Coreference Resolution for Structured Drug Product Labels

Halil Kilicoglu and Dina Demner-Fushman
National Library of Medicine
National Institutes of Health
Bethesda, MD, 20894
{kilicogluh,ddemner}@mail.nih.gov

Abstract

FDA drug package inserts provide comprehensive and authoritative information about drugs. DailyMed database is a repository of structured product labels extracted from these package inserts. Most salient information about drugs remains in free text portions of these labels. Extracting information from these portions can improve the safety and quality of drug prescription. In this paper, we present a study that focuses on resolution of coreferential information from drug labels contained in DailyMed. We generalized and expanded an existing rule-based coreference resolution module for this purpose. Enhancements include resolution of set/instance anaphora, recognition of appositive constructions and wider use of UMLS semantic knowledge. We obtained an improvement of 40% over the baseline with unweighted average F_1-measure using B-CUBED, MUC, and CEAF metrics. The results underscore the importance of set/instance anaphora and appositive constructions in this type of text and point out the shortcomings in coreference annotation in the dataset.

1 Introduction

Almost half of the US population uses at least one prescription drug and over 75% of physician office visits involve drug therapy[1]. Knowing how these drugs will affect the patient is very important, particularly, to over 20% of the patients that are on three or more prescription drugs[1]. FDA drug package inserts (drug labels or Structured Product Labels (SPLs)) provide curated information about the prescription drugs and many over-the-counter drugs. The drug labels for most drugs are publicly available in XML format through DailyMed [2]. Some information in these labels, such as the drug identifiers and ingredients, could be easily extracted from the structured fields of the XML documents. However, the salient content about indications, side effects and drug-drug interactions, among others, is buried in the free text of the corresponding sections of the labels. Extracting this information with natural language processing techniques can facilitate automatic timely updates to databases that support Electronic Health Records in alerting physicians to potential drug interactions, recommended doses, and contraindications.

Natural language processing methods are increasingly used to support various clinical and biomedical applications (Demner-Fushman et al., 2009). Extraction of drug information is playing a prominent role in these applications and research. In addition to earlier research in extraction of medications and relations involving medications from clinical text and the biomedical literature (Rindflesch et al., 2000; Cimino et al., 2007), in the third i2b2 shared task (Uzuner et al., 2010), 23 organizations have explored extraction of medications, their dosages, routes of administration, frequencies, durations, and reasons for administration from clinical text. The best performing systems used rule-based and machine learning techniques to achieve over 0.8 F-measure for extraction of medication names; however, the remaining information was harder to extract. Researchers have also tackled extraction of drug-drug interactions (Herrero-Zazo et al., 2013), side effects (Xu and Wang, 2014), and indications (Fung et al., 2013) from various biomedical resources.

As for many other information extraction tasks,

[1]Centers for Disease Control and Prevention: FASTSTATS - Therapeutic Drug Use: http://www.cdc.gov/nchs/fastats/drugs.htm

[2]DailyMed: http://dailymed.nlm.nih.gov/dailymed/about.cfm

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 45–53,
Baltimore, Maryland USA, June 26-27 2014.

extracting drug information is often made more difficult by coreference. Coreference is defined as the relation between linguistic expressions that are referring to the same entity (Zheng et al., 2011). Coreference resolution is a fundamental task in NLP and can benefit many downstream applications, such as relation extraction, summarization, and question answering. Difficulty of the task is due to the fact that various levels of linguistic information (lexical, syntactic, semantic, and discourse contextual features) generally play a role.

Coreference occurs frequently in all types of biomedical text, including the drug package inserts. Consider the example below:

(1) *Since amiodarone is a substrate for CYP3A and CYP2C8, drugs/substances that inhibit* ***these isoenzymes*** *may decrease the metabolism . . . .*

In this example, the expression *these isoenzymes* refer to *CYP3A* and *CYP2C8*. Resolving this coreference instance would allow us to capture the following drug interactions mentioned in the sentence: *inhibitors of CYP3A POTENTIATE amiodarone* and *inhibitors of CYP2C8 POTENTIATE amiodarone*.

