Leveraging coreference to identify arms in medical abstracts: An experimental study

Elisa Ferracane
Department of Linguistics
The University of Texas at Austin
elisa@ferracane.com

Iain Marshall
Dept. of Primary Care and Public Health Sciences
Kings College London
iain.marshall@kcl.ac.uk

Byron C. Wallace
College of Computer and Information Science
Northeastern University
byron@ccs.neu.edu

Katrin Erk
Department of Linguistics
The University of Texas at Austin
katrin.erk@mail.utexas.edu

Abstract

Performing systematic reviews is a critical yet manual, labor-intensive step in evidence-based medicine. Automating systematic reviews is an active area of research, requiring innovations in machine learning and computational linguistics. We examine how coreference resolution can aid in identifying the arms of a study, an often overlooked piece of information needed to synthesize the results in a systematic review. A classification model[1] that performs better with the coreference features supports the intuition that coreference is able to capture the discourse salience of arms. We note that control arms do not benefit as much from these features.

1 Introduction

Evidence-based medicine (EBM) is a paradigm that seeks to inform medical practitioners of the optimal treatment, based on the totality of the available evidence (i.e., the results of all relevant clinical trials). To this end, teams of medical experts often conduct *systematic reviews*, which synthesize all published medical literature pertaining to a specific clinical question. The first step in a systematic review is to formulate the research question to be investigated, and then find all of the relevant citations. Abstracts and then full texts are screened to exclude irrelevant trials. Once a set of trials pertinent to the research question are identified (typically 10-20 trials), key pieces of information are extracted from each trial. This information generally consists of the patient Population under study, the Intervention(s) being tested, the Comparison and the Outcomes (abbreviated as PICO). Results from all identified trials are typically statistically combined via meta-analysis to produce an aggregated result.

Producing systematic reviews is a time-consuming, largely manual process. This is exacerbated by the rapidly growing evidence base: PubMed[2] contains 800,000+ publications on clinical trials in humans (Wallace et al., 2013), and on average reports of 75 new trials are published daily. A single systematic review can take over a year to produce – at which point it risks becoming outdated. Therefore, automating evidence synthesis poses an enormous yet enticing challenge for automation.

A crucial step towards automating synthesis is identifying the *arms*, or groups, in trials. A clinical trial consists of one control arm, and one or more intervention arms. For example, a study comparing the efficacy of aspirin versus a placebo would consist of two arms: those taking *aspirin* (the intervention group), and those taking the *placebo* (the control group). Previous work has mostly focused on identifying the PICO elements. However, the PICO elements alone are insufficient to convey the design of the study, a key piece of evidence necessary in the downstream task of data synthesis and analysis. Thus, the present study focuses on improving the automated identification of arms. We observed that arms are often salient in the discourse of the abstract, in that they corefer more often than other to-

[1] https://github.com/elisaF/extractGroups

[2] publicly available resource for accessing medical references and abstracts https://www.ncbi.nlm.nih.gov/pubmed/

Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 86–95,
Austin, TX, November 5, 2016.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group (arm1) had significantly lower mean BMI SDS at 12 months compared with standard care (arm2). The mean meal size in the Mandometer group (arm1) fell by 45 g. Those in the Mandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Table 1: Excerpt from medical abstract illustrating the discourse salience of the intervention arm, *arm1*, where the control arm is *arm2* (note that not all mentions of the arms are annotated in the gold data, as discussed in section 5.3).

Randomised controlled trial with 12 month intervention (chain1) . Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention (chain1) . Using the last available data on all participants (n=106), those in the Mandometer group (chain2, chain3; arm1) had significantly lower mean BMI SDS at 12 months compared with standard care (arm2). The mean meal size in the Mandometer group (chain3; arm1) fell by 45 g. Those in the Mandometer group (chain2, chain3) also had greater improvement in concentration of high density lipoprotein cholesterol.

