Relying on Discourse Trees to Extract Medical Ontologies from Text

Boris Galitsky[1], Dmitry Ilvovsky[2], and Elizaveta Goncharova[2(✉)]

[1] Oracle Inc., Redwood Shores, Redwood City, CA, USA
boris.galitsky@oracle.com
[2] National Research University Higher School of Economics, Moscow, Russia
{dilvovsky,egoncharova}@hse.ru

Abstract. We explore the role of discourse analysis in ontology construction. While extracting candidate phrases to form ontology entries from text, it is important to pay attention to which discourse units these phrases occur in. It turns out that not all discourse units are equal in terms of their contribution to forming ontology entries. We survey text mining and ontology information extraction techniques in medical domain and select the ones where advanced linguistic analysis including the discourse level is leveraged the most to produce a robust and efficient ontology. We evaluate the consistency of the resultant ontology and its role in assuring high search relevance using several real-life medical datasets and prove the importance of introducing discourse information into the ontology construction.

Keywords: Medical ontology · Discourse analysis · Text mining · Question answering

1 Introduction

Building and adapting medical ontologies is a complex task that requires substantial human effort and close collaboration between domain experts (e.g. health professionals) and knowledge engineers. Even if automatic ontology construction techniques are mature enough to support this task [22], they provide only partial solutions, and manual interventions from healthcare professionals will always be necessary if high quality is expected. The use of ontologies in medicine is mainly focused on the representation of medical terminologies. Healthcare professionals use them to represent knowledge about symptoms and treatments of diseases. Pharmaceutical enterprises use ontologies to represent information about drugs, dosages, and allergies.

Ontologies are a foundation for numerous Decision Support Systems (DSS) used to support medical activities; therefore, the quality of the underlying ontologies affects the performance of DSSs that rely on them. In consequence, automatically-built medical ontologies (including schema knowledge and individuals descriptions) must be validated by domain experts. However, healthcare professionals are usually not fluent in ontology management and must be

S. M. Kovalev et al. (Eds.): RCAI 2021, LNAI 12948, pp. 215–231, 2021.
https://doi.org/10.1007/978-3-030-86855-0_15

assisted by knowledge engineers during the validation process, which can potentially extend errors and inconsistencies.

Extracting medical or clinical information from health records, which contain important items such as eligibility criteria, a summary of diagnosing results, and prescribed drugs, is an important task, especially with the adoption of electronic health records (EHR). These records are normally stored as free-form text documents and contain valuable unstructured information that is essential for better decision-making for a patient's treatment. Gaining insight from a tremendous amount of unstructured clinical data has been a critical and challenging issue for medical organizations. Having an automated system that is able to read patients' medical reports, extract medical entities, and analyze the extracted data is not only desirable but also a necessary component of medical organizations' routine. The challenging part of this task is how to extract and encode the unstructured data to improve an overall healthcare system.

Information extraction (IE) and text mining (TM) is a potentially suitable technique here. There are three major elements that should be extracted from these clinical records: entities, attributes, and relations between them [20]. Automatic recognition of medical entities in the unstructured text is a key component of biomedical information retrieval systems. Its applications include analysis of unstructured medical text, such as the one presented in EHR [2] or obtained from the medical social networks, and knowledge discovery from biomedical literature [15]. The extracted medical terminologies are often structured as ontologies, with the relations connecting the entities and a list of synonyms for each term.

Ontologies are a critical component for these tasks, and the quality and consistency of an ontology automatically extracted from text determine the overall DSS accuracy. The bottleneck of building concise, robust, and complete ontologies is the lack of a mechanism to extract ontology entries from reliable, authoritative parts of documents. Building ontologies, one needs to use reliable text fragments expressing the central point of a text and avoid constructing entries from additional comments, clarifications, examples, instances, and other less significant parts of the text. To overcome this challenge, we rely on discourse analysis (that has been proven useful for a number of natural language processing tasks, such as argumentation mining [13], text classification [12], and summarization [33]) to select discourse units that yield relevant ontology entries. In this paper, we present the ontology construction system that consists of several text mining blocks significant for providing efficient and complete medical ontology.

The paper is organized as follows. The first section is devoted to reviewing existing techniques for ontology construction and their limitation. Then the discourse structure, which is introduced to overcome these limitations, is defined and explained. Section 3 shows the overall system architecture for the ontology construction, which is followed by a more detailed description of the main system's components (Sects. 4–7). In Sect. 8, we provide the experiment scenarios and analyze the obtained results. We conclude in Sect. 9.

2 Related Work

2.1 Ontology Extraction in Medical Domain

Usually, an ontology presents the information as the sets of entities bound by a relation [4]. Information presented in this format is useful for many applications (mining biomedical text, ontology learning, and question answering). Ontologies structure knowledge as a set of terms with edges between them labeled by the type of relation to evoke meaningful information.

Retrieving relevant phrases that can be considered as an ontology entry is a critical component in medical ontologies construction. Initially, rule-based methods that apply lexical or syntactical templates to form the ontology entities have been used for this task. Further, they have been replaced with machine learning approaches, and lately, deep learning (DL) models have become the most popular in this domain.

