Inducing Terminologies From Text: A Case Study for the Consumer Health Domain

Smaranda Muresan
Department of Library and Information Science, School of Communication and Information, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901. E-mail: smuresan@rutgers.edu

Judith L. Klavans
Institute for Advanced Computer Studies, University of Maryland, A.V. Williams Building, Room 3117, College Park, MD, 20742-3251. E-mail: jklavans@umd.edu

Specialized medical ontologies and terminologies, such as SNOMED CT and the Unified Medical Language System (UMLS), have been successfully leveraged in medical information systems to provide a standard web-accessible medium for interoperability, access, and reuse. However, these clinically oriented terminologies and ontologies cannot provide sufficient support when integrated into consumer-oriented applications, because these applications must "understand" both technical and lay vocabulary. The latter is not part of these specialized terminologies and ontologies. In this article, we propose a two-step approach for building consumer health terminologies from text: 1) *automatic extraction* of definitions from consumer-oriented articles and web documents, which reflects language in use, rather than relying solely on dictionaries, and 2) *learning* to map definitions expressed in natural language to terminological knowledge by inducing a syntactic-semantic grammar rather than using hand-written patterns or grammars. We present quantitative and qualitative evaluations of our two-step approach, which show that our framework could be used to induce consumer health terminologies from text.

Introduction

The purpose of medical ontologies and terminologies is to facilitate the development of information systems that can "understand" the meaning of language in the biomedical domain. To that end, there has been a constant effort to develop biomedical ontologies, such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), OpenGALEN (Nowlan, Rector, Rush, & Solomon, 1994), and Foundational Model Anatomy (FMA) (Rosse & Mejino, 2003), to name just a few. These ontologies have been used to build or enhance electronic information systems that create, process, retrieve, and integrate biomedical data and information.

However, these resources cannot be used in consumer-oriented applications, because consumer-oriented information systems must be able to "understand" lay vocabulary used by consumers of health information. Studies have shown that currently there is a mismatch between general terminologies, such as WordNet (Miller, 1990), and technical medical terminologies, such as the UMLS (Burgun & Bodenreider, 2001). Recent work has been proposed to address this issue by developing methods for building consumer health vocabularies (Cardillo, 2011; Elhadad, 2006; Elhadad & Sutaria, 2007; Smith & Fellbaum, 2004; Zeng & Tony, 2006). One of the recent largest initiatives is the Consumer Health Vocabulary Initiative at the Harvard Medical School, which has developed the Open Source Collaborative Consumer Health Vocabulary (OSC CHV), and which aims to link lay medical terms to their corresponding technical concepts from the UMLS Metathesaurus (Keselman, Logan, Smith, Leroy, Zeng-Treitler, Q., 2008a, Keselman, Smith, Divita, Kim, Browne, Leroy, Zeng-Treitler, 2008b; Zeng & Tony, 2006). This approach leverages existing medical terminologies, such as the UMLS, by mapping lay vocabulary to their technical equivalents.

In this article, we present a complementary approach for creating consumer health terminologies—automatically building terminologies from *consumer-oriented text*. When talking about *terminologies* in this article, we mean entities (concepts) and relations among entities (concepts). We talk about terminologies and not ontologies, because ontologies would require a stricter formalization, as discussed by Bodenreider (2006) and Ceusters, Capolupo, Moor, Devlies, and Smith (2011).

In our approach, we rely on the assumption that a key source for terminological knowledge is the definition of a term, an assumption that has been used by a large body of

Received May 21, 2012; revised July 23, 2012; accepted July 24, 2012

© 2013 ASIS&T • Published online 1 March 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22787

research on processing dictionary/glossary definitions (Chodorow, et al., 1985; Klavans, Chodorow, & Wacholder, 1992; Moldovan & Rus, 2001; Navigli & Velardi, 2008; Richardson, Dollan, & Vanderwende, 1998; Rus, 2002; Wilks, Slator, & Guthrie, 1996). However, dictionaries are inherently incomplete in rapidly evolving scientific and technical domains, such as the medical domain. Thus, unlike this previous work, we present a method for *automatically extracting definitions from consumer-oriented medical text*, which consists of medical articles or manuals written by medical specialists for a general audience, where medical terminology must be defined. Then, we *automatically map the definitions extracted from text to terminological knowledge*, by *learning* a syntactic-semantic grammar, rather than using hand-written patterns or grammars. In this article, we present our grammar learning framework together with a semantic interpreter targeted for terminological knowledge. We show that our method is suitable for direct acquisition of terminological knowledge from natural language definitions automatically extracted from text as well as for querying this knowledge using both precise and vague questions, obtaining precise answers at the concept level.

An example of input and output for our two-step approach is presented in Figure 1. Starting with a set of consumer-oriented articles, we automatically extract definitions and their medical terms using our system DEFINDER (Klavans & Muresan, 2000, 2001; Muresan & Klavans, 2002), and then we apply machine-learning techniques to map these definitions to terminological knowledge—a graph-based representation that encodes concepts (e.g., #hepatitisA, #hepatitis, #blood), instances of concepts (#virus25, #virus33, #cause24), and relations (e.g., sub, which is the inverse of isa; loc, which is given by the semantic role of the preposition *in*; and ag, th, which are semantic roles of the verb *cause*). These roles can be general or domain specific.

In the next section, we present our system architecture and related work. In the DEFINDER: Extracting Definitions From Consumer-Oriented Text section, we present our system DEFINDER and the characteristics of the acquired corpus of definitions. In the section Learning to Map Definitions to Terminological Knowledge, we present our grammar formalism, grammar learning model, and our semantic parser and interpreter used to map definitions to terminological knowledge. In Results we present our experiments and results. In the Discussion section we discuss our findings. Our conclusions and plans for future work are given in Conclusions and Future Work.

System Architecture and Related Work

Our method for building and querying terminological knowledge bases is a two-step approach. The architecture of our system is given in Figure 2. The backbone of this architecture can be seen as a standard natural language understanding system: given an input text (corpus) and a grammar, a parser produces syntactic/semantic representations of the input text (text-level representations), which are then transformed into knowledge by a semantic interpreter. There are two key contributions of our approach: (1) our input text is a corpus of definitions automatically extracted from consumer-oriented articles using our system DEFINDER and (2) the grammar is learned from a small amount of training data (training corpus in Figure 2) rather than hand-written, as in most traditional deep language understanding systems. We briefly highlight these two components of our architecture below, with pointers to related work.

DEFINDER: Extracting Definitions From Consumer-Oriented Medical Articles

A key source for terminological knowledge is the definition of a term. However, dictionaries are inherently incomplete in the rapidly evolving scientific domains, such as the medical domain. In our work, we developed a rule-based system DEFINDER to automatically extract definitions from *consumer-oriented medical text*, which consists of *medical articles or manuals written by medical specialists for a general audience*, where terminology must be defined (Klavans & Muresan, 2000, 2001). We give more details in DEFINDER: Extracting Definitions From Consumer-Oriented Text. Other studies have used alternative sources for consumer-health vocabulary: (1) elicitation of terms directly from lay users to build a lexi-ontological resource for consumer healthcare for Italian (Cardillo, 2011) and (2) combination of elicitation data and text from bulletin boards to collect medical facts and beliefs used to enhance WordNet with consumer health information (Smith & Fellbaum, 2004). Our choice of corpus is based on two desiderata: *scalability* and *presence of definitions* of consumer heath terms.

The DEFINDER system has been the building block for research in extracting definitions from vetted web documents. Most of this subsequent work on definition extraction has been developed since TREC-2003, which included a track of answering definitional questions (*Who is X?, What is X?*). Several lines of research on definition extraction have enhanced and adapted our pattern matching approach for different domains (Liu, Chin, & Ng, 2003) and also developed hybrid approaches that combine manually written patterns and machine-learning techniques (Blair-Goldensohn, McKeown, & Schlaikjer, 2004; Fahmi & Bouma, 2006). Applications of definition extraction work range from mining definitions for topic-specific concepts (Liu et al., 2003) to answering definitional questions in the context of a question-answering task.

One aspect is worth mentioning here: Even if we use full syntactic parsers in our definition extraction system, we only use pattern matching of limited depth over the parse trees. Thus, we are able to extract fairly accurate definitions (~86.95% precision and 75.47% recall), even if the deep parse tree structure is not fully correct (see Results for a detailed evaluation of DEFINDER). However, it is not

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hepatitis is a disease caused by infectious or toxic agents and characterized by jaundice, fever and liver enlargement. [...]

Hepatitis A is an acute but benign viral hepatitis caused by a virus that does not persist in the blood serum. [.....] Hepatitis B is an acute viral hepatitis caused by a virus that tends to persist in the blood serum. [.....]
*[....]*About a third of the world's population, have been infected with the hepatitis B virus.[...]

AUTOMATIC EXTRACTION OF DEFINITIONS FROM TEXT (DEFINDER)

Hepatitis is a disease caused by infectious or toxic agents and characterized by jaundice, fever and liver enlargement.

Hepatitis A is an acute but benign viral hepatitis caused by a virus that does not persist in the blood serum.

Hepatitis B is an accute viral hepatitis caused by a virus that tends to persist in the blood serum.

