Methodology to Develop and Evaluate a Semantic Representation for NLP

Jeannie Y. Irwin, MS[1], Henk Harkema, PhD[1], Lee M. Christensen, MS[1], Titus Schleyer, DMD, PhD[1], Peter J. Haug, MD[2] Wendy W. Chapman, PhD[1]
[1]University of Pittsburgh, Pittsburgh, PA; [2]University of Utah, SLC, UT

Abstract

Natural language processing applications that extract information from text rely on semantic representations. The objective of this paper is to describe a methodology for creating a semantic representation for information that will be automatically extracted from textual clinical records. We illustrate two of the four steps of the methodology in this paper using the case study of encoding information from dictated dental exams: (1) develop an initial representation from a set of training documents and (2) iteratively evaluate and evolve the representation while developing annotation guidelines. Our approach for developing and evaluating a semantic representation is based on standard principles and approaches that are not dependent on any particular domain or type of semantic representation.

Introduction

Natural Language Processing (NLP) applications that extract information from text rely on semantic representations, like semantic networks[1], to guide the information extraction (IE) process and provide a structure for representing the extracted information. Semantic representations model the concepts and relationships that are important for the target domain and that appear in the relevant document collections. The structure of semantic representations must support further processing of the extracted text required by the final NLP application and is thus constrained by the capabilities of the NLP engine driving the application.

Since the content of a semantic representation depends largely on a document set and an application, it is usually not possible to "plug in" a previously developed semantic model. Also, existing domain ontologies are less useful as a model for structuring the information found in actual text because they tend to focus on abstract descriptions of knowledge organization. Therefore, it is often necessary to build a new semantic representation as part of an IE project.

Although there is some documentation about the evaluation of semantic networks[2], there is no widespread literature concerning the detailed process of constructing semantic representations for NLP applications. In the context of an IE project, we devised a four-step methodology for developing and evaluating semantic representations. The methodology integrates principles and techniques in semantic modeling, annotation schema development, and human inter-annotator evaluation.

Background

While providing care, dentists are restricted in their use of a keyboard and mouse, primarily due to infection control[3]. Therefore, dentists generally record patient data either by dictating findings to an assistant or personally entering the data after an exam. A survey of U.S. general dentists on clinical computer use singled out speech recognition, a way to facilitate direct charting, as one of the most desirable improvements in current applications[4]. Current systems using speech recognition lack a flexible, robust, and accurate natural language interface[3].

The long-term goal of our research is to develop a system that uses speech input and NLP to automatically enter patient data into electronic dental records. While developing this system, we created semantic models to represent the information that a dentist would chart during an exam. Our NLP system will ultimately extract information from a transcribed exam and instantiate the models. The semantic models both guide the IE process and store the extracted information in a format that can be automatically converted to a detailed dental chart.

The NLP system we are developing, called ONYX, is based on MPLUS, which has been used to encode clinical information from radiology reports[5] and chief complaints[6]. ONYX uses concept models (CMs) to represent the relationship between words in text and the concepts the words represent. CMs have two types of nodes—terminal nodes for slotting relevant words from the text and non-terminal nodes for inferring higher-level concepts from the words. Based on training cases, a Bayesian joint probability distribution is calculated for the variables in the CMs so that the probability of values in one node can be calculated based on value assignments in other nodes. One advantage of probabilistic over purely symbolic

representations like first order logic is graceful degradation of performance in the presence of noise and uncertainty.

Methods

We describe a four-step methodology for developing and evaluating a semantic representation that integrates: 1. principles for the creation of semantic representations; 2. methods for the development of annotation guidelines and schema and; 3. methods for evaluating semantic representations base on inter-annotator agreement. The four steps include: (1) develop an initial representation from a set of training texts; (2) iteratively evaluate and evolve the representation while developing annotation guidelines; (3) evaluate the ability of domain experts to use the representation for structuring the content of new texts according to the guidelines; (4) evaluate the expressiveness of the representation for information needed by the final application.

In creating and evaluating our representation, we wanted to address five standard requirements for a semantic representation[7]: 1. verifiability: the ability to validate statements from the represented knowledge; 2. unambiguous representations: a representation with only one valid interpretation that is able to supports vagueness; 3. canonical form: inputs that have the same meanings should have the same representation; 4. inference: the ability to infer information not explicitly modeled; and 5. expressiveness: the ability to model unseen but relevant information.

In this paper, we describe the first two steps of the methodology, using a case study from our experience of modeling chartable information from a dictated dental exam.

Step 1: Develop an Initial Semantic Representation

The first step in developing an initial semantic representation is to determine which concepts to extract and model. This decision is largely driven by the desired end application and the feasibility of automated extraction. For our study, we identified the 13 most frequently occurring dental conditions, including filling, crown, caries, and missing tooth.