The rule-based approaches provide a high level of control on the entities added into the ontology, as the syntactic or lexical templates are usually proposed by the domain expert. In [31], the authors utilize syntactical patterns to retrieve medical key phrases. These phrases are, first, grouped w.r.t. their informativeness with different weights assigned to them, and then the most relevant ones form the ontology entries. The weight of each key phrase is calculated via pair-wise mutual information. In [29], the authors propose a novel method to retrieve significant key phrases based on a naive Bayesian learning algorithm. The specificity of the approach is that it requires many statistical features and several domain-specific features to extract medical key phrases. More recent [1] introduces a method to retrieve key phrases based on heuristics that collaborate natural language processing (NLP) techniques, statistical knowledge, and the internal structural pattern of terms. In addition, DBpedia is utilized to align the terms that may be relevant to the candidate key phrases extracted from the original document. The candidate key phrases are ranked in accordance with several metrics, including the term frequency, then the candidates with the highest rank are treated as the ontology entry. The fact that expert knowledge is required to create relevant templates or generate informative features is the main disadvantage of these techniques.

In [27], the authors compare the performance of statistical and semantic approaches to medical concept extraction, and key phrases identification, specifically. They have implemented conditional random fields (CRF) for clinical named entities extraction and used MetaMap [3], an automotive system that utilizes external medical knowledge to get crucial features from texts. The authors have noticed that the use of only CRF classifier performs much better than rule-based MetaMap relying on external knowledge. However, they also have mentioned that the machine-learning method is highly dependent on the annotated training corpora, therefore, better results are obtained for well-represented classes that the model has seen during the training procedure. Finally, the authors have shown that the best performance in entity extraction is obtained from the combination of a CRF classifier with some lexical features and semantic features obtained from the domain knowledge-based method using MetaMap.

Recently, state-of-the-art results for a number of NLP tasks have been achieved by DL models. As ontology construction requires processing textual data, DL models have been adjusted to this domain. Named entity recognition (NER) models are widely applied to medical documents to retrieve candidates for ontology entries. For example, in [16], the authors combine BiLSTM and CRF models (BiLSTM-CRF) to retrieve named entities on Chinese electronic medical records. They notice that medical entities retrieval is still a big challenge for the medical domain as, first, there is no uniform standard to name medical entities; second, there may be several names for one entity, and, third, new entities are constantly being created, which is hard to follow with the predefined set of rules. The combination of BiLSTM model joining with a CRF layer introduced in the work improves the performance of NER for medical texts. In [24], the authors propose a modification of the well-known transformer-based BERT architecture to better combine general and clinical knowledge learned in the pre-training phase, and show that this model provides good performance on various medical datasets. We should mention that all these approaches are data-driven and require huge labeled medical datasets for models training that are not always available. Besides, the authors have noticed that DL models trained on some highly specialized datasets are failed to be generalized for other domains. In [2] the authors developed the novel hybrid DL-based approach, called Neural Concept Recognizer (NCR) which includes an additional neural dictionary manager that learns to generalize to novel synonyms for concepts to overcome this challenge.

The system introduced in this work includes several modules for constructing and validating medical ontologies. We apply discourse analysis to the entry recognition component of the ontology construction system. Analysis of the text discourse structure allows the system to pay more attention to relevant text fragments that yield ontology entries.

2.2 Discourse Organization of the Text

To construct the ontology from a large amount of unstructured data, which is the common way to represent information, one should be able to retrieve relevant entities from the text and identify the type of relations connecting them. We believe that this goal could be achieved by processing the discourse structure of the text.

The discourse organization of the text shows how discourse units (text spans) are related to each other. Discourse analysis reveals this structure and describes the relations that hold between text units in the document. One of the most popular theories that describe the discourse structure is Rhetorical Structure Theory (RST) [23]. RST divides a text into minimal atomic units, called Elementary Discourse Units (EDUs), and retrieves the rhetorical relation, such as *Elaboration*, *Explanation*, *Causes*, etc., that holds between these atomic text spans. RST forms a tree representation of discourse called a Discourse Tree (DT). In DT, the EDUs are the leaves, and rhetorical relations are edges. EDUs linked by a rhetorical relation are also distinguished based on their relative importance in

conveying the author's idea: the nucleus is the central part, whereas the satellite is a supportive part.

Discourse analysis leverages language features, which allow speakers to specify that they are:

- talking about something they have talked about before in the same discourse;
- indicating a relation that holds between the states, events, beliefs, etc. presented in the discourse;
- changing to a new topic or resuming one from earlier in the discourse.

Discourse can be structured by its topics, each comprising a set of entities and a limited range of things being said about them. The topic structure is common in the expository text found in schoolbooks, encyclopedias, and reference materials. A topic can be characterized by the question it addresses. Each topic involves a set of entities, which may (but do not have to) change from topic to topic. This aspect of structure has been modeled as entity chains [5]: each a sequence of expressions that refer to the same entity. A place, where a sequence of entity chains terminates and another set begins can be used as an indicator that the discourse has moved from one topically oriented segment to another. This is important for tuple extraction logic in the process of ontology formation from the text. Thus, it seems reasonable to leverage such information in the ontology construction.

Modern discourse parsers that construct DT are DL-based. Due to the availability of the large annotated discourse corpora for many languages, especially English, discourse parsers [17,19,21] provide reliable and correct DT for the text. Manually annotated Ru-RSTreebank corpus [26] has been recently introduced which resulted in the creation of discourse parser for Russian [9]. The availability of state-of-the-art discourse parsers for different languages makes the discourse-based models universal, so they could be applied to different texts without modifications.

