Journal of Biomedical Informatics 45 (2012) 507–521

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Anaphoric reference in clinical reports: Characteristics of an annotated corpus

Wendy W. Chapman [a,*], Guergana K. Savova [b], Jiaping Zheng [c], Melissa Tharp [a], Rebecca Crowley [d]

[a] University of California, San Diego, Division of Biomedical Informatics, 9500 Gillman Drive #0505, La Jolla, CA 92093, United States
[b] Children's Hospital Boston and Harvard Medical School, Boston, MA 02114, United States
[c] University of Massachusetts Amherst, 140 Governors Drive, Amherst, MA 01003-9264, United States
[d] University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, PA 15260, United States

ARTICLE INFO

Article history:
Received 4 August 2011
Accepted 30 January 2012
Available online 9 February 2012

Keywords:
Natural language processing
Clinical reports

ABSTRACT

Motivation: Expressions that refer to a real-world entity already mentioned in a narrative are often considered anaphoric. For example, in the sentence "The pain comes and goes," the expression "the pain" is probably referring to a previous mention of pain. Interpretation of meaning involves resolving the anaphoric reference: deciding which expression in the text is the correct antecedent of the referring expression, also called an anaphor. We annotated a set of 180 clinical reports (surgical pathology, radiology, discharge summaries, and emergency department) from two institutions to indicate all anaphor–antecedent pairs.

Objective: The objective of this study is to describe the characteristics of the corpus in terms of the frequency of anaphoric relations, the syntactic and semantic nature of the members of the pairs, and the types of anaphoric relations that occur. Understanding how anaphoric reference is exhibited in clinical reports is critical to developing reference resolution algorithms and to identifying peculiarities of clinical text that may alter the features and methodologies that will be successful for automated anaphora resolution.

Results: We found that anaphoric reference is prevalent in all types of clinical reports, that annotations of noun phrases, semantic type, and section headings may be especially important for automated resolution of anaphoric reference, and that separate modules for reference resolution may be required for different report types, different institutions, and different types of anaphors. Accurate resolution will probably require extensive domain knowledge—especially for pathology and radiology reports with more part/whole and set/subset relations.

Conclusion: We hope researchers will leverage the annotations in this corpus to develop automated algorithms and will add to the annotations to generate a more extensive corpus.

1. Introduction

A number of natural language processing (NLP) systems are being developed to classify, extract, and summarize information described in narrative reports stored in electronic health records. Most systems identify diseases, findings, or medications and their modifiers one sentence at a time. However, understanding the meaning of a narrative report requires interpretation of not only the individual concepts described in the report but also their relationships with each other. The coreference relation represents one of the most important relations in narrative for valid information extraction. Linguistic expressions that refer to the same real-world entity are considered coreferential. When coreference relations are not resolved, incomplete output and spurious concepts may result. For example, in (1) (originally cited in Hahn et al. [1]), if an NLP system does not recognize that the nodule and the tumor are referring to the same real-world entity, two separate objects will be created—indicating two separate findings—and their modifying information will not be merged, resulting in incomplete knowledge about the location and the size of the entity.

(1) Chest X-ray again shows **a well-circumscribed nodule** located in the left upper lobe. **The tumor** has increased in size since the last exam with a diameter of approximately 2 cm.

One reason the clinical NLP community has been slow to address coreference resolution is the lack of an annotated corpus for developing and evaluating coreference resolution algorithms. We annotated a corpus of 180 clinical notes for coreference and two anaphoric relations (set/subset and part/whole) [2].

* Corresponding author.
E-mail addresses: wwchapman@ucsd.edu (W.W. Chapman), Guergana.Savova@childrens.harvard.edu (G.K. Savova), jzheng@cs.umass.edu (J. Zheng), m.tharp813@gmail.com (M. Tharp), crowleyrs@upmc.edu (R. Crowley).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2012.01.010

Annotators highlighted clinically relevant concepts we call markables, including disorders, procedures, and medications, and connected pairs of markables that are anaphoric. In this paper, we describe the characteristics of the annotated corpus, positioning our findings in relation to linguistic theory and computational experience in anaphoric reference resolution.

Our corpus is annotated in such a way as to facilitate theoretical investigations of anaphoric relations and discourse models in clinical reports and to provide practical training cases for computational methods research. Understanding how anaphoric reference is exhibited in clinical reports is critical to developing reference resolution algorithms and to identifying peculiarities of clinical text that may alter the features and methodologies that will be successful for this task. In a separate manuscript [3], we review the existing computational methodologies for and scientific advances in coreference resolution in the general natural language processing community.

In Section 2, we review linguistic theories of anaphoric reference and the features used by computational algorithms in an attempt to model the knowledge we as humans use to resolve anaphoric reference. In Section 3, we describe the annotations we performed on the corpus, which were motivated by features in resolution algorithms. In Section 4, we describe the distribution of the annotations in the overall corpus and describe differences among sub-genres of clinical narratives. Finally, we compare the characteristics we discovered in clinical texts against those exhibited in general English texts and discuss the implications to the design of anaphoric resolution algorithms for clinical narratives.

2. Background

2.1. Definition of anaphoric relations

As Olsson wrote [4], "the phenomenon of anaphora is sensitive to context." Anaphoric relations are relations between linguistic expressions in which the interpretation of one of the linguistic expressions relies on the interpretation of another linguistic expression. For instance, understanding the expression "the pain" in a history of present illness section may rely on having read an earlier description of the patient's chief complaint of "neck pain." In this example, the anaphoric relation between "the pain" and "neck pain" is that of identity: both linguistic expressions refer to the same real world entity. Coreferring expressions are sometimes but not always anaphoric [5]. For instance, interpretation of a later mention of "the neck pain" may not rely on the previous mention of "neck pain." In this project, we followed the MUC-7 task definition [6] and annotated all coreferential relations, regardless of whether they were anaphoric. Similarly, not all anaphoric relations are coreferential—the identity relation is only one of many types of anaphoric relations. In this project, in addition to coreferential relations, we annotated two types of anaphoric relations: part/whole and set/subset.

Mapping from the linguistic expressions in the discourse to the actual entities being discussed is called reference resolution. For instance, in a report describing a 55-year-old woman's visit to the emergency room, a physician dictated the following:

(2) **The patient** presents complaining of shortness of breath, which **she** has experienced for 3 days.

In (2), "she" is an *anaphor* referring to "the patient," which is the *antecedent*. The interpretation of the anaphor depends on our knowledge of a previous linguistic expression, the antecedent. Therefore "she" is considered to be anaphoric. For an excellent tutorial on anaphoric resolution, see [7].

2.2. Discourse model

A hearer's (or reader's) mental model of an ongoing discourse is called a discourse model [8], which contains entities referred to in the discourse and the relationships in which the entities participate. Some of the most frequent entities in a discourse model of a clinical report include patient, caregivers, diseases, symptoms, procedures, anatomical sites, medications, and hospitals. When an entity is first referenced, a representation for the entity is evoked in the discourse model. In (2), a representation for the 55-year-old woman who is the patient is evoked through the expression "the patient." Subsequent mentions of the patient serve to access the patient from the discourse model. The expression evokes the referent in a variety of ways. In our example, the anaphor "she" is related to the antecedent "the patient" with an identity relationship, meaning that the expressions both refer to the same entity and are therefore coreferential. Anaphoric expressions can exhibit other relationships, such as part/whole or set/subset. We annotated these three types of relations in our clinical corpus, the Ontology Development and Information Extraction (ODIE) corpus (for details see [2]). The scope of the annotation task included referring expressions that refer to an entity that has been explicitly evoked in the text; we did not include inferrables, bridging inferences, generics, speech acts, propositions [9], or other referents that are only inferentially related to an evoked entity. For instance, consider the following example (3) describing a patient's thoughts on treatments for sleep apnea:

(3) The patient was shown **a CPAP machine**, but thought **a mask** would interfere with her sleeping and **the noise** would be distracting. She would be interested in learning about **a surgical procedure**, if **they** would provide better relief.

In this example, "a mask" is an inferrable, because "a mask" is not referring to any type of mask, but one can infer it is a mask belonging to "a CPAP machine." Along the same line, one can infer "the noise" is not just any noise but the noise coming from the CPAP machine. Thus, "the noise" and "a CPAP machine" provide an example of bridging inference. Additionally, "a surgical procedure" and "they" are generic references, because the expressions do not refer to any particular surgical procedure. We suspect speech acts and propositions are extremely rare in clinical notes.

2.3. Salience

Humans refer to entities in many ways. For example, chest pain could be referred to as "chest pain," "the pain," "this pain," "a pain in her chest," or "it." The most common referring expressions are indefinite noun phrases, definite noun phrases, pronouns, demonstratives, and proper names, but we can also refer to entities in a discourse with more complex constructs such as clauses and temporal expressions [9]. The expressions we use to refer to an entity are not always interchangeable, because referring expressions encode different signals about the location of referent within the hearer's mental model of the discourse—the referent's *cognitive status*. An important question about cognitive status is which entities are activated or in focus in the discourse model. An entity is *activated* if there is a representation of that entity in working memory. The activated entity is *in focus* if it is at the center of the hearer's attention [10]. Because discourse is often structured around a central topic, the topic is usually the focus for a few sentences before the focal point shifts to a new topic [11]. Therefore, entities that have been mentioned in more recent utterances tend to be more salient or accessible than entities that were mentioned earlier. An anaphor's distance from the antecedent (i.e., its *recency*)

is often used as a surrogate for the salience of the entity, because recency is more straightforward to measure. In our corpus, we measured the recency of anaphor–antecedent pairs (i.e., the distance between the two markables in the pair).

