Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus
Olivier Steichen, Christel Daniel-Le Bozec, Maxime Thieu, Eric Zapletal, Marie-Christine Jaulent

To cite this version:

HAL Id: hal-01171766
http://hal.upmc.fr/hal-01171766
Submitted on 6 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus

Olivier Steichen\textsuperscript{a}, Christel Daniel - Le Bozec\textsuperscript{ab} MD PhD, Maxime Thieu\textsuperscript{a}, Eric Zapletal\textsuperscript{ab}, Marie-Christine Jaulent\textsuperscript{a} PhD

\textit{a} INSERM ERM 202, SPIM laboratory, Paris.
\textit{b} Medical Informatics Department, Georges Pompidou Hospital, Paris
Abstract

Computer-assisted consensus in medical imaging involves automatic comparison of morphological abnormalities observed by physicians in images. We built an ontology of morphological abnormalities in breast pathology to assist inter-observer consensus. Concepts of morphological abnormalities extracted from published grading systems, medical reports and existing terminologies were organized in an is-a hierarchy and linked by the relation “is a diagnostic criterion of” according to diagnostic meaning. We implemented position-based, content-based and mixed semantic similarity measures between concepts in this ontology and compared the results with experts’ judgment. The position-based similarity measure using both taxonomic and non-taxonomic relations performed as well as the other measures and was used for automatic comparison of morphological abnormalities within the IDEM computer-assisted consensus platform.

Keywords
multimedia knowledge representation; ontology; semantic similarity; similarity measures; computer-assisted consensus; breast pathology
I. Introduction

Pathologists describe specimens to provide relevant diagnostic and prognostic information to guide patient care. Many studies have shown that intra- and inter-pathologist variability is large and therefore question the accuracy of the reports [1;2]. Although diagnostic and prognostic criteria have been defined and validated for many fields in pathology, differences in training and experience between pathologists can cause residual variability. Consensus between pathologists is a cooperative means to reduce this variability [3]. However, consensus sessions are not easy to organize since pathologists often have a heavy workload and are unavailable to participate in the sessions [4]. The aim of the IDEM (Images and Diagnosis from Evidence in Medicine) project is to assist a panel of experts in the cooperative creation of consensual multimedia structured reports in pathology [5]. During an IDEM computer-assisted consensus session over the Internet, experts can 1) edit virtual slides of the case, 2) describe morphological abnormalities observed in virtual slides and link these to the corresponding regions of interest and 3) debate disagreements about descriptions to construct a consensual description of the case. We have already shown that this last point is drastically simplified if based on the automatic detection of concordance between the experts’ descriptions [5].

Providing automatic comparison of different descriptions raises several knowledge engineering issues:

- To formalize the domain knowledge required for image interpretation and the diagnostic process in pathology
- To define semantic similarity measures between morphological abnormalities to compare several descriptions of the same case
- To validate both the knowledge model and the similarity measures to enable computer-assisted inter-observer consensus.

To address these issues, our objective was to develop a computer environment that allowed experts to build an ontology of morphological abnormalities, to compute similarity scores between the concepts of the ontology, and to evaluate the usefulness of these scores for inter-observer consensus.

The paper is organized as follows. Section II presents the issues of knowledge representation and ontology building in pathology. Section III examines the different types of semantic similarity measures between concepts. Section IV describes the method used to build an ontology in breast pathology and to compute three different similarity measures within this ontology thanks to new functionalities added to an ontology editing tool. An evaluation study was conducted to compare the three similarity measures to human judgment in order to choose one of them to perform semantic similarity between concepts within the IDEM internet platform. Results are presented in section V, and section VI concludes with discussion and future work.

II. Building an ontology in pathology

II.1. Methods and tools to build an ontology

Several publications about building an ontology have been published. Uschold and Grüninger [6] proposed a method with three successive phases:

1. Identifying the purpose and the scope of the ontology
2. Building the ontology
   a. Ontology capture involves acquiring the knowledge in accordance with the purpose and scope that has been identified. From the knowledge engineering perspective, two main approaches are used for this knowledge acquisition process. The aim of a top-down approach (model-driven) is instantiation of preexisting models of the general task and the domain, and a bottom-up approach (data-driven), leads to the design of a conceptual model of the knowledge used in the domain based on available data [7].
   b. Ontology coding and integration of existing ontologies consists of structuring the domain knowledge in a formalized conceptual model.

3. Verifying and validating the resulting ontology.
   Many ontology editing tools have been developed but differ in user friendliness, knowledge representation and updating performance [8;9]. Protégé® is an open source software and can be customized to enable computer-based applications to make use of terminological inferences such as conceptual distance [10].