3 System Architecture

In this section, we present the overall description of discourse-enhanced ontology extractor from texts. The architecture of the system is shown in Fig. 1. For a corpus of texts, we apply Candidate Ontology Entry Extractor (COEE), which is the first block of the introduced system. It first performs discourse parsing and yields DT. This DT is then subject to a rule-based extractor of EDUs appropriate for tuple formation.

It happens in multiple steps, first, the EDU with the central entity is extracted and other associated EDUs are labeled as appropriate for tuple formation. Then, all nucleus EDUs are considered and if they constitute a too short phrase, they are merged with the respective nucleus to form a single DT node. Finally, all nucleus EDUs outside of EDUs associated with the central entity is included in the list of EDUs appropriate for tuple formation. As a result, we obtain the list of phrases from which the tuples will be formed as candidates

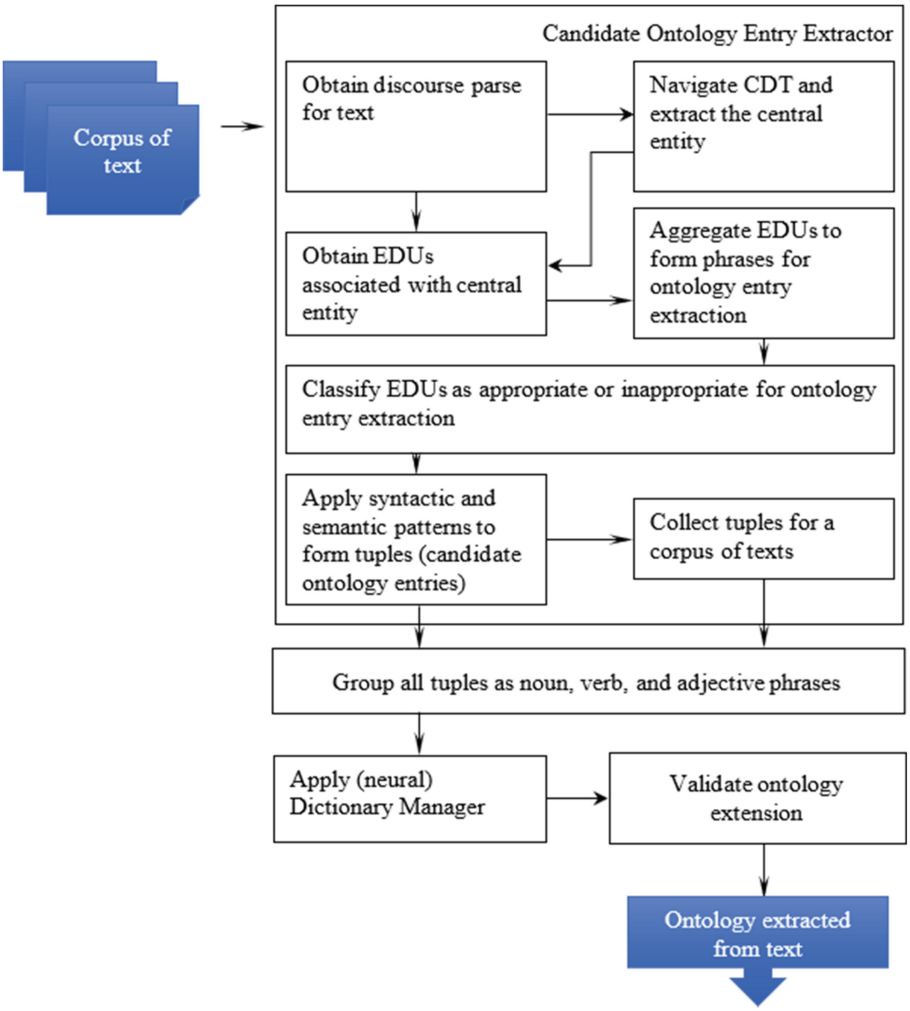

Fig. 1. An architecture of discourse-enhanced ontology builder.

for inclusion into the ontology. Hence, the COEE pipeline includes the following transformations:

Text ⟶ DT ⟶ list-of-EDUs ⟶ list-of-phrases ⟶ list-of-tuples.

We apply syntactic templates to extract a tuple such as ⟨*predicate, subject, object*⟩ from a phrase. As a result, the output of the COEE is a list of tuples for a given text. Before grouping, these tuples need to be accumulated for all texts in a corpus.

The grouping component combines tuples of the same sort so that tuples can be matched to each other to produce reliable, informative ontology entries, minimizing inconsistencies. Noun phrases are grouped with the noun ones, verb with the verb, and propositional with the prepositional phrases.

The aggregation component that follows next performs tuples generalization to avoid too specific, noisy entries that cannot be reliably applied with sufficient confidence. Dictionary manager that includes identification of synonyms helps in generalizing tuples that have the same meaning but different words expressing it. Also, specific ontology types have certain generalization rules for values like space or proper names which are generalized in a different way than entities expressed in words (see Sect. 5). Finally, we validate the ontology to keep the ontology up to date. Let us consider all of these stages in more detail.