In this paper, we present a study that focuses on identification of coreference links in drug labels, with the view that these relations will facilitate the downstream task of drug interaction recognition. The rule-based system presented is an extension of the previous work reported in Kilicoglu et al. (2013). The main focus of the dataset, based on SPLs, is drug interaction information. Coreference is only annotated when it is relevant to extracting such information. In addition to evaluating the system against a baseline, we also manually assessed the system output for precision. Furthermore, we also evaluated the system on a similarly drug-focused corpus annotated for anaphora (DrugNerAR) (Segura-Bedmar et al., 2010). Our results demonstrate that set/instance anaphora resolution and appositive recognition can play a significant role in this type of text and highlight some of the major areas of difficulty and potential enhancements.

2 Related Work

We discuss two areas of research related to this study in this section: processing of drug labels and coreference resolution focusing on biomedical text. Drug labels, despite their availability and the wealth of information contained within them, remain underutilized. One of the reasons might be the complexity of the text in the labels: in a review of publicly available text sources that could be used to augment a repository of drug indications and adverse effects (ADEs), Smith et al. (2011) concluded that many indication and adverse drug event relationships in the drug labels are too complex to be captured in the existing databases of interactions and ADEs. Despite the complexity, the labels were used to extract indications for drugs in several studies. Elkin et al. (2011) automatically extracted indications, mapped them to SNOMED-CT and then automatically derived rules in the form ("Drug" HasIndication "SNOMED CT"). Fung et al. (2013) used MetaMap (Aronson and Lang, 2010) to extract indications and map them to the UMLS (Lindberg et al., 1993), and then manually validated the quality of the mappings. Oprea et al. (2011) used information extracted from the adverse reactions sections of 988 drugs for computer-aided drug repurposing. Duke et al. (2011) have developed a rule-based system that extracted 534,125 ADEs from 5602 SPLs. Zhu et al. (2013) extracted disease terms from five SPL sections (indication, contraindication, ADE, precaution, and warning) and combined the extracted terms with the drug and disease relationships in NDF-RT to disambiguate the PharmGKB drug and disease associations. A hybrid NLP system, AutoMCExtractor, uses conditional random fields and post-processing rules to extract medical conditions from SPLs and build triplets in the form of([drug name]-[medical condition]-[LOINC section header]) (Li et al., 2013).

Coreference resolution in the biomedical domain was addressed in the 2011 i2b2/VA shared task (Uzuner et al., 2012), and the 2011 BioNLP Shared Task (Kim et al., 2012); however these community-wide evaluations did not change much the observation in the 2011 review by Zheng et al. (2011) that only a handful of systems were developed for handling anaphora and coreference in clinical text and biomedical publications. Since this comprehensive article was published, Yoshikawa et al. (2011) proposed two coference resolution models based on support vector machine and joint Markov logic network to aid the task of biological event extraction. Similarly, Miwa et al. (2012) and Kilicoglu and Bergler (2012) extended their biological event

extraction pipelines using rule-based coreference systems that rely on syntactic information and predicate argument structures. Nguyen et al. (2012) evaluated contribution of discourse preference, number agreement, and domain-specific semantic information in capturing pronominal and nominal anaphora referring to proteins. An effort similar to ours is that of Segura-Bedmar et al. (2010), who resolve anaphora to support drug-drug interaction extraction. They created a corpus of 49 interactions sections extracted from the DrugBank database, having on average 40 sentences and 716 tokens. They then manually annotated pronominal and nominal anaphora, and developed a rule-based approach that achieve 0.76 F_1-measure in anaphora resolution.

3 Methods

3.1 The dataset

We used a dataset extracted from FDA drug package labels by our collaborators at FDA interested in extracting interactions between cardiovascular drugs. The dataset consists of 159 drug labels, with an average of 105 sentences and 1787 tokens per label. It is annotated for three entity types (Drug, Drug Class, and Substance) and four drug interaction types (Caution, Decrease, Increase, and Specific). 377 instances of coreference were annotated. Two annotators separately annotated the labels and one of the authors performed the adjudication. The relatively low number of coreference instances is due to the fact that coreference was annotated only when it would be relevant to drug interaction recognition task. This parsimonious approach to annotation presents difficulty in automatically evaluating the system, and to mitigate this, we present an assessment of the precision of our end-to-end coreference system, as well. We split the dataset into training and test sets by random sampling. Training data consists of 79 documents and the test set has 80 documents. We used the training data for analysis and as the basis of our enhancements.