II.2. Knowledge sources available for ontology capture in pathology

This section presents the sources of knowledge in pathology and the main available medical terminologies including ontologies in this field, focusing on the necessary and sufficient knowledge to describe the morphological abnormalities observed in virtual slides.

II.2.a Top-down approach

A top-down approach relies on models of the domain formalized by experts. These models are available as published guidelines. Best practice or “evidence-based pathology” is the ability to recognize morphological abnormalities in images (“from observation to features”) and to interpret these features according to published guidelines and consensus among recognized experts (“from features to diagnosis”) [11]. Current published grading and scoring systems provide pathologists with rules to make decisions “from features to diagnosis” [1]. For instance, Table 1 shows the rules to reach ductal carcinoma in situ (DCIS) of high, intermediate and low grade respectively from a list of morphological abnormalities. These rules are from the most consensual grading systems in this field [12;13].

Table 1

| Efforts to standardize pathology have also resulted in several terminological resources available in electronic form. SNOMED or ADICAP in France are already used in pathology laboratories for coding diagnoses [14-16] [17]. The three main ontological resources including pathological concepts are the GALEN ontology, GeneOntology and SNOMED CT [18-20]. There are many studies on the modeling of anatomy in the literature [21-23]. In SNOMED CT, concepts that describe normal anatomy are in the subcategory “anatomical concepts” of “body structure”. There are fewer studies on the modeling of morphological abnormalities. In SNOMED CT, concepts describing altered body structures are in a specific hierarchy, the “morphologically abnormal structure” or “morphological abnormalities” (MA) of “body structure”. In this hierarchy, MAs are classified by site (i.e. anatomical concepts), which may... |
be subcellular, cellular, ductal, or lobular. Current ontologies in pathology are limited because they do not provide links between morphological diagnoses and diagnostic criteria i.e the attributes precisely defining the morphology of these diagnoses. For example, there is no relation to express that elementary morphological abnormalities such as “comedo necrosis” and “high nuclear grade” are associated attributes (i.e. diagnostic criteria) of the morphological diagnosis “high grade ductal carcinoma in situ”.

II.2.b Bottom-up approach

Domain knowledge in pathology is available in textual reports of daily practice. Ontology capture using a bottom-up approach may rely on natural language processing (NLP) tools which provide experts with morphological, syntactical and statistical analysis of texts to extract concepts and ontological relations from structured and unstructured data [24;25]. For example, NLP tools based on the differential linguistic approach have been developed to define classes of concepts and relations using distributional analysis of the contexts of the terms in a corpus of text [26-28].

III. Similarity measures between concepts in semantic nets

Similarity measures between concepts are manifold and can be classified according to the information they use. Position-based similarity measures use the distance between two concepts in the net, which is defined as the number of links on the shortest path between the concepts [29;30]. However, as links do not always convey the same distance (proximity of meaning), they are often weighted according to type, hierarchical depth and local density [31-34].

Content-based similarity measures mostly rely on comparison of the properties of the concepts. As this requires a very rigorous and costly formalization of the concepts [35], the information content ($IC$) of the terms is commonly used instead [36]. For a given term $T$, $IC(T) = -\log[p(T)]$, where $p(T)$ stands for the probability of encountering an instance of $T$ in a given text corpus. For the computation, instances of $T$ itself, its synonyms, and all of its taxonomic descendants, including synonyms of the descendants, are pooled. For example, occurrence of “comedo-type necrosis” will also count as an occurrence of “necrosis”, assuming that "comedo-type necrosis" is subsumed by “necrosis” in the ontology. Here, the costly step is to record occurrences of terms in a corpus, and to compute the frequency of each one. Information content relies on the notion that the less frequently used terms are more informative because they are more discriminating; therefore their $IC$ is close to 1. Only the $IC$ of two taxonomically related terms can be compared. In this case, for example if $T_2$ is a taxonomic descendant of $T_1$, each instance of $T_2$ is also an instance of $T_1$, and the $IC$ difference quantifies the relative part of instances of $T_2$ among instances of $T_1$:

$$IC(T_2) - IC(T_1) = -\log[p(T_2)] + \log[p(T_1)] = -\log[p(T_2)/p(T_1)]$$

Mixed measures are computed as position-based, but the links are weighted according to the $IC$ difference between the concepts they link. The greater the $IC$ difference, the higher the weight (and therefore the lower the similarity) [37].

IV. Methods

We adapted the Uschold and Grüninger method [6] to our specific domain and purpose.
IV.1. Identifying purpose and scope

We built an ontology for breast pathology to enable computer-assisted inter-observer consensus in breast pathology. This ontology had to provide pathologists with both the standardized terms and the fundamental knowledge needed for case description and inter-observer consensus.