There are several theories expressing the relationship between the surface form of a referring expression and the accessibility or saliency of the referent in the discourse [12,13]. Centering theory [14] examines interactions between local coherence and the choice of referring expressions. The amount of linguistic description required to call a referent to mind is related to the saliency of the referent in the discourse—less salient referents tend to have more linguistic material and are often longer, whereas more salient referents can be mentioned using shorter forms like pronouns. Determining the saliency of a particular referent is at the core of many anaphoric resolution algorithms.

2.4. Linguistic form and salience

There are linguistic clues to whether an entity is in focus. Gundel proposed the Givenness Hierarchy [12], relating six cognitive statuses to the form of a referring expression in discourse. The Givenness Hierarchy arranges cognitive statuses in order of accessibility, from most accessible on the left (i.e., currently in focus) to least accessible on the right (i.e., not in focus) and shows examples of referring expressions that map to the amount of accessibility on the hierarchy:

$$\underset{it}{\text{In focus}} > \underset{\substack{that\\this\\this\ N}}{\text{activated}} > \underset{That\ N}{\text{familiar}} > \underset{the\ N}{\text{uniquely identifiable}} > \underset{indefinite\ this\ N}{\text{referential}} > \underset{a\ N}{\text{type identifiable}}$$

The particular form a linguistic expression takes signals the assumed cognitive status of the referent in the hearer's discourse model. According to the Givenness Hierarchy, pronouns like "it" and "she" are only used when an entity is in focus. Moving towards the right of the hierarchy, demonstratives like "that" and "this pain" are used when an entity is activated or in focus, and definite noun phrases, like "the pain" are only used to refer to a uniquely identifiable entity. At the far right are type identifiable expressions that are used when an entity has not been evoked yet in the discourse (e.g., "She felt 'a pain' in her chest"). Syntactic position can also offer a clue about the cognitive status of an entity; many theories specify a salience hierarchy of entities that orders referring expressions by their grammatical position in a sentence [9]. Referring expressions in syntactically prominent positions, like the subject of a sentence, are more salient than those introduced in less prominent positions, like a direct object, which are in turn more salient than expressions in other positions [10].

Because of the differences in linguistic form and saliency, anaphoric and coreference resolution algorithms are often specialized based on the form of the referring expression. For example, Denis and Baldridge [15] learn separate ranking models for 3rd person pronouns, 1st/2nd person pronouns, proper names, and definite noun phrases. We discuss the various computational methodologies in a separate manuscript [3].

In our corpus, we annotated the form of the referring expression and its syntactic position in a sentence to investigate the relationship between saliency and linguistic form in anaphoric reference in clinical reports.

2.5. Types of information considered in anaphoric resolution algorithms

Performing anaphoric reference requires a wide range of linguistic knowledge [3], and automated algorithms try to capture that knowledge as rules and features used to train a system. These linguistic features motivated the elements of our annotation scheme and the measurements we performed to characterize the corpus.

2.5.1. Lexical and morphological attributes

Some anaphors can be successfully resolved based only on the number and gender of the antecedent. For instance, in (4) the correct antecedent is the one that matches the number and gender of the pronoun (pronominal) anaphor.

(4) The surgeon photographed **the tumor** after removing **it**.

In non-pronominal coreference, one important indicator of the identity relationship is overlapping or identical phrases in the antecedent and anaphor, as with "pain" and "chest pain" in (5).

(5) Patient complains of **chest pain**. **The pain** radiates down her right arm.

2.5.2. Syntactic and grammatical attributes

Syntactic information plays a central role in anaphoric resolution—especially for intrasentential anaphora [16]. Although world knowledge must often be applied to select from among possible antecedents, the vast majority of possible antecedents for pronouns can be derived by purely syntactic considerations. Rule-based anaphora resolution algorithms incorporate syntactic preferences in a variety of ways, including the order in which the algorithms search for antecedents [17], weights assigned to syntactic classes [18], and ranking of antecedent candidates [14]. Machine learning classifiers for anaphoric reference also include a variety of syntactic attributes [3,19,20].

We captured the syntactic form of the anaphor and antecedent by annotating the *phrasal tag* of the anaphors in our corpus with the values noun phrase, pronoun, clause, or other. Noun phrases were further categorized as indefinite (e.g., "a nodule"), definite (e.g., "the surgery"), bare (e.g., "headache"), demonstrative (e.g., "this pain"), and proper (e.g., "Dr. XXX"); pronouns were further categorized as personal (e.g., "she"), possessive (e.g., "her"), demonstrative (e.g., "this"), and relative (e.g., "which") [2]. To help distinguish proper noun phrases from other types of noun phrases, if a markable was classified in the UMLS as having a semantic type of Disease or Syndrome, it was not annotated as a proper noun phrase. However, if the markable belonged to the semantic type People, the markable was annotated as a proper noun phrase.

In many theories, the grammatical positions of the antecedent are ordered hierarchically so that entities in a subject position are the most salient, those in the object position are less salient, and those in other positions are the least salient. We annotated all markables with a *function tag* that indicated the grammatical function of the markable as subject, object, modifier, or section heading, a function specific to clinical reports. Subjects were further categorized as surface subject, logical subject, or predicate nominal subject; objects were further categorized as direct, indirect, or prepositional.

2.5.3. Semantic attributes

There are many cases in which the morphological, lexical, and syntactic information are not sufficient for resolving anaphors. In those cases, semantic and pragmatic information are essential [21]. Semantic knowledge is particularly important for intersentential anaphora and for indirect noun phrase anaphora [11].

The selectional restrictions a verb places on its arguments can help filter candidate referents. For example, in (4), candidate antecedents for "it" include "surgeon" and "tumor." Selectional restrictions for the verb "remove" help the reader select the correct

antecedent "tumor" since a tumor is more likely to be removed than a person.

Lexical semantics (i.e., who and what the words of a language denote [22,23]) has been demonstrated to be one of the most important features of successful reference resolution for pronouns [24]. We annotated the *semantic type* of the markables in our corpus, using a subset of the UMLS semantic types [25]: People, Anatomical Site, Disease or Syndrome, Sign or Symptom, Procedure, Lab or Test Result, Indicator, Reagent, or Diagnostic Aid, Organ or Tissue Function, Other, and None.

2.6. Existing annotated corpora

The NLP community focusing on general English has used two coreference data sets [6,26] to develop and evaluate algorithms for coreference resolution. The annotation schema for MUC-7 includes annotated entities with an identity relation (i.e., coreference). The GNOME project [27] extended the annotations to include set/subset and part/whole relations. The ACE [28] annotation schema added appositive and predicative phrases to the identity relation links. Furthermore, the C3 project used a set of ACE guidelines to allow entities of unknown types to be included in the annotation of 135 files from the Discourse GraphBank coherence corpus [29]. Through a multi-institution collaboration, the OntoNotes project created a large-scale coreference corpus across three languages (English, Chinese, and Arabic) and across various genres of text such as news articles, conversational telephone speech, weblogs, broadcast and talk shows. The annotations included entities and events and were not limited to noun phrases or a limited set of entity types [30].

In the biomedical literature domain, the GENIA corpus contains almost 2000 Medline abstracts that were collected using the MeSH terms "human," "blood cells," and "transcription factors" [31]. The GENIA-MedCo coreference corpus annotated coreference information in the GENIA collection and in full biology papers in the same domain [32]. The BioNLP Shared Task 2011, Protein/Gene Coreference Task used the GENIA-MedCo coreference corpus to address the issue of finding anaphoric references to proteins or genes [33].

2.7. Anaphoric resolution algorithms

Most algorithms attempt to capture syntactic, semantic, and pragmatic constraints for pruning the number of potential antecedents that could be paired with an anaphor and preferences for selecting one antecedent over another. Many of the attributes we annotated in our corpus were motivated by the constraints and preferences commonly incorporated in reference resolution algorithms. In Zheng et al. [3], we detail the advances in anaphoric resolution algorithms. Here we present a high-level description. Automated anaphoric reference resolution has been a focus of research since the 1960s. Initially, several heuristic algorithms were developed, but in the last 15 years the focus has shifted to statistical algorithms. All algorithms have a common goal of identifying candidate antecedents for an anaphor and selecting from that list the best antecedent. A candidate antecedent could be a single entity or a chain of entities that have already been linked together.

In the next section, we briefly describe our corpus of clinical notes and the annotations we captured and analyzed in this paper. A very detailed description of the annotation schema and inter-annotator agreement scores is presented in [2].

3. Methods: corpus and annotation schema

Three human experts annotated instances of anaphoric reference in a set of clinical reports. Two of the experts are trained ICD coders with extensive experience annotating clinical reports for NLP; the third expert is a knowledge engineer in pathology with a degree in linguistics. A detailed description of the corpus, the annotation task and inter-annotator agreement can be found in [2]. We provide here a summary of the annotated corpus.

To characterize anaphoric reference in clinical reports, we compiled 180 reports (105,082 tokens) from a variety of University of Pittsburgh Medical Center (UPMC) hospitals and from Mayo Clinic. From UPMC notes, we randomly selected 20 notes each of emergency department notes (er), discharge summaries (ds), surgical pathology notes (sp), radiology notes (rad) from an existing corpus that had been annotated for symptoms, signs, findings, and diagnoses as part of a previous study [34].

The Mayo Clinic set comprised 100 notes, 50 each of pathology notes (pa) and clinical notes (cc), which were selected randomly from the Mayo Clinic Electronic Medical Record system to represent a mix of report types. CC notes represent a random mix of clinical report types. The Mayo Clinic set is a subest of a pre-existing corpus annotated with named entities of type disorders, signs/symptoms, procedures and anatomy [35].