3.2 The system

The work described in this paper extends and refines earlier work, described in Kilicoglu et al. (2013), which focused on disease anaphora and ellipsis in the context of consumer health questions. We briefly recap that system here. The system begins by mapping named entities to UMLS Metathesaurus concepts (CUIs). Next, it identifies anaphoric expressions in text, which include personal (e.g., *it*, *they*) and demonstrative pronouns (e.g., *this*, *those*), as well as sortal anaphora (definite (e.g., with *the*) and demonstrative (e.g., with *that*) noun phrases). The candidate antecedents are then recognized using syntactic (person, gender and number agreement, head word matching) and semantic (hypernym and UMLS semantic type matching) constraints. Finally, the co-referent is then selected as the *focus* of the question, which is taken as the first disease mention in the question.

The coreference resolution pipeline used in the current work, while enhanced significantly, follows the same basic sequence. The relatively simple approach of earlier work is generally sufficient for consumer health questions; however, we found it insufficient when it comes to drug labels. Aside from the obvious point that the approach was limited to diseases, there are other stylistic differences that have an impact on coreference resolution. In contrast to informal and casual style of consumer health questions, drug labels are curated and provide complex indication and ADE information in a formal style, more akin to biomedical literature. Our analysis of the training data highlighted several facts regarding coreference in drug labels: (1) the set/instance anaphora (including those involving distributive anaphora such as *both*, *each*, *either*) instances are prominent, (2) demonstrative pronominal anaphora is non-existent in contrast to consumer health questions, (3) the *focus*-based salience scoring is simplistic for longer texts. We describe the system enhancements below.

3.2.1 Generalizing from diseases to drugs and beyond

We generalized from resolution of disease coreference only to resolution of coreference involving other entity types. For this purpose, we paramaterized semantic groups and hypernym lists associated with each semantic group. We generalized the system in the sense that new semantic types and hypernyms can be easily defined and used by the system. In addition to Disorder semantic group and Disorder hypernym list defined in earlier work, we used Drug, Intervention, Population, Procedure, Anatomy, and Gene/Protein semantic groups and hypernym lists. Semantic group classification largely mimics coarse-grained UMLS semantic groups (McCray et al., 2001). For example, UMLS semantic types Pharmaco-

logic Substance and Clinical Drug are aggregated into both Drug and Intervention semantic groups, while Therapeutic or Preventive Procedure is assigned to Procedure group only. Drug hypernyms, such as *medication*, *drug*, *agent*, were derived from the training data.

3.2.2 Set/instance anaphora

Set/instance anaphora instances are prevalent in drug labels. In our dataset, 19% of all annotated anaphoric expressions indicate set/instance anaphora (co-referring with 29% of antecedent terms). An example was provided earlier (Example 1). While recognizing anaphoric expressions that indicate set/instance anaphora is not necessarily difficult (i.e., recognizing *these isoenzymes* in the example), linking them to their antecedents can be difficult, since it generally involves correctly identifying syntactic coordination, a challenging syntactic parsing task (Ogren, 2010). Our identification of these structures relies on collapsed Stanford dependency output (de Marneffe et al., 2006) and uses syntactic and semantic constraints. We examine all the dependency relations extracted from a sentence and only consider those with the type *conj_** (e.g., *conj_and*, *conj_or*). For increased accuracy, we then check the tokens involved in the dependency (conjuncts) and ensure that there is a coordinating conjunction (e.g., *and*, *or*, , (comma), & (ampersand)) between them. Once such a conjunction is identified, we then examine the semantic compatibility of the conjuncts. In the case of entities, the compatibility involves that at the semantic group level. In the current work, we also began recognizing distributive anaphora, such as *either*, *each* as anaphoric expressions. When the recognized anaphoric expression is plural (as in *they*, *these agents* or *either drug*), we allow the coordinated structures previously identified in this fashion as candidate antecedents. The current work does not address a more complex kind of set/instance anaphora, in which the instances are not syntactically coordinated, such as in Example (2), where *such agents* refer to *thiazide diuretics*, in the preceding sentence, as well as *Potassium-sparing diuretics* and *potassium supplements*.