Table 2: Medical abstract with annotated arms and coreference chains. The chains were automatically determined as described in section 4.3. All phrases with the same chain label are judged to co-refer.

kens. This study is exploratory work that focuses on investigating the effectiveness of using coreference features for identifying arms.

The remainder of this paper is organized as follows. We motivate the choice of coreference features for arm identification. We then examine prior work in identifying the arms in medical texts, and how coreference resolution has been applied to the medical field. Next, we present an experiment to classify whether tokens in annotated medical abstracts are part of an arm. We propose features that take advantage of the discourse salience of arms, and we discuss the results with and without the coreference features.

2 Motivation

Identifying the arms is not a simple information extraction task. The arms in a study consist of one control group, and one or more intervention groups. Often, the control group is never explicitly mentioned in the abstract. In the following excerpt, only the intervention arm is mentioned:

> *To determine whether modifying eating behaviour with use of a feedback device facilitates weight loss in obese adolescents.*

An arm in a study is typically a noun phrase (NP), where this NP is repeated, either verbatim or anaphorically, throughout the abstract. An example of the discourse salience of arms in a medical abstract is in Table 1. The intervention arm, *Mandometer group*, is repeated several times verbatim throughout the abstract.

Given this recurring linguistic pattern in medical abstracts, we investigated the use of *coreference resolution* to help identify arms. The goal of coreference resolution is to determine which mentions in a text refer to the same entity. A referring expression, or *mention*, is the natural language expression used by discourse participants to refer to entities. Two or more mentions that refer to the same entity are

coreferent, and together form a *coreference chain*. An anaphor and its antecedent (or cataphor and its postcedent) will form a coreference chain. Mentions can be indefinite noun phrases, definite noun phrases, proper names and pronouns, where clinical trial abstracts contain mostly NP's. Using an off-the-shelf coreference tool (to be discussed in more detail in section 4.3) yields the mentions and coreference chains illustrated in Table 2.

Note that the token *intervention*, which is not part of an arm, appears at most 2 times within a single coreference chain, whereas *Mandometer*, part of the experimental arm, appears 3 times. Further, *intervention* is found only in 1 chain, whereas *Mandometer* appears in 2 chains. More generally, we hypothesize a token forming part of an arm is more salient in two ways: (i) an arm token appears more often within a single coreference chain, and (ii) an arm token appears more frequently across different chains (within the same abstract). These observations motivate the coreference features presented in section 4.3. In Table 2, *standard care* is not a member of any chains. More generally, we can expect salience to help more with intervention arms than control.[3]

3 Related work

3.1 Automated Identification of Arms

Previous work has identified PICO elements either at the word or sentence level. Most research has extracted information from medical abstracts, although some studies have used the full text of the articles (De Bruijn et al., 2008; Zhao et al., 2012; Wallace et al., 2016). One of the seminal studies in PICO extraction (Demner-Fushman and Lin, 2007) collapsed intervention and comparator, where interventions were short noun phrases based largely on recognition of semantic types (mapped to UMLS concepts) and a few manually constructed rules. The intervention/comparator extractor returned a list of all the interventions under study, and the extractor was evaluated at the sentence level. However, it is important to distinguish between experimental and control treatments as the bias for the experimental group must be accounted for in the data synthesis step (Lumley, 2002).

Beyond PICO, De Bruijn et al. (2008) extracted data from full-text articles based on the CONSORT Plus Guideline,[4] a list of required, recommended and optional items to include in a systematic review compiled by medical experts. The study found that one of the most difficult items to identify was the experimental treatment, which varied widely beyond just drug names. Elsewhere, Chung (2009) identified interventions as a coordinating structure in a single sentence, and found the major weakness in this approach was parsing errors when identifying the boundaries of the conjuncts. And Summerscales et al. (2011) focused on the downstream task of calculating the absolute risk reduction (ARR), identifying the number of bad outcomes for the control and experimental treatment groups, along with the sizes of both treatment groups. This study found outcomes hardest to detect because of their variability, but also had an overall poor recall partly because coreference was not taken into account.