LEARNING TO MAP DEFINITIONS TO TERMINOLOGICAL KNOWLEDGE

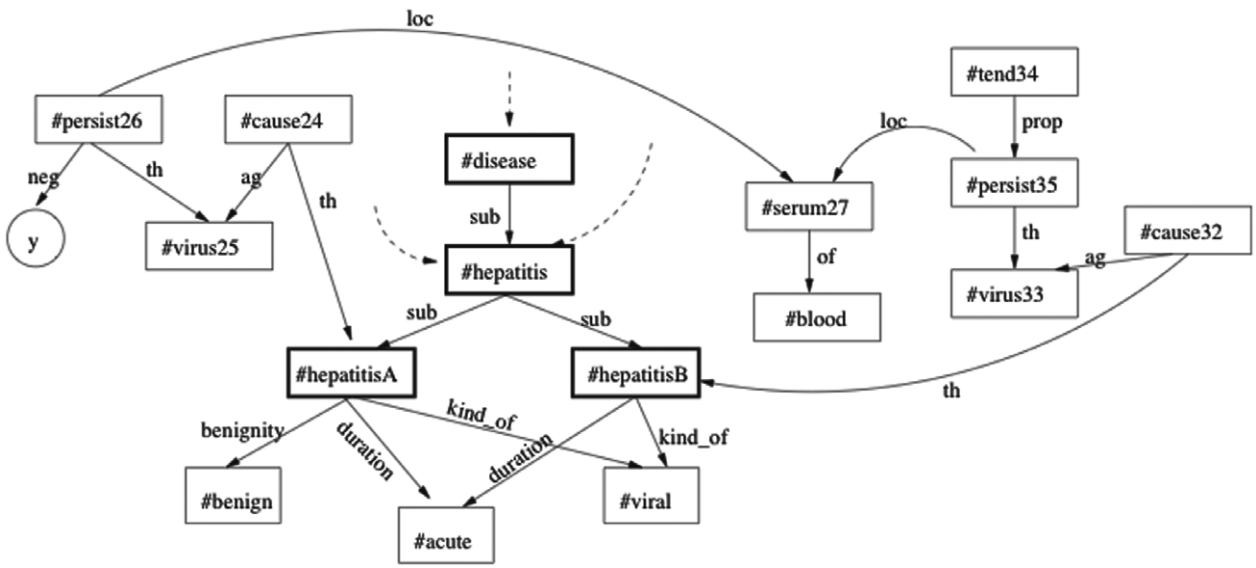


FIG. 1. Input and output. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

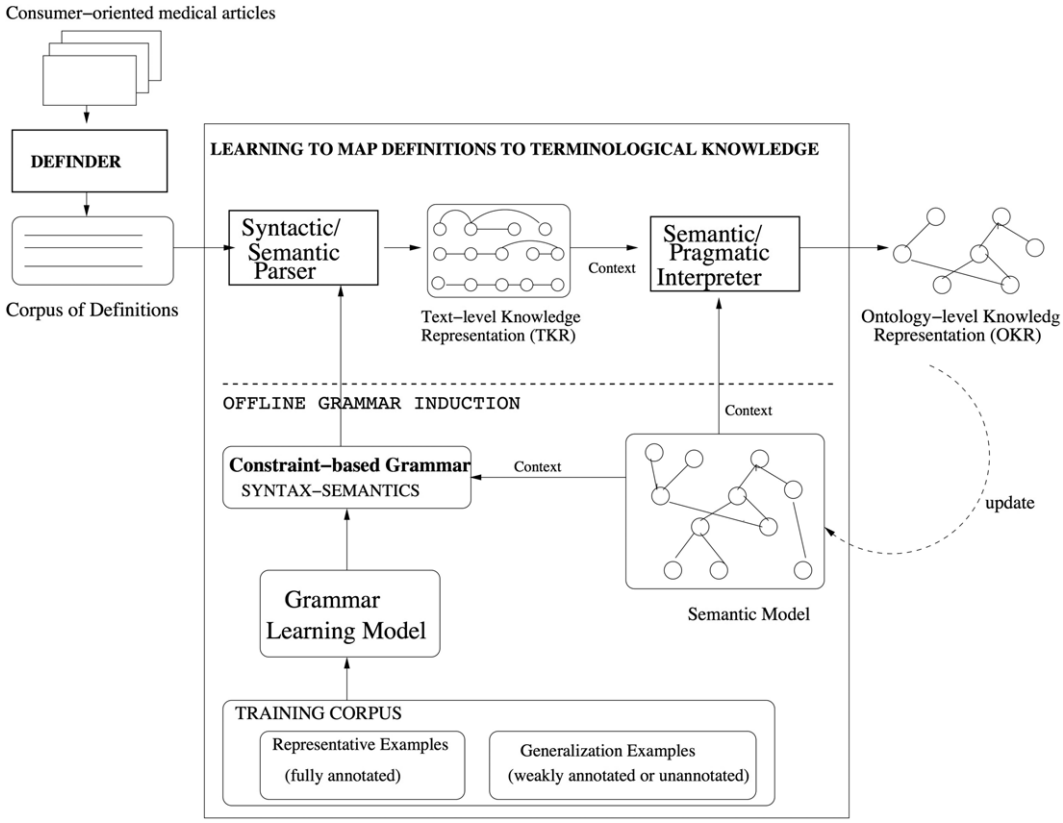


FIG. 2. System architecture.

adequate to use these parsers to acquire the semantics of these definitions (incorrect syntactic parsing will lead inevitably to incorrect semantics). Figure 1 shows some definitions found in consumer-oriented text extracted by DEFINDER. These definitions contain complex constructions, which require both syntactic and semantic information to be interpreted correctly, including noun compounds (e.g., *liver enlargement*), coordinations and prepositions (e.g., *caused **by** infectious **or** toxic agents **and** characterized **by** jaundice, fever **and** liver enlargement)*, raising and control constructions (e.g., *virus that **tends** to persist*), embedded relative clauses (e.g., *disease caused by a virus **that tends to persist***), and negation (e.g., ***does not*** persist). The inability of current statistical syntactic parsers trained on the Penn Treebank (*Wall Street Journal* articles) (Marcus, et al.,1994) to accurately parse medical text is one of the motivations for our grammar learning framework that links syntax to semantics.

Learning to Map Definitions to Terminological Knowledge

Most work on acquiring knowledge from dictionary-based definitions has relied on hand-written patterns, grammars, or semantic transfer rules used on top of syntactic parses (Chodorow et al., 1985; Klavans et al., 1992; Moldovan & Rus, 2001; Richardson et al., 1998; Rus, 2002; Wilks et al., 1996). These methods lack scalability when going beyond the dictionary-like format or when moving to new domains.

In this article, we present a methodology for *learning to map the definitions extracted from text to terminological knowledge*. First, we *learn* a grammar able to capture the syntactic and semantic information of linguistic constructions found in medical definitions (Figure 2). We rely on our recently developed grammar formalism for deep language understanding, Lexicalized Well-Founded Grammar (LWFG), which captures syntax and semantics and can be efficiently learned from data (Muresan, 2006, 2010; 2011; Muresan & Rambow, 2007).

Unlike grammar learning for statistical syntactic parsing (e.g., Charniak, 2000; Clark & Curran, 2007; Collins, 1999), our grammar can be learned from a small amount of data. The input to the grammar learning model consists of a small set of utterances paired with their semantic representations. Learning from a small amount of annotated data are essential for rapid domain adaptation. In Learning to Map Definitions to Terminological Knowledge, we present our grammar formalism and grammar learning model. Our LWFG learning belongs to recent efforts in the natural language processing community to develop machine learning algorithms for mapping utterances to formal representations of their underlying meaning (Poon & Domingos, 2009; Wong & Mooney, 2007; Zettlemoyer & Collins, 2005).

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

These techniques have been mostly applied to very specific domains, such as GeoQuery and RoboCup events. A notable exception is the work of Poon and Domingos (2009) who applied their semantic parser to PubMed abstracts. In addition, our work complements efforts on adapting statistical syntactic parsers trained on Penn Treebank to the medical domain (Lease & Charniak, 2005).

A key property of our grammar formalism is the use of constraints to capture grammatical aspects (e.g., agreement) and to encode a direct link between natural language expressions and *a semantic model* (e.g., domain context). These constraints are important for the disambiguation required for some phenomena (e.g., prepositional phrase attachment, coordination), as well as for the semantic interpretation of phenomena not usually analyzed by current syntactic parsers (e.g., prepositions, noun-noun compounds). Consider the following example: *the two endocrine glands [located above the kidney] [that secrete hormones and epinephrine]*. The second relative clause can be attached to the noun *kidney* or the noun *glands*. Since using LWFGs, we can model agreement between the head noun and the verb in the relative clause, and we have that the relative clause is attached to the noun *glands* (plural). In addition, the semantic interpretation constraints will give us the semantics of the utterance (e.g., the meaning of the preposition *above* is location, the semantic roles of the verb *secrete* are agent and theme).

Once a LWFG grammar for definitions has been learned, we use a parser to produce semantic representations for each definition in our corpus (Figure 2). These representations form the text-level knowledge representation (TKR). To build terminological knowledge we need a semantic interpreter that takes into account task-specific constraints (e.g., for terminological knowledge we do not interpret determiners and the tense and aspect of verbs, but we do interpret modals and negation). To be consistent with our previous work (Muresan, 2006, 2008), we refer to the representation produced by the semantic interpreter as the ontology-level knowledge representation (OKR), which is a graph-based representation where vertices are concepts or instances of concepts, whereas edges represent relations among concepts or instances of concepts. An example was given in Figure 1, and more detail will be given below in the section Learning to Map Definitions to Terminological Knowledge.

The framework proposed in this article allows us to start with a weak semantic model (e.g., admissibility relations at the level of lexical items, such as thematic roles of verbs—agent, theme), which can be gradually enhanced to become a stronger semantic model (e.g., hierarchy of concepts and roles) by automatically acquiring terminological knowledge from natural language definitions (dotted line in Figure 2). This stronger semantic model can then be used in conjunction with our grammar and parser to interpret consumer health text beyond just definitions. This is especially useful for interpreting domain-specific text, such as the medical text (Aronson, 2001; Friedman, Borlawsky, Shagina, Xing, & Lussier, 2006; Hahn et al., 2000). Although we mention this bootstrapping ability of our framework, this is beyond the scope of this paper.

DEFINDER: Extracting Definitions From Consumer-Oriented Text

The acquisition of definitions from textual corpora is a challenging task because the structure of definitions in text is not always similar to those in online dictionaries. The algorithm for the extraction of definitions from text is a rule-based method implemented in the DEFINDER system (Klavans & Muresan, 2000, 2011). Our corpus consists of consumer-oriented medical articles containing more than one million words from different authoritative sources on the web.

We have developed DEFINDER in the context of a medical digital library project, PERSIVAL (McKeown, et al. 2001). First, a development set of articles was analyzed and patterns that occur frequently and reliably in many text genres (e.g., articles, book chapters, health newsletters) were identified. We grouped these patterns into three categories: cue-phrases (e.g., *is the medical term for*), text markers (e.g., "—", "(", ")"), and syntactic patterns (e.g., syntactic complement of the link verb ***to be*** in sentences such as *Acne **is** a skin disease caused by overactive oil glands*, and appositional patterns in sentences such as *Angina, the chest pain that occurs when an area of your heart muscle doesn't get enough oxygen-rich blood*, . . .). The identification of definitions from these initial contexts was performed in two steps: (1) shallow syntactic parsing for identification of simple definitions and candidate complex definitions, and (2) full syntactic parsing of these candidate definitions using a statistical syntactic parser (Charniak, 2000).

In the first step, we use Brill's part-of-speech tagger (Brill, 1992) and a noun phrase (NP) chunker (Ramshaw & Marcus, 1995) in conjunction with a simple finite-state grammar. We have augmented Brill's tagger lexicon with medical terms from the UMLS lexicon (Lindberg, Humphreys, & McCray, 1993) to increase accuracy. An automatic filtering step is performed to remove patterns for enumerations, or explanations. As a result of shallow analysis, we have both *<term> (<definition>)* and *<definition> (<term>)*. When length was not sufficient, a simple statistical measure based on frequency counts was used to discriminate between the term and its definition. This is usually the case for technical/lay pairs, such as "tachycardia/irregular heartbeat." In addition, we select candidate definitions that cannot be easily identified by shallow processing and for which full parsing is more reliable.