(2) *. . . can attenuate potassium loss caused by thiazide diuretics. Potassium-sparing diuretics . . . or potassium supplements can increase . . . . if concomitant use of* ***such agents*** *is indicated . . .*

3.2.3 Appositive constructions

Coreference involving appositive constructions[3] are annotated in some corpora, including the BioNLP shared task coreference dataset (Kim et al., 2012) and DrugNerAR corpus (Segura-Bedmar et al., 2010). An example is given below, in which the indefinite noun phrase *a drug* and the drug *lovastatin* are appositives.

(3) *PLETAL does not, however, appear to cause increased blood levels of drugs metabolized by CYP3A4, as it had no effect on lovastatin,* ***a drug*** *with metabolism very sensitive to CYP3A4 inhibition.*

In our dataset, coreference involving appositive constructions were generally left unannotated. However, it was consistently the case that when one of the items in the construction is annotated as the antecedent for an anaphoric expression, the other item in the construction was also annotated as such. Therefore, we identified appositive constructions in text to aid the antecedent selection task. We used dependency relations for this task, as well. Identifying appositives is relatively straightforward using syntactic dependency relations. We adapted the following rule from Kilicoglu and Bergler (2012):

$$APPOS(Antecedent,Anaphor) \vee APPOS(Anaphor,Antecedent) \Rightarrow COREF(Anaphor,Antecedent)$$

where $APPOS \in \{appos, abbrev, prep\_including, prep\_such\_as\}$. In our case, this rule becomes

$$(APPOS(Antecedent1,Antecedent2) \vee APPOS(Antecedent2,Antecedent1)) \wedge COREF(Anaphor,Antecedent1) \Rightarrow COREF(Anaphor,Antecedent2)$$

which essentially states that a candidate is taken as an antecedent, only if its appositive has been recognized as an antecedent. Additionally, semantic compatibility between the items is required.

This allows us to identify *their* and *Class Ia antiarrhythmic drugs* as co-referents in the following example, due to the fact that the exemplification indicated by the appositive construction between *Class Ia antiarrythmic drugs* and *disopyramide* is recognized, the latter previously identified as an antecedent for *their*.

[3]We use the term "appositive" to cover exemplifications, as well.

(4) *Class Ia antiarrhythmic drugs, such as disopyramide, quinidine and procainamide and other Class III drugs (e.g., amiodarone) are not recommended . . . because of **their** potential to prolong refractoriness.*

3.2.4 Relative pronouns

Similar to appositive constructions, relative pronouns are annotated as anaphoric expressions in some corpora (same as those for appositives), but not in our dataset. In the example below, the relative pronoun *which* refers to *potassium-containing salt substitutes*.

(5) *. . . the concomitant use of potassium-sparing diuretics, potassium supplements, and/or potassium-containing salt substitutes, **which** should be used cautiously. . .*

Since we aim for generality and this type of anaphora can be important for downstream applications, we implemented a rule, again taken from Kilicoglu and Bergler (2012), which simply states that the antecedent of a relative pronominal anaphora is the noun phrase head it modifies.

$$rel(X,Anaphor) \wedge rcmod(Antecedent,X) \Rightarrow COREF(Anaphor,Antecedent)$$

where *rel* indicates a *relative* dependency, and *rcmod* a *relative clause modifier* dependency. We extended this in the current work to include the following rules:

(6) (a) $LEFT(Antecedent,Anaphor) \wedge NO\_INT\_WORD(Antecedent,Anaphor) \Rightarrow COREF(Anaphor,Antecedent)$

(b) $LEFT(Antecedent,Anaphor) \wedge rcmod(Antecedent,X) \wedge LEFT(Anaphor,X) \Rightarrow COREF(Anaphor,Antecedent)$

where *LEFT* indicates that the first argument is to the left of the second and *NO_INT_WORD* indicates that the arguments have no intervening words between them.