Most recently, Trenta et al. (2015) proposed a novel approach for identifying the arms and PICO elements that does not rely on a first stage of sentence classification, but instead classifies each token directly, followed by an inference process to constrain the labels to more accurate results. As with previous studies, outcome results were the hardest because they are more variable. A significant limitation of this study is that the abstracts were limited to two-arm trials, and in a specific domain.

3.2 Automated Coreference Resolution

Coreference resolution is a long-studied task that remains a challenging problem. Most recent work on coreference resolution builds mainly on one of four models.

- The first and most widely-used approach is the *mention-pair* model (Soon et al., 2001; Ng and Cardie, 2002b). A classifier first identifies all the pairs of mentions which are coreferent. These pairs are then grouped into coreferent chains by clustering techniques such as closest-first (Soon et al., 2001) or best-first (Ng and Cardie, 2002b; Ng and Cardie, 2002a).

[3] Cases of joint coreference such as *all participants* referring to both arms in the example abstract are not addressed in this paper, but pose an interesting problem for identifying PICO elements such as population and outcome.

[4] `http://rctbank.ucsf.edu/home/cplus`

In closest-first, you link to the closest preceding mention, whereas in best-first, you choose the likeliest one. Common features in these models include distance between the two mentions, syntactic features (e.g., POS tags), semantic features (e.g., named entity type), lexical features (e.g., head word of the mention), and string matching.

- The *mention-ranking* model (Denis and Baldridge, 2008), reframes the task as a ranking function rather than a classification function, ranking all the candidate antecedents of a mention to determine which candidate antecedent is the most probable.

- The *entity-centric* model makes use of entity-level information, focusing on features of mention *clusters*, and not just pairs (Raghunathan et al., 2010). The coreference clusters are built up incrementally, using information from partially-completed coreference chains to guide later decisions. Features include whether a mention head word matches any of the head words in the antecedent cluster.

- The *antecedent tree model* (Yu and Joachims, 2009) builds a graph from a document, where the nodes are the mentions and arcs are the links between mention pairs that are coreferent candidates. The coreference chains are then modeled as latent trees in the graph.

Constraints are imposed on these models for improved results, such as enforcing a transitive closure to guarantee you end up with legal assignments (Finkel and Manning, 2008). For example, if *John Smith* is coreferent with *Smith*, and *Smith* with *Jane Smith*, then it should not follow that *John Smith* and *Jane Smith* are coreferent. Other work has shown that joint models improve performance. Denis et al. (2007) recognized that anaphoricity (whether an entity is the first mention) and coreference should be treated as a joint task since one informs the other. Durrett and Klein (2014) models coreference together with named entity recognition and linking named entities to Wikipedia entities. Combinations of these models have also yielded improved results, such as Clark and Manning (2015) stacking *mention-pair* and *entity-centric* systems (which the current paper uses as its off-the-shelf coreference resolver).

Many coreference resolvers exploit deeper linguistic knowledge, beyond the features mentioned above. Chowdhury and Zweigenbaum (2013) eliminated less-informative training instances prior to model training by creating a list of criteria based on semantic and syntactic intuitions such as a mismatch in semantic types. Peng et al. (2015) created predicate schemas to constrain inference, such as two predicates with a semantically shared argument. Yang et al. (2015) used semantic role labeling to link the time and locations for event mentions, and for verbal mentions they linked their participants. More recently, Kilicoglu et al. (2016) focused on sortal anaphoras which they found to commonly occur in biomedical literature, resolving anaphors that carry a specific semantic type, or sort, such as *these drugs*. Many of these studies take advantage of linguistic resources such as WordNet[5] and FrameNet[6].