IV.1.a Case description and diagnosis

According to evidence-based pathology, only clinically reproducible and relevant features with a demonstrated diagnostic or prognostic significance features should be reported in descriptions [11]. Table 1 provides an example of the list of morphological abnormalities that are relevant for grading ductal carcinoma in situ (DCIS).

A case description is a list of morphological diagnoses observed in images. Morphological diagnoses are complex morphological abnormalities (MAs) such as “low grade ductal carcinoma in situ” and are inferred from other features such as complex MAs e.g “low nuclear grade” or elementary MAs e.g “moderate nuclear enlargement”.

Case description in the IDEM platform followed this model. Figure 1 shows an example of a case description. Furthermore, in IDEM, morphological diagnoses are unambiguously located the images and if a morphological diagnosis is inferred from other morphological abnormalities, the latter are also associated with the same region of interest (ROI) in the image.

Figure 1

IV.1.b Inter-observer consensus

A consensus session began with the detection of concordance between experts’ descriptions. Experts agreed if they observed either similar morphological diagnoses or correlated features in the same region of interest (ROI). Two features (e.g. “comedo necrosis” and “high nuclear grade”) were correlated if they were diagnostic criteria of the same morphological diagnosis (e.g. “high grade DCIS”).

In the IDEM platform, a consensus session is an interactive phase, directed by a moderator, leading pathologists to a consensual description from a set of initial individual descriptions. According to the consensus model, an automatic phase of agreement evaluation between descriptions preceds this interactive phase. To compare two descriptions, the system automatically checks that experts observe the same morphological abnormalities (MAs) at the same place in the virtual slides. It performs a two-by-two comparison of the MAs, computing both geographic agreement between regions of interest (ROIs) and semantic agreement between labels. By convention, two ROIs are in geographic agreement if they overlap. Two labels are in semantic agreement if the semantic similarity measure between the corresponding concepts in the breast pathology ontology is above a set threshold. At the end of the two-by-two comparison of MAs, each pair is classified as either in whole agreement (same or similar labels, same or overlapping ROIs), in geographic agreement only (same or overlapping ROIs), in semantic agreement only (same or similar labels) or in disagreement. Once the consensus module has computed initial agreement between descriptions, the results are used as the starting point for the second phase of inter-observer consensus.
IV.2. Building the ontology

An ontology of breast pathology was built for IDEM using the aforementioned sources of knowledge (cf. II.2).

IV.2.a Ontology capture

Top-down approach from published evidence and existing ontologies

A top-down approach was used for the instantiation of the domain model according to the purpose and scope of the ontology (consensus on case description and diagnosis). The domain model was acquired from published evidence and existing ontologies. Experts selected consensual published grading and scoring systems in breast pathology [12;13]. According to these recommendations, they built standardized reports for 28 of the main breast pathology diagnoses. Useful morphological abnormalities were then extracted from these 28 prototypical descriptions. Morphological abnormalities were organized using the Protégé® ontology editing tool [10]. The Ontoviz plug-in was used for graphic representation.

- Taxonomic relations
In accordance with the SNOMED CT organization, the hierarchy of morphological abnormalities was divided by anatomical site. We distinguished subcellular, cellular, ductal and lobular abnormalities.

- Non-taxonomic relations
Complex morphological abnormalities such as “high grade DCIS” and “high nuclear grade” were linked to the morphological abnormalities from which they were inferred. The relation “is a diagnostic criterion of” (abbreviated as part of diag) was defined according to rules in published grading systems that distinguish “optional diagnostic criteria” from “mandatory diagnostic criteria”. For example, although “comedo-type necrosis” and “high nuclear grade” were very distant in the morphological abnormalities hierarchy (6 links between them) both were linked to the same morphological diagnosis (“high grade DCIS”) thanks to the relation “is a diagnostic criterion of” (see Figure 4).

Bottom up approach from reports

A bottom-up approach was used to enrich the ontology using a corpus of pathological reports. Domain terms were acquired from 710 reports written by 10 pathologists from 2 French pathology departments. Terminological extraction and distributional analysis were performed on a corpus of 280,000 words using the Syntex® and Upéry® software [Chodkiewicz] to obtain a list of all the terms with the number of occurrences of each term and the textual context.

IV.2.b Ontology coding and integration of existing terminologies and ontologies

The ontology was coded using the Protégé® ontology editor which uses frame-based knowledge representation [10]. In this representation, the values of the attributes of a concept are inherited to all its descendants. The attribute of the frequency of occurrence (IC value) cannot be inherited like standard attributes. For instance, the IC of “papillary architecture” (=3.41) cannot be inherited from the IC of “ductal architecture differentiation” (= 1.92). To overcome this, we defined a custom meta-class template to create new classes. Concepts were
considered to be instances of the meta-class satisfying the constraint IC inherited from the meta-class and were also considered to be regular classes with their own templates slots.

