Annotating Named Entities in Consumer Health Questions

Halil Kilicoglu*, Asma Ben Abacha*, Yassine Mrabet*, Kirk Roberts†
Laritza Rodriguez*, Sonya E. Shooshan*, Dina Demner-Fushman*
*Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health,
Bethesda, MD, USA
{kilicogluh,mrabety,ddemner}@mail.nih.gov
{asma.benabacha,laritza.rodriguez,sonya.shooshan}@nih.gov

†University of Texas Health Science Center at Houston, Houston, TX, USA
kirk.roberts@uth.tmc.edu

Abstract

We describe a corpus of consumer health questions annotated with named entities. The corpus consists of 1548 de-identified questions about diseases and drugs, written in English. We defined 15 broad categories of biomedical named entities for annotation. A pilot annotation phase in which a small portion of the corpus was double-annotated by four annotators was followed by a main phase in which double annotation was carried out by six annotators, and a reconciliation phase in which all annotations were reconciled by an expert. We conducted the annotation in two modes, manual and assisted, to assess the effect of automatic pre-annotation and calculated inter-annotator agreement. We obtained moderate inter-annotator agreement; assisted annotation yielded slightly better agreement and fewer missed annotations than manual annotation. Due to complex nature of biomedical entities, we paid particular attention to nested entities for which we obtained slightly lower inter-annotator agreement, confirming that annotating nested entities is somewhat more challenging. To our knowledge, the corpus is the first of its kind for consumer health text and is publicly available.

Keywords: consumer health questions, biomedical named entities, assisted annotation, nested entities

1. Introduction

The general public is increasingly turning to online resources for health information needs (Tustin, 2010). The National Library of Medicine (NLM) receives health-related questions from a wide range of consumers worldwide. Most health-related questions are concerned with disease information, such as diagnosis, treatment, and prognosis, and drug information, including ingredients, generic names, and adverse effects. In 2014, NLM received more than 2,500 questions that were classified as disease-related and more than 2,100 questions that were categorized as drug-related by the customer service staff. We have been building a system to assist customer service staff in answering such questions. Focusing on disease questions only, we previously reported an end-to-end question understanding system that extracts question frames from questions (Kilicoglu et al., 2013), which form the basis for search engine queries. Our previous work in question understanding also involved more intermediate tasks, such as question decomposition and focus recognition (Roberts et al., 2014b), question type recognition (Roberts et al., 2014a), anaphora and ellipsis resolution (Kilicoglu et al., 2013), and spelling correction (Kilicoglu et al., 2015).

Named entity recognition and normalization is a core aspect of most of these tasks (e.g., frame extraction and focus recognition). So far, we have used relatively simple dictionary lookup methods to identify named entities in questions and normalize them to UMLS Metathesaurus concepts (Lindberg et al., 1993). However, it has become increasingly clear that more sophisticated methods are needed, since the methods we explored assume well-written questions, whereas consumer health questions are rife with misspellings, informal abbreviations, and non-canonical forms of referring to medical terms. The following request illustrates some of these points:

> (1) *I am taking Amlodipine and it has caused my pause rate to be very high. Is there a weaning process when you stop taking Amlodipine and start atenolol? I am taking 5 mg of amlodipine and will be taking 50 mg of atenolol?*

One of the named entities, the diagnostic procedure *pulse rate* is misspelled as *pause rate*, while *weaning process* is used instead of the more typical term used for drugs, *tapering*. Furthermore, the last sentence contains nested entities. Should the annotation be *5 mg of amlodipine* or *amlodipine*? Such cases are likely to cause problems in question understanding, leading to impoverished question answering performance.

Biomedical named entity recognition (NER) has been studied for various text types, including clinical narratives (Uzuner et al., 2011) and biomedical literature (Kim et al., 2003a; Doğan et al., 2014). However, using tools developed on these types of corpora often perform poorly on consumer health questions. For example, a clinical NER system (Ben Abacha and Zweigenbaum, 2011) trained on i2b2 corpus (Uzuner et al., 2011) yielded a F_1 score of 0.491 in recognizing the named entities in a small set of consumer health questions, while the F_1 score on i2b2 test corpus was 0.875.