We created our semantic representation using a bottom-up, data-driven approach. In this approach, one uses the textual source of information—in our case dictated dental exams—to design a representation for the mappings from words to concepts, as well as the relationships among the concepts.

To create our representation, we read a single transcribed dental exam, containing 551 words, and identified the information in the text related to the 13 target conditions. To represent the information in the exam, we created two types of semantic representations: a semantic network and concept models.

For each statement in the exam, we identified any concepts related to one of the 13 dental conditions. For example, for the sentence "There is a cavity on tooth 2," we identified two concepts: a dental condition of caries and an anatomic location of tooth 2. We developed a CM with non-terminal nodes for the concepts and terminal nodes for the words from the text that indicated the concepts, as shown in Figure 1. We then labeled relationships among the nodes.

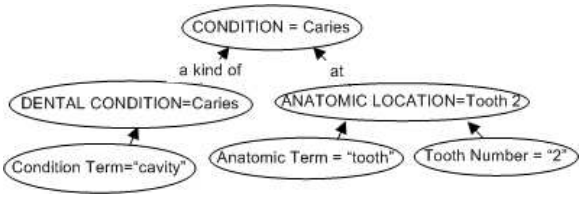

Figure 1. Initial network from training sentence "There is a cavity on tooth 2."

It became clear that we did not only need a CM for the way words are used to describe concepts but we also needed a mechanism for relating concepts to each other. For instance, a sentence describing a "crack on the crown of tooth 2" describes two concepts: a DENTAL CONDITION called fracture and a RESTORATIVE CONDITION called crown. Understanding the relationship between the crack and the crown is critical to our ability to chart the information. Therefore, we developed a semantic network encoding general domain knowledge to represent allowable relationships among dental concepts (Figure 2).

Terminal (white) nodes in the semantic network represent the root of individual CMs. Nonterminal (gray) nodes represent abstract types with no associated CMs that are useful for indirect relations and discourse processing. The semantic network allows different types of relationships between concepts. For instance, the network expresses the relations *at*(CONDITION, ANATOMIC LOCATION) and *has*(ANATOMIC LOCATION SURFACE). The semantic network also represents taxonomic relationships, via the *a kind of* label. A type may have multiple parent types. For instance, RESTORATIVE CONDITION is a subtype of both CONDITION and LOCATION.

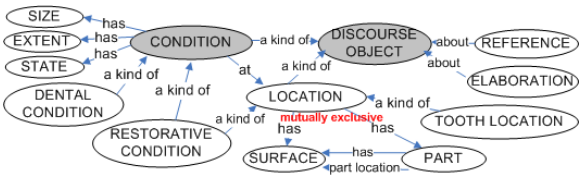

Figure 2. Semantic network for our domain. White nodes represent the top node in an independent concept model. Arrows represent relationships among the nodes.

Figure three shows how we use both the semantic network and CMs to interpret the sentence "Fifteen has one occlusal amalgam". We infer concepts from values in the leaf nodes of the CMs and then use the semantic network to model the relationships among the inferred concepts.

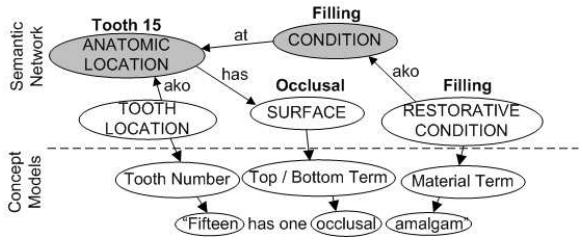

Figure 3. Example of the ideal interpretation of the sentence "Fifteen has one occlusal amalgam." Words above nodes are the inferred concepts.

Step 2: Evaluate and Evolve the Representation and Develop Annotation Guidelines

Step 2 is an iterative process involving structuring information from new documents to evaluate the coverage of the current representation, to evolve the representation based on new data, and to develop or enrich guidelines to ensure consistency among annotators.

We selected 12 exams of new patients: one exam from our original dentist, six from a new dentist and five from a hygienist. We developed a training tool for assisting human annotators in structuring information from a report into the concept networks. Three of the authors (JI—an informaticist, HH—a linguist, and LC—a developer), with input from dental experts, independently reviewed two exams identifying any instances of the 13 target conditions and related concepts found in the exams. The annotators entered the terms from the exam into the terminal nodes of the CMs. For instance, for the sentence in Figure 1, the word "cavity" was slotted in the condition term node, the word "tooth" in the anatomic location node, and the word "2" in the tooth number node. The annotators created values for the non-terminal nodes (i.e., implied concepts). For example, in Figure 1, the dental condition node received the value Caries, and the anatomic location node Tooth Two. According to the semantic network, the training tool generated all allowable relationships between instantiated CMs for that sentence, and each annotator selected the semantic relationships for each related pair of CMs. The sentence in Figure 1 has two relevant relations: *at*(CONDITION, ANATOMIC LOCATION) and *akindof*(DENTAL CONDITION, CONDITION).