3.3 Drug ingredient/brand name synonymy

A specific, non-anaphoric type of coreference, between drug ingredient name and drug's brand name, is commonly annotated in our dataset. An example is provided below, where *COREG CR* is the brand name for *carvedilol*.

(7) *The concomitant administration of amiodarone or other CYP2C9 inhibitors such as fluconazole with COREG CR may enhance the -blocking properties of **carvedilol** . . . .*

To identify this type of coreference, we use semantic information from UMLS Metathesaurus. We stipulate that, to qualify as co-referents, both terms under consideration should map to the same UMLS concept (i.e., that they are considered synonyms). If the terms are within the same sentence, we further require that they are appositive.

3.3.1 Demonstrative pronouns

Anaphoric expressions of demonstrative pronoun type generally have discourse-deictic use; in other words, they often refer to events, propositions described in prior discourse or even to the full sentences or paragraphs, rather than concrete objects or entities (Webber, 1988). This fact was implicitly exploited in consumer health questions, since the coreference resolution focused on diseases only, which are essentially processes. However, in drug labels, discourse-deictic use of demonstratives is much more overt. Consider the sentence below, where the demonstrative *This* refers to the event of *increasing the exposure to lovastatin*.

(8) *Co-administration of lovastatin and SAMSCA increases the exposure to lovastatin and . . . . **This** is not a clinically relevant change.*

To handle such cases, we blocked entity antecedents (such as drugs) for demonstrative pronouns and only allowed predicates (verbs, nominalizations) as candidate antecedents.

3.3.2 Pleonastic *it*

We recognized pleonastic instances of the pronoun *it* to disqualify them as anaphoric expressions (for instance, *it* in *It may be necessary to . . .*). Generally, lexical patterns involving sequence of tokens are used to recognize such instances (e.g., (Segura-Bedmar et al., 2010). We used a simple dependency-based rule that mimics these patterns, given below.

$$nsubj*(X,it) \wedge DEP(X,Y) \Rightarrow PLEONASTIC(it)$$

where *nsubj** refers to *nsubj* or *nsubjpass* dependencies and *DEP* is any dependency, where $DEP \notin \{infmod, ccomp, xcomp\}$.

3.3.3 Discourse-based constraints

Previously, we did not impose limits on how far the co-referents could be from each other, since the entire discourse was generally short and the salient antecedent (often the topic of the question) appeared early in discourse. This is often not the

case in drug labels, especially because often intricate interactions between the drug of interest and other medications are discussed. Therefore, we limit the discourse window from which candidate antecedents are identified. Generally, the search space for the antecedents is limited to the current sentence as well as the two preceding sentences (Segura-Bedmar et al., 2010; Nguyen et al., 2012). In our dataset, we found that 98% of antecedents occurred within this discourse window and, thus, use the same search space. We make an exception for the cases in which the anaphoric expression appear in the first sentence of a paragraph and no compatible antecedent is found in the same sentence. In this case, the search space is expanded to the entire preceding paragraph.

We also extended the system to include different types of salience scoring methods. For drug labels, we use linear distance between the co-referents (in terms of surface elements) as the salience score; the lower this score, the better candidate the antecedent is. Additionally, we implemented syntactic tree distance between the co-referents as a potential salience measure, even though this type of salience scoring did not have an effect on our results on drug labels.

Finally, we block candidate antecedents that are in a direct syntactic dependency with the anaphoric expression, except when the anaphor is reflexive (e.g., *itself*).