The pre-existing corpus was manually reviewed to select notes exhibiting anaphoric relations. Three human annotators were asked to follow guidelines we developed for this project (guidelines are available as an online supplement to [2] and at http://nlp-ecosystem.ucsd.edu) and to identify anaphoric relations between pre-annotated named entities. Annotators were allowed to add missing markables if the missing markable participated in an anaphoric relation. The annotation schema we developed (Fig. 1) was modeled after the MUC-7 coreference task definition [6]. Each coreferential markable is linked in an anaphor–antecedent pair. Pairs consisting of markables referring to the same entity represent an anaphoric chain. We supplemented the schema for identity relations with instructions for annotating part/whole and set/subset relations to capture all three pair relation types. We also modified the schema slightly based on the content of clinical reports: we added section heading as a function tag, added predicate nominal as a type of subject, and specified semantic types found in clinical reports. Human experts applied the schema to capture markables, pairs of markables with an anaphoric relation, and chains of pairs so that the annotations could be used to train pair-based and chain-based reference resolution algorithms. The schema identifies lexical, syntactic, and semantic information critical in performing anaphoric reference resolution, along with a surrogate measurement of processing complexity called the Bagga class.

In a 1998 paper, Bagga [36] presented a framework for evaluating coreference resolution algorithms that considers the relationship of an antecedent to its anaphor. The framework breaks the coreference task into eleven classes and orders the classes by the amount of processing required to resolve the references. According to Bagga, appositives, in which the title of the person is listed immediately after his/her name separated by a comma, takes the least processing power to resolve, whereas pairs that require external knowledge to resolve take the most. We included one additional class indicating that resolution of the pair requires knowledge contained within an ontology, which could be considered a special type of world knowledge. Our reasoning for creating this additional category was to determine to what extent implementing ontological knowledge in a reference resolution algorithm would be required. Table 1 describes the Bagga classes with examples from our corpus. Quoted pronouns did not occur in our corpus.

From the guidelines, we created a schema for the Knowtator annotation tool [37] (schema can be downloaded at http://nlp-ecosystem.ucsd.edu). Knowtator provided a user interface for the annotators to find and highlight the appropriate markables, pairs, and chains for the annotation task and to fill in the appropriate

Fig. 1. Annotation scheme, including markables, pairs of markables, and chains of pairs. Markables and pairs were annotated with properties like semantic type and pair relation, respectively.

Table 1
Bagga classes and examples from our corpus. Classes that did not occur in the corpus are marked as N/A.

Number	Bagga class	Example
1	Appositives	Her **primary care physician**, **Dr. **NAME[ZZZ]**, was contacted...
2	Syntactic equatives	**The patient** is **a **AGE[in 70s]-year-old female** who came to the emergency room...
3	Proper Names	...the attending physician, Dr. ****NAME[ZZZ]**, felt she was stable... She would therefore be discharged today with followup with Dr. ****NAME[ZZZ]** in 1 week
4	Pronouns	...she had a workup that included **a CT scan of the abdomen which** showed abscesses...
5	Quoted speech pronouns	N/A
6	Demonstratives	...**a endovascular recanalization** of his SVC... He underwent **this procedure** without complication
7	Exact matches	...**hepatic duct stricture** seen on prior ERCP... she has had findings of **hepatic duct stricture**...
8	Substring matches	Fusion of **the left C3–4 facet**... APPARENT FUSION OF **THE LEFT C3–4 FACET JOINT**
9	Identical lexical heads	**The current marrow** findings... The phenotype of **the present marrow** evaluation is similar...
10	Synonyms	Status post PCI to **the RCA**... **Right coronary artery** is a large dominant vessel...
11	External world knowledge	**Mrs. Smith** presents with complaints of chest pain. PAST HISTORY: **The patient** suffered from an MI in...
12	Ontological knowledge	...a chief complaint of **abdominal** pain... it started in her **epigastric area**...

metadata for each type of annotation. Inter-annotator agreement on the corpus varied, depending on the type of annotation, and was moderate for the UPMC dataset (0.41) and high for the Mayo dataset (0.66). The annotations were compared and reviewed within Knowatator. Any annotations that were not identical were reviewed and discussed by two annotators until consensus was reached. After consensus, all of the annotations were exported from Knowtator into an XML file in a lossless fashion for further analysis. Using these annotations, we developed a machine learning module [38] to resolve coreference markables. Annotations from the UPMC reports are available for research purposes at http://www.dbmi.pitt.edu/nlpfront, and those from the Mayo

notes are available on an individual basis through a Data Use Agreement.

The schema and guidelines have since been applied by a new group of annotators to another dataset for annotation of coreference relations as part of the 2011 i2b2/VA challenge [39]. Several research groups have used the additional annotations, along with the annotations we developed, to train and evaluate coreference resolution systems [38].

In the following section, we describe the distribution of annotations in our corpus of clinical reports.

4. Results

Three experts annotated a set of 180 clinical reports for anaphoric reference. In our presentation of results, we distinguish between notes that describe findings from a procedure (i.e., surgical pathology (sp and pa) and radiology (rad)) and notes that describe a narrative story of the patient's visit (cc, ds, and er). Table 2 shows the number of annotations performed on the corpus, which resulted in 7214 markables (average 40 per report), 5992 pairs (average 33 per report), and 1304 identity chains (average 7 per report).

The length of the reports differed by report type, ranging from eight sentences (pa) to 100 sentences (er) per report, as demonstrated in Fig. 2.

Median sentence length for most report types ranged from 9 to 14 tokens, with a notable exception of sp reports, with a median of 25 tokens (Fig. 3).

As shown in Table 3, the mean number of anaphoric markables per sentence ranged from 0.70 (sp) to 1.5 (pa), with high standard deviations, indicating that some sentences have few or no anaphoric markables (minimum of 0.09–0.8) and others have several (maximum of 1.3–2.8). Reports from Mayo (cc and pa) were more dense with anaphoric markables but also showed higher standard deviations.

4.1. Markables

4.1.1. Semantic type

According to Table 4, the most prevalent semantic type for anaphoric markables was People, which accounted for 51% of all markables. The next most prevalent types were Anatomical Site (14%) and Disease or Syndrome (14%). Because most mentions of people in a clinical report refer to the patient, and because identifying non-patient mentions can probably be accomplished in a simpler way than with anaphoric reference, anaphoric reference for the semantic type People may be of lower priority than other semantic types. If we remove markables with the type People, Anatomical Site and Disease or Syndrome each comprise 30% of the markables, with Sign or Symptom comprising 16% and Procedure 15%. Markables with type Anatomical Site are especially prevalent in procedural notes (pa, rad, sp), which focus on descriptions of pathological and radiological findings of anatomical structures. Very few markables were labeled with the type Other and included non-specific references, such as those shown in (6)and (7).

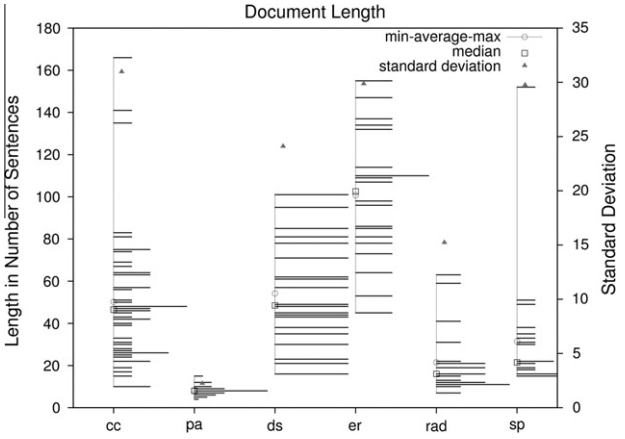

Fig. 2. Length of reports in the corpus measured by number of sentences.

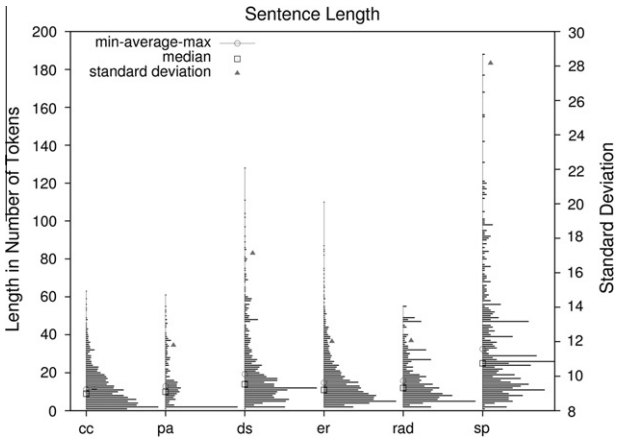

Fig. 3. Length of sentences by type of report and measured by number of tokens.

Table 3
Mean number of anaphoric markables per sentence.

Report type	Mean	Median	Minimum	Maximum	Standard deviation
cc	0.98	0.85	0.09	2.39	0.52
pa	1.51	1.47	0.50	2.75	0.45
ds	1.21	1.11	0.81	2.17	0.31
er	1.15	1.08	0.59	2.09	0.34
rad	0.84	0.82	0.20	1.57	0.37
sp	0.71	0.67	0.19	1.33	0.29

(6) Motor stimulation of **each site** was performed using 5 Hz and sensory stimulation of **each site** was performed using 50 Hz.

(7) As a final note, we briefly discussed back extensor strengthening exercises for osteoporosis, and I think she would be an excellent candidate for a home program of **that variety**.

Table 2
Number of annotations performed on corpus.

	cc	pa	ds	er	rad	sp	Total
Number of markables	2309	633	1279	2228	352	413	7214
Average number of markables per report	46.18	12.66	63.95	111.40	17.60	20.65	40.08[a]
Number of pairs	1962	440	1068	1953	255	314	5992
Average number of pairs per report	39.24	8.80	53.40	97.65	12.75	15.70	33.29[a]
Number of identity chains	409	173	212	289	112	109	1304
Average number of chains per report	8.18	3.46	10.60	14.45	5.60	5.45	7.24[a]

[a] This is an average over the entire corpus, not a sum of the individual averages.