After structuring the information from two exams the three annotators met to discuss disagreements, to come to consensus on the best instantiations, to change the CMs or semantic network in order to successfully model the information in the two exams, and to clarify the guidelines. The annotators iterated through the set of 12 reports in six cycles, annotating two reports independently before each meeting.

After each iteration, we measured agreement between pairs of annotators. Because it is not possible to quantify the number of true negatives in text annotation, we could not use Kappa. Therefore, we calculated agreement via inter-annotator agreement (IAA)[8]. IAA= matches / matches + non-matches, where matches = 2 x correct, and non-matches = spurious + missing. We calculated IAA separately for words, concepts, and relationships. Step 2 can be repeated until agreement reaches a threshold level or plateaus and the models appear stable and complete.

Results

We developed initial models using a single report of 551 words and evolved the models through iterative cycles of independent annotation and consensus meetings. Our final model resulted from annotations of 289 sentences in 13 reports.

Development of Initial Models

We identified 33 sentences containing relevant conditions (hereafter called cases) in the training exam. From those 33 cases we instantiated 125 words (73 unique) and 160 concepts (74 unique) into the CMs. Our initial semantic network had 11 nodes, eight of which represented individual concept models. After annotating the 12 exams in six iterations, changing the semantic model and concept models to accommodate all relevant information in the exams, the semantic model contained 13 nodes, 11 of which were concept models and 15 relationships. (see Figure 2).

Because we used a data-driven approach to design the initial models, we revised them several times to

account for new concepts described in unseen exams. One type of change was modularizing the CMs. Having a semantic network removed the need to link related concepts within large CMs, so we, for instance, split the ANATOMIC LOCATION and DENTAL CONDITIONS networks shown in Figure 1.

We added nodes to CMs and the semantic network and added new CMs. For instance, although initially we attempted to use the same CM for dental conditions, such as caries and fractures, and restorative conditions, such as crowns and fillings, we ultimately created separate DENTAL CONDITIONS and RESTORATIVE CONDITIONS networks, because we found these conditions have different properties.

We also added new relationships to the semantic network to capture the different roles the same concept can assume in different contexts. For example, the word "crown" can indicate a restorative condition ("crown on 16") or the location of a dental condition ("fracture on the crown").

Evaluating and Evolving the Model

Generally, as annotators instantiated cases, they found that a case consisted of a dental or restorative condition at an anatomic location. In the 12 exams two or more annotators identified a total of 256 cases for an average of 21 cases per exam. Further, for the 256 cases, each annotator slotted an average of 783 words and 1,018 concepts and defined an average of 394 relationships.

The average agreement for the three annotators for all iterations was 88%: 88% for words, 90% for concepts, and 86% for relationships. Figure 4 shows the average IAA for each iteration. All changes to the CMs and semantic network occurred after iterations one through four, but we made no changes after iterations five or six.

Disagreements among annotators can reveal lack of expressiveness and ambiguity in the semantic representations. For example, annotators slotted "some" in "22 has some incisal wear" in the severity term node, which is a modifier in the CONDITION CM. However, annotators disagreed on where to slot the similar word "small." In the end, we created a new CM for size.

Disagreements can also indicate inadequate annotation guidelines. After each iteration, we changed the annotation guidelines based on our discussions of how to best model the concepts in the text. IAA dropped in the second iteration due to multiple cases in which the annotators disagreed on how to slot the words "not missing" and "not present" as seen in the sentence "tooth number one is not present". We made almost half (8/20) of the changes to the guidelines during the discussion after iteration 2.

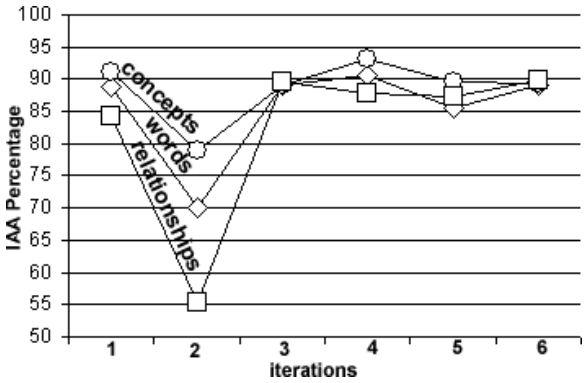

Figure 4. Graph of average IAAs for each iteration.