We used the Prompt plug-in of Protégé® to integrate existing ontologies because it provided users with a uniform framework to compare, align and merge ontologies [38]. The morphological abnormalities of breast pathology previously acquired using both a top-down approach (from published evidence) and a bottom-up approach (using automatic processing of textual reports) were enriched by those in SNOMED v3.0 which was imported in Protégé®.

IV.3. Verifying and validating the resulting ontology

As the ontology was built to assist description and consensus on diagnosis between experts, validation of the ontology involved implementing similarity measures between concepts and evaluating the results according to their effectiveness for inter-observer consensus assistance.

IV.3.a Similarity measures

We developed a Protégé® plug-in to implement Leacock’s position-based measure [30], Lin’s content-based measure [36] and an adjusted version of Jiang’s mixed-measure, which was not limited to the specialization/generalization relation [37].

Leacock’s position-based measure

For any two given terms \( T_1 \) and \( T_2 \), the distance between them according to Leacock is:

\[
\text{Dist}_{\text{Leacock}}(T_1, T_2) = N_{\text{max}}[T_1, \text{Anc}(T_1, T_2)] + N_{\text{max}}[T_2, \text{Anc}(T_1)]
\]

with \( \text{Anc}(T_1, T_2) \) being the common ancestor of \( T_1 \) and \( T_2 \) minimizing \( \text{Dist}_{\text{Leacock}}(T_1, T_2) \), then Leacock’s measure is:

\[
\text{Sim}_{\text{Leacock}}(T_1, T_2) = -\log \frac{1 + \text{Dist}_{\text{Leacock}}(T_1, T_2)}{d_{\text{max}}}
\]

where \( d_{\text{max}} \) is the maximal depth in the hierarchy.

Lin’s content-based measure

Lin’s measure is:

\[
\text{Sim}_{\text{Lin}}(T_1, T_2) = 2 \times \frac{\text{IC}(\text{Anc}_{\text{tax}}(T_1, T_2))}{\text{IC}(T_1) + \text{IC}(T_2)}
\]

where \( \text{Anc}_{\text{tax}}(T_1, T_2) \) is the lowest common taxonomic ancestor of \( T_1 \) and \( T_2 \). To compute information content, we counted the occurrences of the instances of each concept in the pathology reports using Syntex ® software.

Jiang’s mixed measure

Jiang’s measure is computed as follows:

\[
\text{Sim}_{\text{Jiang}}(T_1, T_2) = \frac{1}{1 + \text{Dist}_{\text{Jiang}}(T_1, T_2)} \quad \text{with} \quad \text{Dist}_{\text{Jiang}}(T_1, T_2) = \sum_{\text{link}(T_1, T_2) \in \text{Path}(T_1, T_2)} W_{\text{link}(T_1, T_2)} \]

where \( W \) stands for Weight, and \( \text{Path}(T_1, T_2) \) is the path between \( T_1 \) and \( T_2 \) minimizing \( \text{Dist}_{\text{Jiang}}(T_1, T_2) \). The original Jiang’s measure weights the links using IC differences, but this difference is only meaningful when one of the terms is a taxonomic descendant of the other. In other cases, the computation leads to incoherent results, even to negative distances. So, for taxonomic links we kept the original weighting:
\[ W[\text{link}_{\text{tax}}(T_i, T_j)] = \left( \beta + (1 - \beta) \frac{LD_{\text{tax}}}{LD_{\text{avg}}(T_j)} \right) \left( \frac{d(T_j) + 1}{d(T_i)} \right)^{\alpha} \Delta IC(T_i, T_j) \]

where \( LD_{\text{tax}}(T_j) \) is the local density of taxonomic links (the number of taxonomic links having \( T_j \) as child), \( LD_{\text{avg}} \) is the average thereof in the whole hierarchy, \( d(T_j) \) is the hierarchical depth of \( T_j \), \( \alpha \) and \( \beta \) are parameters (in our case set to 0.5 and 0.7 respectively, as in Jiang’s publication) and \( \Delta IC(T_i, T_j) \) is the information content difference between \( T_i \) and \( T_j \).

But for non-taxonomic links we had to adapt the weighting:

\[ W[\text{link}_{\text{type}}(T_i, T_j)] = k_{\text{type}} \left( \beta + (1 - \beta) \frac{LD_{\text{type}}}{LD_{\text{type}}(T_i)} \right) \left( \frac{d(T_j) + 1}{d(T_i)} \right)^{\alpha} \overline{IC(T_i)} \]

where \( \text{type} \) stands for the relation type (diagnostic participation in our case), \( k_{\text{type}} \) is a type specific weighting factor and \( \overline{IC(T_i)} \) is the average of the information content difference of all taxonomic links having \( T_i \) as child.