Table 4
Proportion of markables by semantic type.

	cc (%)		pa (%)		ds (%)		er (%)		rad (%)		sp (%)		Total (%)	
People	58	(1336/2309)	0	(0/633)	61	(775/1279)	65	(1454/2228)	2	(7/352)	29	(121/413)	51	(3693/7214)
Disease or syndrome	14	(324/2309)	29	(184/633)	14	(182/1279)	8	(182/2228)	25	(89/352)	18	(75/413)	14	(1036/7214)
Anatomical site	6	(132/2309)	61	(388/633)	7	(89/1279)	5	(118/2228)	51	(181/352)	29	(121/413)	14	(1029/7214)
Sign or symptom	9	(207/2309)	0	0/633)	6	(75/1279)	13	(280/2228)	1	(4/352)	0	(0/413)	8	(566/7214)
Procedure	8	(188/2309)	6	(35/633)	8	(100/1279)	4	(80/2228)	14	(49/352)	19	(79/413)	7	(531/7214)
None	4	(92/2309)	0	(1/633)	4	(50/1279)	4	(82/2228)	3	(12/352)	2	(7/413)	3	(244/7214)
Other	1	20/2309)	1	(7/633)	0	(1/1279)	0	(1/2228)	1	(3/352)	0	(2/413)	0	(34/7214)
Laboratory or test result	0	(6/2309)	0	(0/633)	1	(7/1279)	0	(8/2228)	0	(0/352)	0	(1/413)	0	(22/7214)
Organ or tissue function	0	(2/2309)	0	(0/633)	0	(0/1279)	1	(17/2228)	0	(0/352)	1	(3/413)	0	(22/7214)
Missing type	0	(2/2309)	0	(0/633)	0	(0/1279)	0	(6/2228)	2	(7/352)	1	(4/413)	0	(19/7214)
Indicator reagent diagnostic aid	0	(0/2309)	3	(18/633)	0	(0/1279)	0	(0/2228)	0	(0/352)	0	(0/413)	0	(18/7214)

4.1.2. Phrasal tags

As shown in Table 5, the most prevalent phrasal tag in the corpus was Bare NP (34%), followed by Pronoun Personal (21%), NP Definite (17%), and Pronoun Possessive (9%). This finding supports previous studies in clinical narratives demonstrating missing determiners, which makes part-of-speech tagging and parsing more difficult [40]. The abundance of bare noun phrases may also pose a challenge due to the fact that many anaphoric resolution algorithms target definite noun phrases and personal pronouns as markables to be resolved. Bare noun phrases pose specific challenges for determining whether the noun phrase is an anaphor candidate (also known as anaphoricity discovery).

Examining phrasal tags in the context of semantic types, we observed that Sign or Symptom is almost always a Bare NP (469/566), People are seldom Pronoun Demonstrative, and Anatomical Site is almost always Bare NP or NP Definite. Most Pronoun Demonstratives were of the semantic type None, as in (8) and (9), because non-human pronouns by themselves do not have a semantic type but inherit one through their antecedents.

(8) **This** is a ##-year-old female with chronic neck pain for 20 years.
(9) While **these** appear to be due to face arthrosis, no reformats were obtained...

4.1.3. Function tags

Half of the anaphoric markables in the corpus performed the function of SubjectSurface (Table 6). SubjectSurface, ObjectPrepositional, and ObjectDirect combined account for 72% of the anaphoric markables. This result is consistent with theories of saliency that claim subjects are more salient than objects, which are in turn are more salient than other syntactic functions.

A notable exception to the trend for an anaphoric markable to be a SubjectSurface is markables of type Sign or Symptom, which are more often ObjectDirect or ObjectPreposition (see example (10)).

(10) The patient also experienced **associated chest pain** with these episodes.

We found several dependent relationships between function tag and semantic type, as shown in Table 7.

Only 2% of the markables were SubjectPredicateNominals, but nearly all of those were of semantic type People or Disease or Syndrome, which makes sense when we examine examples *in* (11), (12), and (13):

(11) This is a ****AGE[in 50s]-year-old male** who comes in complaining of several days of voice hoarseness.
(12) There may be **a tiny right pleural effusion**.
(13) ...this is **a viral pharyngitis**...

The function tag SectionHeading was usually of semantic type Anatomical Site or Procedure. An ObjectIndirect function tag was overwhelmingly of the semantic type People. An example of this is "Therefore, we continued ***her*** on her home dose of Cortef."

Table 5
Proportion of markables by phrasal tag.

	cc (%)	pa (%)	ds (%)	er (%)	rad (%)	sp (%)	mean (%)
Bare NP	29	18	14	23	8	9	34
Clause	32	0	16	32	16	3	0
Missing phrasal tag	48	14	0	6	17	16	1
NP definite	19	9	21	40	8	3	17
NP demonstrative	41	1	14	29	4	11	1
NP indefinite	26	10	21	33	4	5	5
Noun proper	31	0	31	24	0	14	4
Pronoun demonstrative	32	0	14	46	8	0	1
Pronoun personal	49	0	15	35	0	2	21
Pronoun possessive	33	0	26	37	0	4	9
Pronoun relative	27	0	25	42	3	2	3
Other	27	10	25	27	5	5	2

Table 6
Function tags for anaphoric markables.

	cc (%)		pa (%)		ds (%)		er (%)		rad (%)		sp (%)		Total (%)	
Missing function tag	2	(52/2309)	2	(15/633)	0	(0/1279)	0	(6/2228)	5	(18/352)	4	(17/413)	1	(108/7214)
Modifier to object direct	6	(140/2309)	1	(7/633)	6	(78/1279)	4	(79/2228)	2	(6/352)	1	(6/413)	4	(316/7214)
Modifier to object indirect	0	(3/2309)	0	(0/633)	0	(0/1279)	0	(0/2228)	0	(1/352)	0	(0/413)	0	(4/7214)
Modifier to object prepositional	7	(162/2309)	4	(23/633)	12	(156/1279)	9	(194/2228)	6	(21/352)	8	(31/413)	8	(587/7214)
Modifier to subject nominal predicate	0	(3/2309)	0	(1/633)	0	(3/1279)	0	(7/2228)	5	(16/352)	0	(0/413)	0	(30/7214)
modifier to subject surface	6	(128/2309)	10	(66/633)	9	(112/1279)	6	(138/2228)	16	(58/352)	18	(74/413)	8	(576/7214)
Object direct	13	(290/2309)	10	(61/633)	10	(126/1279)	11	(241/2228)	4	(15/352)	4	(18/413)	10	(751/7214)
Object indirect	2	(41/2309)	0	(0/633)	1	(8/1279)	1	(15/2228)	0	(0/352)	0	(0/413)	1	(64/7214)
Object prepositional	11	(258/2309)	17	(110/633)	13	(167/1279)	11	(246/2228)	22	(77/352)	8	(31/413)	12	(889/7214)
Other	0	(6/2309)	0	(1/633)	0	(2/1279)	1	(18/2228)	0	(0/352)	0	(1/413)	0	(28/7214)
Section heading	0	(2/2309)	0	(0/633)	1	(7/1279)	1	(22/2228)	9	(32/352)	1	(6/413)	1	(69/7214)
Subject logical passives	0	(10/2309)	0	(2/633)	1	(17/1279)	1	(18/2228)	0	(0/352)	0	(2/413)	1	(49/7214)
Subject predicative nominal	2	(37/2309)	0	(0/633)	2	(31/1279)	3	(68/2228)	6	(21/352)	3	(12/413)	2	(169/7214)
Subject surface	51	(1177/2309)	55	(347/633)	45	(572/1279)	53	(1176/2228)	25	(87/352)	52	(215/413)	50	(3574/7214)

4.2. Pairs

From the 7214 markables, annotators created 5992 pairs. The main type of relation was that of identity (91%). However, set/subset (5%) and part/whole (4%) relations also occurred in every type of report (see Table 8). We included two additional categories—Separate Instances Of Same Concept (see (14)) and Other—and used the label Missing Pair Relation for annotation omissions. All of these categories were extremely rare.

(14) The patient on **CT scan** showed to have the right hilar mass with collapse of right lower lobe that was seen in **DATE[Jan 01 2008]... On **DATE[Jan 27 2008], ***CT of the chest*** and also a right pleural effusion...

Procedural notes had a higher prevalence of non-identity relations than narrative notes. For instance, pa notes had a prevalence of 25% for part/whole relations and 11% for set/subset. The pa notes consistently listed in the patient history the organ being examined (i.e., the "whole"), then systematically described findings for "parts" of that organ that were examined. For example, in (15), three part/whole pairs were generated from the four markables in bold (we allowed overlapping annotations and also allowed disjoint annotations, which are indicated by "..."): Colon—Colon, right; Colon—Colon... transverse; Colon—the submucosa.

(15) **Colon**, **right** and **transverse**, resection: Residual invasive grade 3 (of 4) adenocarcinoma is present in an ulcerated region of previous biopsy. The carcinoma invades into **the submucosa** only.

In addition, immunohistochemistry panels and other ancillary tests were ordered ("set") and the results analyzed in the text ("subset"). For instance, in (16), seven subsets were paired with Immunohistochemical studies as the set.

(16) **Immunohistochemical studies** reveal that the tumor cells do not react with antibodies to **CD117**, **CD34**, **desmin**, **actin**, keratin (**AE1/AE3** and **wide spectrum)** or **S100 protein.**

The consistency of these patterns in the pa reports seemed largely due to a sort of dictation template or synoptic that pathologists employed when writing the reports.

The vast majority of antecedent–anaphor pairs were of the same semantic type (see Table 9), suggesting that semantic type match could be used as a filtering criteria or as a feature for anaphoric reference resolution development. Exceptions included anaphors with semantic type Other and None (i.e., non-human pronouns), and antecedent–anaphor pairs between Sign or Symptom and Disease or Syndrome, as in example (17) in which "increasing left-sided facial weakness" is a Sign or Symptom and "complete left-sided facial paresis" is a Disease or Syndrome. A rare exception occurred when a Disease or Syndrome referred to an Anatomical Site (see example (18)).