In this paper, we present a corpus of consumer health questions annotated with 15 broadly defined categories of biomedical named entities. The corpus is intended to serve for training and testing NER methods as well as to serve

as the basis for annotation of other layers, such as question types, concepts, semantic relations, and question frames. To our knowledge, this corpus is the first concerned with named entities in consumer health text. In this paper, we also aim to present a principled approach to nested biomedical entities and their evaluation.

2. Background

In this section, we focus on corpora annotated for biomedical named entities. Several annotated corpora addressed named entities, including problems and treatments, in clinical narratives, such as discharge summaries. For instance, the i2b2 corpus and the related 2010 clinical NLP challenge (Uzuner et al., 2011) focused on recognition of three categories of biomedical concepts (Problem, Treatment, and Test) as well as the extraction of relations and assertions that involve these concepts. The ShARe corpus (Saeed et al., 2002) was used for the 2013 ShARe/CLEF challenge (Suominen et al., 2013) as well as for the SemEval 2014 (Pradhan et al., 2014) and SemEval 2015 (Elhadad et al., 2015) tasks on the analysis of clinical text. This corpus focuses on disorder mentions and their unique UMLS concept identifiers (CUIs). In contrast to the i2b2 corpus, the discontinuous entities are annotated in the ShARe corpus. There are also annotation efforts in languages other than English. For example, *Quaero* French Medical Corpus (Névéol et al., 2014) addresses clinically relevant entities in French and was used for the CLEF eHealth 2015 challenge on clinical NER (Névéol et al., 2015). The *Quaero* corpus allowed annotation of nested entities.

Several biomedical corpora focused on named entities in biomedical literature. The GENIA corpus (Kim et al., 2003b) consists of 2000 abstracts annotated for named entities from the molecular biology domain, including *proteins*, *protein complexes*, *amino acids*, and *DNA domains and regions*, and was used for the JNLPBA shared task (Kim et al., 2004). A similar corpus is BioInfer (Pyysalo et al., 2007), in which both named entities and their relations are annotated. Both corpora allow nested entities, although their annotation is relatively restricted from a semantic perspective. The NCBI disease corpus (Doğan et al., 2014) provides disease annotations in 793 MEDLINE abstracts. Only disease mentions referring to a unique UMLS concept with specific UMLS semantic types are annotated. Nested and discontinuous mentions were not annotated. Pre-annotations from an automatic classifier were used as the starting point. The CHEMDNER corpus (Krallinger et al., 2015) provides annotations of chemical entities from 10,000 MEDLINE abstracts into one of seven structure-associated chemical entity mention (SACEM) classes, including *abbreviation*, *family*, *systematic*, and *trivial*. Nested entities and overlapping mentions were not annotated. Pre-annotation with an automated tool was not conclusive and led to the choice of manual annotation.

3. Methods

In this section, we first describe our data collection, question selection, and pre-processing for protected health information (PHI). Next, we discuss the annotation scheme/guidelines and provide details on the annotation and reconciliation process.

3.1. Question selection and PHI pre-processing

We collected all the consumer requests that were manually labeled as disease or drug questions by the NLM customer service staff in 2014 and the first half of 2015. A total of 6,166 requests were collected (3,412 disease and 2,754 drug questions, 55.3% and 44.7%, respectively). We discarded requests that were duplicates or near duplicates, relying on exact string matches between requests.

The first phase of our study was concerned with determining whether a request contained an answerable question and, if so, annotating protected health information (PHI) in the request so as to remove it automatically before continuing with the biomedical named entity annotation task. Five of the authors participated in this phase and we continued question selection until we obtained a set of 1,548 answerable requests. Some unanswerable question types are the following:

- Questions asking for a diagnosis based on symptoms
- Questions asking for financial support for treatment of a disease
- Questions about where to get a particular drug
- Requests for free drugs/samples
- Requests with no clear question

For example, the following request was discarded as unanswerable.

(2) *I am looking for a Stem Cell transplant for my Multiple Sclerosis. Can you help me find one??*

The answerable questions were then annotated for PHI. We used the top level PHI categories used in the recent i2b2 challenge on de-identification of clinical narratives (Stubbs and Uzuner, 2015): NAME, PROFESSION, LOCATION, AGE, DATE, CONTACT, and ID. We diverged from the i2b2 guidelines with regards to age, and followed the HIPAA[1] guidelines, by annotating only ages over 89. With respect to locations, we did not annotate country names, US state names, or large city names (e.g., *San Francisco, Dhaka*). The mentions annotated as PHI were then replaced with surrogates. For example, all mentions annotated as LOCATION were replaced with the token [*LOCATION*].