IV.3.b Evaluation of the similarity measures

Similarity scores compared to expert’s judgment

To choose the most appropriate similarity measure for computer-assisted inter-observer consensus, we compared the similarity scores with three experts’ similarity judgments on a random sample of 200 pairs of concepts. Each expert was asked to rate the similarity between concepts. Pairs of terms had to be rated as “similar” or “dissimilar”. Inter-experts variability in the similarity ratings was evaluated with kappa statistics. A gold standard was built considering a pair as “similar” when there was unanimous agreement among the experts and “dissimilar” otherwise. This gold standard was used to draw a receiver operating characteristics (ROC) curve for each similarity measure. All statistical analyses were performed using the Stata® software package.

Consensus support based on similarity measures

To make the semantic similarity measures available within the IDEM platform, we developed a Protégé® plug-in to export similarity tables computed within the breast pathology ontology. Therefore, the IDEM platform could rely on semantic similarity scores to compute the initial agreement between descriptions (first step). We designed specific interfaces to assist inter-observer consensus based on this initial agreement between descriptions (second step of the consensus process).

V. Results

V.1. Domain ontology

The breast pathology ontology comprised 209 morphological abnormalities which were connected by 176 taxonomic links and 66 “is a diagnostic criterion of” links.
V.2. Similarity measures implementation

Computation of the three chosen similarity measures was implemented as a plug-in for Protégé®. For each one, the corresponding similarity table of all the pairs of concepts was launched and browsed through (Figure 3).

The OntoViz plug-in for Protégé® provided a graphical representation of the ontology which was extended to highlight all concepts similar to a given one in a specific color. For each pair of concepts, the computation path of the similarity score was visible (Figure 4).

V.3. Evaluation of the similarity measures

Similarity scores compared to expert’s judgment

The ROC curves showed the accuracy of the three measures compared with the experts' unanimity, with the area under the curve ranging from 0.76 [0.64-0.87] for Leacock to 0.81 [0.70-0.91] for Lin, without statistical difference between the two measures (Figure 5).

Consensus support based on similarity measures

The similarity measures, available within the IDEM platform, are used in the first step of the consensus derivation process. Once the pathologists of the panel have described all cases, the moderator displays the computed results of the initial agreement between descriptions. Then, he selects the cases to be consensually described during the interactive phase.

For each selected case, a list of clusters of MAs is displayed (Box 1a in figure 6). A cluster is either a unique MA described by only one pathologist or a set of different MAs described by several pathologists and found in agreement by the system (same or overlapping ROIs and same or similar labels). For a given cluster, MAs in partial agreement with at least one of the MAs of the cluster are displayed (Boxes 3 and 4 in Figure 6).

To reach a consensual description of a case, the moderator selects each cluster of MAs and the panel of pathologists either validates or rejects the computed agreement. For each selected cluster, the moderator selects the most appropriate label and the most illustrative ROI. The pathologists can see 1) the most relevant label and ROI according to the moderator and 2) all the MAs (labels and ROI) found by IDEM in whole or partial agreement with the selected cluster (figure 7). For each MA, each pathologist states 1) whether the label is the same, different or more appropriate than the label selected by the moderator and 2) whether
the corresponding ROI is the same, different or more appropriate than the ROI selected by the moderator.

*Figure 7*

Then, the moderator analyzes the results of the procedure and selects the label and ROI which is agreed to characterize the current cluster. The set of the consensual MAs of the case, unambiguously localized in the images, constitutes the consensual case description.

## VI. Discussion

Within the context of the IDEM project, we used a knowledge engineering approach to assist consensus between experts in breast pathology. The discussion is organized around the steps of the methodology we followed.

### VI.1. Defining the purpose and scope of the ontology

Two experts defined a model of case description and diagnosis in accordance with published evidence in breast pathology. This model focused on clinically useful and scientifically validated description terms and therefore fitted the paradigm of evidence-based pathology [1;11]. According to the consensus task model, to build a consensual case description, pathologists must ensure that they observe the same relevant morphological abnormalities at the same place in the images of the case. To support the consensus, the system had to recognize both similar morphological abnormalities and also correlated morphological abnormalities which were diagnostic criteria for the same morphological diagnosis. Therefore, the ontology on which the system relied had to provide both the standardized terms for breast morphological abnormalities, i.e concepts describing altered body structures and also the diagnostic knowledge in breast pathology, i.e the morphological relationships between these abnormalities. It is not unusual to embed expert knowledge in medical ontologies, often considered as knowledge bases codifying biomedical concepts [39]. For example, the Foundational Model of Anatomy provides knowledge representation of canonical macroscopy based on partitive and other spatial relationships among anatomical entities [21]. In SNOMED CT, although some semantic relationships express medical expert knowledge such as, for example “due to”, “has associated morphology”, “has interpretation”, “has definitional manifestation”, none of them is used to associate morphological diagnoses and the diagnostic criteria that precisely define these diagnoses. For example, there is no relation to express that elementary morphological abnormalities such as “comedo necrosis” and “high nuclear grade” are associated attributes (i.e. diagnostic criteria) of a given morphological diagnosis such as high grade ductal carcinoma in situ”. As no pre-existing ontology contains this information, we had to build a specific ontology with a specific non-taxonomic relationship (“is a diagnostic criterion of”) to compare descriptions of pathologic cases.