(17) The patient reports a 3-day history of **increasing left-sided facial weakness**... The patient has **complete left-sided facial paresis** involving his forehead.

(18) It was noted that she had **a perforated septum**... I do have the operative report for the **septal** reconstruction.

4.3. Identity Chains

From 5992 pairs of coreferring markables, annotators created 1304 identity chains for pairs referring to the same concept. Fig. 4 is a visualization of the chains annotated in one report. Non-identity links are shown with dotted lines. In this report, a chain for otitis externa contains six markables with surface strings "Otitis externa", "the seborrheic otitis externa", "Right otitis externa", and "a rather stubborn otitis externa." The chain indicating the patient contains a number of markables throughout the entire

Table 7
Relationships between function tag and semantic type.

	Anatomical site	Disease or syndrome	Indicator reagent diagnostic aid	Laboratory or test result	Missing type	None	Organ or tissue function	Other	People	Procedure	Sign or symptom
Missing function Tag	22	66	0	89	17	193	53	0	162	1	51
Modifier to object direct	13	54	0	77	3	60	193	0	224	7	3
Modifier to object indirect	0	1	0	6	0	4	0	0	4	0	0
Modifier to object prepositional	1	0	0	0	0	0	8	0	3	0	0
Modifier to subject nominal predicate	19	0	0	0	0	0	0	0	0	0	0
Modifier to subject surface	3	3	0	1	0	1	15	0	25	0	0
Object direct	3	2	0	2	0	0	3	0	6	0	0
Object indirect	3	0	0	1	0	2	8	0	10	0	0
Object prepositional	14	142	4	340	6	274	145	63	250	1	0
Other	22	18	0	27	4	33	132	1	73	4	13
Section heading	8	30	0	44	0	9	194	0	132	15	2
Subject logical passives	22	66	0	89	17	193	53	0	162	1	51
Subject predicative nominal	13	54	0	77	3	60	193	0	224	7	3
Subject surface	0	1	0	6	0	4	0	0	4	0	0

Table 8
Type of pairs and the proportion with which they occurred in the corpus.

	cc (%)		pa (%)		ds (%)		er (%)		rad (%)		sp (%)		Total (%)	
Identity	91	(1794/1962)	63	(278/440)	95	(1012/1068)	98	(1905/1953)	75	(191/255)	88	(275/314)	91	(5455/5992)
Part/whole	1	(26/1962)	25	(111/440)	2	(21/1068)	1	(15/1953)	15	(39/255)	7	(21/314)	4	(233/5992)
Set/subset	7	(140/1962)	11	(50/440)	3	(32/1068)	2	(30/1953)	10	(25/255)	6	(18/314)	5	(295/5992)
Other	0	(1/1962)	0	(0/440)	0	(0/1068)	0	(0/1953)	0	(0/255)	0	(0/314)	0	(1/5992)
Separate instances	0	(1/1962)	0	(0/440)	0	(3/1068)	0	(2/1953)	0	(0/255)	0	(0/314)	0	(6/5992)
Missing relation	0	(0/1962)	0	(1/440)	0	(0/1068)	0	(1/1953)	0	(0/255)	0	(0/314)	0	(2/5992)

Table 9
Semantic types of antecedent–anaphor pairs. Anaphors are columns, and antecedents are rows.

	Anatomical site	Disease or syndrome	Indicator reagent diagnostic aid	Laboratory or test result	Missing type	None	Organ or tissue function	Other	People	Procedure	Sign or symptom
Anatomical Site	*741*	3	0	0	1	4	0	8	0	1	0
Disease or syndrome	6	*665*	0	0	2	61	0	11	0	0	14
Indicator reagent diagnostic aid	0	0	*15*	0	0	0	0	1	0	0	0
Laboratory or test result	0	1	0	*6*	3	1	0	0	0	1	1
Missing type	0	6	0	3	*2*	3	0	0	0	1	0
None	1	30	0	0	1	*26*	1	0	36	25	25
Organ or tissue function	0	0	0	0	0	3	*13*	0	0	1	0
Other	3	3	0	0	0	3	0	*5*	0	3	0
People	0	0	0	0	0	28	0	0	*3407*	0	0
Procedure	1	0	0	6	0	72	1	7	0	*305*	0
Sign or symptom	0	7	0	2	0	43	0	4	0	1	*368*

Italicized values represent pairs with the same semantic type.

report. The first mention (“the patient”) also has a set/subset relationship with the markable “them,” as does a markable with the surface string “wife.” A mention of “the ear canal” has a set/subset relationship with an identity chain for “the ear canals,” and a chain for “the right ear” has a part/whole relationship with the mention of “the ear canal.”

Table 10 shows the average number of unique anaphoric chains per report, which ranged from 3.46 (pa) to 14.45 (er). Narratives, which are longer, contained more unique chains.

Length of chains, measured in markables, also varied by report type. As shown in Fig. 5, almost half (607/1304) of all chains comprised only two markables, and 85% (1109/1304) of chains comprised five or fewer. Chains in narratives were sometimes very long, with 4% (51/1304) of chains comprising twenty or more markables. On inspection of chains, almost all of the very long chains refer to the patient. Appendix A contains visualizations of chains for several reports that include some long patient chains.

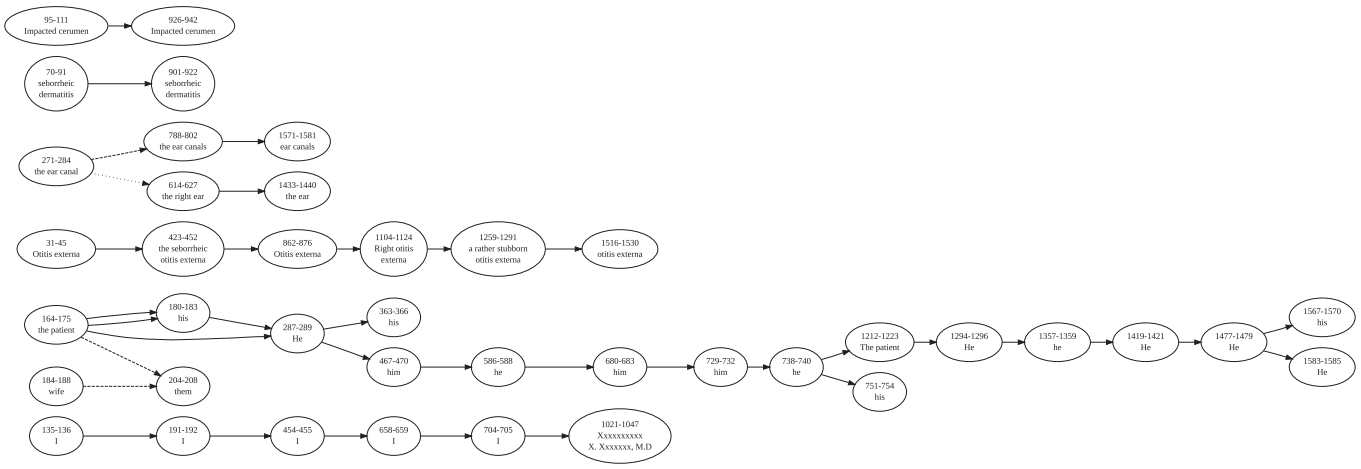

Fig. 4. A graphical visualization of all anaphoric markables and chains in a report from the corpus. Solid lines indicate identity relations, dashed indicate set/subset, and dotted indicate part/whole.

Table 10
Average number of unique anaphoric chains per report.

cc	pa	ds	er	rad	sp
8.18	3.46	10.6	14.45	5.6	5.45

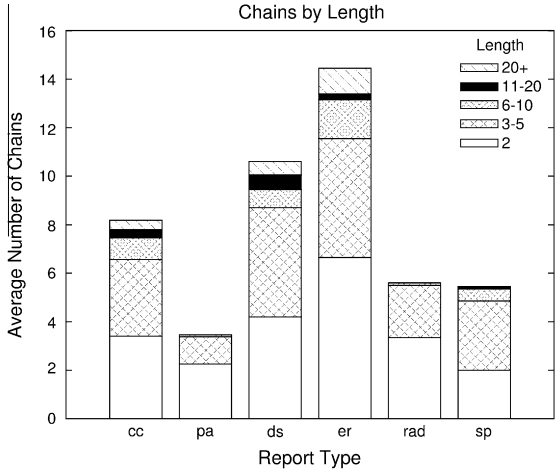

Fig. 5. Number of chains in each report by type of report. Bars are broken down by length of the chains measured by markables.

A handful of chains comprised as many as 100 markables, as shown in Fig. 6. However the vast majority of chains comprised 10 or fewer markables.

As discussed previously, the majority of pairs were semantically homogeneous (i.e., containing markables of the same semantic type), and Table 11 shows that chains were also quite homogeneous; 84% of chains (1091/1304) contained markables of only one semantic type. The proportion of homogeneity was higher for procedural notes (0.82–0.97) than for narratives (0.77–0.81).

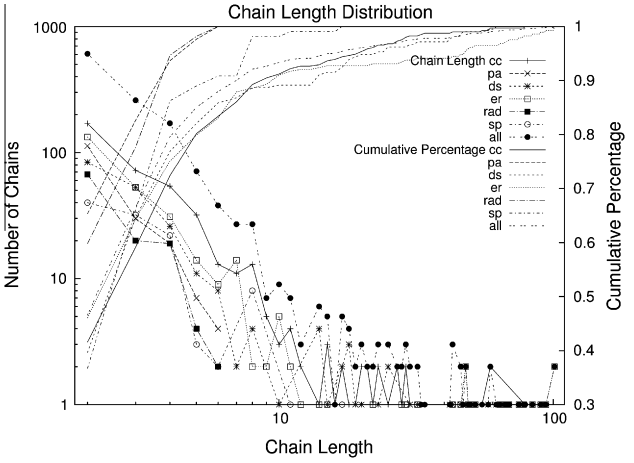

Fig. 6. Distributions of chains by length in number of markables. Left *y*-axis indicates the number of chains with that length, and the number deceases quickly to ten. Right *y*-axis indicates the cumulative percentage, which rises steadily until the length of ten and then tapers off to the maximum length of 100.