3.2. Named entity annotation scheme

In previous work, to understand and answer disease-related questions, we automatically recognized entities of four very broad categories: disease, intervention (drugs, procedures), anatomy (including genes, proteins, molecular entities), and population groups. Based on our previous experience and the types of questions identified as answerable, we

[1]Health Insurance Portability Accountability Act; 45 CFR 164.514

Entity Type	Brief Definition	Examples	UMLS semantic types
ANATOMY	Includes organs, body parts, and tissues.	*head*, *neck*, *gum*	Body System, Anatomical Structure
CELLULAR_ENTITY	Includes anatomical entities at the molecular or cellular level.	*hemoglobin*, *giant cell*	Cell, Cell Component
DIAGNOSTIC_ PROCEDURE	Includes tests and procedures used for diagnosis.	*biopsy*, *hemoglobin*, *iron levels*	Diagnostic Procedure, Laboratory Procedure
DRUG_SUPPLEMENT	Includes substances used for therapeutic purposes.	*atenolol*, *atenolol 50 mg, campho-phenique*	Clinical Drug, Vitamin
FOOD	Refers to specific nutritional substances.	*eggs*, *breads*, *meat*	Food
GENE_PROTEIN	Includes specific genes and gene products.	*BRCA1*, *BRCA1 gene*, *GLUT4 protein*	Gene or Genome, Enzyme
GEOGRAPHIC_ LOCATION	Includes countries, cities, etc.	*India*, *Singapore*	Geographic Area
LIFESTYLE	Refers to daily and recreational activities.	*smoking*, *yoga*	Daily or Recreational Activity
MEASUREMENT	A quantity that is a core attribute of a named entity, such as dosage of a drug.	*10mg*, *2%*	
ORGANIZATION	Includes institutions as well as their subparts.	*navy*, *California hospitals*	Organization
PERSON_ POPULATION	Includes individuals (gender, age group, etc.) and population groups.	*daughter*, *war veteran*, *16 year old*, *female*	Age Group, Population Group
PROBLEM	Includes disorders, symptoms, abnormalities, and complications.	*autoimmune disease*, *broke*, *cholesterol*, *HIV*	Disease or Syndrome, Neoplastic Process
PROCEDURE_DEVICE	Refers to procedures or medical devices used for therapeutic purposes as well as unspecific interventions.	*shingles treatment*, *nephrolithotomy*, *implants*	Medical Device, Therapeutic or Preventive Procedure
PROFESSION	Includes occupations, disciplines, or areas of expertise.	*dermatologist*, *dr*, *surgeon*	Professional or Occupational Group
SUBSTANCE	Includes chemicals, hazardous substances, and body substances.	*iron*, *cholesterol*, *blood*, *alcohol*	Inorganic Chemical, Biologically Active Substance
OTHER	Includes entities that are relevant to question understanding, but do not fit in one of the categories above.	*pregnancy*	Organism Function

Table 1: Named entity categories

opted for a more fine-grained annotation scheme. The entity types were finalized following a practice annotation of 20 questions by four of the authors. The entity types, their definitions, relevant examples and UMLS semantic types are presented in Table 1. Note that OTHER type is used with the view that, if there is a critical mass of certain categories, they can be added as new categories.

3.3. Guidelines

In this subsection, we discuss several issues that came up in the course of pilot annotation and had to be clarified in annotation guidelines.

Entity ambiguity Annotators are allowed to assign multiple types to mentions. For example, *L-leucine* can be annotated as both DRUG_SUPPLEMENT and SUBSTANCE.

Generic mentions Annotators are instructed to avoid annotating generic mentions, such as *problem* or *organization*. As a principle, it was agreed that words that are used in the names of entity categories above or those that are used in the names of UMLS semantic types are likely to be generic. With this principle, *syndrome* is considered generic, because it is in the name of a UMLS semantic type (Disease or Syndrome). On the other hand, some terms that are more specific than these examples but not very specific on their own can be annotated (such as *surgery*, *operation*, *cancer*, and *trauma*). The annotators were also provided a list of 56 overly generic Problem terms from a previous study and were asked to contribute to a collaborative list of generic terms for other categories in the course of annotation.

