cancer is found, the report often includes its histopathological stage.
- **Prognostic factor**. Reports often include an assessment of a few prognostic factors, such as the expression of estrogen and progesterone receptors.

To reuse the event-attribute ontology structure to analyze the anatomic pathology reports, the four identified entities were modeled as ontology events, and for each of them a set of attributes of interest was identified. For example, analyzed specimens can be characterized by their size (e.g., "6x6x2 cm"), while prognostic factors can be linked to a test result (positive or negative). Moreover, both specimens and diagnoses can be related to a specimen number.

In the automatic processing of anatomic pathology reports, it is important to keep the relation between each extracted diagnosis and the specimen it refers to. As both diagnoses and specimens were represented as ontology Events, an extension of the ontology was required to allow creating diagnosis-specimen links. In particular, relations between pair of events were added, too. In the proposed IE process, the link between each diagnosis and its related specimen is derived in two different ways. First, a specimen mention is searched for in the same sentence containing the diagnosis. Second, for those diagnoses that are linked to a specific specimen number within the text (e.g., "*1- invasive ductal carcinoma*"), this number is used to retrieve the associated specimens.

## 3. Results

### *3.1. Ontology Development*

To develop and refine the ontology and the annotation process, an iterative approach was followed. The first version of the ontology (*version 1*) was manually built on the *set for ontology design*, considering the information written in reports and the available domain knowledge. Then, both the ontology and the IE system were iteratively refined, evaluating the performance through several discussions with the domain expert. This process led to the creation of a *version 2*, which we evaluated on an independent test set.

The final ontology contains 44 events and 16 attributes. Figure 2 shows the complete class structure, implemented in the Protégé framework [4]. Both events and attributes are arranged into four main classes: Specimen, Diagnosis, Histopathological Stage, and Prognostic Factors. Specimens are grouped into biopsies and surgical resections, which are divided into organs (e.g., left breast) or organ portions (e.g., left nipple). As an interesting characteristic, all specimens are specific to an organ, and some specimens can be linked to an organ portion or a nodule, too. Therefore, *Localized Organs*, *Organ Portions*, and *Nodules* can represent both events and attributes.

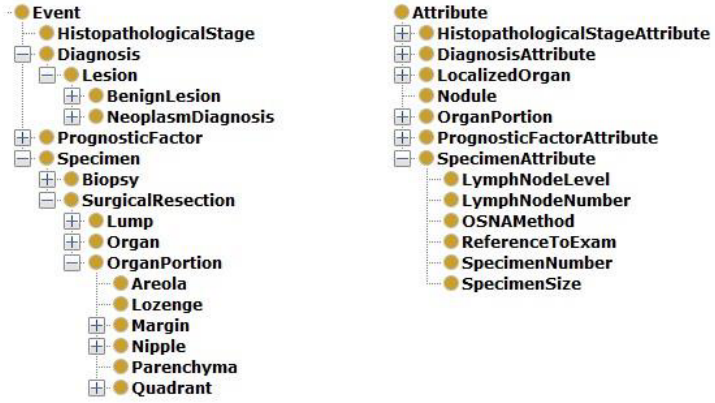


**Figure 2.** Domain ontology class structure: events (left) and attributes (right).

### *3.2. Validation with Expert*

To evaluate the performance of the pipeline, *version 2* was run on a test set made up of 34 documents. To give a sense of the task complexity, we computed the number of items that were automatically extracted from each report (*system items*): a total of 476 system items were identified, corresponding to an average of 14 items per report.

To enable the evaluation of the IE system, the information extracted from each report was written on an output file, including both the original report and the system items. This output was manually reviewed by a domain expert, who was trained to identify three types of errors:

- *Missing items*, i.e., relevant information that was not considered and thus not included in the ontology.
- *False negatives (FN)*, i.e., information that should have been extracted but was not found in the system's output.
- *False positives (FP)*, i.e., errors found in the system's output, such as incorrect specimen-number associations or attributes linked to the wrong event.

In Table 1, the total number of missing items, false negatives, and false positives is shown ("raw count" column). These three groups were further analyzed by removing duplicates or similar entries ("distinct count" column); for example, the string "c-erbB-2" was marked as a missing item in several reports, but it was counted only once in the "distinct count" column. As it can be noticed from the table, most errors were due to items that are currently not searched for (38 distinct items). As regards false negatives and false positives, which are instead a more direct measure of the performance of the IE system itself, the raw counts were 15 and 26, respectively.

**Table 1.** Evaluation results: error types (test set).

| Items | Raw count | Distinct count |
|---|---|---|
| Missing items | 57 | 38 |
| FN | 15 | 11 |
| FP | 26 | 21 |

Starting from the identified error types, it was possible to compute the precision (P), the recall (R), and the F1 score (F1) of the IE system:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

In these formulas, true positives (TP) were computed by substracting FPs from the total number of system items. The finally computed values were 94.5%, for precision, 96.8%, for recall, and 95.6% for the F1 score.

## 4. Discussion and Conclusions

In this work, we adapted an ontology-driven IE approach (originally developed for molecular cardiology reports) to analyze a set of anatomic pathology reports on breast cancer. Despite the differences between the cardiology and the oncology domains, the proposed ontology structure was reused without major modifications, exploiting the event-attribute framework to model the relevant entities to be extracted.

The major adaptation that was performed was the inclusion of a new IE task, which is the extraction of Event-Event relations. Although the proposed approach for linking diagnoses and specimens does not allow reconstructing relations that are reported in a complex way, it performs well when the relation to be extracted is clearly stated within the text (for example with the specimen number or name).

The developed IE system was manually evaluated by a domain expert, with promising results. In particular, most relevant items were extracted in the correct way, leading to an F1 score of 95.6%. Moreover, the evaluation allowed identifying 38 relevant items that were not previously considered in the existing ontology. In a future version of the system, these items will be added, and the ontology will be enriched to enable the processing of reports related to multiple cancer types. As a final step, the extracted information will be integrated into the i2b2 data warehouse of the Papa Giovanni XXIII hospital. According to the positive results of the conducted validation, the IE system could be effectively used to help retrieve useful information for research.

Future developments will address the limitations of the proposed IE approach. First, the extraction of specimen-number links was not trivial: specimen sizes were often mistaken for specimen numbers, leading to the construction of an incorrect link. To address this issue, future work will focus on how to disambiguate these items. Another limitation regards the small size of the test set. In future work, all the available anatomic pathology reports will be processed and evaluated. As a matter of fact, the domain expert is currently validating more documents, which will allow gathering further suggestions to improve the system. Finally, while in this work the most frequent n-grams were manually analyzed to support the ontology creation, in the future it would be interesting to automatically propose the ontology structure starting from the extracted n-grams.

## References

[1] I.S. Kohane, S.E. Churchill, S.N. Murphy, A translational engine at the national scale: informatics for integrating biology and the bedside, *J Am Med Inform Assoc* **19** (2012), 181–5.

[2] S. Velupillai, D. Mowery, B.R. South, M. Kvist, H. Dalianis, Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis, *Yearb Med Inform* **10** (2015), 183–93.

[3] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, L. Sacchi. Information Extraction from Italian Medical Reports: An Ontology-driven Approach. *Int J Med Inf* **111** (2018), 140-8.

[4] Protégé. Available at: http://protege.stanford.edu/ (accessed 7 Jan 2017).