4 Candidate Ontology Entry Extractor

4.1 Discourse-Level Support for Ontology Construction

The main novel component of the introduced system is the discourse-aware entry extractor. We take a text and its discourse tree and explore which phrases can potentially form an ontology entry. Let us consider a piece of text where discourse relations are crucial for ontology construction. We take a paragraph from the

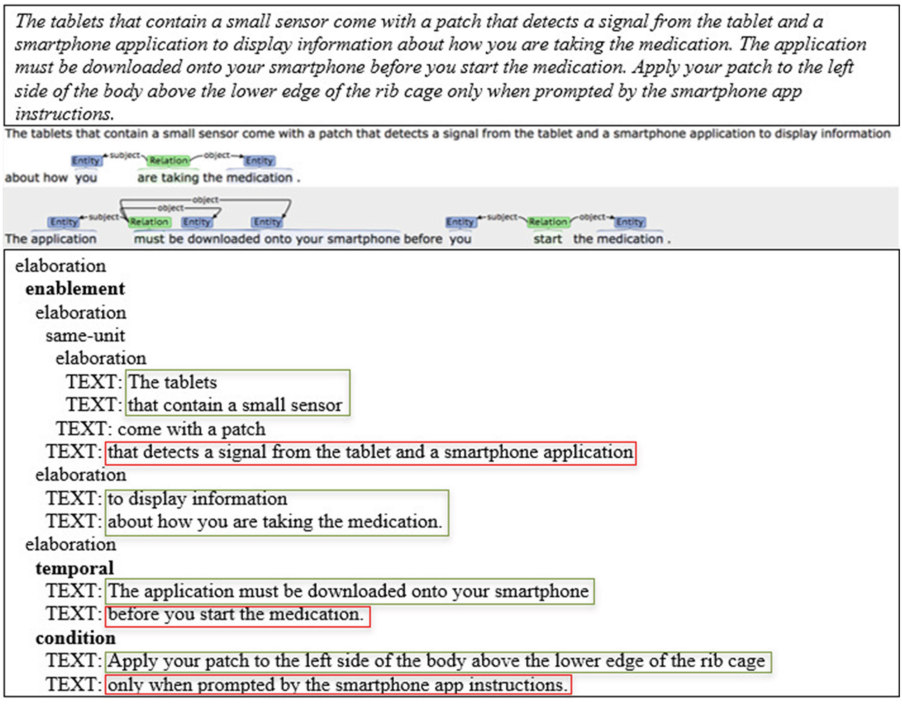

Fig. 2. Text paragraph, its DT, and entity graph showing that not all phrases are equally good to extract tuples to form ontology entries. (Color figure online)

MedlinePlus website[1] and show that not all phrases are good to extract tuples to form ontology entries (see Fig. 2).

A typical syntactic and semantic approach to the entity/tuple extraction considers all highlighted phrases equally. However, some of these phrases are central to this text and, therefore, should serve as a source for relation extraction. At the same time, the rest of the phrases are meaningful in the context of central phrases and should not be used for relation extraction in a stand-alone mode to avoid extracting relations that should not be generalized [11,14].

In Fig. 2, green bolded rectangles show central phrases where extracted relations are informative and express the central topic of this text. Red rectangles show the rest of the phrases which should not yield entity tuples as they are informative only being attached to the central phrases.

In the constructed discourse tree, we see that the central phrase *"tablet-contain-sensor"* corresponds to the nucleus EDU of the top-level rhetorical relation of *Enablement*. This phrase talks about the *"tablet"* which is a central topic of this text, and its predicate *"contain a small sensor"*. Another important phrase associated with the main entity node *"The tablets"* is *"to display information about how you are taking the medication"*.

The satellite EDUs contain phrases that cannot be properly interpreted in a stand-alone mode. For example, *"Come with a patch that detects a signal"* must be interpreted in the context of the *"tablet"*. Otherwise, a hypothetical ontology entry *detect (patch, signal)* is too general and does not necessarily hold on its own. A consistent ontology should not generalize from this expression.

As we proceed from the central entity, navigating the discourse tree, we observe that nucleus EDUs are interpretable on their own and can form an ontology entry, while the satellite EDUs depend on the nuclei and should not form an entry.

Thus, having analyzed the discourse structure of the observed text paragraph, we are able to extract the following entries:

contain (tablet, sensor (small));

display (information (take (people, medications)).

4.2 Rhetorical Relations Determining Informative Text Spans

Let us now look closer at how each type of rhetorical relation defines whether or not the part of the input text contains a candidate ontology entry. For nucleus-satellite *Elaboration* relation, we index the nucleus part as informative and assume that the satellite part is too specific to be mentioned in the ontology. For *Enablement* relation, we have the following template *"To achieve some state [NUCLEUS]—do this and that [SATELLITE]"*. A query that should be asked with the constructed ontology may be of the form *"how to achieve some state?"* but less likely be of the form *"what can I achieve doing this and that?"*. Therefore, we treat the nucleus of *Enablement* relation as a relevant part.

[1] https://medlineplus.gov/druginfo/meds/a603012.html.