Nested entities Nested entity annotation is allowed in order to capture the inner structure of some named entities. For example, in the noun phrase *left hand ring finger*, four nested annotations (*hand*, *finger*, *left hand*, and *ring fin-*

ger), in addition to the top-level annotation *left hand ring finger*, can be annotated. Determining when and how to annotate nested entities and what constitutes a top-level entity can be challenging. The basic principle we applied is that the nested annotations and the related top-level annotation should refer to *different but related* entities at a conceptual level and the nested annotations should not be generic. Using this principle, we do not annotate *BCG* in *BCG treatment*, since it refers to the same entity as *BCG treatment*, which is annotated. The type of relation that can hold between the top-level entity and the nested entity is defined broadly and includes specialization and attribute, among other types of semantic modification. For example, the relation between *50 mg of atenolol* and the nested *atenolol* is one of specialization, while the relation between *67 years old female* and the nested *67 years old* and *female* is one of attribute. Even though a top-level annotation and the related nested annotations are generally contained within a simple noun phrase, the former example also shows that they can involve certain types of prepositional phrases. The main motivation for annotating nested entities was to provide the means for a more precise semantic evaluation of NER systems, regardless of whether they can generate nested annotations or flat annotations.

Part-of-speech While named entities are often contained within simple noun phrases, there is no annotation restriction based on part-of-speech; for example, *sensitive* was annotated as a PROBLEM, in the following sentence: *I am still sensitive in my back and head*.

Multi-part entities Annotated named entities can be multi-part and non-contiguous. For example, in *hair loss/thinning*, both *hair loss* and the non-contiguous entity *hair…thinning* can be annotated.

Measurements MEASUREMENT entities are annotated only when they can be seen as a core attribute of a named entity, such as dosage in Example (1). On the other hand, *4&8 hrs* in *onset of autoimmune disease between 4&8 hrs. later* was not annotated.

PHI PHI surrogates (e.g., [*PROFESSION*]) are not annotated.

3.4. Manual vs. assisted annotation

We performed annotation in two modes: manual and assisted. In the manual annotation mode, questions were annotated with the aforementioned categories manually. In the assisted annotation mode, the annotators were provided pre-annotations generated by five NER systems. The motivation for assisted annotation is that pre-annotations can both speed up the annotation process and reduce missed annotations. The five systems used for pre-annotation are briefly described below.

- MetaMap (Aronson and Lang, 2010) is a widely-used tool that maps biomedical free text to UMLS Metathesaurus concepts. MetaMap is a linguistically-based system that relies on lexical analysis and shallow parsing to identify noun phrases, which in turn are used to generate noun phrase variants and candidate mappings. To identify the best mapping, candidates are scored, based on several principles, such as centrality, variation, coverage, and cohesiveness. MetaMap performs acronym expansion, as well.

- Essie (Ide et al., 2007) is a concept-based search engine that supports several NLM websites, including ClinicalTrials.gov. It maps free text to UMLS Metathesaurus concepts, using synonymy information in UMLS and addressing some inflectional variants.

- KODA (Mrabet et al., 2015) is a named entity annotator that relies on the relationships between the entities specified in a knowledge base to perform a global disambiguation of all noun phrases in the target text. KODA is knowledge base-agnostic and has been evaluated on several "wikification" benchmarks, on which it outperformed machine-learning based systems significantly. For pre-annotation, KODA relied on the DBpedia (Auer et al., 2007) open-domain knowledge base, extracted from Wikipedia infoboxes.

- Customized UMLS dictionary lookup relies on a set of four customized dictionaries (Demner-Fushman and Lin, 2007) created using UMLS concepts belonging to pre-defined semantic types. The dictionaries contain Problem, Intervention, Anatomy, and Population Group terms. For example, the Problem dictionary includes UMLS terms from semantic types, such as Congenital Abnormality, Injury or Poisoning, Neoplastic Process, and Bacteria. Some common false positive UMLS terms were filtered from the lexicons (e.g., *age* as an intervention due to its being an abbreviation for *advanced end glycosylation*). Exact string matching based on UMLS Metathesaurus synonyms is performed. This method only considers the longest-matching string.