In the second step, these candidate definitions are syntactically parsed using a statistical parser (Charniak, 2000). We perform a shallow-level pattern matching over full parse trees to identify complex appositives, syntactic complements of the link verb *to be*, and complex definitions found in the context of text markers. Using just shallow-level pattern matching allows us achieve high accuracy on extracting definitional sentences, even if the full syntactic

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1. Examples of DEFINDER's output: Multiple definitions of the term "acne."

1. Acne is a skin disease characterized by papules and pustules on the face and neck.
2. Acne is an inflammatory skin disease characterized by pimples that can appear on any part of the body.
3. Acne is a skin disease caused by overactive oil glands.
4. Acne is an inflammatory disease involving the sebaceous glands of the skin.

TABLE 2. Examples of DEFINDER's output: Linguistic complexity.

1. Atherosclerosis is the progressive narrowing of arteries from cholesterol plaque deposits.
2. Acne is an inflammatory skin disease characterized by pimples that can appear on any part of the body.
3. Addison's disease is a degenerative disease caused by a deficiency in adrenocortical hormones and characterized by weight loss, brown pigmentation of the skin, and low blood pressure.

parse tree was not correct. In Results we present the quantitative and qualitative evaluation of DEFINDER.

We analyzed a sample of the DEFINDER corpus to observe the characteristics of definitions, both at the conceptual level and linguistic level. This analysis has informed our development of training data for grammar learning as well as our semantic interpreter for terminological knowledge.

Conceptual dimension. Our corpus of definitions automatically extracted from text has the advantage of containing multiple definitions of the same term, which enables us to analyze the conceptual nature of definitions. Referring to Table 1, which presents multiple definitions of the term *acne*, it can be observed that each definition has additional properties (e.g., definition 1 specifies the symptoms (*papules and pustules*) and their location (*the face and neck*), whereas definition 3 specifies the cause (*caused by overactive oil glands*); definition 1 specifies that *acne* is a *skin disease*, whereas definition 4 specifies that *acne* is an *inflammatory disease*). Thus, none of these definitions can be considered complete. This incompleteness questions the core assumption of the logical definition (Weisman, 1992): the equation between the definiendum (term to be defined) and the definiens (the definition of the term). How can we be sure that equivalence can be obtained? How can we be sure that there is no other definition, which contains an additional property required to fully determine the "acne" concept? In the lexicography literature, the issue of incompleteness has been attributed to the educational role of definitions: they explain the essence of a concept to different types of users, in different contexts (Eck & Meyer, 1995). The above observation motivates our choice of interpreting copula *be* as predicative-be instead of equative-be for terminological knowledge.

Linguistic dimension. Table 2 shows several definitions extracted by DEFINDER illustrating their complex linguistic constructions (both from a syntactic and a semantic point of view): nominalization (*narrowing of arteries from cholesterol plaque deposits*) and noun compounds (e.g., *cholesterol plaque deposits*, *weight loss*); wh-relative and that-relative clauses (*that can appear on any part of the body*); reduced relative clauses (*caused by a deficiency in . . .*), and embedded relative clauses (*characterized by pimples that can appear on any part of the body*). Verbal constructions contain active and passive voice, modals (*can appear on any part of the body*), negation, and sometimes raising and control verbs (e.g., *tends*). One characteristic of the definitional corpus is the presence of prepositions and coordinating conjunctions. An analysis of our corpus shows that on average there are three prepositions per definition. Regarding coordination, definition 1 in Table 1 and definition 3 in Table 2 are eloquent examples of the complexity of coordination constructions. We can have coordination between all categories at lexical-level (e.g., *papules or pustules*), phrase-level (e.g., *weight loss, brown pigmentation of the skin, and low blood pressure*) and clause-level (e.g., *caused by . . . and characterized by . . .*).

Learning to Map Definitions to Terminological Knowledge

As mentioned above, our research showed that we cannot rely on existing syntactic parsers trained on the Penn Treebank to obtain accurate parses for the definitions in the medical domain. Incorrect syntactic parsing inevitably leads to incorrect semantic analysis. Moreover, from the analysis of a sample of our corpus of definitions, we noticed that many linguistic constructions present in definitions require access to both syntactic and semantic information to be analyzed correctly. To transform these definitions into knowledge, we use our recently developed grammar formalism that captures syntax and semantics and is learnable from a small set of representative examples. In this article, we summarize the description of our grammar formalism and grammar learning model in Learning Grammars for Deep Language Understanding and instruct the interested reader to consult Muresan and Rambow (2007) and Muresan (2008, 2010, 2011) for full technical details. Below, we present our semantic parser and semantic interpreter used in conjunction with the learned grammar to transform definitions into terminological knowledge.

Learning Grammars for Deep Language Understanding

In this paper, we use our recently developed grammar formalism for deep language understanding, Lexicalized Well-Founded Grammar (LWFG), which captures syntax and semantics and can be learned from data (Muresan, 2006, 2010, 2011; Muresan & Rambow, 2007). In LWFG, every word/phrase/clause/utterance is associated with a

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

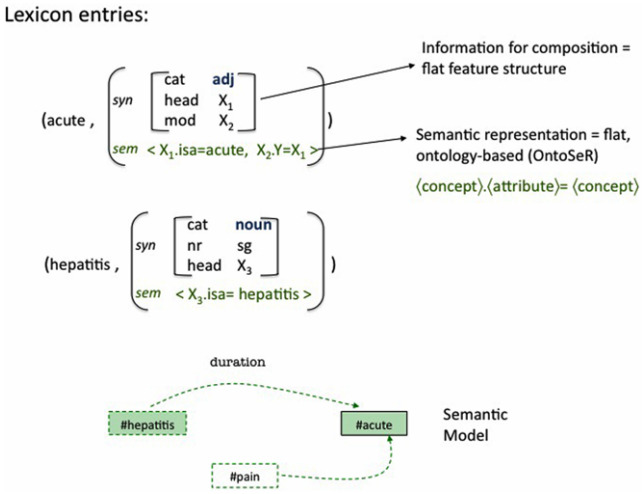


FIG. 3. Examples of LWFG's lexicon entries for the adjective "acute" and the noun "hepatitis." [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

syntactic-semantic representation, and grammar rules can have two types of constraints: one for semantic composition (C_c)—defines how the meaning of a natural language utterance is composed from the meaning of its words and phrases—and one for semantic interpretation (C_i)—validates the semantic constructions based on a given semantic model (i.e., a domain ontology). These two properties make LWFGs a type of syntactic-semantic grammars.

The lexicon in LWFG consists of words paired with *elementary* syntactic-semantic representations. The lexicon in LWFG is not learned. Unlike other lexicalized grammar formalisms for deep understanding, such as Combinatorial Categorial Grammar (Steedman, 1996, 2000) and Lexicalized Tree-Adjoining Grammars (Joshi, 1985; Joshi & Schabes, 1997), the lexicon in LWFG does not specify the syntactic context in which the word is anchored. That context will be learned from examples, by learning grammar rules and compositional constraints. In Figure 3 we show examples of lexicon entries for the adjective *acute* and the noun *hepatitis*.

Our syntactic-semantic representation is denoted $\left(\frac{syn}{sem}\right)$, where ***syn*** encodes the information required for semantic composition (e.g., the syntactic category of the word or phrase), and ***sem*** is the actual semantic representation of the string. This representation is simple enough to allow learning and tractable inferences, but expressive enough for natural language (Muresan, 2006, 2008). Formally, ***syn*** is a one-level feature structure—that is, the values are atomic—and has at least two attributes: (1) *cat*, which encodes the syntactic category of the associated word or phrase (e.g., adjective, noun, noun phrase), and 2) *head*, which is the index of the head word of a phrase/sentence/clause/word (i.e., the head word of a noun phrase is the noun, the head word of a prepositional phrase is the preposition, the head word of a sentence is the verb). In addition, feature attributes for agreement and other grammatical features can be present (e.g., number [singular of plural], person [first, second, or third]). The actual semantic representation, ***sem*** is a flat logical form, built as a conjunction of atomic predicates *<concept >.<attr> = <concept>*, where variables are either concepts or slot identifiers in the semantic model. The richness of the semantic model can range from just a lightweight ontology—encoding the admissibility relations that we can find at the level of lexical entries, such as thematic roles of verbs and prepositions—to a heavyweight ontology with hierarchy of concepts and roles. This semantic representation, called OntoSeR (Ontology-based Semantic Representation) can serve as an ontology-query language (Muresan, 2006, 2008). For example, the adjective *acute* is represented as $<X_1.isa = acute,\ X_2.Y = X_1>$, which says that the meaning of an adjective is a concept ($X_1.isa = acute$ maps to the concept #acute in the semantic model), which is a value of a property of another concept ($X_2.Y = X_1$) in the semantic model. Variable X_2 will be instantiated through composition, when the adjective *acute* will be combined with a noun, for example, *hepatitis* to build a noun phrase *acute hepatitis*. Variable Y will be instantiated after the semantic interpretation based on a given semantic model (e.g., $Y = duration$).

To combine words to build phrases, clauses, and sentences we need grammar rules. LWFG has a set of constraint grammar rules, which can be recursive and where the nonterminals are augmented with pairs of strings and their syntactic-semantic representations. For example, a simple grammar rule for noun phrases, such as *acute hepatitis* is given in Figure 4.

Grammar rules have two types of constraints: one for semantic composition C_c, and one for semantic interpretation C_i. The semantic composition constraints C_c define how the meaning of a natural language expression (e.g., phrase, clause, sentence) is composed from the meaning of its parts (e.g., words, phrases). These constraints are applied to

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

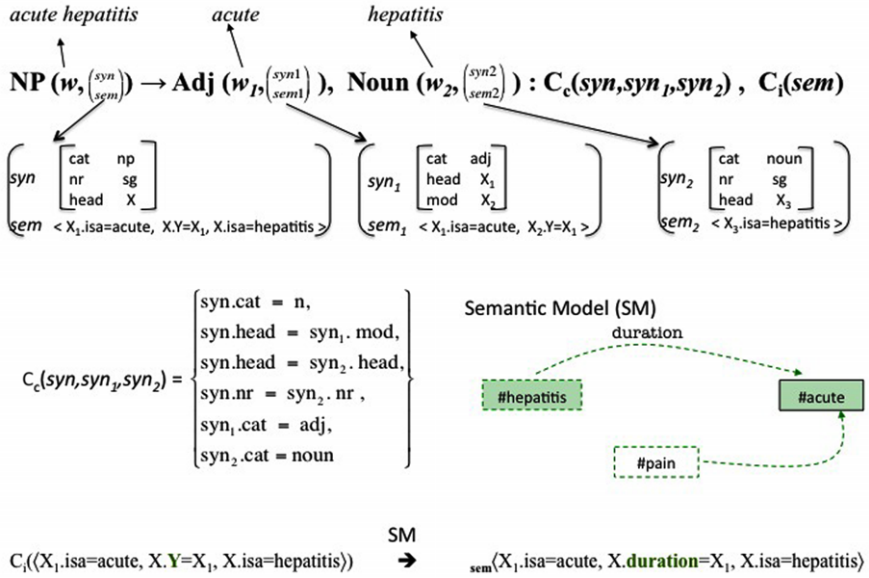


FIG. 4. LWFG grammar rule and semantic composition and semantic interpretation constraints. [Color figure can be viewed in the online issue, which is available atwileyonlinelibrary.com.]

the ***syn*** part of the syntactic-semantic representations, the semantic representations (***sem***) being just concatenated. The semantic composition constraints C_c are learned together with the grammar rules (Muresan, 2006, 2010).