Moreover, similarity measures between morphological abnormalities rely on both taxonomic and non-taxonomic relationships that are relevant in the specific context of the consensus application. Semantic similarity usually relies on taxonomic relations. Semantic relatedness is a more general relation because dissimilar entities may be semantically related by non-taxonomic lexical relations such as antonymy and meronymy [40]. Computational applications often require semantic relatedness rather than strict similarity. In our context, although taxonomic relations were convenient to recognize similar morphological abnormalities, they did not provide semantic relatedness between two correlated morphological abnormalities. Experts defined an appropriate kind of semantic relatedness that
relied on the relationship “is a diagnostic criterion of” to compute useful similarity scores to assist consensus.

VI.2. Building the ontology

We chose to acquire the morphological abnormalities mainly from validated scoring or grading systems that synthesize the efforts of experts to express the most relevant pathology knowledge (top-down approach). These published guidelines include templates and checklists of morphological abnormalities with explicit rules to infer morphological abnormalities from features observed in images. This top-down approach was strengthened by the verification that no obvious knowledge in SNOMED was lacking in the resulting ontology. This verification relied on SNOMED merging with the Prompt plug-in of Protégé®.

Bottom-up approaches consist in acquiring expert knowledge from data in textual pathology reports. Natural language processing has already contributed extensively to the construction of ontologies but several studies have shown the need to pay great attention to the choice of corpus [26;27]. In this work, using a purely bottom-up approach to extract the description terms from textual reports remained insufficient because reports in daily pathology practice were not always consistent and complete. Although pathologists have initiated efforts to establish guidelines for reports on the most commonly occurring human lesions [41], objective morphological abnormalities on which the interpretation is based is not always included in the reports; this is contrary to the demands of evidence-based pathology [11].

The ontology edition with Protégé® has many advantages. Graphical tools allowed us to manipulate the ontology in a user-friendly way. Because of its extensible user interface, it was adapted to the semantic similarity computation and validation that required specific interaction.

VI.3. Computing similarity measures

Available comparisons of similarity measures in semantic networks are equivocal even if they tend to favor content-based measures [40;42]. The results seem to depend on the implemented semantic net, the chosen gold standard and the intended use. Therefore, because these evaluation contexts differ from our own, we chose to compare one measure of each type – position-based, content-based and mixed - with a relevant gold-standard. Three previously published measures of semantic similarity were implemented. Lin’s and Jiang’s measures need information content computation, which requires an extremely high human investment. The results of information content computation are theoretically corpus dependent. However, varied and broad corpuses are likely to produce equivalent results. For example, Patwardhan showed that there was no significant difference between similarities evaluated on the same net but computing information content with two different corpuses [42]. Most of the general language corpuses are 200,000-300,000 words, and there is no gain with larger corpuses. Our highly specialized and redundant corpus comprised 280,000 words, so it is unlikely that the study was limited in this respect.

Until recently, Jiang’s measure was mostly implemented in purely taxonomic hierarchies [37;40;43;44], but our study showed that it could be adapted for non-taxonomic links. Our adaptation respected the notion of information content in that the occurrence of a term could only count as an occurrence of another if the first term was a taxonomic descendant of the second [45]. This was not the case with other recent adaptations of Jiang’s measure that counted the occurrences of all descendants of taxonomic and non-taxonomic terms to compute information content [46].
VI.4. Evaluating similarity measures

The accuracy of the three measures compared with unanimous expert opinion shown in the ROC curves demonstrated the adequacy of the ontological organization of concepts. None of the three measures was better than the other. Leacock’s position-based measure using both taxonomic and non-taxonomic links performed as well as Lin’s content-based measure. In most other studies, position-based measures involving only taxonomic relations performed worse than content-based measures [40;42]. This suggests that non-taxonomic links improve the performance of position-based measures and make it possible to avoid the use of information content. Furthermore, as results in a semantic net cannot be generalized to others, this finding confirmed the need to evaluate the measures in specific application settings.

Our evaluation of the three similarity measures allowed us to choose Leacock’s position-based measure and to integrate the results into the IDEM platform. Leacock’s similarity table was used in the first step of the consensus derivation process to compute initial agreement between descriptions. Comparison with the experts’ judgment also helped us to define the threshold of semantic agreement. Two morphological abnormalities were considered to be in semantic agreement if the similarity measure between their labels was above the best cut-off point according to the ROC curve.