- A NER method based on Conditional Random Fields (CRF) (Lafferty et al., 2001) trained on the i2b2 corpus (Uzuner et al., 2011). This method is an extension of Ben Abacha and Zweigenbaum (2011) and uses lexical, morphosyntactic, and orthographic features, including the word itself, preceding and following words, their lemmas and part-of-speech tags, presence of special characters (e.g., hyphen, ampersand), word capitalization, and prefixes/suffixes.

3.5. Annotation phases and inter-annotator agreement

The practice annotation of 20 questions by four annotators was followed by a discussion to clarify the annotation guidelines and make them more precise. We conducted the next phase of annotation (*pilot phase*) to determine whether assisted annotation would be beneficial in reducing missed annotations and providing higher annotation consistency than manual annotation. In this phase, two pairs of annotators annotated 25 questions manually and 25 questions with assistance, resulting in annotation of a total 100 questions. Each pair then reconciled their annotations. We calculated inter-annotator agreement between the annotators forming a pair. As the inter-annotator agreement measure, we used

F_1 score of one set of annotations, with the other set of annotations taken as the reference standard. We calculated inter-annotator agreement using exact match of named entities as well as partial match (span overlap). We also calculated agreement measures ignoring entities of OTHER type and ignoring nested entities and only focusing on top-level entities, to measure their effect. We calculated the overall inter-annotator agreement as the average F_1 score among all pairs.

In the *main annotation phase*, we double-annotated the remaining 1,428 questions. Six authors participated in this phase, annotating 476 questions each. The results of the pilot phase were inconclusive with respect to the use of manual vs. assisted annotation; therefore, each annotator annotated half of their questions in assisted mode. Annotators were paired such that each pair had around 95 (476/5) questions in common, half of which were annotated in assisted mode. After an annotator completed her/his annotations, we automatically analyzed their annotations and provided feedback with respect to their annotation consistency. The feedback included the entropy for each annotated string as well as the list of all instances of the string with the corresponding entity types (including NULL for unannotated instances). For a string X annotated with $\{x_1, \ldots, x_n\}$ different entity types, the entropy was calculated as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

The strings with high entropy were expected to be the chief sources of errors. The annotators were then allowed to incorporate the feedback into their annotations.

The *reconciliation phase* of the double-annotated 1,428 questions was conducted by the remaining author of the paper (SS). After she reconciled the annotations, she was also provided feedback, which she incorporated into the final annotations. *brat* annotation tool was used for all phases of annotation and reconciliation (Stenetorp et al., 2012).

4. Results and Discussion

The corpus consists of 1548 requests and 86,135 tokens (an average of 55.6 tokens per request). The number of tokens per request ranges from 2 to 427. 645 mentions were marked as PHI (0.42 PHI mentions per request) in the entire corpus. The distribution of PHI types is given in Table 2. The generic PHI category indicates cases in which the mention does not fit neatly into one of the categories but it can be used to recover protected health information. The predominance of NAME, LOCATION, and CONTACT can partially be explained by the fact that some NLM web forms allow the consumers to specify such information.

The inter-annotator agreement results for the pilot phase are given in Table 3. Overall, moderate to good agreement was obtained in this phase (average of 0.78 with exact span matching and 0.811 with partial span matching). However, the comparison between the pairs did not reveal clear trends. One pair (pair 1) agreed more on manual annotation and the other (pair 2) on assisted annotation. Similarly, pair 1 had more disagreement on entity boundaries in assisted annotation and pair 2 in unassisted annotation. Some of pair 1 disagreements were due to OTHER type, while pair 2 did not annotate any. Ignoring nested entities largely improved agreement for pair 1, while it mostly reduced agreement for pair 2.