A key benefit of the iterative annotation phase is to enrich the guidelines while developers perform annotations so that the guidelines presented to experts in Step 3 will be as clear and useful as possible.

Discussion

As we began developing a semantic representation for our NLP system, we searched the literature for advice on how to best create a semantic model and on how to determine its quality. Although we could not find articles directly addressing development and evaluation of semantic models, we found relevant techniques in related areas, which we integrated in a four-step methodology we have begun to implement.

The methodology addresses principles for the creation of semantic representations[7], including a model's expressivity, its ability to represent information unambiguously, and the ability to map information to canonical form. The methodology incorporates techniques used in training annotators to develop training and testing sets for assessing output of an NLP system. Our method is similar to Roberts and colleagues[8] who compiled an annotated corpus of clinical reports, trained annotators on a semantic network they developed and iteratively evaluated agreement.

The first step of the methodology—creating the representation from example documents—allows developers to design models that relate the words in the text to the meaning conveyed by the words. To our surprise, creating our initial representations from a single document took several months as our models changed multiple times in an attempt to facilitate what the dentist said in the exam.

The second step—iteratively evaluating the representation by annotating new documents—is a critical step for ensuring generalizability of the models and for writing annotation guidelines to help non-developer annotators. This step is a quantitative step that allows developers to measure agreement and reveals deficiencies in the existing models. While slotting cases in Step 2, annotators test the representation's expressiveness and ability to support unambiguous representations while assigning words to canonical form.

The third step—evaluating agreement among expert annotators who follow the guidelines—is a familiar step in assessing the quality of training and test set annotations that serves a second purpose—to determine how usable the models are by non-developers. Our representation is quite complex, and we look forward to measuring its usability by dentists.

The fourth step—evaluating the expressiveness of the representation for information needed by the final application—is important for determining whether the models really convey the same information conveyed by the text. We plan to use the methodology described by Rocha et al.[2]. For this step, we will present domain experts—dentists, in our case—with two types of exams: transcriptions of dental exams for one set of patients and semantic models with manually instantiated information from the exams for another set of patients. We will test the ability of the domain experts to answer questions based on the two exam formats (in our case, the experts will graphically chart the exam). If the semantic representation successfully conveys relevant information from the text, the experts should answer questions from the semantic representation as well as from the text itself.

Our approach is a largely bottom-up approach, which can be an effective method for designing models for representation of ideas expressed in text. Disadvantages of a bottom-up approach include not leveraging expert knowledge contained in existing models and the possibility of designing a model that can only be used for a specific task. When we began development, we explored the UMLS and the Foundational Model of Anatomy as potential models; however the UMLS dental entries were limited, and existing dental concepts did not map well to what we saw in dental exams. In spite of using the text to drive our model development, we frequently consulted with dentists to ensure our models were consistent with domain expertise.

Conclusion

We described a process for developing and evaluating a semantic representation for an NLP application and illustrated the process in the domain of spoken dental exams. The methodology we describe explicitly addresses general requirements for semantic representations using a data-driven and iterative approach that can be replicated by others. In this study, we carried out the first two steps of the methodology, illustrating the types of changes we made to our models through our approach. Although we applied the methodology to a single domain, the methodology is based on standard principles and approaches that are not dependent on any particular domain or type of semantic representation.

This work was funded by the NIDCR R21DE018158-01A1 Feasibility of a Natural Language Processing-based Dental Charting Application grant.

References

1. Quillian, M. Semantic memory. In M. Minsky, editor, Semantic Information Processing. MIT Press, Cambridge, MA, 1968.
2. Rocha RA, Huff SM, Haug PJ, Evans DA, Bray BE. Evaluation of a semantic data model for chest radiology: application of a new methodology. METHOD INFORM MED. 1998 Nov; 37:477-90.
3. Irwin JY, Fernando S, Schleyer T, Spallek H. Speech recognition in dental software systems: features and functionality. Medinfo 2007;12:1127-31.
4. Schleyer TK, Thyvalikakath TP, Spallek H, et al. Clinical computing in general dentistry. JAMIA. 2006 May-Jun;13(3):344-52.
5. Christensen L, Haug P, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proceedings and Workshop on NLP in the Biomedical Domain; 2002. p. 29–36
6. Chapman WW, Christensen LM, Wagner, MM, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. ARTIF INTELL MED. 2005 Jan;33(1):31-40.
7. Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. 2nd ed. Upper Saddle River NJ: Prentice Hall; 2008.
8. Roberts A, Gaizauskas R, Hepple M, et al. The clef corpus: semantic annotation of clinical text. AMIA Annu Symp Proc.; 2007. p. 625-9.