In the medical area, coreference resolution has been most closely studied for analyzing clinical narrative text such as that found in Electronic Health Records (EHRs), and biomolecular studies. In fact, there have been corpora (i2b2/VA Corpus(Uzuner et al., 2012), GENIA Event Corpus(Kim et al., 2008)) and shared tasks (SemEval-2015 shared task on Analysis of Clinical Text (Task 14)(Elhadad et al., 2015), BioNLP09 shared task(Kim et al., 2009), ShARe/CLEF eHealth 2013 Evaluation Lab Task 1(Pradhan et al., 2013)) created specifically to advance this area. Given that resources such as FrameNet and WordNet are based mostly on news (e.g. British National Corpus, U.S. newswire), a large number of resources have been created to aid in natural language processing of medical texts. By far the largest and most complex is the Unified Medical Language System (UMLS)[7], consisting of three main components: Metathesaurus with terms and codes from many vocabularies (including CPT, ICD-10-CM, MeSH, RxNorm, and SNOMED CT), Semantic Network with semantic types and semantic relations, and the SPECIALIST Lexicon, which contains syntactic, morpholog-

[5] http://wordnet.princeton.edu
[6] https://framenet.icsi.berkeley.edu
[7] https://www.nlm.nih.gov/research/umls/

ical and orthographic information on terms, along with NLP tools such as POS tagger and word sense disambiguator. Other tools include MetaMap[8], a tool for recognizing UMLS concepts, DrugBank[9], a database of drug names, BANNER[10], a named entity recognizer for biomedical texts, BioText for identifying entities and relations in bioscience texts, and BioFrameNet[11], an extension of FrameNet for molecular biology (and BioWordNet(Poprat et al., 2008) was a failed attempt at extending WordNet also to the biomolecular field). However, when applied to clinical trial texts, these tools prove useful mainly for identifying only medical terms and drug names, and thus more linguistically-motivated resources are still lacking for clinical trial texts.

In the area of clinical narratives, Raghavan et al. (2012) took advantage of the temporal features present in these texts to help determine whether two medical concepts corefer with each other. Their 2014 paper (Raghavan et al., 2014) expanded on this idea to identify medical events spanning across narratives, such as admission notes, medical reports, and discharge notes. Yoshikawa et al. (2011) exploited coreference information for extracting event-argument relations from biomedical texts in the Genia Event Corpus. Jindal and Roth (2013) used very specific domain knowledge to resolve coreference in clinical narratives, such as creating a specific discourse model (i.e. a single patient, several doctors and a few family members) to resolve entities of type "person". Despite the active interest in coreference resolution, there has been much less research investigating its application to clinical trial texts. Most of the literature that does exist is applied to the bio-medical field, focusing more on full-text articles (Gasperin and Briscoe, 2008; Huang et al., 2010; Kilicoglu et al., 2016) than on abstracts (Castano et al., 2002; Yang et al., 2004). To the best of the authors' knowledge, there have been no papers using coreference features to identify arms in clinical trial abstracts.

[8] https://metamap.nlm.nih.gov

[9] http://www.drugbank.ca

[10] http://banner.sourceforge.net

[11] http://biotext.berkeley.edu

4 Experiment

The goal of this experiment is to explore empirically whether incorporating coreference features improves the performance of a classifier for arm identification, as compared to a baseline model without coref features (note that we do not aim to necessarily achieve state-of-the-art results on this task). The task of the classifier is to label a token as either part of an arm or not.

4.1 The corpus

The corpus[12] consists of 263 abstracts from the British Medical Journal (BMJ) annotated with the experimental and control groups (and other PICO elements) by Summerscales (2013). The BMJ requires structured input, and the number of sections varies with some abstracts only containing a few sections such as BACKGROUND, METHODS, FINDINGS and INTERPRETATION. These structured abstracts usually consist of short phrases and incomplete sentences.