Type	**# of occurrences**	**%**
AGE	4	0.6
CONTACT	106	16.5
DATE	37	5.7
ID	1	0.2
LOCATION	157	24.3
NAME	327	50.7
PROFESSION	9	1.4
PHI	4	0.6
TOTAL	645	100.0

Table 2: PHI distribution in the corpus

Inter-annotator agreement results for the main annotation phase are also given in Table 3. In this phase, we obtained moderate agreement (average of 0.706 with exact span matching and 0.754 with partial span matching). Two annotators who did not participate in the previous phases had noticeably lower agreement with the rest of the annotators. With exact span matching, their average inter-annotator agreement was 0.67, while that of the annotators participating in previous phases was 0.761. Similarly, with partial span matching, the averages were 0.728 and 0.794, respectively. This suggests that active contribution to guideline development is more useful for subsequent annotation than simply following the guidelines developed by others. In contrast to the pilot phase, the main phase showed clearer trends. In assisted annotation mode, overall agreement was higher between annotator pairs than in manual annotation, even though the differences between the annotation modes were generally not large. The agreement on entity boundaries, on average, increased in assisted mode, even though this was not the case for all annotator pairs. As expected, ignoring OTHER annotations consistently increased inter-annotator agreement, indicating that clearly defined semantic classes will lead to better inter-annotator agreement. Similarly, ignoring nested entities and only considering top-level entities increased inter-annotator agreement with partial matching and lowered it with exact matching. This indicates that the main difficulty with nested annotation lies in recognizing the exact boundaries of the top level entities rather than recognizing the individual nested entities.

After reconciling the annotations between annotator pairs, we also calculated agreement with the reference standard. These results are provided in Table 4. The overall agreement with the reference standard was higher in assisted mode than in manual mode, even though the differences were often relatively small, a trend similar to that seen for inter-annotator agreement. In manual annotation, the agreement with the reference standard ranged from 0.797 to 0.917 and, in assisted mode, from 0.814 to 0.933. One of the motivations for assisted annotation was to assess whether the pre-annotations would lead to fewer complete

	All		Ignore OTHER		Ignore Nesting	
Annotation Mode	**Exact**	**Partial**	**Exact**	**Partial**	**Exact**	**Partial**
Pilot phase						
Manual	0.782	0.810	0.784	0.812	0.765	0.828
Assisted	0.763	0.796	0.762	0.803	0.707	0.790
Overall	0.780	0.811	0.785	0.817	0.739	0.813
Main annotation phase						
Manual	0.696	0.749	0.712	0.765	0.664	0.773
Assisted	0.715	0.758	0.716	0.812	0.696	0.793
Overall	0.706	0.754	0.722	0.771	0.682	0.784

Table 3: Inter-annotator agreement (average F_1 score among all pairs)

misses by the annotators. These results are also shown in Table 4. On average, more reference annotations were missed in manual mode than in assisted mode. Among those missed in assisted mode, an average of 18% had not been pre-annotated (1.1% of all annotations in assisted mode), suggesting that annotators could benefit from more attention to the pre-annotations. The completely missed pre-annotations were generally common terms, such as *broke*, *thin*, and *tired*.

Mode	**Agreement w/ reference standard**	**Missed (%)**	**Missed & not pre-annotated (%)**
Manual	0.849	6.9%	N/A
Assisted	0.856	5.9%	1.1%
Overall	0.853	6.2%	N/A

Table 4: Annotator agreement with the reference annotations and the effect of pre-annotation (M=manual, A=assisted)

The distribution of entity categories annotated in this study is provided in Table 5. Overall, more than 15K entities were annotated. Unsurprisingly, PROBLEM is the most commonly annotated category, followed with a significant margin by ANATOMY, PERSON_POPULATION, and DRUG_SUPPLEMENT categories.

An average of 9.7 entities were annotated per request. 35.4% of annotations are nested and nesting level goes as high as 3. A mention with 3 levels of nesting is *glucose tolerance test drink*, where *glucose*, *glucose tolerance*, and *glucose tolerance test* in addition to the full phrase is annotated. 0.4% of the entities annotated were assigned multiple types (i.e., they were ambiguous). For example, *diabetic* was sometimes annotated as both PROBLEM and PERSON_POPULATION.

The number of pre-annotations generated by the five tools described earlier are given in Table 6. MetaMap generates the highest number of annotations due to its broad coverage, while the CRF method generates the fewest due to its limited focus on i2b2 categories. Interestingly, KODA generated pre-annotations for misspellings.