Table 11
Number of chains per report type with all markables of the same semantic type (homogeneous) and with markables of more than one semantic type (heterogeneous).

	cc	pa	ds	er	rad	sp	Total
Heterogeneous	79	6	44	67	9	8	213
Homogeneous	330	167	168	222	103	101	1091
	409	173	212	289	112	109	1304

Most of the heterogeneity was due to markables of the semantic type None (i.e., non-human pronouns like "it" that do not have a semantic type). A handful of heterogeneous chains contained combinations like Sign/Symptom and Disease/Disorder (e.g., "headache" and "migraine headaches"), Anatomical Site and Disease/

Table 12
Number of homogeneous and heterogeneous chains by length of chain in markables.

	2	3–5	6–10	11–20	20+	Total
Heterogeneous	71 (12)	78 (16)	33 (31)	13 (36)	18 (35)	213 (16)
Homogeneous	535 (88)	425 (84)	75 (69)	23 (64)	33 (65)	1091 (84)
Total	606	503	108	36	51	1304

Disorder (e.g., "an ulcerated area" and "tumor"), and Procedure and Lab Result (e.g., "blood pressure" and "elevated blood pressure").

As shown in Table 12, the amount of heterogeneity within the chains increased as the chain size increased beyond 5; however, sample size for longer chains is smaller, and more variance would be expected.

4.4. Bagga classes

Table 13 lists the distribution of pairs in our corpus annotated with each Bagga class and references the distribution for the classes from the Wall Street Journal (WSJ) corpus, as described in [36]. On the one hand, our corpus showed substantially fewer appositives (0.4% compared to 4.5%) and proper names (3.3% compared to 27.8%) than pairs in the WSJ. On the other hand, exact matches (26.9% compared to 12.6%) and external world knowledge (11.0% compared to 5.9%) were more prevalent in the clinical note corpus. If we combine external world knowledge and ontological knowledge, nearly 16% of pairs in the clinical corpus require some type of world knowledge to resolve.

Bagga class distribution differed substantially by report type (see Table 14). According to annotators, 35% of pa notes required ontological knowledge for resolution, whereas only 0.77% of pairs in er notes did. Pronouns were most prevalent in cc notes (52%) but were also frequent in ds (41%) and er (44%) notes. However pronouns accounted for only 0.23% of pairs in pa notes. Identical lexical heads also varied across report types, showing higher prevalence in procedural reports and ranging from 9% in er notes to 30% in rad reports.

4.5. Distance

We measured the distance between an antecedent and its anaphor in several different ways. Table 15. shows that the median number of intervening tokens between the pair was much longer for rad and sp notes (51 tokens and 61 tokens, respectively) than it was for the four other types of notes (cc 17, pa 17, ds 20, er 18). The mean distance was much larger than the median distance for all report types, showing differences between 9 (pa) and 79 (sp) tokens, suggesting that there were outlying anaphor–antecedent pairs with extremely long distances from each other (also evident in the minimum distance of 0 and maximum around 1500 for four report types shown in the figure). Reports showing the smallest and largest difference between the median and the mean were both pathology reports but were from different institutions, indicating that institutional practices for report dictation may differ in a way that affects anaphoric reference. Removing chains involving people increased the median antecedent–anaphor distance from a range of 17–61 to a range of 42–109. When measuring distance in terms of sentence breaks rather than tokens, findings were similar, except that rad and sp had longer median sentence breaks between antecedent–anaphor pairs (4 and 2, respectively) than other report types. The mean number of sentence breaks ranged from 1.6 (pa) to 7 (rad).

A potential heuristic for finding an antecedent could be the most recent markable of the same semantic type. For Mayo reports, which were pre-annotated for all markables, including non-anaphoric markables, we measured the number of intervening markables between an antecedent and anaphor that were of the same semantic type as the anaphor (Fig. 7). For cc notes, almost half of the pairs had no intervening markables of the same semantic type. The maximum number of intervening markables for cc notes was 33 for Disease or Syndrome, 9 for Procedure, and 18 for Sign or Symptom. This could be due to the summary nature of the last sections in a narrative report. For pa reports, 91% of the Disease or Syndrome pairs had 0 intervening markables, and only two pairs had more than 1. Procedure and Anatomical Site pairs showed more variety, with a maximum of 16 intervening markables of the same semantic type.

Table 13
Distribution of Bagga classes in our corpus and in the Wall Street Journal (WSJ) corpus as described in Bagga [36].

Number	Bagga class	Percentage in our corpus (%)	Percentage in WSJ corpus (%)
1	Appositives	0.4	4.5
2	Syntactic equatives	1.7	1.7
3	Proper names	3.3	27.8
4	Pronouns	39.4	21.0
5	Quoted speech pronouns	N/A	1.4
6	Demonstratives	2.3	2.0
7	Exact matches	16.9	12.6
8	Substring matches	4.1	7.5
9	Identical lexical heads	12.3	10.3
10	Synonyms	3.6	5.3
11	External world knowledge	11.0	5.9
12	Ontological knowledge	4.9	N/A

Table 14
Distribution of Bagga classes for each report type.

	cc (%)	pa (%)	ds (%)	er (%)	rad (%)	sp (%)
Appositives	0.20	0.00	0.66	0.61	0.00	0.00
Syntactic equatives	1.17	0.00	1.87	2.97	0.00	0.96
Proper names	4.03	0.00	4.12	2.41	0.00	9.24
Pronouns	51.63	0.23	40.73	43.57	2.35	18.15
Demonstratives	2.50	0.00	1.87	2.71	3.53	1.27
Exact matches	12.64	17.95	16.48	19.00	23.92	24.52
Substring matches	3.06	5.45	4.59	1.74	18.43	10.83
Identical lexical heads	9.17	27.73	10.67	8.70	29.80	23.57
Synonyms	2.50	5.45	4.68	4.51	1.96	0.64
External world knowledge	10.24	7.27	12.45	13.01	7.84	6.37
Ontology knowledge	2.85	35.45	1.87	0.77	12.16	4.46
Missing Bagga class	0.00	0.45	0.00	0.00	0.00	0.00

Table 15
Number of tokens between an anaphor and antecedent.

	Mean	Median	Min	Max	Standard deviation
cc	62	17	0	1469	133
pa	26	17	0	183	29
ds	77	20	0	1574	150
er	97	18	0	1453	201
rad	90	51	0	450	97
sp	137	61	1	1259	173

5. Discussion

We annotated a corpus of clinical notes with clinically relevant markables that participate in anaphoric reference or coreference, assigned syntactic and semantic attributes to the markables, and identified anaphoric pairs and chains within the corpus. The annotations illustrate the extent to which anaphoric reference occurs in clinical reports and reveal in part its nature in the clinical domain. The characteristics we measured provide some insight into how one might address automated resolution of anaphora in a similar corpus.

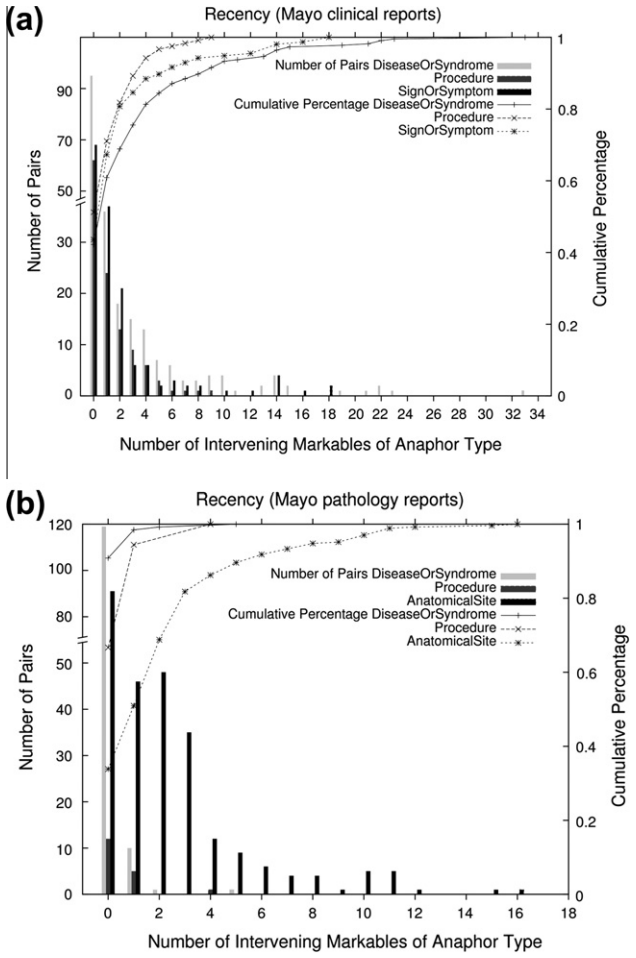

Fig. 7. Number of intervening markables of some semantic type as anaphor for Mayo cc notes (a) and pa notes (b). The bars correlate with the left *y*-axis (number of pairs) and the line plots correlate with the right *y*-axis (cumulative percentage).

This study showed that anaphoric relations were quite frequent in clinical notes—on average, every sentence in the corpus had an anaphoric markable. By far the most prevalent type of reference was that of identity, but part/whole and set/subset relations were quite prevalent in procedural notes. There were a variety of semantic types participating in anaphoric relations, and the small number of anaphoric markables labeled with the catch-all class Other (34/7214) indicates that the semantic types we included from the UMLS were comprehensive.