The semantic interpretation constraints C_i represent the validation based on a semantic model and are not learned. These constraints serve as a local semantic interpreter at the grammar rule level. Currently, C_i is a predicate that can succeed or fail—when it succeeds, it instantiates the variables of the semantic representation with concepts/slots in the semantic model (Muresan, 2006, 2008). For example, given the phrase *acute hepatitis*, C_i succeeds and returns $<X_1.isa = acute, X.\mathbf{duration} = X_1, X.isa = hepatitis>$, whereas given the phrase *blonde hepatitis* it fails.

The constraints C_c and C_i are important for the disambiguation required for some phenomena (e.g., prepositional phrase attachment, coordination) and for the semantic interpretation of phenomena not usually analyzed by current broad-coverage grammars or statistical syntactic parsers (e.g., prepositions, noun-noun compounds).

Before describing how a LWFG grammar in association with a semantic parser and semantic interpreter can be used to map utterances to knowledge, we briefly present the grammar learning model (how can we learn grammars from data).

Grammar learning model. LWFG's learning is a relational learning framework, which characterizes the "importance" of substructures in the model not simply by frequency, as in most previous work (Charniak, 2000; Collins 1999; Klein & Manning, 2003), but rather linguistically, by defining a notion of "representative examples" that drives the acquisition process. Informally, representative examples are "building blocks" from which larger structures can be inferred via reference to a larger unannotated or weakly annotated corpus (called the generalization corpus, see Figure 2). For example, *effect*, *the effect*, and *adverse effect*, annotated similarly to *hepatitis* and *acute hepatitis* shown above, might all be representative examples for the English nominal system; *adverse* annotated similarly to *acute*, might be a representative example for English adjectives; and the unannotated generalization corpus might contain *the major adverse effect*. With this information, it is possible to learn grammar rules permitting English noun heads to be modified by a determiner and multiple adjectives (learning recursive grammar rules).

We have implement three algorithms for LWFG learning and have studied their efficiency and annotation effort required for the training data. In this article, we only give a summary of our findings, referring the reader to Muresan (2006, 2011) for details of these algorithms.

Learning from representative examples. In this case, the learning algorithm is presented with an ordered set of representative examples, that is, learning from simpler examples first, and then gradually from more complex examples (similar to how a child acquires language being exposed first to simpler utterances). The annotation effort is reduced because only the representative examples need to be annotated, whereas the generalization corpus can be unannotated. The order of magnitude for the representative examples is hundreds of examples, whereas the generalization corpus can be several thousands (see Results where we discuss the data used for learning a grammar of definitions).

Learning from unordered representative examples. A practical problem with the previous algorithm is that in some cases it is hard to determine a priori the right order of the

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

utterance level representation OntoSeR$^-$ (before local semantic interpretation C_i.)

$\langle$(A.det=a)$_a$, (A.isa=virus)$_{virus}$, (A.isa=that)$_{that}$, (B.tense=pr)$_{does}$, (B.neg=y)$_{not}$, (B.isa=persist, B.P1=A)$_{persist}$, (P2.isa=in, B.P2=C)$_{in}$, (C.det=the)$_{the}$, (D.isa=blood, C.P3=D)$_{blood}$, (C.isa=serum)$_{serum}$ $\rangle$

utterance level representation OntoSeR$^+$ (after local semantic interpretation C_i.)

$\langle$(A.det=a)$_a$, (A.isa=virus)$_{virus}$, (A.isa=that)$_{that}$, (B.tense=pr)$_{does}$, (B.neg=y)$_{not}$, (B.isa=persist, B.th=A)$_{persist}$, (loc.isa=in, B.loc=C)$_{in}$, (C.det=the)$_{the}$, (D.isa=blood, C.of=D)$_{blood}$, (C.isa=serum)$_{serum}$ $\rangle$

text level knowledge representation TKR (after assertion.)

$\langle$(1.det=a)$_a$, (1.isa=virus)$_{virus}$, (1.isa=that)$_{that}$, (2.tense=pr)$_{does}$, (2.neg=y)$_{not}$, (2.isa=persist, 2.th=1)$_{persist}$, (loc.isa=in, 2.loc=3)$_{in}$, (3.det=the)$_{the}$, (4.isa=blood, 3.of=4)$_{blood}$, (3.isa=serum)$_{serum}$ $\rangle$

ontology-level knowledge representation OKR (after global semantic interpretation)

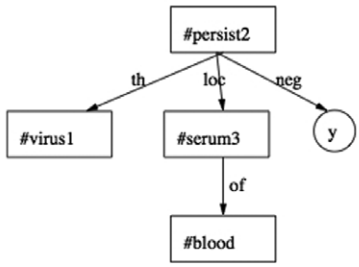


FIG. 5. Levels of representation for the utterance *a virus that does not persist in the blood serum.*

representative examples (should relative clauses be learned before noun phrases?). Thus, we implemented a second algorithm, which learns a grammar from unordered representative examples using an iterative method with theory revision. We proved that the grammar converges to the same target grammar as the previous algorithm (Muresan, 2006). This algorithm is polynomial, but less efficient than the first algorithm.

Learning from the entire generalization corpus. A potential problem with the previous two algorithms is that sometimes it is hard to know a priori which are the representative examples. We implemented an algorithm for learning from the entire generalization corpus, proving that the algorithm is still polynomial, but less efficient than the previous two (Muresan, 2011). The annotation effort is bigger than for the previous two algorithms, because now we need to annotate the entire generalization corpus with syntactic-semantic information.

Annotation of training data. To learn a LWFG, annotations for phrases, clauses, and sentences are required, in the form of syntactic-semantic representation discussed above. It is clear that even for a small corpus, which our learning model needs, writing by hand these annotations might be a difficult task. We have developed an annotation tool that, through interaction with the LWFG parser and lexicon, replaces manual assignment of full semantic representations with the manual specification of unlabeled dependencies between words (or chunks). This could be accomplished because in our framework the lexicon is given and the semantic representation of a phrase is just a concatenation of the semantic representations of its words together with variable bindings that indicate dependencies. Description of the annotation tool is left for a future publication.

In DEFINDER Evaluation we give details of learning a grammar for definitions that is used in our knowledge acquisition and querying experiments.

Semantic Parsing and Semantic Interpretation

Once a grammar for definitions is learned, we use our semantic parser and semantic interpreter to transform definitional sentences to terminological knowledge. We have the following three levels of semantic representation: the utterance level, the text level, and the ontology/terminology level (see Figure 5).

The semantic parser in conjunction with the learned grammar gives us directly the semantic representation (OntoSeR) of each utterance. This is the ***utterance level*** representation. During parsing, we have two types of representations: OntoSeR$^-$—the semantic representation obtained before the semantic interpretation constraint C_i is applied; and OntoSeR$^+$—the semantic representation after the semantic interpretation constraint C_i is applied. Thus, the semantic interpretation constraints C_i can be seen as providing a ***local level semantic interpretation***. In Figure 5, we show an example of OntoSeR$^-$ and OntoSeR$^+$ for the

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

utterance *a virus that does not persist in the blood serum*. At OntoSeR$^-$ the semantic roles of the verb *persist*, the meaning of the preposition *in*, and the relations among the nouns *blood* and *serum* are still unknown (i.e., they are variables): P1, P2, and P3, respectively. At OntoSeR$^+$, the semantic interpretation constraint C_i instantiates these variables with roles from the semantic model—theme, loc and the dummy of, respectively.

This example shows the semantic representation of several linguistic phenomena, such as relative clauses (*virus that . . .*), negation (*does not persist*), and noun compounds (*blood serum*). For readability, we indicate what part of OntoSeR corresponds to each lexical item. It can be noticed that OntoSeR contains representations of both ontological meaning (concepts and relations among concepts) as well as extra-ontological meaning, such as the verb's tense (*B.tense = pr*).

After parsing each definition, their semantic representations form the ***text level knowledge*** representation (TKR). The variables become constants, and no composition can happen at this level. In Figure 5, we see that TKR is the same as OntoSeR$^+$ except that the variables are constants (e.g., A becomes 1, B becomes 2).

To transform these representations to knowledge (***ontology-level knowledge*** representation (OKR), we use a semantic/pragmatic interpreter that implements task-specific interpretation and filtering. Although the semantic interpretation given by the constraints C_i can be seen as local semantic interpretation, the interpretation from TKR to OKR can be seen as a ***global semantic interpretation***. For our task, the task-specific interpretation is geared toward terminological interpretation. OKR is a directed acyclic graph (DAG) *G = (V,E)*. Vertices *V* are concepts (corresponding to nouns, verbs, adjectives, adverbs, pronouns, cf. Quine's criterion [Sowa, 1999]), or values of extra-ontological properties, such as y corresponding to the neg property (which in our example means that the verb is negated). Edges *E*, are semantic roles given by verbs, prepositions, adjectives and adverbs, or are extra-ontological properties, such as neg (negation). In Figure 5, we give an example of OKR for the same utterance *a virus that does not persist in the blood serum* obtained using our semantic interpreter.

In this article, the semantic interpretation (both local and global) is based only on a weak semantic model. This model is given by the admissibility relations that can be found at the level of lexical entries (i.e., we do not use synonymy, anaphora, and predefined hierarchies of concepts and roles). For the verb thematic roles we considered the thematic roles derived from Dorr's LCS Database (e.g., ag for agent, th for theme, prop for proposition) (Dorr, 1997). For adjectives and adverbs we took the roles (properties) from WordNet (Miller, 1990). For prepositions we considered the LCS Database. We also have added specific/dummy semantic roles when they were not present in these resources (e.g., of between *blood* and *serum*).