Category	**# of annotations**	**%**
ANATOMY	2,360	15.7
CELLULAR_ENTITY	105	0.7
DIAGNOSTIC_PROCEDURE	418	2.8
DRUG_SUPPLEMENT	1,554	10.3
FOOD	185	1.2
GENE_PROTEIN	78	0.5
GEOGRAPHIC_LOCATION	189	1.2
LIFESTYLE	178	1.2
MEASUREMENT	163	1.1
ORGANIZATION	259	1.7
PERSON_POPULATION	1,699	11.3
PROBLEM	5,134	34.1
PROCEDURE_DEVICE	1,102	7.3
PROFESSION	507	3.4
SUBSTANCE	688	4.6
OTHER	433	2.9
TOTAL	15,052	100.0

Table 5: The distribution of annotated entity categories

Tool	**# of pre-annotations**
MetaMap	9,498
Essie	4,272
KODA	7,163
UMLS dictionary lookup	6,947
i2b2 CRF	1,653
TOTAL	29,533

Table 6: Pre-annotation counts for 764 pre-annotated requests

4.1. Nested Entities and Their Evaluation

In this study, we considered nested entity annotation to enable a more accurate and precise semantic evaluation of NER methods that use the corpus as benchmark. We briefly discuss the shortcomings of flat (non-overlapping) entity annotations and the principles involved in evaluating nested entities below.

Named entities are commonly evaluated in two modes: in *exact span matching*, character offsets and the semantic type of a named entity generated by the system (*S*) is expected to match those of a reference entity (*R*), while in

partial span matching (or relaxed span matching), a character offset overlap between *S* and *R*, in addition to semantic type match, is considered sufficient. In the latter case, the semantic relevance of *S* to *R* may be difficult to establish, possibly leading to a somewhat inaccurate evaluation.

For example, let us consider a reference standard based on flat annotations and evaluation of a NER system that also generates flat annotations against this reference standard. Let us also assume that the phrase *lung cancer* has been annotated as a PROBLEM in the reference standard. If the NER system recognizes the full phrase as a PROBLEM, the evaluation is straightforward; the annotation counts as a true positive in both evaluation modes. If the NER system recognizes two separate entities instead, *lung* (ANATOMY) and *cancer* (PROBLEM), evaluation with partial span matching considers *lung* a false positive because its semantic type does not match that of the reference annotation and *cancer* a true positive because its semantic type matches that of the reference annotation. In evaluating these two entities, the reference annotation *lung cancer* is used twice, leading to a false negative in one case. Overall, the evaluation yields 0.5 precision and 0.5 recall. On the other hand, if the system only recognizes *cancer* (PROBLEM), evaluation with partial span matching generates a true positive (there is span overlap and semantic types match) and yields perfect precision and recall. The former NER output is clearly preferable to the latter; because, in the former case, *lung* and *cancer* are semantically related to the reference annotation; the meaning of the the top level entity (*lung cancer*) can be composed from the annotated entities. On the other hand, only recognizing *cancer* should arguably be penalized, since it is not as informative.

Allowing nested entity annotation, all three entities above, *lung* (ANATOMY), *cancer* (PROBLEM), and *lung cancer* (PROBLEM), can be annotated in the reference standard, indicating that the former two are semantically relevant to the third, the top level entity, and the system should not be penalized for annotating them. If the system generates nested entities, as well, the evaluation is not any different from what is outlined above: all entities (nested or not) are simply considered flat annotations and evaluated as such. However, if the system generates only flat annotations, the situation becomes more complex. In such evaluation, we use the notion of *annotation coverage*, which we define as the ratio of the number of tokens in a given annotation to the number of tokens in the corresponding top level entity. In the case above, for example, *lung* and *cancer* both have annotation coverage of 0.5, while *lung cancer*, the top level entity, has annotation coverage of 1. Because all semantically relevant entities are precisely annotated, only evaluation with exact span matching is performed and evaluation takes place relative to the top level annotation (*lung cancer*). We find that the system has recognized two named entities in the span of this top level annotation, *lung* and *cancer*. The first (*lung* (ANATOMY)) yields 0.5 points, since its annotation coverage is 0.5 and it exactly matches a nested entity. Similarly, the second yields 0.5 points. Relative to the top level annotation *lung cancer*, the scores are summed to yield 1, meaning that we get a true positive. If only *lung* (ANATOMY) or *cancer* (PROBLEM) were recognized, the evaluation would yield only 0.5 true positives.