3.4 Evaluation

To evaluate our approach, we used a baseline similar to that reported in Segura-Bedmar et al. (2010), which consists of selecting the closest preceding nominal phrase for the anaphoric expressions annotated in their corpus. These expressions include pronominal (personal, relative, demonstrative, etc.) and nominal (definite, possessive, etc.) anaphora. We compared our system to this baseline using the unweighted average of F_1-measure over B-CUBED (Bagga and Baldwin, 1998), MUC (Vilain et al., 1995), and CEAF (Luo, 2005) metrics, the standard evaluation metrics for coreference resolution. We used the scripts provided by i2b2 shared task organizers for this purpose. Since coreference annotation was parsimonious in our dataset, we also manually examined a subset of the coreference relations extracted by the system for precision. Additionally, we tested our system on DrugNerAR corpus (Segura-Bedmar et al., 2010), which similarly focuses on drug interactions. We compared our results to theirs, using as evaluation metrics precision, recall, and F_1-measure, the metrics that were used in their evaluation.

4 Results and Discussion

With the drug label dataset, we obtained the best results without relative pronominal anaphora resolution and drug ingredient/brand name synonymy strategies (OPTIMAL) and with linear distance as the salience measure. In this setting, using gold entity annotations, we recognized 318 coreference chains, 54 of which were annotated in the corpus. The baseline identified 1415 coreference chains, only 10 of which were annotated. The improvement provided by the system over the baseline is clear; however, the low precision/recall/F_1-measure, given in Table 1, should be taken with caution due to the sparse coreference annotation in the dataset. To get a better sense of how well our system performs, we also performed end-to-end coreference resolution and manually assessed a subset of the system output (22 randomly selected drug labels with 249 coreference instances). Of these 249, 181 were deemed correct, yielding a precision of 0.73. The baseline method extracted 1439 instances, 56 of which were deemed correct, yielding a precision of 0.04. The precision of our method is more in line with what has been reported in the literature (Segura-Bedmar et al., 2010; Nguyen et al., 2012). For i2b2-style evaluation using the unweighted average F_1 measure over B-CUBED, MUC, and CEAF metrics, we considered both exact and partial mention overlap. These results, provided in Table 1, also indicate that the system provides a clear improvement over the baseline.

Metric	**Baseline**	**OPTIMAL**
With gold entity annotations		
Unweighted F_1 Partial	0.55	0.77
Unweighted F_1 Exact	0.66	0.78
Precision	0.01	0.17
Recall	0.04	0.26
F_1-measure	0.01	0.21
End-to-end coreference resolution		
Precision	0.04	0.73

Table 1: Evaluation results on drug labels

We also assessed the effect of various resolution strategies on results. These results are presented in Table 2.

Strategy	F_1-measure
OPTIMAL	0.21
OPTIMAL - SIA	0.21
OPTIMAL - APPOS	0.15
OPTIMAL + DIBS	0.16 (0.39 recall)

Table 2: Effect of coreference strategies

Disregarding set/instance anaphora resolution (SIA) does not appear to affect the results by much; however, this is mostly due to the fact that the "instance" mentions are generally exemplifications of a particular drug class which also appear in text. In the absence of set/instance anaphora resolution, the system often defaults to these drug class mentions, which were annotated more often than not, unlike the "instance" mentions. Take the following example:

(9) *Use of ZESTRIL with potassium-sparing diuretics (e.g., spironolactone, eplerenone, triamterene or amiloride) . . . may lead to significant increases . . . if concomitant use of* ***these agents*** *. . .*

Without set-instance anaphora resolution, the system links *these agents* to *potassium-sparing diuretics*, an annotated relation. With set-instance anaphora resolution, the same expression is linked to individual drug names (*spironolactone*, etc.) as well as the the drug class, creating a number of false positives, which, in effect, offsets the improvement provided by this strategy.

On the other hand, recognizing appositive constructions (APPOS) appears to have a larger impact; however, it should be noted that this is mostly because it helps us expand the antecedent mention list in the case of set/instance anaphora. For instance, in Example (9), this strategy allows us to establish the link between the anaphora and the drug class (*diuretics*), since the drug class and individual drug name (*spironolactone*) are identified earlier as appositive. We can conclude that, in general, set/instance anaphora benefits from recognition of appositive constructions.