We expect the relations such as *Contrast* or *Condition* to be processed as follows: the EDU which expresses facts that actually hold (and not the satellite part facts which are unusual, unexpected, unanticipated) are considered as relevant. *Attribution* acts in a similar way: the nucleus fact is important and may occur in a factoid question, and the satellite part (on whom this is attributed to) is usually a detail.

The *Same-Unit* and *Joint* relations are symmetric and should not affect our selection of relevant text portions.

For *Contrast*, a satellite is good because it is an expression with elevated importance. For *Evidence*, just nucleus is good because the statement is important but its backup is unlikely to be queried.

If the second discourse unit in the *Elaboration* relation describes the same state of affairs as the first one (in different words), or, at a certain level of abstraction, says the same thing, then both nucleus and satellite would form meaningful answers. In the original formulation of RST, usually, an additional requirement for *Elaboration* is imposed that the satellite is more detailed and longer. The broadest definition *Elaboration* also includes in special cases such relation as *Reformulation* or *Restatement*, *Summary*, *Specification* and *Generalization*.

Explanation gives the cause or reason why the state of affairs presented in the context sentence takes place, or why the speaker believes the content of that sentence holds, or why the speaker chose to share information with us; these cases correspond to the three types of causal relations identified. For the cases of content level causality, epistemic causality, and speech act causality satellite should not form an entry [14].

Rhetorical relations *Evidence*, *Justify*, *Motivation*, *Enablement*, *Evaluation*, *Background* all overlap in their function with *Explanation*, but vary in goals and means of giving reasons. For example, *Evidence* is given in order to increase the hearer's belief in a claim.

There are the most popular rhetoric relations that could be identified by state-of-the-art discourse parsers. Following the introduced rules we analyze DT obtained from the discourse parser proposed by Joty et al. [19], and retrieve the candidate text fragments from the nucleus or satellite EDUs based on the type of rhetoric relation they are connected by.

4.3 Relation Extractor Based on Syntactic Parsing

Having revealed the candidate text spans to form the ontology entries, we then aim to process them to extract relevant tuples. This task could be achieved by applying dependency parser. Dependency parser reveals syntactical structure of the input text and presents it in the tree format. This parser analyzes the grammatical structure of the text and provides the relations that hold between "root" words and their dependents. The relations are standard grammatical relations existing in the observed language, such as *subject*, *object*, etc.

In the work, we use open information extraction library ClausIE [10] to derive knowledge tuples for ontology engineering. This library relies primarily on Stanford dependency parser [8] and analyzes grammatical sentence structure to evoke

knowledge triples. Not only can explicit knowledge triples be derived from this method, but also implied, embedded knowledge can also be evoked. ClausIE is domain-independent and, compared to other well-known domain-independent open IE approaches, performs significantly better.

ClausIE is applied to the candidate text fragments, which are the combination of relevant EDUs retrieved from the previous COEE block. The knowledge tuples are derived in RDF-like format, i.e. *predicate(subject, object)* triplets. For example, a sentence *"The human papillomavirus virus (HPV) leads to cervical cancer"* would produce an explicit triple *"leads to" ("The human papillomavirus virus", "cervical cancer")* and an implicit triple *"is" ("human papillomavirus virus", "HPV")*.

5 Phrase Aggregator

To accumulate the obtained tuples for all the texts in a corpus and form meaningful ontology entries, we imply a phrase aggregator. This component takes a list of tuples, where the subject and object are represented by the words or phrases, and merges synonymous and related phrases to form concise ontology entries. The aggregator outputs a hierarchical structure of phrase entities obtained by means of generalization of phrase instances. We use the following phrase filtering rules:

1 Only extract noun, verb, and prepositional phrases.
2 Exclude phrases with sentiments because they can occur in opinionated context.
3 Exclude name entities since they cannot be generalized across properties. However, we include a specific type of such proper nouns in connection with relation specific to health domain such as *affect/cure/drug-for/followed-by* and others.
4 Numbers and prepositions are excluded.
5 There is a limit on phrase length.
6 Too frequent phrases and too rare phrases are removed.
7 Phrases that start with an article if they are short are avoided.
8 Strings which are not words are cleaned/normalized (e.g., *is* $\longrightarrow$ *to be*).

For sentiment analysis, we use Stanford CoreNLP pipeline to perform the rules introduced above. Stanford CoreNLP sentiment component [30] is utilized to assess the sentimental power of each word in the phrase.

Once the phrases are extracted, they are clustered and aggregated to obtain reliable, repetitive instances. Phrases which only occur once are unreliable and considered to be the *noise*. For example, let us consider the following phrases: *"insulin-dependent diabetes mellitus"*, *"adult-onset dependent diabetes mellitus"*, *"diabetes with almost complete insulin deficiency"*, and *"diabetes with almost complete insulin deficiency and strong hereditary component"*, the hierarchy obtained for them is shown in Fig. 3.

Head noun extraction occurs as follows. If two phrases have the same head noun, we combine them into a category. If two phrases within a category have other nouns or adjectives in common besides the head noun, we form a subcategory from these common nouns. In this regard, we follow the cognitive procedure of induction, finding a commonality between data samples retaining the head noun, such as *diabetes*. Hence we have the following class, subclasses and sub-subclasses: *diabetes* ⟶ *mellitus* ⟶ *insulin-dependent*.