The global (task-specific) semantic interpretation is geared toward terminological interpretation. We filter determiners, and some verb forms, such as aspect, because temporal relations appear less in terminological knowledge than in factual knowledge. However, we treat modals and negation, as they are relevant for terminological knowledge. The semantic interpreter considers both concepts (e.g., #blood), and instances of concepts (e.g., #virus1, #persist2), which is key for definition understanding. Concepts are denoted in OKR by #name_concept. An instance of a concept is denoted by the name of a concept followed by the instance number (e.g., #virus1). A concept and an instance of this concept are two different vertices in OKR, having the same name. Concepts form a hierarchy based on the subsume relation (sub), which is the inverse of the isa relation. At the OKR level we have the ***principle of concept identity***, which means that there is a bijection between a vertex in OKR and a referent. For example, if we do not have pronoun resolution, the pronoun and the noun it refers to will be represented as two separate vertices in the graph. Currently, our semantic interpreter implements only a ***weak concept identity principle*** that facilitates structure sharing and inheritance (we do not have anaphora resolution, for example).

In DEFINDER Evaluation we discuss how starting with a weak semantic model and using our learned grammar and semantic interpreter, which implements the weak concept identity principle, we can get a step closer to building consumer health terminologies from definitions automatically extracted from consumer-oriented text.

Results

In this section, we present the results of our two-step approach. First, we present quantitative and qualitative evaluations of DEFINDER. Second, we present a case study for building and querying a medical terminological knowledge base using our learned grammar and our semantic parser and interpreter for terminological knowledge.

DEFINDER Evaluation

We thoroughly evaluated the DEFINDER system on several dimensions: performance of the definition extraction algorithm in terms of precision and recall; quality of the generated dictionary as judged both by nonspecialists and by medical specialists; coverage of online dictionaries (Klavans & Muresan, 2001; Muresan & Klavans, 2002). The qualitative evaluation builds on the work of Smith and Fellbaum (2004) who validate the output sentences in terms of *usefulness* (for lay users) and *correctness* (as judged by medical experts). In our study, we add two other dimensions: *readability* (for lay users) and *completeness* (as judged by medical experts).

Quantitative evaluation. A standard approach in any system evaluation is to compare the results against human performance. We selected four subjects, not trained in the medical domain and who did not participate in the development of the system. Each was provided with a set of 10

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3. Example of false positives (bold indicates the medical term, and italics its definition identified by DEFINDER).

Marked also by one human judge	The **dye**, or *contrast medium*, filters into all parts of the heart or arteries . . . A recent survey conducted by the **National Abortion Federation (NAF)**, *the professional association for abortion providers*, found that . . .
Marked by none of the human judges	This **laser therapy**, which has been touted as something that improves angina and increases exercise capacity, *is a very potent placebo*. Among the most unpleasant of dysautonomia's effects are panic attacks and **anxiety**—*mental states appropriate to a life-threatening situation*, but unhelpful in the extreme . . .

TABLE 4. Examples of missed hits (bold indicates the medical term, and italics its definition annotated by human judges).

Anaphora	**ANGIOGRAPHY** [headline] *This x-ray visualization of the arteries and veins* is used to detect abnormalities, such as aneurysms . . . **Amiadrone**. This is a *very toxic drug that can control ventricular tachycardia or fibrillation after all other drugs have failed*. [part of a bulleted list]
Complex sentence structure	If cellulitis recurs, *an underlying condition*—such as **athlete's foot**—*that predisposes a person to cellulitis* is likely and must be treated.

articles from different genres—medical articles (e.g., *Cardiovascular Institute of the South*), book chapters (e.g, *Merk Manual Home Edition, Columbia University College of Physicians and Surgeons Complete Home Medical Guide*), and newspapers (e.g., *Reuthers Health*). Each subject was asked to annotate the definitions and their defined terms in text. Instructions, including examples of how definitions can be introduced in text, were given to each subject. Because of this annotation effort, we limited the length of the articles to two pages. The annotation task was performed in an hour, on average, by each subject.

The gold-standard against which we compared our system was determined by the set of definitions marked up by at least three of the four subjects and consists of 53 definitions. Our system extracted 46 definitions, out of which 40 were present in the gold standard, leading to a performance of 86.95% precision and 75.47% recall, and 80.8% F-measure.

To perform a detailed error analysis of DEFINDER's output, we first analyzed the human performance on identifying definitions and their associated terms from text. Besides the 53 definitions, which were marked by at least three of the four subjects and which constitute our gold standard, 10 definitions were identified by two of the four subjects and eight definitions by only one subject.

Four of the six false positives identified by DEFINDER were also marked by one human subject, whereas two false positives were marked by none of the human judges. Examples of these false positives are given in Table 3.

In addition, we also encountered cases of partial matches between the gold standard and DEFINDER's output. For example, in the sentence below DEFINDER identified as definition for atherosclerosis: *the progressive narrowing of the heart's own arteries by cholesterol plaque buildups, which starves the heart itself for oxygen and nutrients*.

> The most frequent cause of the condition in older patients is atherosclerosis—the progressive narrowing of the heart's own arteries by cholesterol plaque buildups, which starves the heart itself for oxygen and nutrients.

However, only two of the four subjects marked up the underlined part, which was not retained as part of the gold standard definition for *atherosclerosis*. In computing the performance of our system, we considered such cases as correct matches.

The decrease in recall was because several definitions identified by human judges contain complex linguistic phenomena, such as anaphora, not currently handled by our system. Table 4 presents some examples of missed hits by DEFINDER when compared with the gold standard. As can be seen, the definitions introduced through anaphora in our corpus show a relatively formulaic pattern: use of the pronoun *this* after a headline or list item where a medical term appeared. Our system could be enhanced to add such patterns. However, to perform anaphora resolution outside these patterns would require specialized modules for anaphora resolution. Existing systems for anaphora resolution (supervised and unsupervised) have been developed mostly for newswire, requiring adaptation techniques to be successfully applied to the medical text (Haghighi & Klein, 2009; Ng, 2008).

The error analysis performed on DEFINDER's output shows that the task of identifying definitions in text is not a trivial task even for humans to perform. Although we considered the gold standard to consist only of definitions marked by at least three of the four annotators, we saw that two thirds of the false positives identified by DEFINDER were also annotated by one human judge. For missed hits, most of the issues were related to anaphora (the medical term was not part of the same sentence as its definition, the demonstrative pronoun *this* being used in its place).

Qualitative evaluation: lay user perspective. As discussed in Sager (1990) an important aspect of the need for definitions is the user requirements. Satisfying both the specialist and the layman with a single definition of a technical term is perhaps an unachievable task. Thus, in our next evaluation, our aim was to compare the quality of our lay dictionary against existing specialized dictionaries from the perspective of nonspecialist users.

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We selected the UMLS Metathesaurus and the On-line Medical Dictionary (OMD).[1] UMLS is one of the leading knowledge sources in the medical domain developed by the National Library of Medicine, whereas OMD is a widely used specialized medical dictionary. Eight subjects, not medical experts, were provided with a list of 15 randomly chosen medical terms and their definitions from the UMLS, OMD, and the definitions extracted by DEFINDER from online consumer-oriented medical text. The terms were among those in our gold standard. The source of each definition was not given in order not to bias the experiment. Also, the order of definitions was randomly changed for each term. The task was to assign each definition a quality rating (QR) for *usefulness* and *readability* on a scale of 1 to 7 (1 *very poor*, 7 *excellent*). *Usefulness* means that the definition can help the user understand the term, whereas *readability* means that the definition is not technical, thus it is easy to read. We want to mention that when we asked about "understanding," we did not administer a comprehension test.

Statistical significance tests were performed for subjects and terms using Kendall's coefficient of concordance, W (Siegel & Castellan, 1988) and the sign test (Siegel & Castellan, 1988). We first measured the Average Quality Rating (AQR) for each definition source on these two criteria (see Figure 6(a)). Our hypotheses were that DEFINDER outperforms both UMLS and OMD in terms of usefulness and readability. For usefulness, our system was rated **6.17**, whereas OMD was rated **4.9** and UMLS was rated **3.93**. In terms of readability, the difference was slightly higher: **6.63** for DEFINDER compared with **5.3** for OMD and **4.18** for UMLS. To statistically validate our results, we applied the sign test (Siegel & Castellan, 1988). Our results were statistically significant for both Usefulness and Readability ($p = 0.0003$).

One question that arises in computing the AQR is whether the high scores given by one subject can compensate for the lower values given by other subject, thus introducing noise in comparing the results. To address this issue we performed a second analysis to evaluate the relative ranking of the three definitional sources. Using Kendall's coefficient of correlation, W (Siegel & Castellan, 1988), we first measured the interjudge agreement on each term and for terms with significant agreement we compute the level of correlation between them. If W was significant, we compared the overall mean ranks of the three sources. We tested the same hypotheses: DEFINDER is better than UMLS and OMD both in terms of usefulness and readability. As Figure 6(b) shows, DEFINDER indeed outperformed the specialized dictionaries. We obtained statistically significant W values (W = 0.54 and W = 0.45 at $p = 0.01$ and $p = 0.05$, respectively).

Qualitative evaluation: Medical specialist perspective. The results of the previous section show that the definitions extracted from consumer-oriented text by DEFINDER are more readable and useful for the lay user than are specialized dictionaries. However, one additional question that arises is if they are also accurate and complete from a medical point of view.

To answer this question, we performed a follow-up user-based evaluation in which we asked medical specialists to rate DEFINDER's definitions in terms of *accuracy* (same as *correctness* in Smith & Fellbaum, [2004]), and *completeness*. Fifteen medical specialists (physician assistants, nurse practitioners, residents, and medical students) were provided with the same set of 15 medical terms and the definitions extracted by DEFINDER from text as the one given in the previous section. They were asked to judge the accuracy and completeness of the definitions on a scale from 1 to 7 (1 *very poor*, 7 *excellent*).

DEFINDER definitions were rated on average 5.87 for accuracy and 5.38 for completeness. The results indicate that consumer-oriented text can be a valuable source of high-accuracy definitions.

Coverage of existing dictionaries. In the introduction, we claimed that online dictionaries are incomplete and our system can be used to fill in the gaps. To evaluate this, we selected two specialized dictionaries (the UMLS Metathesaurus and the On-line Medical Dictionary) and one popular glossary (the Multilingual Glossary of Technical and

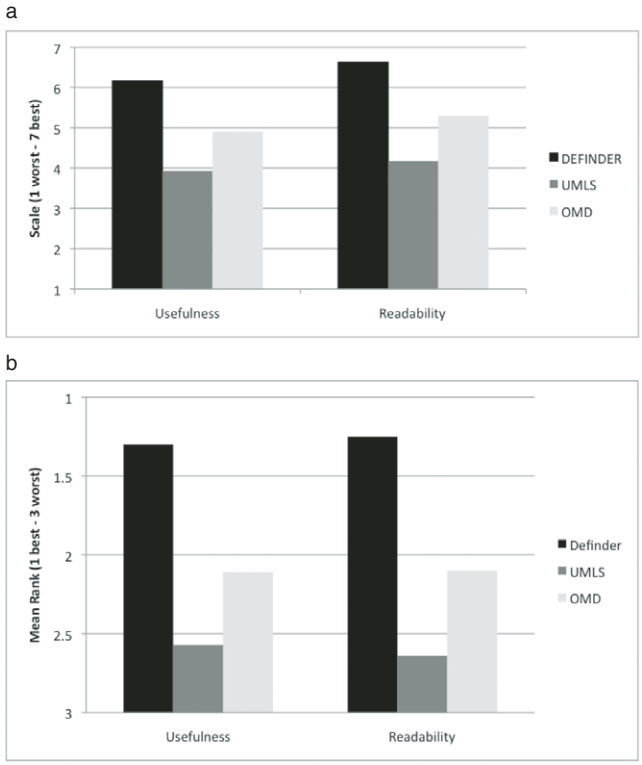


FIG. 6. Qualitative evaluations in terms of Usefulness and Readability: a) Average Quality Rating; b) Ranking.