Recognizing drug ingredient/brand name synonymy (DIBS) improved the recall and hurt the precision significantly, the overall effect being negative. Since this non-anaphoric type of coreference is strictly semantic in nature and resources from which this type of semantic information can be derived already exist (UMLS, among others), it is perhaps not of utmost importance that a coreference resolution system recognizes such coreference.

We additionally processed the DrugNerAR corpus with our system. The optimal setting for this corpus was disregarding the drug ingredient/brand name synonymy but using relative pronoun anaphora resolution, based on the discussion in Segura-Bedmar et al. (2010). Somewhat to our surprise, our system did not fare well on this corpus. We extracted 524 chains, 327 of which (out of 669) were annotated in the corpus, yielding a precision of 0.71, recall of 0.56, and F_1-measure of 0.63. This is about 20% lower than their reported results. When we used their baseline method (explained earlier), we obtained similarly lower scores (precision of 0.18, recall of 0.45, F_1-measure of 0.26, about 40% lower than their reported results). In light of this apparent discrepancy, which clearly warrants further investigation, it is perhaps more sensible to focus on "improvement over baseline" (reported as 73% in their paper and is 140% in our case).

We analyzed some of the annotations more closely to get a better sense of the shortcomings of the system. The majority of errors were due to using linear distance as the salience score. For instance, in the following example, *they* is linked to *ACE inhibitors* due to proximity, whereas the true antecedent is *these reactions* (itself an anaphor and is presumably linked to another antecedent). It could be possible to recover this link using principles of Centering Theory (Grosz et al., 1995), which suggests that subjects are more central than objects and adjuncts in an utterance. Following this principle, the subject (*these reactions*) would be preferred to *ACE inhibitors* as the antecedent.

(10) *In the same patients, these reactions were avoided when ACE inhibitors were temporarily withheld, but* ***they*** *reappeared upon inadvertent rechallenge.*

Semantic (but not syntactic) coordination sometimes leads to number disagreement between the anaphora and a true antecedent, as shown in Example (11), leading to false negatives. In this example, *such diuretics* refers to both *ALDACTONE*

and *a second diuretic*; however, we are unable to identify the link between them and the number disagreement between the anaphora and either of the antecedents blocks a potential coreference relation between these items.

(11) *If, after five days, an adequate diuretic response to ALDACTONE has not occurred, a second diuretic that acts more proximally in the renal tubule may be added to the regimen. Because of the additive effect of ALDACTONE when administered concurrently with* ***such diuretics*** *...*

5 Conclusion

We presented a coreference resolution system enhanced based on insights from a dataset of FDA drug package inserts. Sparse coreference annotation in the dataset presented difficulties in evaluating the results; however, based on various evaluation strategies, the performance improvement due to the enhancements seems evident. Our results show that recognizing coordination and appositive constructions are particularly useful and that non-anaphoric cases of coreference can be identified using synonymy in semantic resources, such as UMLS. However, whether this is a task for a coreference resolution system or a concept normalization system is debatable. We experimented with using hierarchical domain knowledge in UMLS (for example, the knowledge that *lisinopril* ISA *angiotensin converting enzyme inhibitor*) to resolve some cases of sortal anaphora. Even though we did not see an improvement due to using this type of information on our dataset, further work is needed to assess its usefulness. While the enhancements were evaluated on drug labels only, they are not specific to this type of text. Their portability to different text types is limited only by the accuracy of underlying tools, such as parsers, for the text type of interest and the availability of domain knowledge in the form of relevant semantic types, groups, hypernyms for the entity types under consideration. The results also indicate that a more rigorous application of syntactic constraints in the spirit of Centering Theory (Grosz et al., 1995) could be beneficial. Event (or clausal) anaphora and anaphora indicating discourse deixis, while rarely annotated in our dataset, appear to occur fairly often in biomedical text. These types of anaphora are known to be particularly challenging, and we plan to investigate them in future research, as well.