Number of documents	263
Number of tokens	63,488
Number of [abstract, token] pairs	35,650
Average no. tokens per document	241
Positive labels	5,757 (9%)

Table 3: Corpus statistics

4.2 Experimental setup

Sentences were tokenized, lower-cased and stop words were removed . Each token was paired with its abstract to form an *[abstract, token]* pair to uniquely correlate the token with the medical abstract where it appeared (e.g. *[abstract_3, "intervention"], [abstract_129, "intervention"]*). A binary classifier was implemented to label each token as belonging to an arm or not (scikit-learn implementation of Support Vector Machine, Pedregosa et al. (2011)). Due to the imbalance of classes (9% positive), the class weights in the model were adjusted to be inversely proportional to the class frequencies in the corpus. We performed five-fold cross validation.

[12] https://github.com/rlsummerscales/bibm2011corpus

Model	Precision (var)	Recall (var)	F1 (var)
baseline	12.9 (2.7e-04)	**88.6** (5.6e-04)	22.5 (6.2e-04)
coref	**19.7** (7.5e-04)	82.7 (8.4e-04)	**31.8** (14.4e-04)

Table 4: Results averaged across 5-folds on the two models with their variances in parentheses.

Feature	Mean	Range	Variance
b-o-w	1.78	1-24	2.71
drugbank	0.09	0-1	0.08
tf-idf	6.06	1-141.1	42.67
coref max_counts	0.14	0-15	0.31
coref num_chains	0.10	0-6	0.11

Table 5: Feature statistics

4.3 Features

The following features, summarized in Table 5, were used in the machine learning algorithm.

bag-of-words The number of times the token occurs within its medical abstract (i.e., the count of *[abstract, token]* pairs for the given token and abstract). As evident in Table 5, abstracts can be quite repetitive in their vocabulary, but on average a token appears only a couple of times within the same abstract.

drugbank Whether the token exists in the DrugBank database version 4.3[13]. The clinical trials often compare the efficacy of different drugs, such that intervention arms would contain drug names. However, note from Table 5 that most words are not drugs, keeping in mind that interventions also consist of therapies, behavior changes and other non-drug-related treatments.

tf-idf: Term frequency-inverse document frequency for term t in document d for corpus D:

$$tf\text{-}idf_{t,d} = tf_{t,d} * (idf_{t,D} + 1), \quad (1)$$

where:

$$tf_{t,d} = f_{t,d}$$

$$idf_{t,D} = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

One is added in the equation (1) so that terms with zero idf (those that occur in all documents of a training set) are not entirely ignored. The goal of this metric is to capture how informative a word is. For example, the token *mandometer* (an arm) from the abstract in Table 2 has a tf-idf measure of 26.29, whereas *intervention* (not an arm) has a value of 3.7. On average, the tokens are slightly more informative than common words such as *intervention*.

coreference:

The Coreference Resolution annotator packaged in Stanford Core NLP 3.0[14] (a model that stacks *mention-pair* and *entity-centric* systems) is used to calculate the maximum number of times the token occurs in a single coreference chain within the same medical abstract (**max_counts**) and the number of chains the token appears in the same medical abstract (**num_chains**). This tool was chosen because it is publicly available and yields state-of-the-art results on the 2012 CoNLL data set. The coreference features aim to capture the discourse salience of arms in medical abstracts. As mentioned before, the (max_counts, num_chains) values for *mandometer* are (3,2), but for *intervention* are (2,1). Note from Table 5 that although a token can occur very frequently in a single chain (*max_counts*) and across many chains (*max_chains*), a token on average is not part of a chain at all. This observed statistic lends weight to the use of coreference features as a measure of salience. Previous work has employed other features such as dependency trees and other predicate argument structures to capture this discourse salience. Summerscales (2013) implemented a form of post-hoc coreference resolution as a way to cluster labeled words into groups, for example into a control group versus an intervention group. However, the present study uses the coreference features at the front end to detect the mentions, and is presently not concerned with differentiating among the different arms.