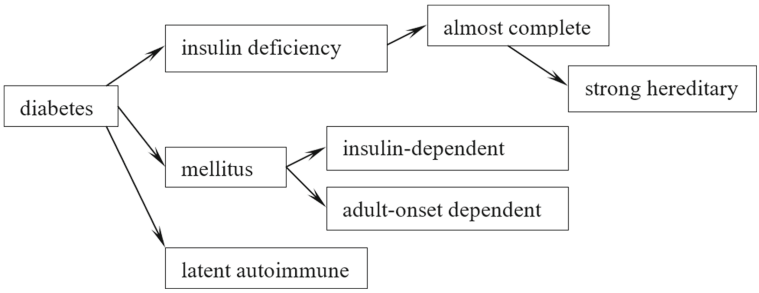

Fig. 3. Phrase hierarchy formed by the aggregator.

6 Neural Dictionary Manager

Neural dictionary manager (NDM) is launched on the final step of ontology construction. NDM is applied to the entity recognition in large unstructured text, which optimizes the use of ontological structures and can identify previously unobserved synonyms for concepts in the ontology. The input of the neural dictionary manager is a word or a phrase. The manager computes the probability of an entity in the ontology matching it. The manager includes a text encoder, which is a neural network that maps the query phrase into a vector representation, and an embedding matrix with rows corresponding to the ontology concepts.

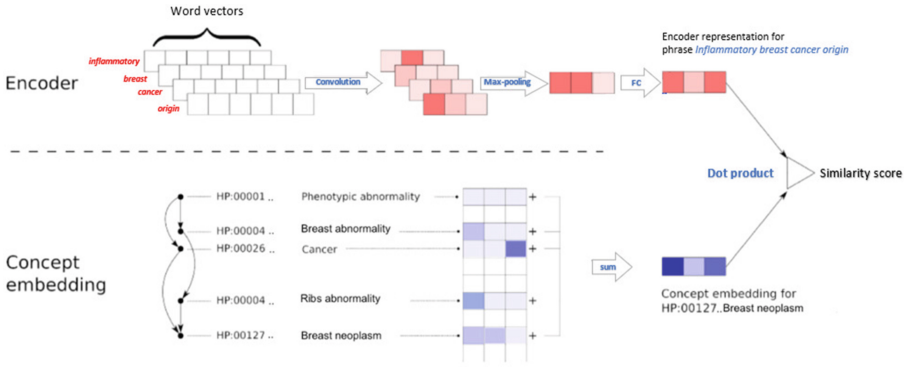

Fig. 4. Architecture of the neural dictionary model.

The architecture of the neural dictionary model that we imply in the system is shown in Fig. 4 [2]. The Encoder flow is at the top, and the flow for computing the embedding for a concept is shown at the bottom. A phrase is first represented in the Encoder by its word vectors, which are then processed by a convolution layer into a new space. A max-over-time pooling layer is employed to merge the set of vectors into a single one. After that, a fully connected layer maps this vector into the final representation of the phrase. To use the NDM for entity recognition in a sentence or larger text, all n-grams of one to seven words are extracted from the text. The neural dictionary manager is used to match each n-gram to an entity. Irrelevant n-grams are removed from the list of candidates when their matching score (the softmax probability provided by the neural dictionary model) is lower than a threshold.

7 Validating Ontology

As a final block of the introduced system, we observe the validation procedure. This procedure is relevant for validating ontologies constructed automatically from medical texts (e.g., clinical guidelines) or re-validating ontologies (constructed manually or automatically) since medical knowledge evolves quickly over time. The following relations can be validated:

- a class A is a subclass of B;
- property P is a sub-property of Q;
- D is the domain class for property P;
- R is the range class for property P;
- I is an individual of class A;
- the property P links the individuals I and J.

To validate the ontology, we utilize question answering (Q/A) schema relying on the domain expert knowledge. The first step of the introduced schema consists of auto-generation of NL questions list from the ontology to be validated. These questions are submitted to domain experts who provide an agreement decision (*Yes/No*) and textual feedback. The next step consists of interpreting expert's feedback to validate or modify the ontology.

Following the idea introduced in [6], we construct manually question templates associated with each type of ontological element. A question template consists of a regular textual expression with the appropriate variables over ontology nodes. For instance, the pattern "Do the SYMPTOMS that PATIENT has correspond to DISEASE?" is a textual pattern with three variables: {SYMPTOMS, PATIENT, and DISEASE}. This question template aims to validate a specific illness with the patient's symptoms.

8 Evaluation

8.1 Dataset

Evaluation of the ontology construction procedure is quite a challenging task as criteria vary from the domain and application areas. There are several complex

domains-specific medical Q/A datasets such as MCTest [28], biological process modeling [7], BioASQ [32], etc. which are available for the analysis, but limited in scale (500-10K). To track the contribution of each ontology construction step, we combine five datasets of varying complexity of questions, texts, and their associations. The utilized Q/A datasets are described in Table 1.