Acknowledgments

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

References

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

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

James J. Cimino, Tiffani J. Bright, and Jianhua Li. 2007. Medication reconciliation using natural language processing and controlled terminologies. In Klaus A. Kuhn, James R. Warren, and Tze-Yun Leong, editors, *MedInfo*, volume 129 of *Studies in Health Technology and Informatics*, pages 679–683. IOS Press.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 449–454.

Dina Demner-Fushman, Wendy W. Chapman, and Clem J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 5(42):760–762.

Jon Duke, Jeff Friedlin, and Patrick Ryan. 2011. A quantitative analysis of adverse events and "overwarning" in drug labeling. *Archives of internal medicine*, 10(171):944–946.

Peter L. Elkin, John S. Carter, Manasi Nabar, Mark Tuttle, Michael Lincoln, and Steven H. Brown. 2011. Drug knowledge expressed as computable semantic triples. *Studies in health technology and informatics*, (166):38–47.

Kin Wah Fung, Chiang S. Jao, and Dina Demner-Fushman. 2013. Extracting drug indication information from structured product labels using natural language processing. *JAMIA*, 20(3):482–488.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Halil Kilicoglu and Sabine Bergler. 2012. Biological Event Composition. *BMC Bioinformatics*, 13 (Suppl 11):S7.

Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 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.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, 13(Suppl 11):S1.

Qi Li, Louise Deleger, Todd Lingren, Haijun Zhai, Megan Kaiser, Laura Stoutenborough, Anil G. Jegga, Kevin B. Cohen, and Imre Solti. 2013. Mining FDA drug labels for medical conditions. *BMC medical informatics and decision making*, 13(1):53.

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

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *In Proc. of HLT/EMNLP*, pages 25–32.

Alexa T. McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Proceedings of Medinfo*, 10(pt 1):216–20.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

Ngan L. T. Nguyen, Jin-Dong Kim, Makoto Miwa, Takuya Matsuzaki, and Junichi Tsujii. 2012. Improving protein coreference resolution by simple semantic classification. *BMC Bioinformatics*, 13:304.

Philip V. Ogren. 2010. Improving Syntactic Coordination Resolution using Language Modeling. In *NAACL (Student Research Workshop)*, pages 1–6. The Association for Computational Linguistics.

T.I. Oprea, S.K. Nielsen, O. Ursu, J.J. Yang, O. Taboureau, S.L. Mathias, L. Kouskoumvekaki, L.A. Sklar, and C.G. Bologa. 2011. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Molecular informatics*, 2-3(30):100–111.

Thomas C. Rindflesch, Lorrie Tanabe, John N. Weinstein, and Lawrence Hunter. 2000. EDGAR: Extraction of drugs, genes, and relations from the biomedical literature. In *Proceedings of Pacific Symposium on Biocomputing*, pages 514–525.

Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez, and Paloma Martínez. 2010. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC Bioinformatics*, 11 (Suppl 2):S1.

J.C. Smith, J.C. Denny, Q. Chen, H. Nian, A. 3rd Spickard, S.T. Rosenbloom, and R. A. Miller. 2011. Lessons learned from developing a drug evidence base to support pharmacovigilance. *Applied clinical informatics*, 4(4):596–617.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *JAMIA*, 17(5):514–518.

Özlem Uzuner, Andrea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R. South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *JAMIA*, 19(5):786–791.

Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC*, pages 45–52.

Bonnie L. Webber. 1988. Discourse Deixis: Reference to Discourse Segments. In *ACL*, pages 113–122.

Rong Xu and QuanQiu Wang. 2014. Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection. *BMC Bioinformatics*, 15:17.

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, and Yuji Matsumoto. 2011. Coreference Based Event-Argument Relation Extraction on Biomedical Text. *Journal of Biomedical Semantics*, 2 (Suppl 5):S6.

Jiaping Zheng, Wendy W. Chapman, Rebecca S. Crowley, and Guergana K. Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6):1113–1122.

Qian Zhu, Robert R. Freimuth, Jyotishman Pathak, Matthew J. Durski, and Christopher G. Chute. 2013. Disambiguation of PharmGKB drug-disease relations with NDF-RT and SPL. *Journal of Biomedical Informatics*, 46(4):690–696.