[13] http://www.drugbank.ca/system/downloads/4.3/drugbank.xml.zip

[14] http://nlp.stanford.edu/software/stanford-corenlp-full-2015-12-09.zip

5 Evaluation

Table 4 summarizes the evaluation scores. The results of the classifier are evaluated against the spans of text that were annotated as arms, following Summerscales (2013). Because an arm consists of several contiguous words (e.g. *mandometer group*), we want to ensure the classifier is able to correctly label the more informative words in that span (*mandometer* vs. *group*). A labeled group of words is considered a match for an annotated group if they consist of the same set of words, ignoring *had, group(s)*, and *arm*. For example, a labeled span of *mandometer* for the annotated span *mandometer group* is a true positive. On the other hand, a labeled span of only *group* is a false positive. Although the scores are relatively low for both models, we emphasize the goal of this experiment is not to achieve state-of-the-art results but to investigate the viability of salience for arm identification. Further, we are being strict in our evaluation, compared to prior work (e.g., Summerscales (2013)).

5.1 Baseline

The **baseline** model includes the features for how many times a token appears in a single abstract (**b-o-w**), whether the token exists in the DrugBank (**drugbank**), and the term-frequency inverse-document-frequency measure for the token (**tf-idf**).

5.2 With Coreference

The **coref** model additionally includes the maximum number of times the token appears in a single coreference chain for a given abstract (**max_counts**), and the number of coreference chains the tokens appears in for a given abstract (**num_chains**).

5.3 Error Analysis

The coref model performed better than the baseline model in almost all the metrics: precision (improved 6.8 points) and F1 (+9.3). Additionally, these improvements are consistent across all the cross-validation runs, as illustrated in Figure 1. Adding the coreference features lowers recall by 5.9 points. To understand the results in more detail, we compare the confusion matrices of the two models. The raw counts in Figure 2 illustrate the class imbalance of the data, giving the impression that a false positive is more likely than a false negative. The normalized confusion matrices in Figure 3 show that false negatives are a higher percentage of the errors than false positives, so that the positive class is the harder one to label.

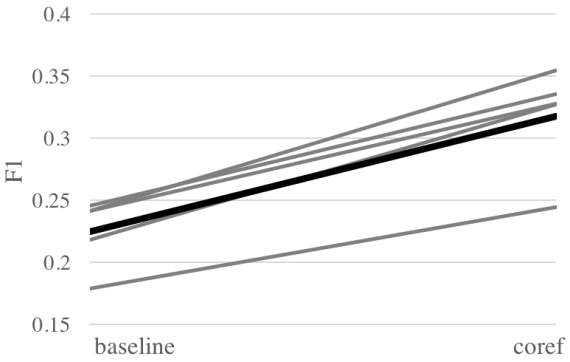

Figure 1: F1 score across the 5 runs in gray, with mean in the thick black line. The lines connect results in the baseline model to results of the the same folds in the coref model.

Given that false negatives are the most common errors across both models, we analyze their occurrences first. The control arm is the most susceptible to this type of error, as it is not as salient in the discourse as the experimental arms. The control words are typically drawn from a finite and small vocabulary (e.g. *control, placebo, sham, standard*), so their tf-idf scores are usually low. The false negative rate worsens in the coref model partly because it places more weight on discourse salience, and control arms are often not part of a coreference chain, compared with experimental arms. We refer back to the abstract presented in Table 1. A small ablation study was conducted to determine that the b-o-w feature is able to correctly label *standard* (count=4) as part of an arm. With the coreference features, the word is no longer labeled as an arm, as it does not appear in any coreference chain.

Next, we analyze the false positives across both models. Given that all the features (except drugbank) in both models are aimed at extracting salient words, they also pick out other relevant PICO information. For example, both models incorrectly label *knee* as part of an arm in the following abstract, where each of these mentions is, in fact, annotated as part of an *outcome*:

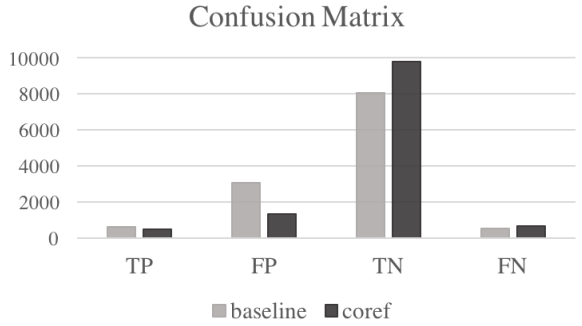

Figure 2: The raw counts of the confusion matrices for the baseline and coref models.