Table 1. The Q/A datasets used for evaluation of the ontology construction

Name	Description	Number of Q/A pairs
Medical-Question-Answer-Datasets	Several sources for medical question and answer datasets from HealthTap.com	1600000
MedQuAD	Medical Q/A pairs created from 12 NIH websites. The collection covers 37 question types associated with diseases, drugs, and other medical entities such as tests	50000
Medical Q/A data	Medical Q/A datasets gathered from eHealth Forum and HealthTap websites	–
PubMedQA [18]	Biomedical question answering dataset collected from PubMed abstracts produced by re-purposing existing annotations on clinical notes	275000
emrQA [25]	Generated domain-specific large-scale electronic medical records datasets produced by re-purposing existing annotations on clinical notes	1 million questions-logical form and 400,000+ Q/A evidence pair

8.2 Assessment of Ontology Consistency

When ontology entries are extracted arbitrarily from noisy data, some entries contradict each other. The frequency of contradiction indirectly indicates the error rate of tuple extraction and overall ontology formation. An example of contradicting entries are ⟨*bird, penguin, fly*⟩ vs ⟨*bird, penguin, not fly*⟩ and ⟨*frog, crawl, water*⟩ vs ⟨*frog, swim, water*⟩ (the third argument should be distinct).

We extract ontology entries from the answers. Then, in the resultant ontology, given each entry, we attempt to find other entries which contradict the given one. If at least one entry is found, we consider the given entry *inconsistent*. The portion of inconsistent entries for the whole ontology is counted and shown as a percentage of all ontology entries. As a baseline, we evaluate an ontology whose entries are extracted from all text parts and left without any refinement. Then we apply various enhancement steps presented in the paper and track if they affect the ontology consistency.

Table 2 presents the percentage of the inconsistent entries in the ontology, thus, we assess how each ontology improvement step affects its resultant consistency. The inconsistency values are normalized for the total number of ontology entries as each refinement step reduces the number of entries, pruning ones determined to be unreliable.

Table 2. Assessment of ontology consistency.

Dataset	Baseline: individual entries extraction	Syntactic parser	w. dictionary manager	w. discourse	w. ontology validation
Medical-Question-Answer-Datasets	7.6	2.8	2.4	1.7	0.8
MedQuAD	6.2	2.3	1.9	1.4	1.1
Medical Q/A data	7.0	2.4	1.7	1.7	1.0
PubMedQA	6.9	3.8	2.9	1.6	0.8
emrQA	11.1	4.9	3.2	2.2	1.3

We observe that adding rules for extracting ontology entries make the resultant ontology cleaner, more robust, and consistent. Employing all means to reduce inconsistencies achieves the contradiction rate of less than 1% of inconsistent ontology entries in most domains. The hardest domains to achieve inconsistency are MedQuAD and emrQA. The worst performance occurs for electronic medical records (the bottom row).

8.3 Assessment of Search Improvement Due to Ontology

We also evaluate the accuracy of search (in percent) on several health-related datasets when this search is supported by an ontology. We vary the complexity of ontological support, steps employed to improve/validate it, and ontological sources (see Table 3). As we have the single best answer for each evaluation dataset, search relevance is measured with the F1 metric. Our baseline search is a default tf-idf method without ontology involvement. We add the ontology at the various construction steps according to the ontology construction system architecture (Fig. 1).

One can observe that there is a small improvement in search relevance (F1) with each enhancement in ontology construction. Such an improvement in the range of 2% may be hard to differentiate from a random deviation. However, the overall improvement due to ontologies is significant and accounts for above 10%. Our ablation experiments show that each step in discourse processing, aggregation, matching, and validation is important and should not be skipped.

Table 3. Assessment of ontology quality via search relevance.

Dataset	Baseline: individual entries extraction	Syntactic parser	w. dictionary manager	w. discourse	w. ontology validation
Medical-Question-Answer-Datasets	78.3	82.3	84.1	85.3	86.1
MedQuAD	75.1	80.4	81.6	83.0	85.0
Medical Q/A data	80.2	83.1	85.8	86.7	86.3
PubMedQA	77.5	82.0	84.2	86.0	87.2
emrQA	76.0	81.2	82.9	83.9	86.4
Improvement		**5.7**	**8.1**	**9.8**	**11.3**

9 Conclusion

Advanced systems for supporting clinical decision is especially enticing in the emergency department. These systems require highly accurate solutions due to the situation is crucial. The use of text mining has played an important role in the development of medical ontologies that support decision-making in emergency services, and its application is already an incipient reality. Despite the rapid development of TM techniques that support the extraction of relevant data from electronic medical records and ontology construction procedures, the latter still suffers from the redundancy and inconsistency of the data retrieved.

In this paper, we introduced the system for automated ontology construction. We reviewed major text mining techniques leveraged for this task and observed an ontology construction bottleneck as selecting portions of documents good for ontology construction. We explored how discourse analysis helps in retrieving the relevant text spans that could be comprised into the ontology as the entry.

Our evaluation showed that relying on discourse analysis indeed improves the quality of an ontology with respect to a lower number of inconsistencies and higher relevance of the resultant search. We conclude that once we extract ontology entries from important and informative parts of text instead of extracting them from all text, the reliability of the resultant ontology for search and decision-making grows.