Nevertheless, the impact of the ontology and similarity measures on consensus derivation needs further investigation. Factors include the completeness of the ontology to report all possible situations in grading DCIS, the adequacy of the ontological organization of concepts, and the ease of preliminary automatic comparison of MAs in the global consensus derivation process.

VII. Conclusion and perspectives

Defining the desired scope and purpose of an ontology of morphological abnormalities for use in computer-assisted inter-observer consensus led us to build a new ontology in breast pathology. Reusing existing ontologies including morphological abnormalities without adaptation was impossible since they do not embed sufficient expert knowledge for the intended task. To build the ontology, a top-down approach was performed rather than a bottom-up approach based on natural-language processing of routine pathology reports because of the low quality of these reports.

Defining the relation “is a diagnostic criterion of” to formalize the semantic rules of published grading systems within the ontology allowed us to compute semantic similarities between morphological abnormalities which were well-matched with the experts’ judgment. The IDEM platform was developed to rely on this semantic similarity to assist pathologists in description and diagnosis consensus. A computer-assisted consensus study on 60 cases is currently underway to quantify the impact of computing semantic similarity between morphological abnormalities on the consensus process. More research is needed to define a methodology to use the feedback of consensus sessions to update the ontology of morphological abnormalities and to instantiate canonical pathology entities of the ontology with illustrative regions of interest. Building a Digital Pathological Atlas that formalizes interpretation terms and rules with acceptable reproducibility among experts remains a serious issue because this may assist researchers to characterize the natural history of lesions and evaluate treatment effectiveness.

VIII. Acknowledgments

This work was supported in part by grants from the French League Against Cancer, the French Research Ministry (RNTS 2002) and the research fund of the Assistance Publique-
Hôpitaux de Paris. We are indebted to D. Bourigault (Toulouse-le Mirail University), who kindly provided us with the results of the corpus analyzing tools Syntex® and Upéry®, and A. Kettaneh (Jean Verdier Hospital, Bondy) for statistical advice.
IX. Bibliography


[28] L. Gillam and M. Tariq, Ontology via Terminology ?, Workshop on Terminology,


Summary

Computer-assisted consensus in medical imaging involves automatic comparison of morphological abnormalities observed in images. We built an ontology of morphological abnormalities in breast cancer pathology to assist inter-observer consensus using a three step method: identifying the purpose and scope of the ontology, building the ontology using Protégé®, and verifying and validating the ontology by implementing position-based, content-based and mixed similarity measures between concepts, and evaluating their effectiveness for inter-observer consensus assistance.

Defining the scope and purpose of an ontology of morphological abnormalities that would be useful in the context of computer-assisted consensus led us to build a new ontology. Reusing existing ontologies which include morphological abnormalities such as GALEN, GeneOntology or SNOMED CT was impossible because they do not take into account the diagnostic meaning of morphological abnormalities. To build the ontology, a top-down approach was mainly used since most morphological abnormalities were acquired from published diagnostic classifications and SNOMED CT. The bottom-up approach based on natural-language processing of medical reports was less helpful due to the poor quality of these reports.

The three chosen similarity measures were implemented as a plug-in for Protégé®. For each one, it was possible to browse the corresponding similarity table for all pairs of concepts. Furthermore, the OntoViz plug-in, which provides a graphical representation of the ontology, was extended to highlight all concepts similar to any given concept in a specific color and to visualize the computation path of the similarity score for each pair of concepts. All three measures were well matched with the experts' judgment. The position-based similarity measure using the two types of relations (“is-a” and “is a diagnostic criterion of”) performed as well as content-based and mixed similarity. We integrated Leacock’s position-based similarity table in the IDEM platform to assist pathologists with description and diagnosis consensus.