5.1. Implications for automated anaphoric reference resolution

A number of our findings have implications for automated anaphoric reference resolution.

Whereas the majority of resolution algorithms target anaphors that are definite noun phrases or pronouns, the most frequent phrasal type for anaphors in our corpus was bare noun phrase. Bare noun phrases pose specific challenges for determining whether an NP is an anaphor candidate, also known as anaphoricity discovery. Candidates for anaphoric resolution are usually determined by the linguistic form of the noun phrase, such as the presence of a definite article like "the" in the noun phrase. Because bare noun phrases omit the definite article, they therefore lack a strong anaphoricity predictor. The high prevalence of bare noun phrases (34% overall) suggests that determining anaphoricity could be especially challenging in a clinical corpus and that successful noun phrase identification will be an essential component of an anaphoric reference resolution application.

Another essential component will be accurate section header recognition, especially in emergency department, pathology, and radiology notes where many of the anaphoric markables are section headers (see Table 6).

Most anaphoric pairs comprise markables of the same semantic types, so the semantic type of the markables could be a critical feature in a resolution algorithm. However, the number of intervening markables of the same type is sometimes surprisingly high and differs by semantic type, indicating that the semantic type of a candidate antecedent is a necessary but not sufficient feature. An anaphor's semantic type is a feature that cannot be applied to demonstratives and pronouns, which implies the possible requirement for separate resolution modules for anaphoric relations between two noun phrases and for anaphoric relations between a noun phrase and demonstratives or pronouns.

The characteristics of anaphoric reference in our corpus differed sometimes by the genre of the clinical report. Generally, procedural notes (i.e., pathology and radiology notes) showed similar qualities, whereas more narrative descriptions (i.e., emergency department notes and discharge summaries) shared distinctive attributes. This suggests that a resolution algorithm should account for document genre. As mentioned earlier, implementing a coreference resolution algorithm for narrative reports would capture the large majority of reference resolution but would leave unresolved about one-third of the anaphoric pairs in procedural notes, which also require part/whole and set/subset resolution.

Characteristics of anaphoric reference also sometimes differed by institution. Semi-structured reporting templates or particular dictation practices, such as mentioning an organ followed by systematic descriptions of findings for parts of the organ, could influence resolution. Another instance of variation between institutions was the distance between an antecedent and anaphor. Anaphoric pairs in pathology notes from Mayo were twice as close to each other as pairs in pathology notes from Pittsburgh. These institutional differences suggest that discourse knowledge about the report structure (a feature we did not examine) could be useful—especially for part/whole relations, which are prevalent in pathology reports (see Table 8).

The mean distance between an anaphor and its antecedent was much larger than the median (see Table 15), which indicates some very long-distance pairs (e.g., 5% of pairs in er reports had 42 sentences between the anaphor and antecedent) that could be reflective of summary sections at the end of a report referring back to a concept described earlier. Many coreference resolution algorithms—especially pronominal resolution algorithms—assume salience of an anaphor decreases with distance. Our findings suggest for clinical reports that distance may not be the most accurate measure of salience and that long-distance anaphoric relationships are often valid. In fact, 20% of the anaphoric pairs in rad reports had over 145 intervening tokens (11 sentences), and 20% of pairs in sp reports were separated by more than 273 tokens (eight sentences)[1]. It would be informative to explore the relationship between section labels, report structure, and anaphoric relations, which would require discourse structure modeling. Pairs involving markables with the semantic type People were less distant, due probably to more frequent mention of the patient throughout many sections of a report (our guidelines instructed annotators to link an anaphor to its closest antecedent). This finding suggests that semantic type does not only affect whether two markables corefer but also the expected distance between an anaphor and its antecedent.

[1] Although rad reports seem more distant when measured by tokens, the sentence length in sp reports was almost three times as long as sentence length in rad reports.

5.2. Potential complexity of anaphoric resolution of clinical texts

Based on annotations of Bagga classes (see Table 13), our findings also have implications about the expected algorithmic complexity required for anaphoric reference in clinical reports.

The simplest pairs to resolve, appositives, were rare in clinical notes (0.4%). Likewise, syntactic equatives and proper names were infrequent (1.7% and 3.3%, respectively). However, exact matches, substring matches, and identical lexical heads comprised 22% of all pairs. Due largely to patient references, pronouns were more frequent than in the WSJ corpus (39% compared to 21%). Approximately 19% of anaphoric pairs require semantic knowledge to resolve (i.e., synonymy, external world knowledge, or ontological knowledge) compared to 11% of the WSJ corpus; in our corpus, for identity pairs only (which is what was annotated in the WSJ corpus), 16% require semantic knowledge.

If we remove markables with the semantic type people, the distribution of Bagga classes become markedly different. Classes that are presumably easiest to resolve became practically non-existent. However, exact matches increased by half again to account for more than one-quarter of anaphoric pairs, and prevalence of identical lexical heads doubled to 26%. Pronoun prevalence decreased from 39% to 6.5%, and the need for semantic knowledge increased from 19% to 29%. These findings suggest that to be successful, anaphoric reference resolution algorithms for clinical text will require domain knowledge about synonyms and hierarchical relations, in addition to external world knowledge.

What the best approaches are for incorporating this type of semantic knowledge for anaphoric reference is an open research question. In the general domain, successful approaches for integrating semantic knowledge include application of the WordNet hierarchy [41,42] and other online knowledge bases, such as Wikipedia infoboxes and the Freebase entity graph [43]. In the medical domain, there are a number of ontologies and taxonomies that could be leveraged for semantic knowledge [44]. Based on our findings, ontologies or other detailed knowledge sources necessary for the variety of semantic types that occur in the clinical corpus will need to not only address symptoms, findings, and diseases but also anatomy and medical procedures. For some of the part/whole and set/subset relations, knowledge from textbooks or comprehensive ontological models may be necessary to get the required granularity (see examples (15) and (16)). For example in pathology notes, ontological knowledge regarding anatomical relationships may be needed to resolve anaphora involving an organ and its parts as well as anaphora involving a tissue and its constituents. The goal of developing a general anaphoric resolution system for a variety of clinical report types will be complicated by the different types of semantic resources that may be necessary for anaphoric resolution within different report genres.

Research in the general domain has shown that semantic knowledge derived from taxonomies or ontologies is often insufficient due to knowledge gaps in the hierarchy, context-dependent relations that are not modeled in a general hierarchy (e.g., age is a risk factor), inferences only indirectly encoded in the hierarchy, and word sense proliferation (i.e., may choose incorrect word sense resulting in the wrong antecedent) [42,45]. To alleviate this problem, several researchers have developed methods for enriching knowledge bases via (semi)automatic knowledge extraction from text (see [46] for a comprehensive review), including the ODIE project that sponsored this study [47]. A few studies have shown that looking for hyponymic or synonymous relationships in unstructured text on the Web performs as well as using manually crafted ontological knowledge [41,45,48].

Based on our analysis of the corpus characteristics, technologies that will be needed for successfully building an anaphoric resolution application for clinical text include the following:

- NP, pronoun, and demonstrative identification with accurate number and gender agreement.
- Anaphoricity detection that includes bare noun phrases.
- Accurate semantic typing.
- Section identification for resolution of anatomic sites and procedures.
- A deep parser to identify the syntactic role within the sentence (the first publicly available deep parser trained on clinical narratives was released in Fall, 2011 as part of the MiPACQ project [49]).
- Modeling of relevant domain knowledge about synonyms and hierarchical relations, in addition to external world knowledge.

An automated coreference resolution module trained on this corpus has been evaluated [38] and will be made available as a UIMA module as part of the next release of ODIE [47] and as part of the cTAKES system [40].

5.3. Limitations

This study was based on annotations performed on a corpus comprising clinical reports from two healthcare institutions and six different report types. Although many similarities existed, disparities found among report types and across institutions imply that to truly model anaphoric reference in the clinical domain we need annotations from other report genres and from other institutions. Another limitation was our addition of a new Bagga class called Ontological knowledge. Distinguishing between world knowledge and ontological knowledge may be too subtle for annotators and perhaps not necessary in analyzing the corpus characteristics. Finally, the prevalence of anaphoric reference in our corpus is probably inflated due to exclusion of reports that did not contain any anaphoric reference.

5.4. Future work

Research on anaphoric reference in clinical reports is just beginning despite the fact that in the general domain there have been many advancements in the last two decades. It would be extremely valuable to synchronize our efforts in the clinical domain with those performed on the general domain to allow interoperability and methods transfer for automated algorithm development. For example, the ability to leverage annotations in the OntoNotes corpus and algorithms evaluated for the 2011 CoNLL shared task on coreference [50] could advance our efforts in the clinical domain.

Some may question the motivation for anaphoric reference in the clinical domain. We have demonstrated that anaphoric reference is plentiful in clinical reports, but we have not demonstrated that resolution of these references is necessary for accurate interpretation of the text. Future work will involve demonstrating the utility of coreference or anaphoric reference resolution in other NLP tasks applied to clinical reports. A recent study showed that coreference resolution is helpful for discovering implicit arguments [51] in general English text. Finding implicit arguments is a step towards the discovery of higher-level inferencing and implicatures. Another recent study demonstrated that information from a coreference resolver improved performance of event-argument relation extraction on a biomedical corpus [52]. Determining the extent to which anaphoric- or co-reference resolution can improve performance of other NLP tasks will be an important research area for the future. The foundation of this research will be clinical corpora with multiple layers of annotation.