[1]At the time of this evaluation OMD could be found at http://www.graylab.ac.uk/omd as technical dictionaries.

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5. Coverage of online dictionaries.

Term	UMLS	OMD	GPTMT
defined	60% (56)	76% (71)	21.5% (20)
undefined	24% (22)	–	–
absent	16% (15)	24% (22)	78.5% (73)

Popular Medical Terms [MGTPMT]).[2] For this evaluation, we included the terms from our gold standard used in DEFINDER Quantitative Evaluation for which DEFINDER correctly identified the definitions (40 terms) and an additional set of 53 terms from DEFINDER's output run on our development set of consumer-oriented text (these terms and their associated definitions are part of the sample corpus used to assess the conceptual and linguistic properties discussed in DEFINDER: Extracting Definitions from Consumer-Oriented Text).

Three cases were found:

- the term is listed in one of the online dictionaries and is defined in that dictionary (defined)
- the term is listed in one of the online dictionaries but does not have an associated definition (undefined)
- the term is not listed in one of the online dictionaries (absent)

Results are presented in Table 5. Looking at the UMLS results we noticed that 24% of terms were undefined, which in the field of lexicography could mean that these terms are part of the axiomatic vocabulary (i.e., they are basic terms, which are used to define other terms). But the question is whether these terms are really known by the lay users (e.g., *Holter monitor*, or *coumadin*)? In the case of the popular dictionary (MGTPMT), only 20 of the 93 terms were present, thus achieving only 21.5% coverage. This lack of coverage might be explained by the fact that this glossary is focused on terms that pertain to the domain of medication information.

These results encourage us to believe that automatically building dictionaries from high-quality consumer-orientedtext is a valuable enhancement to existing definitional resources in the medical domain.

Acquisition and Querying of Consumer Health Terminologies

In the previous section, we showed that DEFINDER can be used to automatically extract high-quality medical definitions from consumer-oriented medical text, which sometimes are not present in existing dictionaries. However, DEFINDER output is only useful to human consumption (similar to how glosses are used in WordNet, or definitions in UMLS Metathesaurus). To create resources useful for consumer health information systems, we need to transform these definitions into terminological knowledge (to build consumer health terminologies and ontologies as the ultimate goal).

[2]http://allserv.rug.ac.be/\%7Ervdstich/eugloss/welcome.html

We conducted an acquisition and querying experiment whose purpose is two-fold: (1) to show qualitatively that by learning a syntactic-semantic grammar for definitions and by using our parser and semantic interpreter, we can acquire terminological knowledge from natural language definitions extracted by DEFINDER from text, and can query this knowledge using natural language questions, obtaining precise answers at the concept level; and (2) to show that the local semantic interpretation at the grammar rule level C_i could help in disambiguation, even if it is based on a weak semantic model.

Before describing our acquisition and querying experiment, we briefly present our method for learning a syntactic-semantic grammar for definitions. The grammar was learned using the LWFG formalism and grammar learning model described above. We chose the representative examples guided by the type of phenomena we wanted to model and which occurred in the sample corpus of medical definitions we used to assess the conceptual and linguistic properties presented in DEFINDER: Extracting Definitions from Consumer-Oriented Text (80 definitions). The phenomena included complex noun phrases (e.g., noun compounds, nominalization), prepositional phrases, relative clauses and reduced relative clauses, finite and nonfinite verbal constructions (including tense, aspect, negation, and subject-verb agreement), link verb *to be*, raising and control constructions. Because our goal is to query the acquired terminological knowledge using natural language questions, we also learned grammar rules for wh-questions (including long-distance dependencies). To learn the grammar, we annotated 151 representative examples, and 448 examples were used as a generalization corpus. Annotating these examples requires knowledge about categories and their attributes. We used 31 syntactic categories (e.g., NP, ADJP) and 37 attributes (e.g., category, number, person). Regarding the lexical items, we used a total number of 13 lexical categories (i.e., parts of speech) and 46 elementary syntactic-semantic templates. For example, the nouns have three types of elementary syntactic-semantic templates, which correspond to basic nouns, modifier nouns (e.g., in case of noun compounds) and nominalizations (where the semantic representation is similar to the representation of a verb). For grammar learning only a reduced lexicon, not domain specific, is needed (e.g., only a few lexical items are given for every open word class, such as nouns [20], verbs [13, six of which are for raising and control verbs], adjectives [14], adverbs [nine], proper nouns [four]).

Once the grammar has been learned, for the acquisition and querying experiment we automatically built a larger lexicon from COMLEX (Grishman, Macleod, & Meyers, 1994) and the UMLS lexicon (Lindberg, Humphreys, & McCray, 1993), which is a medical lexicon. This is because the definitions extracted from consumer-oriented text contain both general and technical vocabulary.

The corpus of definitions used in the acquisition and querying experiment consists of the correctly extracted definitions by DEFINDER, which were used in DEFINDER's evaluation and which were different from our sample corpus

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1. Hepatitis is a disease caused by infectious or toxic agents and characterized by jaundice, fever and liver enlargement.
2. Hepatitis A is an acute but benign viral hepatitis caused by a virus that does not persist in the blood serum.
3. Hepatitis B is an acute viral hepatitis caused by a virus that tends to persist in the blood serum.

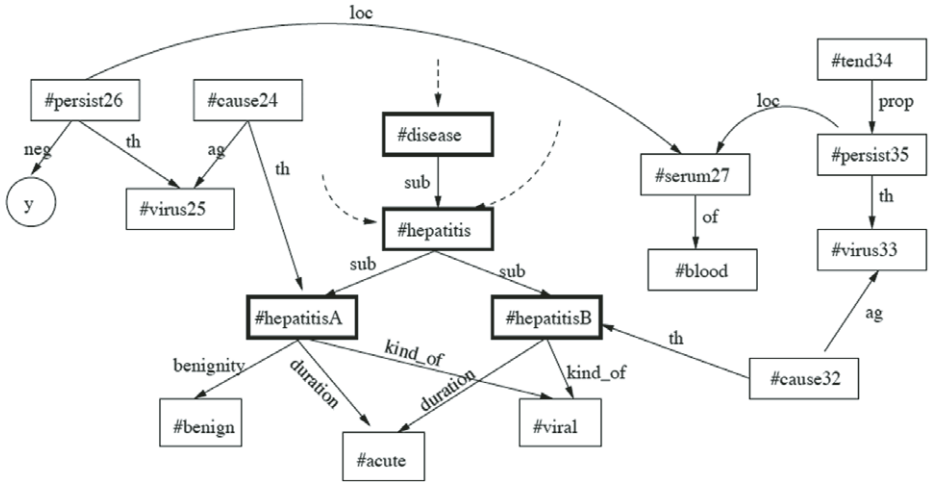


FIG. 7. Terminological knowledge acquired from consumer health definitions.

used to assess the conceptual and linguistic properties (used in building the representative examples for grammar learning). In the next two sections we present and discuss the acquisition and querying experiments.

Acquisition of terminological knowledge from consumer health definitions. In this experiment, we tested the use of the learned grammar, semantic parser, and the semantic interpreter based on a weak semantic model to acquire terminological knowledge from consumer health definitions. Although our grammar covered all the constructions present in the corpus of definitions, we obtain besides the correct semantic representations also incorrect semantic representations, which shows that our weak semantic model is not enough to remove all erroneous parses. To gain further insight, we looked at the number of alternative semantic representations obtained with and without our local semantic interpreter C_i. Without C_i, the average number of representations obtained by the parser is 2.53 per definition. After C_i is applied, the average number of different representations obtained for a definition is 2.00. This result shows that even with a weak semantic model our semantic interpreter helps remove some erroneous parses. However, it is not enough to obtain only the correct semantic analysis in all cases. Thus, we developed the system to allow a user to manually select the correct OKR (i.e., the graph-based representation), which was then added to the knowledge base. The selection of the OKR-level of representation for human validation is because of the fact that this representation is much more readable for a user than the OntoSeR levels (as can be seen from Figure 5).

In order to further discuss the processes of knowledge acquisition, we present an example of constructing a hierarchy of concepts from definitions of *hepatitis*, *Hepatitis A* and *Hepatitis B*. The definitional text and OKRs are presented in Figure 7, OKR being shown only for the last two definitions for readability reasons. The acquisition of knowledge can be done directly, because we consider both concepts (#hepatitis, #blood) and instances of concepts (#virus25, #virus33) in our OKR representation (Nirenburg & Raskin, 2004).

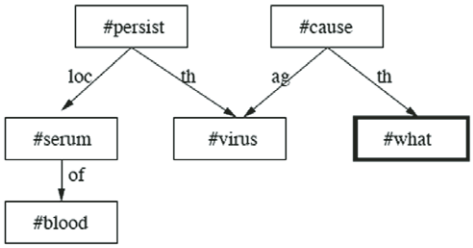
The defined term is always a concept, and it is part of the sub hierarchy. The concepts in the sub hierarchy are presented in bold in Figure 7. All the definitional properties of concepts are directly linked to the concept vertex (facilitated by our interpretation of copula *be*-predicative). For example, even if in the text we have *Hepatitis B is an acute viral hepatitis*}, the properties "acute" and "viral" are linked to the concept #hepatitisB and not to the concept #hepatitis. This is obtained because only #hepatitis was previously part of the sub hierarchy. If the concept #viral_hepatitis is present, then this most specific concept is selected as the direct parent of #hepatitisB.