A computer-assisted consensus study in 60 cases is currently underway to quantify the impact of computing semantic similarity between morphological abnormalities in the consensus process. More research is needed to define a methodology to use the feedback from consensus sessions to update the ontology of morphological abnormalities, and, for example, to instantiate canonical pathology entities of the ontology with illustrative regions of interest. Building a Digital Pathological Atlas that formalizes interpretation terms and rules with acceptable reproducibility among experts remains a serious issue because this may assist researchers to characterize the natural history of lesions and evaluate treatment effectiveness.
<table>
<thead>
<tr>
<th>Complexity of MA</th>
<th>Complex or elementary MA (“Features”); Elementary MA (“Observations”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low grade DCIS</td>
<td>Low nuclear grade Inferred from: Average nuclear size no larger than 1.5 a red blood cell (RBC) (a RBC is 6-7 µm wide)</td>
</tr>
<tr>
<td></td>
<td>Monomorphic nuclei (monomorphic size, oval or round regular outline and spacing)</td>
</tr>
<tr>
<td></td>
<td>Finely distributed chromatin</td>
</tr>
<tr>
<td></td>
<td>Nucleoli, if present, small and indistinct</td>
</tr>
<tr>
<td></td>
<td>Polarization of cells marked</td>
</tr>
<tr>
<td></td>
<td>Rare mitosis</td>
</tr>
<tr>
<td></td>
<td>No necrosis or individual necrotic cells</td>
</tr>
<tr>
<td>Intermediate grade DCIS</td>
<td>Intermediate nuclear grade Inferred from: Moderate nuclear enlargement. Average nuclear size no larger than 1.5 to 2 RBC</td>
</tr>
<tr>
<td></td>
<td>Minimal nuclear pleomorphism (some variation in size, irregular outline and spacing)</td>
</tr>
<tr>
<td></td>
<td>Finely to coarsely dispersed chromatin</td>
</tr>
<tr>
<td></td>
<td>Present but infrequent nucleoli</td>
</tr>
<tr>
<td></td>
<td>Polarization of cells present</td>
</tr>
<tr>
<td></td>
<td>Mitosis occasionally present</td>
</tr>
<tr>
<td></td>
<td>Central necrosis absent or minimal (non-comedo type necrosis)</td>
</tr>
<tr>
<td>High grade DCIS</td>
<td>High nuclear grade Inferred from: Average nuclear size larger than 2 RBC</td>
</tr>
<tr>
<td></td>
<td>Very pleomorphic nuclei (variation in size, irregular outline and spacing)</td>
</tr>
<tr>
<td></td>
<td>Vesicular, coarse, clumped chromatin</td>
</tr>
<tr>
<td></td>
<td>Frequent and prominent nucleoli</td>
</tr>
<tr>
<td></td>
<td>Polarization of cells absent or minimal</td>
</tr>
<tr>
<td></td>
<td>Mitosis often present</td>
</tr>
<tr>
<td></td>
<td>Extensive necrosis, comedo type necrosis</td>
</tr>
</tbody>
</table>

*Table 1: DCIS grading system [13]. The complex MA (“diagnosis”) DCIS grade is inferred from complex and elementary MA (“features”). The most important and consistent features are nuclear grade, polarization, mitosis and necrosis. Some features (complex MA) e.g. “low nuclear grade” may be inferred from other features (elementary MA) e.g. “finely dispersed chromatin”.*
Figure 1: Case description in the IDEM platform. This case is described by a list of morphological diagnoses observed in images. A morphological diagnosis is a complex morphological abnormality such as “High Grade Ductal Carcinoma in Situ” linked to the corresponding ROI. This morphological abnormality is inferred from the following morphological abnormalities: “High Nuclear Grade” and “Comedo-Type Necrosis”. “High Nuclear Grade” itself is inferred from the following elementary morphological abnormalities: “Mean Nuclear Size Larger than 2 RBC”, “Coarsely Dispersed Chromatin” and “Frequent and Prominent Nucleoli”.

Figure 2: View of the ontology for breast pathology
Figure 3: Visualization of the similarity table for each similarity measure
Figure 4: Justification of the computed similarity between two correlated morphological abnormalities (MA). Due to the relation “is a diagnostic criterion of” (abbreviated as “part of diag”), “Comedo Type Necrosis” and “High Nuclear Grade” are both linked to the diagnostic “High Grade Ductal Carcinoma in Situ”. They are both “correlated” features that can be observed in the same region of interest.
Figure 5: ROC curves of the three measures according to the unanimous gold standard for a sample of 200 pairs of concepts.
Figure 6: For each selected case, the list of MA clusters is displayed in Box 1. A MA cluster is a set of one to n MAs described by different pathologists and found to be in whole agreement (same or overlapping ROIs, same or similar labels) by the system. For a given cluster selected in box 1 (for example 4 in the picture), the MAs of the cluster are displayed in Box 2, MAs in partial agreement with at least one of the MAs of the cluster are displayed in Box 3 for MAs in semantic agreement (same or similar labels only) or in Box 4 for MAs in geographic agreement (same or overlapping ROIs only).
Figure 7: For each cluster selected by the moderator, each expert of the panel can see 1) the most relevant label and ROI according to the moderator (“best label” and “best ROI” underlined in blue) and 2) all the MAs (labels and ROI) found by IDEM in whole or partial agreement with the MAs of the selected cluster. For each MA, every pathologist states 1) if the label is in fact the same, different or most appropriate than the one selected by the moderator and 2) if the corresponding ROI is the same, different or most appropriate than the one selected by the moderator.