The annotated corpus described here is part of the dataset for the 5th i2b2 NLP challenge on coreference resolution [39], which has supplemented this corpus with annotations on additional reports. These corpora are a partial response to the urgent appeal within

the clinical NLP community for the development of a clinical corpus annotated with layers of syntactic, semantic, and domain-specific information similar to the general domain corpora available through the Linguistic Data Consortium. Within the last few years, there have been several pioneering efforts to release corpora of de-identified clinical notes to be used for research purposes by the community [53–58]. The next step is to overlay these corpora with much needed annotations and to make the annotated corpora available to NLP researchers through appropriate data use agreements.

5.5. Conclusion

If "[a]utomatic identification of anaphoric reference in text has been an uphill battle for several decades" [50] in the general domain, we could describe our position in the clinical domain as queued up at the starting line. Anaphoric resolution is extremely challenging partly because it requires syntactic, semantic, and world knowledge and partly because we lack substantial annotated data. This paper describes the characteristics of the first clinical corpus annotated with three types of reference: identity, part/whole, and set-subset. In addition to annotated pairs and identity chains, the corpus is annotated with syntactic and semantic features potentially useful in performing automated reference resolution. From our characterization of the corpus, we conclude that anaphoric reference is prevalent in many types of clinical reports, that annotations of noun phrases, semantic type, and section headings may be especially important for automated resolution of anaphoric reference, and that separate modules for reference resolution may be required for different report types, different institutions, and different types of anaphors. We also conclude that accurate resolution will require extensive domain knowledge—especially for pathology and radiology reports with more part/whole and set/subset relations. We hope researchers will leverage the annotations in this corpus to develop automated algorithms and will add to the annotations to generate a more extensive corpus.

Acknowledgments

This work was funded by R01 CA127979. The ODIE toolkit – software for information extraction and ontology development. We are thankful to our annotators – Donna Ihrke, Pauline Funk, and Melissa Tharp – and to Lynette Hirschman, Cheryl Clark, and Kevin Cohen for excellent feedback. The work was conducted under IRB 08-007020 and REN09050055/PRO07070252.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2012.01.010.

References

[1] Hahn U, Romacker M, Schulz S. MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports. Int J Med Inform 2002;67(1–3):63–74.
[2] Savova GK, Chapman WW, Zheng J, Crowley RS. Anaphoric relations in the clinical narrative: corpus creation. J Am Med Inform Assoc: JAMIA 2011;18(4):459–65 (July 1).
[3] Zheng J, Savova GK, Chapman WW, Crowley RS. Coreference resolution: a review of general methodologies and applications in the clinical domain. J Biomed Inform 2011;44:1113–22.
[4] Olsson F. A survey of machine learning for reference resolution in textual, discourse; 2004.
[5] vanDeemter K, Kibble R. On coreferring: coreference in MUC and related annotation schemes. Comput Linguist 2000;26(4).
[6] Hirschman L, Chinchor N. MUC-7 Coreference task definition, version 3.0. In: Proc of MUC-7; 1997.
[7] Webber BL. Anaphora and anaphoric resolution. <http://www.inf.ed.ac.uk/teaching/courses/anlp/slides/anlp23-2x2.pdf> and <http://www.inf.ed.ac.uk/teaching/courses/anlp/slides/anlp24-2x2.pdf> [accessed 25.07.11].
[8] Webber BL. Description formation and discourse model synthesis. In: 1978 Workshop on theoretical issues in natural language processing. Stroudsburg, PA: ACL; 1978. p. 42–50.
[9] Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, New Jersey: Prentice-Hall Inc.; 2000.
[10] Gundel JK, Hegarty M, Borthen K. Cognitive status, information structure, and pronominal reference to clausally introduced entities. J Logic, Lang, Inform 2002;3:299–319.
[11] Mitkov R. Anaphora resolution. New York: Longman; 2002.
[12] Gundel JK, Hedberg N, Zacharski R. Cognitive status and the form of referring expressions in discourse. Language 1993 Jun 1;69(2):274–307.
[13] Ariel M. Accessibility theory: an overview. In: Sanders T, Schilperoord J, Spooren W, editors. Text representation: linguistic and psycholinguistic aspects. Amsterdam: John Benjamins; 2001.
[14] Grosz BJ, Weinstein S, Joshi AK. Centering: a framework for modeling the local coherence of discourse. Comput Linguist 1995;21(2):203–25.
[15] Denis P, Baldridge J. Learning specialized ranking models for coreference resolution. In: Conference on empirical methods in natural language processing; 2008. p. 660–9.
[16] Nash-Webber B, Beranek B, Newman I, Reiter R. Anaphora and logical form: on formal meaning representations for natural language. In: 5th International joint conference on, artificial intelligence; 1977. p. 121–31.
[17] Hobbs J. Resolving pronoun references. Lingua 1978;44:311–38.
[18] Lappin S, Leass H. An algorithm for pronominal anaphora resolution. Comput Linguist 1994;20(4):535–61.
[19] Rahman A, Ng V. Supervised models for coreference resolution. EMNLP; 2010. p. 968–77.
[20] Ng V, Cardie C. Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th annual meeting of the association for, computational linguistics; 2002. p. 104–11.
[21] Tetrault JR. A corpus-based evaluation of centering and pronoun resolution. Comput Linguist 2001;27:507–20.
[22] Pustejovsky J. The generative lexicon. Comput Linguist 1991;17(4): 409–41.
[23] Pustejovsky J. The generative lexicon. Cambridge: MIT Press; 1995.
[24] Tetreault J. Empirical evaluations of pronoun resolution. Computer Science; 2005.
[25] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36(6):414–32.
[26] Girshman A, Sundheim B. Design of the MUC-6 evaluation. In: 6th Conference on message understanding; 1995. p. 1–11.
[27] Poesio M. The MATE/GNOME proposals for anaphoric annotations, revisited. In: Sidner MS, editor. 5th SIGdial workshop on discourse and dialogue, Cambridge, MA; 2004. p. 154–62.
[28] ACE. Cited 2011 July 23. <http://projects.ldc.upenn.edu/ace/annotation/2005TTasks.html>.
[29] Nicolae C, Nicolae G, Roberts K. C-3: Coherence and coreference corpus. In: Seventh conference on international language resources and evaluation (LREC'10), Valetta, Malta; 2010.
[30] OntoNotes. <http://www.bbn.com/ontonotes> [accessed 23.07.11].
[31] GENIA CoreferenceAnnnotation. <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Coreference+Annotation> [accessed 23.07.11].
[32] MEDCo annotation project. <http://nlp.i2r.a-star.edu.sg/medco.html> [accessed 23.07.11].
[33] BioNLP shared task 2011. Protein/gene coreference task. <https://sites.google.com/site/bionlpst/home/protein-gene-coreference-task> [accessed 23.07.11].
[34] Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009 Oct;42(5):839–51.
[35] Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. LREC; 2008. p. 3143–50.
[36] Bagga A. Evaluation of coreferences and coreference resolution systems. In: Second colloquium on discourse anaphora and anaphor resolution (DAARC2); 1998. p. 28–33.
[37] Ogren P. Knowtator : A Protégé plug-in for annotated corpus construction. Proceedings of the 2006 Conference of the, North. 2006;(June): 273–5.
[38] Savova GK, Miller T, Chen L, Chapman WW, Crowley RS. A system for coreference resolution for the clinical narrative. J Am Med Inform Assoc. 2012 [Epub ahead of print].
[39] i2b2 Challenge. <https://www.i2b2.org/NLP/Coreference/> [accessed 24.07.11].
[40] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.
[41] Meyer J, Dale R. Using the WordNet hierarchy for associative anaphora resolution. In: Proc of the Coling 2002 workshop: Semanet'02: building and using semantic, networks; 2002.
[42] Harabagiu S, Bunescu R, Maiorano S. Text and knowledge mining for coreference resolution. In: Proceedings of the second conference of the North American chapter of the ACL; 2001. p. 55–62.
[43] Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M, Jurafsky D. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: 15th conference on computational natural language learning: shared task. Association for, Computational Linguistics, Portland, OR; 2011. p. 28–34.

[44] Bioportal. <http://bioportal.bioontology.org/> [accessed 11.07.11].
[45] Markert K, Nissim M. Comparing knowledge sources for nominal anaphora resolution. Comput Linguist 2005;31(3):367–99.
[46] Liu K, Chapman WW, Savova G, Chute CG, Sioutos N, Crowley RS. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. Methods Inf Med 2010;49(6).
[47] ODIE. <http://www.bioontology.org/ODIE> [accessed 23.07.11].
[48] Poesio M, Viera R, Teufel S. Resolving bridging references in unrestricted text. In: Mitkov R, editor. ACL workshop on operational factors in robust anaphora resolution, Madrid; 1997. p. 1–6.
[49] Nielsen R, Masanz J, Ogren P, Ward W, Martin J, Palmer M, et al. An architecture for complex clinical question answering. In: 1st Annual ACM international conference on health informatics; 2010.
[50] CoNLL-2011 shared task. <http://conll.bbn.com/index.php/introduction.html> [accessed 23.07.11].
[51] Gerber M, Chai JY. Beyond NomBank: a study of implicit arguments for nominal predicates. Association for Computational Linguistics; 2010. p. 1583–92.
[52] Yoshikawa K, Riedel S, Hirao T, Asahara M, Matsumoto Y. Coreference based event-argument relation extraction in biomedical text. In: Fourth international symposium on semantic mining in, biomedicine (SMBM); 2010. p. 93–101.
[53] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550–63.
[54] Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1):14–24.
[55] Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc 2009;16(4):561–70.
[56] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.
[57] Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A shared task involving multi-label classification of clinical free text. In: BioNLP workshop of the association for computational linguistics, Prague, Czech Republic; 2007. p. 97–104.
[58] Chapman WW, Saul M, Irwin JY, Mowery DL, Harkema H, Becich MJ. Creation of a repository of automatically de-identified clinical reports: processes, people, and permission. In: American medical informatics association clinical research informatics summit, San Francisco; 2011.