References

1. Amer, E., Fouad, K.M.: Keyphrase extraction methodology from short abstracts of medical documents. In: 2016 8th Cairo International Biomedical Engineering Conference, CIBEC 2016 (2016)

2. Arbabi, A., Adams, D.R., Fidler, S., Brudno, M.: Identifying clinical terms in medical text using ontology-guided machine learning. JMIR Med. Inform. **7**, e12596 (2019)
3. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA Symposium (2001)
4. Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 2670–2676 (2007)
5. Barzilay, R., Lapata, M.: Modeling local coherence: an entity-based approach. Comput. Linguis. **34**, 1–34 (2008)
6. Ben Abacha, A., Da Silveira, M., Pruski, C.: Medical ontology validation through question answering. In: Peek, N., Marín Morales, R., Peleg, M. (eds.) AIME 2013. LNCS (LNAI), vol. 7885, pp. 196–205. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38326-7_30
7. Berant, J., et al.: Modeling biological processes for reading comprehension. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1499–1510. Association for Computational Linguistics, Doha, Qatar (2014)
8. Chen, D., Manning, C.: A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 740–750. Association for Computational Linguistics, Doha, Qatar (2014)
9. Chistova, E., et al.: RST discourse parser for Russian: an experimental study of deep learning models. In: van der Aalst, W.M.P., et al. (eds.) AIST 2020. LNCS, vol. 12602, pp. 105–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72610-2_8
10. Corro, L., Gemulla, R.: ClausIE: clause-based open information extraction. In: WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web, pp. 355–366 (2013)
11. Galitsky, B.: Improving relevance in a content pipeline via syntactic generalization. Eng. Appl. Artif. Intell. **58**, 1–26 (2017)
12. Galitsky, B., Ilvovsky, D., Kuznetsov, S.O.: Text classification into abstract classes based on discourse structure. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 200–207. Incoma Ltd., Shoumen, Bulgaria, Hissar, Bulgaria (2015)
13. Galitsky, B., Ilvovsky, D., Kuznetsov, S.O.: Detecting logical argumentation in text via communicative discourse tree. J. Exp. Theor. Artif. Intell. **30**, 637–663 (2018)
14. Galitsky, B.A., Dobrocsi, G., de la Rosa, J.L., Kuznetsov, S.O.: Using generalization of syntactic parse trees for taxonomy capture on the web. In: Andrews, S., Polovina, S., Hill, R., Akhgar, B. (eds.) ICCS 2011. LNCS (LNAI), vol. 6828, pp. 104–117. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22688-5_8
15. Gonzalez, G., Tahsin, T., Goodale, B., Greene, A., Greene, C.: Recent advances and emerging applications in text and data mining for biomedical discovery. Briefings Bioinform. **17**, 33–42 (2015)
16. Ji, B., et al.: A hybrid approach for named entity recognition in Chinese electronic medical record. BMC Med. Inform. Decis. Making **19**, 149–158 (2019)
17. Ji, Y., Eisenstein, J.: Representation learning for text-level discourse parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13–24. Association for Computational Linguistics, Baltimore, Maryland (2014)

18. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: PubMedQA: a dataset for biomedical research question answering. CoRR abs/1909.06146 (2019). http://arxiv.org/abs/1909.06146
19. Joty, S., Carenini, G., Ng, R., Mehdad, Y.: Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In: ACL 2013–51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, vol. 1 (2013)
20. Jusoh, S., Awajan, A., Obeid, N.: The use of ontology in clinical information extraction. J. Phys. Conf. Series **1529**, 052083 (2020)
21. Li, J., Li, R., Hovy, E.: Recursive deep models for discourse parsing. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2061–2069. Association for Computational Linguistics, Doha, Qatar, October 2014
22. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. CVPR **2011**, 3337–3344 (2011)
23. Mann, W., Thompson, S.: Rethorical structure theory: toward a functional theory of text organization. Text Talk **8**, 243–281 (1988)
24. Nejadgholi, I., Fraser, K.C., De Bruijn, B., Li, M., LaPlante, A., El Abidine, K.Z.: Recognizing UMLS semantic types with deep learning. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pp. 157–167. Association for Computational Linguistics, Hong Kong (2019)
25. Pampari, A., Raghavan, P., Liang, J., Peng, J.: emrQA: a large corpus for question answering on electronic medical records. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2357–2368. Association for Computational Linguistics, Brussels, Belgium (2018)
26. Pisarevskaya, D., et al.: Towards building a discourse-annotated corpus of Russian. In: Kompjuternaja Lingvistika i Intellektualnye Tehnologii, vol. 1 (2017)
27. Khin, N.P.P., Lynn, K.T.: Medical concept extraction: a comparison of statistical and semantic methods. In: 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 35–38 (2017)
28. Richardson, M., Burges, C., Renshaw, E.: MCTest: a challenge dataset for the open-domain machine comprehension of text. In: EMNLP 2013–2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 193–203 (01 2013)
29. Sarkar, K.: A hybrid approach to extract keyphrases from medical documents. Int. J. Comput. Appl. **63** (2013)
30. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP 2013–2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2013)
31. Song, M., Tanapaisankit, P.: Biokeyspotter: An unsupervised keyphrase extraction technique in the biomedical full-text collection. Intell. Syst. Ref. Libr. **25** (2012). https://doi.org/10.1007/978-3-642-23151-3_3
32. Tsatsaronis, G., et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. **16**, 138 (2015)
33. Wang, X., Yoshida, Y., Hirao, T., Sudoh, K., Nagata, M.: Summarization based on task-oriented discourse parsing. IEEE/ACM Trans. Audio Speech Lang. Process. **23**, 1358–1367 (2015)