In addition to the concepts that are defined, we can also have concepts that are referred (i.e., they are part of the definition of a medical term), if they do not have any modification (e.g., #blood in definition of *Hepatitis A*, and *Hepatitis B*). If a referred concept has modifications, it is represented as an instance of a concept in OKR. As a consequence, various verbalizations of concept properties can be differentiated in OKR, allowing us to obtain direct answers that are specific to each verbalization. For example, the term *virus* appears in the definition of both *Hepatitis A* and *Hepatitis B*. In OKR, they are two different instances of a concept, #virus25 and #virus33, because they have different modifications: *persists in the blood serum*, and *does not persist in the blood serum*, respectively. These modifications are an essential part of the differentia of the two

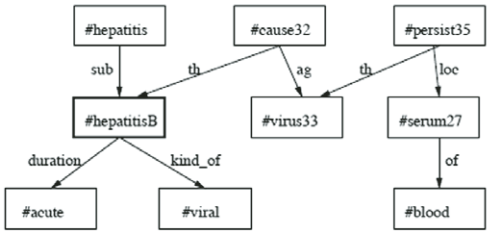
DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

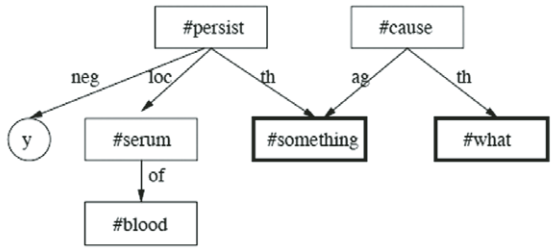
Q1: What is caused by a virus that persists in the blood serum?



A1: #hepatitisB



Q2: What is caused by **something** that does not persist in the blood serum?



A2: #hepatitisA

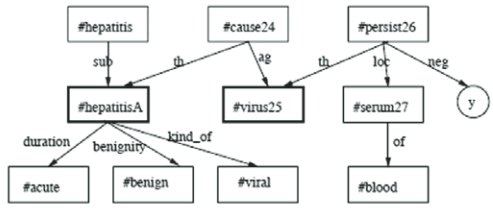


FIG. 8. Examples of precise and vague questions, their OKR representations, and the concept-level answers.

concepts #hepatitisA and #hepatitisB, causing the distinction between the two. When we ask the question: *What is caused by a virus that persists in the blood serum*? (Q1 in Figure 8), we obtain only the correct answer #hepatitisB (A1 in Figure 8).

Another important aspect that shows the adequacy of our representation for direct acquisition and query is the OKR-equivalences that we obtain for different syntactic forms. They are related mainly to verbal constructions. Among OKR-equivalences we have (1) active and passive constructions; (2) *-ed* and *-ing* verb forms in reduced relative clauses are equivalent to passive/active verbal constructions (e.g., the question can be formulated in present tense and active voice: *What causes hepatitis A?*, whereas the answer is obtained from a definitional statement involving the reduced relative clause: *hepatitis A is an acute but benign viral hepatitis caused by a virus . . .*; and 3) constructions involving raising verbs, where we can take advantage of the fact that the controller is not the semantic argument of the raising verb (e.g., in the definition of *Hepatitis B* we have: *". . . caused by a virus that tends to persist in the blood serum* (*virus* is not the argument of the raising verb *tends* but the argument of the verb *persists*), whereas the question can be asked without the raising verb: *What is caused by a virus that persists in the blood serum?*).

A consequence of our weak concept identity principle is that we have structure sharing among OKRs (e.g., the OKRs of *Hepatitis A* and *Hepatitis B* share the representation corresponding to *blood serum* [#serum27, #blood]), as well as hierarchies of concepts and inheritance.

Because in our experiment we only used a weak semantic model given by the admissibility relations that we can find at the level of lexical entries, our qualitative evaluation seems to support the intuition that a lexicon can sometimes be the basis for the development of a practical ontology (Hirst, 2004). However, although the knowledge we obtained (OKR) does have properties such as structure sharing, inheritance, hierarchies of concepts, relations among concepts, we cannot claim at this point that this knowledge is an actual ontology, which will imply a deeper level of formalization, and also application of a strong concept identity principle dealing with synonymy and anaphora. But this experiment shows that our framework allows us to build consumer health terminological knowledge from definitions extracted from consumer-oriented medical text.

Natural language querying. Besides acquisition of terminological knowledge, our grammar and semantic interpreter facilitates natural language querying of the acquired terminological knowledge by treatment of wh-questions. For this experiment, we created a benchmark of 29 questions. The type of questions we used are "Who did what to whom?" that is only questions regarding the verbs' arguments. Because in our knowledge base we obtained a hierarchy of concepts (an example of hierarchy is given in Figure 7), the questions can be related to this hierarchy, for example, the question *Which are viral diseases?* has as answer #hepatitisA and #hepatitisB, even if their direct parent is #hepatitis and not #disease. Because OKR is a direct acyclic graph, the natural language querying is reduced to a graph matching problem. A question is a subgraph of the utterance graph where the wh-word substitutes the answer concept. An answer is a vertex in the OKR of an utterance, together with all the edges incident from/to it. We have experimented both with precise and vague questions. An example of a vague

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

question is the following: *What is caused by something that does not persist in the blood serum?*, where *something* is considered as a variable concept that will match a vertex in the OKR. We obtain a precise answer at the concept level (see example in Figure 8). A practical advantage of being able to handle vague questions is that we can obtain all the concepts that are in a particular relation with other concepts, or that have particular properties. For questions we have an average of 6.06 semantic representations per question without C_i validation. After semantic validation, we have an average of 2.35 semantic representations per question. In this experiment though, even if the weak semantic model is not always enough to eliminate incorrect semantic representations of questions, we only obtain the correct answer(s), because we match the OKRs of these questions against the manually validated knowledge base.

Discussion

This article has introduced a new method to automatically acquire consumer health terminological knowledge from natural language definitions extracted from consumer-oriented medical text.

Lay Versus Technical Terminology

The main reason for focusing on consumer-oriented medical articles for our task was that these articles are written by doctors for lay users and thus terminology must be defined. If we want to adapt DEFINDER to technical text (rather than lay text), the question is whether technical articles, such as medical journals, or books, contain definitions of medical terms, given that their audience consists of medical experts. We analyzed a set of five technical articles in cardiology, oncology, and gastroenterology, each ranging from three pages to 10 pages long (total of 37 pages of technical text). We found two definitions of medical terms. In comparison, our quantitative evaluation of DEFINDER showed that 10 consumer-oriented articles, each limited to two pages, with a total of 17 pages of consumer-oriented text, contained 53 definitions. Thus, the ratio of pages to definitions was 37:2 for technical text, compared to the ratio of 17:53 for consumer-oriented text.

However, we estimate that if larger technical corpora are used, more definitions could be present: either definitions for new terms, or new definitions for existing terms. The techniques described in this article are general and can be applied to technical text. DEFINDER relies on shallow pattern matching, which can be generalized to technical articles.

In order to build medical terminologies from technical definitions—either extracted from text or from existing technical dictionaries—the same methodology described in this paper could be used (learning a syntactic-semantic grammar and using our semantic parser and semantic interpreter). Technical definitions differ from lay definitions both in their syntactic constructions and semantic concepts. As discussed in the next section, our grammar learning method is general and can be used to learn additional constructions by simply annotating a small set of representative examples (using a general lexicon). The difference in vocabulary between technical and lay definitions can be addressed by using a specialized semantic resource (such as the UMLS) instead of general resources, such as WordNet and the LCS Database, which were used in our experiments.

General Grammar Learning Versus Task/Domain Specific Semantic Interpreter

The method for mapping text to knowledge introduced in this article relies on a general grammar learning framework and a task-specific semantic interpreter. Learning is done based on annotated examples that do not contain domain-specific roles or concepts, and thus our learning framework is general. We can use any semantic model (domain ontology), depending on the application.

The semantic interpreter used in this article is targeted for terminological knowledge and currently uses a weak semantic model. As we deal only with weak semantic context given by the admissibility relations that we can find at the level of lexical entries, our qualitative evaluation supports the claim in Hirst (2004) that a lexicon can often serve as a useful basis for the development of a practical ontology. The weak concept identity guarantees the same OKR representation for different syntactic forms (e.g., nominalizations vs. verbal forms; active vs. passive voice; *-ing* and *–ed* forms of reduced relative clauses vs. active/passive forms of verbs), or different forms of tense and aspect, which are filtered. Because we focus on terminological knowledge, modals and negation are important, whereas temporal reasoning is not. However, if we would not filter tense and aspect, the semantic interpreter could be further developed towards temporal reasoning needed for factual knowledge bases. The weak concept identity principle (structure sharing and use of both concepts and instances of concepts) allows us to merge multiple definitions of the same term (e.g., Table 1). Merging will be a key element in building a terminological knowledge base from DEFINDER's output on a large scale. The reason is that terms can be defined differently depending on the context of the article, and thus we need to merge all the information related to this term when creating the knowledge base.

Conclusions and Future Work

In this article, we presented a two-stage architecture for building consumer health terminologies from text: (1) *automatic extraction of definitions from consumer-oriented articles and web documents*, which reflects language in use, rather than relying solely on dictionaries, and (2) *learning* a grammar that directly maps natural language to graph-based meaning representations rather than using hand-written patterns or grammars. Grammar learning is done based only on a small set of annotated examples and a generic lexicon, which makes our approach appealing for domain adaption.

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We presented a qualitative evaluation that shows that our learned grammar in conjunction with a semantic interpreter targeted to terminological interpretation allows us to acquire consumer-health terminological knowledge from definitions automatically extracted from consumer-oriented text, and to query this knowledge using natural language questions, obtaining precise answers at the concept level. We showed that the definitions extracted by DEFINDER are fairly accurate and complete as judged by medical specialists and also more understandable and readable to lay users than technical medical definitions.

We plan to extend this work in two main directions. The first direction is to build a larger corpus of annotated definitions in consumer-oriented medical articles using crowdsourcing techniques, such as Amazon Mechanical Turk. One challenge of using crowdsourcing is the quality of annotations. We will rely on majority voting and additional techniques used in the literature to filter noisy annotations. This larger gold standard corpus will allow us to enhance DEFINDER with machine learning techniques, as well as to evaluate our grammar learning, semantic parser and interpreter on a larger scale. The second direction is related to our semantic interpreter. In this article, we used a weak semantic model for role/property assignment and a weak concept identity principle, which did not take into account synonymy or anaphora. We plan to develop the interpreter to handle semantic equivalences based on synonymy and anaphora (either by using resources such as WordNet, when appropriate, or by learning these semantic equivalences). We also plan to enhance the grammar and semantic model with probabilities in order to further remove ambiguities that we noticed in our experiments.

Based on these two main future steps, our two-step methodology could be further exploited to build a terminological knowledge base for lay users on a larger scale, thus filling gaps in existing terminological resources for consumer health information systems.

Acknowledgments

This research was initially supported by the National Science Foundation under Digital Library Initiative Phase II (IIS-98-17434.0415865), while the authors were at Columbia University. We thank the medical residents, nurses and patients at the New York Presbyterian Hospital who participated in the evaluation of our DEFINDER system. Any opinions, findings and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Aronson, A. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the Metamap program. In Proceedings of AMIA Symposium (pp. 17–21). Washington, DC.