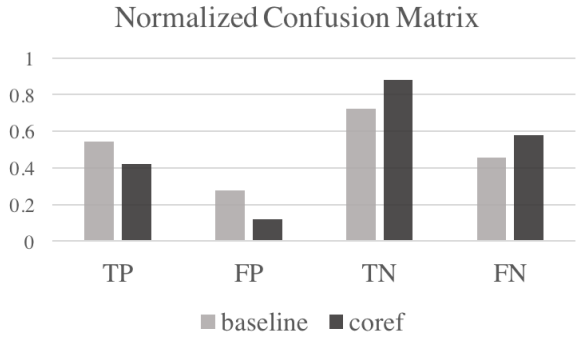

Figure 3: The normalized confusion matrices for the baseline and coref models.

...reduce the incidence of knee and ankle injuries in young people participating in sports. The rate of acute injuries to the knee or ankle. A structured programme of warm-up exercises can prevent knee and ankle injuries...

Another issue with false positives is that the gold data is not comprehensively annotated. Note that in Table 2, the annotator failed to label the third occurrence of *mandometer* as an arm, although both models attempt to classify it as such. However, striving for a thoroughly annotated data set is not realistic, and so the models should be more robust to these gaps and inconsistencies. The false positive rate improves in the coref model partly because the coreference features prove to be a better measure of discourse salience for the intervention arms. As noted earlier, repetition in medical abstracts is not limited to the words describing the arm. For example, in the abstract from Table 1, the baseline model incorrectly labels the high-frequency tokens *eating, months* and *mean* as parts of an arm. The coref model instead correctly labels these as negative, given that they do not occur in a coreference chain.

Finally, we note that the coreference features help in grouping together words with conflicting tf-idf measures. In the abstract from Table 1, the baseline model correctly labels *mandometer* (tf-idf=26.3), but misses *group* (tf-idf=4.2). However, the coref model correctly labels the entire span *mandometer group* as an arm, because both of these tokens appear together in a mention and have the same coreference features.

6 Conclusion

We introduced a new approach to identify the arms in a clinical trial abstract by creating coreference features aimed at capturing the discourse salience of arms. The coreference features were shown to help in classifying a word as part of an arm, confirming the intuition that mentions of arms throughout the abstract often corefer. However, we note this pattern holds more for the experimental than control arms. The error analysis also revealed that arms are not the only concepts that are coreferent: other PICO elements such as the outcome often have the same features. This observation could motivate a model that jointly labels these PICO elements along with the arms, since one would inform the other. There are several other recurring linguistic patterns yet to be explored that could further aid in arm identification, such as apposition:

A computerised device, Mandometer, providing real time feedback...

and paraphrasing:

..half were produced automatically with a larger volume of material...The larger booklets produced automatically were...

Another avenue of research is to investigate how these linguistic features pattern across abstracts in the same review. For example, finding the paraphrases across all abstracts that study the same treatment (as defined in a systematic review) could yield finer-grained information on the language used to describe that intervention. To compensate for the inconsistent and small number of annotations, label propagation might be used to retrieve clusters of relations and find the structure in the data.

As noted earlier, the present study focused on the effect of salience on arm identification. In a future study, we plan to implement Summerscales (2013)

as a strong baseline (which achieved an F-score of 0.69) to understand whether coreference can still yield improved results when compared to a model that nears state-of-the-art performance.

Acknowledgments

We thank Dr. Rodney Summerscales for providing us with the annotated corpus, and the anonymous reviewers for their helpful feedback.