If the system recognizes an entity subsumed by a top level entity annotation but not annotated as a nested entity, its contribution can be subtracted from the overall score, as well. For example, let us consider the following reference annotations: *glucose tolerance test drink* (SUBSTANCE) as the top level entity and the nested entities, *glucose tolerance test* (DIAGNOSTIC_PROCEDURE) and *glucose* (SUBSTANCE). Let us also assume that the NER system generates *glucose* (SUBSTANCE), *tolerance test* (DIAGNOSTIC_PROCEDURE) and *drink* (SUBSTANCE). With the steps outlined above, *glucose* (SUBSTANCE) will yield 0.25 points (annotation coverage=0.25). On the other hand, *tolerance test* (DIAGNOSTIC_PROCEDURE) and *drink* (SUBSTANCE) are not annotated in the reference standard, yielding -0.5 pand -0.25 points respectively, and the overall score for the phrase will be -0.5 points (=0.25-0.5-0.25), indicating a half false positive. To summarize, we believe nested entity annotation can lead to a more semantically specific and precise evaluation, mostly alleviating the need for evaluation with partial span matching, while also addressing its shortcomings.

5. Conclusion

We presented the first consumer health corpus focusing on named entities. We discussed our multi-step annotation process, which involved manual and assisted annotation. We also described a principled way of annotating and evaluating nested entities. We plan to annotate other semantic layers on top of named entities annotated in this study, such as question types and concepts. The corpus is fully deidentified and is publicly available for research purposes at our project webpage[2].

6. Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

7. Bibliographical References

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association (JAMIA)*, 17(3):229–236.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In *6th International Semantic Web Conference*, pages 11–15.

Ben Abacha, A. and Zweigenbaum, P. (2011). Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, pages 56–64.

Demner-Fushman, D. and Lin, J. J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

[2] https://lhncbc.nlm.nih.gov/project/consumer-health-question-answering

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Elhadad, N., Pradhan, S., Gorman, S., Manandhar, S., Chapman, W., and Savova, G. (2015). SemEval-2015 Task 14: Analysis of Clinical Text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Ide, N. C., Loane, R. F., and Demner-Fushman, D. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of Americal Medical Informatics Association*, 14(3):253–263.

Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2013). Interpreting consumer health questions: The role of anaphora and ellipsis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 54–62.

Kilicoglu, H., Fiszman, M., Roberts, K., and Demner-Fushman, D. (2015). An ensemble method for spelling correction in consumer health questions. In *AMIA Annual Symposium Proceedings*.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003a). GENIA corpus - a semantically annotated corpus for bio-textmining. In *Bioinformatics*, volume 19, Suppl. 1, pages 180–182.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003b). GENIA corpus - semantically annotated corpus for bio-text mining. *Bioinformatics*, 19 Suppl 1.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bioentity recognition task at JNLPBA. In *Proceedings of Joint Workshop on Natural Language Processing in Biomedicine and its applications*.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Akhondi, S., Kors, J., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., Batista-Navarro, R., Rak, R., Huber, T., Rocktäschel, T., Campos, D., and et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(Suppl 1):S2.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning*, pages 282–289.

Lindberg, D. A. B., Humphreys, B. L., and McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, 32:281–291.

Mrabet, Y., Gardent, C., Foulonneau, M., Simperl, E., and Ras, E. (2015). Towards knowledge-driven annotation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2425–2431.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization. In *Proceedings of LREC 2014 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 24–30.

Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., and Zweigenbaum, P. (2015). CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition. In *Working Notes of CLEF 2015*.

Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., and Savova, G. (2014). SemEval-2014 Task 7: Analysis of Clinical Text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

Roberts, K., Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2014a). Automatically classifying question types for consumer health questions. In *AMIA Annual Symposium Proceedings*, pages 1018–1027.

Roberts, K., Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2014b). Decomposing consumer health questions. In *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing*, pages 54–62.

Saeed, M., Lieu, G., and Mark, R. G. (2002). MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Computers in Cardiology*, 29:641–644.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Stubbs, A. and Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G. K., Elhadad, N., Pradhan, S., South, B. R., Mowery, D., Jones, G. J. F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D., and Zuccon, G. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *CLEF*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231.

Tustin, N. (2010). The role of patient satisfaction in online health information seeking. *Journal of health communication*, 15(1):3–17.

Uzuner, O., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556.