Blair-Goldensohn, S., McKeown, K., & Schlaikjer, A. (2004). Answering definitional questions: A hybrid approach. In New Directions in Question Answering.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research, 32(No. Database issue), D267–D270.

Bodenreider, O. (2006). Lexical, terminological and ontological resources for biological text mining. In Text mining for biology and biomedicine (pp. 43–66). Boston: Artech House.

Brill, E. (1992). A simple rule-based part of speech tagger. In Proceedings of the 3rd Applied Natural Language Processing Conference (ANLP '92). Trento, Italy.

Burgun, A., & Bodenreider, O. (2001). Comparing terms, concepts and semantic classes in WordNet and unified medical language system. In Proceedings of the Workshop on WordNet and other Lexical Resources: Applications, Extensions and Customizations (pp. 77–82). Pittsburgh, PA.

Cardillo, E (2011). A lexi-ontological resource for consumer healthcare: The Italian consumer medical. Vocabulary (PhD thesis). University of Trento, Italy.

Ceusters, W., Capolupo, M., Moor, G.D., Devlies, J., & Smith, B. (2011). An Evolutionary Approach to Realism-Based Adverse Event Representations. Methods of Information in Medicine, 50(1), 62–73.

Charniak, E. (2000). A maximum-entropy-inspired parser. In Proceedings of the first conference on North American chapter of the Association for Computational Linguistics (NAACL-2000). Seattle, WA.

Chodorow, M.S., Byrd, R.J., & Heidorn, G. (1985). Extracting semantic hierarchies from a large on-line dictionary. In Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (pp. 299–304). Chicago.

Clark, S., & Curran, J.R. (2007). Wide-coverage efficient statistical parsing with ccg and log-linear models. Computational Linguistics, 33(4), 493–552.

Collins, M. (1999). Head-driven statistical models for natural language parsing (Ph.D. thesis). University of Pennsylvania, Philadelphia, PA.

Dorr, B.J. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. Machine Translation, 12(4), 271–322.

Eck, K.E., & Meyer, I.E. (1995). Bringing Aristotle into the 20th century: Computer-assisted definition construction in a terminological knowledge base. In S.E. Wright & R.A. Strehlow (Eds.), Standardizing and harmonizing terminology: Theory and practice (pp. 83–101). Philadelphia, PA: ASTM.

Elhadad, N. (2006). Comprehending technical texts: Predicting and defining unfamiliar terms. In Proceedings of AMIA Annual Symposium (pp. 239–243). Washington, DC.

Elhadad, N., & Sutaria, K. (2007). Mining a Lexicon of Technical Terms and Lay Equivalents. In Proceedings of ACL BioNLP Workshop (pp. 49–56). Prague, Czech Republic.

Fahmi, I., & Bouma, G. (2006). Learning to identify definitions using syntactic features. In Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications. Trento, Italy.

Friedman, C., Borlawsky, T., Shagina, L., Xing, R., & Lussier, Y. (2006). Bio-ontology and text: Bridging the modeling gap. Bioinformatics, 22(19), 2421–2429.

Grishman, R., Macleod, C., & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. In Proceeding of 15th International Conference on Computational Linguistics (COLING-1994). Kyoto, Japan.

Haghighi, A., & Klein, D. 2009. Simple coreference resolution with rich syntactic and semantic features. In Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP) (pp. 1152–1161). Singapore.

Hahn, U., Romacker, M., & Schulz, S. (2000). Medsyndikate–design considerations for an ontology-based medical text understanding system. In Proceedings of The American Medical Informatics Association (AMIA) Symposium (pp. 330–334). Los Angeles, CA.

Hirst, G. (2004). Ontology and the lexicon. In S. Staab, & R. Studer (Eds.), Handbook on ontologies. (pp. 209–229). Berlin, Germany: Springer.

Joshi, A. (1985). How much context-sensitivity is necessary for characterizing structural descriptions. In D. Dowty, L. Karttunen, & A. Zwicky

DOI: 10.1002/asi

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Eds.), Natural language processing: Theoretical, computational, and psychological perspectives (pp. 206–250). New York, NY: Cambridge University Press.

Joshi, A., & Schabes, Y. (1997). Tree-adjoining grammars. In G. Rozenberg & A. Salomaa (Eds.), Handbook of formal languages, Vol. 3 (pp. 69–124). Berlin, New York: Springer.

Keselman, A., Logan, R., Smith, C., Leroy, G., & Zeng-Treitler, Q. (2008a). Developing informatics tools and strategies for consumer-centered health communication. Journal of American Medical Informatics Association, 15(4), 473–483.

Keselman, A., Smith, C.A., Divita, G., Kim, H., Browne, A.C., Leroy, G., & Zeng-Treitler, Q. (2008b). Consumer health concepts that do not map to the UMLS: Where do they fit? Journal of American Medical Informatics Assoc, 15, 496–505.

Klavans, J., Chodorow, M., & Wacholder, N. (1992). Building a knowledge base from parsed definitions. In G. Heidorn, K. Jensen, & S. Richardson (Eds.), Natural language processing: The PLNLP approach (pp. 119–133). New York: Kluwer.

Klavans, J., & Muresan, S. (2000). DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In Proceedings of the American Medical Informatics Association Symposium (AMIA-2000).

Klavans, J., & Muresan, S. (2001). Evaluation of DEFINDER: A system to mine definitions from consumer-oriented medical text. In Proceedings of The First ACM+IEEE Joint Conference on Digital Libraries.

Klein, D., & Manning, C.D. (2003). Accurate Unlexicalized Parsing. In Proceedings of the 41st Meeting of the Association for Computational Linguistics (pp. 423–430). Sapporo, Japan.

Lease, M. & Charniak, E. (2005). Parsing biomedical literature. In Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP '05). Jeju Island, Korea.

Lindberg, D., Humphreys, B., & McCray, A. (1993). The unified medical language system. Methods of Information in Medicine, 32, 281–291.

Liu, B., Chin, C., & Ng, H. (2003). Mining topic-specific concepts and definitions on the web. In Proceedings of the 12th International Conference on World Wide Web. Budapest, Hungary.

Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., & Schasberger, B. (1994). The Penn Treebank: annotating predicate argument structure. In Proceedings of the Workshop on Human Language Technology (pp. 114–119). Plainsboro, NJ.

McKeown, K., Chang, S.-F., Cimino, J., Feiner, S., Friedman, C., Gravano, L., et al. (2001). Persival, a system for personalized search and summarization over multimedia healthcare information. In Proceedings of The First ACM+IEEE Joint Conference on Digital Libraries (JCDL). Roanoke, VA.

Miller, G. (1990). WordNet: An on-line lexical database. Journal of Lexicography, 3(4), 235–312.

Moldovan, D.I., & Rus, V. (2001). Logic form transformation of WordNet and its applicability to question answering. In Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL 2001) (pp. 394–401).

Muresan, S. (2006). Learning constraint-based grammars from representative examples: Theory and applications (PhD Thesis). Columbia University, New York.

Muresan, S. (2008). Learning to map text to graph-based meaning representations via grammar induction. In Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing (pp. 9–16). Manchester, UK.

Muresan, S. (2010). A learnable constraint-based grammar formalism. In Proceedings of 23rd International Conference on Computation Linguistics (COLING). Beijing, China.

Muresan, S. (2011). Learning for Deep Language Understanding. In Proceedings of International Joint Conference on Artificial Intelligence, IJCAI-11. Barcelona, Spain.

Muresan, S., & Klavans, J.L. (2002). A method for automatically building and evaluating dictionary resources. In Proceedings of the Language Resources and Evaluation Conference (LREC-2002).

Muresan, S., & Rambow, O. (2007). Grammar approximation by representative sublanguage: A new model for language learning. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic.

Navigli, R., & Velardi, P. (2008). From glossaries to ontologies: Extracting semantic structure from textual definitions. Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Series information for Frontiers in Artificial Intelligence and Applications (pp. 71–87).

Nirenburg, S., & Raskin, V. (2004). Ontological semantics. Cambridge, MA: MIT Press.

Nowlan, W.A., Rector, A.L., Rush, T.W., & Solomon W.D. (1994). From terminology to terminology services. Proceedings of the Annual Symposium on Computer Application in Medical Care.

Ng, V. (2008). Unsupervised models for coreference resolution. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 640–649). Edinburgh, Scotland.

Poon, H., & Domingos, P. (2009). Unsupervised semantic parsing. In Proceedings of EMNLP '09. Singapore.

Ramshaw, L.A., & Marcus, M.P. (1995). Text chunking using transformation-based learning. In Proceedings of Third ACLWorkshop on Very Large Corpora. Cambridge, MA: MIT Press.

Richardson, S., Dollan, W., & Vanderwende, L. (1998). MindNet: Acquiring and structuring semantic information from text. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL '98). Montreal, CA.

Rosse, C., & Mejino, J.L. Jr. (2003). A reference ontology for biomedical informatics: The foundational model of anatomy. Journal of Biomedical Informatics, 36(6), 478–500.

Rus, V. (2002). Logic form for wordnet glosses and application to question answering (Ph.D. thesis). Computer Science Department, School of Engineering, Southern Methodist University, Dallas, Texas.

Sager, J. (1990). A practical course in terminology processing. Amsterdam: John Benjamins Publishing Co.

Siegel, S., & Castellan, N. (1988). Non-paprametric statistics for the behavioural sciences (2nd ed.). New York: McGraw Hill.

Smith, B.F., & Fellbaum, C. (2004). Medical Wordnet: a new methodology for the construction and validation of information resources for consumer health. In Proceedings of the 20th International Conference on Computational Linguistics (COLING). Geneva, Switzerland.

Sowa, J.F. (1999). Knowledge Representation: Logical, Philosophical, and Computational Foundations. Pacific Grove, CA: Brooks Cole Publishing Co.

Steedman, M. (1996). Surface structure and interpretation. Cambridge, MA: MIT Press.

Steedman, M. (2000). The syntactic process. Cambridge, MA: MIT Press.

Weisman, H.M. (1992). Basic technical writing. New York: Merrill.

Wilks, Y.A., Slator, B.M., & Guthrie, L. (1996). Electric words: Dictionaries, computers, and meanings. Cambridge, MA: MIT Press.

Wong, Y.W., & Mooney, R. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007). Prague, Czech Republic.

Zeng, Q., & Tony, T. (2006). Exploring and developing consumer health vocabularies. Journal of American Medical Informatics Association, 13(1), 24–9.

Zettlemoyer, L.S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In Proceedings of Uncertainty in Artificial Intelligence UAI-05. Edinburgh, Scotland.

15322890, 2013, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/asi.22787 by Columbia University Libraries, Wiley Online Library on [12/08/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License