Wallace and Marshall were supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under award number R01LM012086. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

José Castano, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature.

Md Faisal Mahbub Chowdhury and Pierre Zweigenbaum. 2013. A controlled greedy supervised approach for co-reference resolution on clinical text. *Journal of biomedical informatics*, 46(3):506–515.

Grace Yuet-Chee Chung. 2009. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *Journal of biomedical informatics*, 42(5):790–800.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Association for Computational Linguistics (ACL)*.

Berry De Bruijn, Simona Carini, Svetlana Kiritchenko, Joel Martin, and Ida Sim. 2008. Automated information extraction of key trial design elements from clinical trial publications. In *AMIA Annual Symposium Proceedings*, volume 2008, page 141. American Medical Informatics Association.

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

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 660–669, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pascal Denis, Jason Baldridge, et al. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL*, pages 236–243. Citeseer.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Noémie Elhadad, Sameer Pradhan, WW Chapman, Suresh Manandhar, and GK Savova. 2015. Semeval-2015 task 14: Analysis of clinical text. In *Proc of Workshop on Semantic Evaluation. Association for Computational Linguistics*, pages 303–10.

Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 45–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Caroline Gasperin and Ted Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 257–264. Association for Computational Linguistics.

Cuili Huang, Yaqiang Wang, Yongmei Zhang, Yu Jin, and Zhonghua Yu. 2010. Coreference resolution in biomedical full-text articles with domain dependent features. In *Computer Technology and Development (ICCTD), 2010 2nd International Conference on*, pages 616–620. IEEE.

Prateek Jindal and Dan Roth. 2013. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. *Journal of the American Medical Informatics Association*, 20(2):356–362.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch. 2016. Sortal anaphora resolution to enhance relation extraction from biomedical literature. *BMC bioinformatics*, 17(1):1.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, 9(1):1.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

Thomas Lumley. 2002. Network meta-analysis for indirect treatment comparisons. *Statistics in medicine*, 21(16):2313–2324.

Vincent Ng and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve

coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. *Urbana*, 51:61801.

Michael Poprat, Elena Beisswanger, and Udo Hahn. 2008. Building a biowordnet by using wordnet's data formats and wordnet's software infrastructure: a failure story. In *Software engineering, testing, and quality assurance for natural language processing*, pages 31–39. Association for Computational Linguistics.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2013. Task 1: Share/clef ehealth evaluation lab.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M Lai. 2012. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–741. Association for Computational Linguistics.

Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M Lai. 2014. Cross-narrative temporal ordering of medical events. In *ACL (1)*, pages 998–1008. Citeseer.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Huperff, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 372–377. IEEE.

Rodney L Summerscales. 2013. *Automatic summarization of clinical abstracts for evidence-based medicine*. Ph.D. thesis, Illinois Institute of Technology.

Antonio Trenta, Anthony Hunter, and Sebastian Riedel. 2015. Extraction of evidence tables from abstracts of randomized clinical trials using a maximum entropy classifier and global constraints. *arXiv preprint arXiv:1509.05209*.

Ozlem Uzuner, Andreea 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. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Byron C Wallace, Issa J Dahabreh, Christopher H Schmid, Joseph Lau, and Thomas A Trikalinos. 2013. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. *Journal of comparative effectiveness research*, 2(3):273–282.

Byron C Wallace, Jol Kuiper, Aakash Sharma, Mingxi (Brian) Zhu, and Iain J. Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research (JMLR)*.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. An np-cluster based approach to coreference resolution. In *Proceedings of the 20th international conference on Computational Linguistics*, page 226. Association for Computational Linguistics.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *arXiv preprint arXiv:1504.05929*.

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(5):1.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, New York, NY, USA. ACM.

Jin Zhao, Praveen Bysani, and Min-Yen Kan. 2012. Exploiting classification correlations for the extraction of evidence-based practice information. In *AMIA*.