Identifying treatments, groups and outcomes in medical abstracts

Rodney L. Summerscales
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA
rsummers@iit.edu

Shlomo Argamon
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA
argamon@iit.edu

Jordan Hupert
Department of Pediatrics
University of Illinois at Chicago
Chicago, IL 60612, USA
jhupert@uic.edu

Alan Schwartz
Departments of Medical Education and Pediatrics
University of Illinois at Chicago
Chicago, IL 60612, USA
alansz@uic.edu

Abstract

Detecting and extracting treatments, treatment groups and outcomes is a key step in generating summaries of medical research papers. We describe initial results in applying named-entity recognition methods to the task of extracting such entities from BMJ abstracts. Results are promising, showing that a conditional random field approach using word and semantic features appears to be more useful for recognizing treatments and outcomes than features based on word shape.

1 Introduction

Evidence Based Medicine (EBM) seeks to help physicians use current medical research results in their practice (Schwartz, 2006). To enable them to make use of research results, EBM researchers create (by hand) summaries of research articles in a form that directly supports physician decision-making. Such summaries include interactive decision tools that calculate the expected benefit or harm of new therapies. The rapid growth of the medical literature demands automated methods for creating such summaries by textual analysis.

In order to compute the expected harm or benefit of a new therapy, it is necessary to identify the control and experimental treatment groups and determine the number of patients in each group that experienced good and bad outcomes. Hence, the first step toward automatic summarization is recognizing mentions of *treatments*, *treatment groups*, and *outcomes* in medical research articles. This task is the focus of this paper.

We approach this task for now as a form of named entity recognition (NER). While much research has looked at finding named entities in biomedical research papers, such work has mainly been concerned with finding the names of genes, proteins and drugs. Relatively little work has been done on the problem of identifying treatments, groups or outcomes, which has its own challenges. Treatments, for instance, may be simple drug names or complex phrases describing procedures such as *conventional coronary artery bypass grafting surgery using cardiopulmonary bypass*. Furthermore, some entities are referred to indirectly via ellipsis, e.g., the *no additional advice on anxiety management* treatment in *half had additional advice on anxiety management and half did not*. (We do not yet address cases of this complexity, leaving them for future work.)

Furthermore, treatments and outcomes lack special orthographic features such as numbers, special characters, or uppercase letters that can help in recognizing other entities like dates, email addresses or genes/proteins.

2 Corpus

No corpus of medical articles was available containing annotated treatments, groups and outcomes, so we created one based on 100 abstracts of controlled trial studies. from BMJ downloaded from PubMed[1]. BMJ was specifically targeted because the full text for all the articles are freely available online in html, which is convenient for automated analysis. Future work will examine extraction from full text.

[1] http://www.ncbi.nlm.nih.gov/pubmed/

	Occurrences	Unique
Treatments	1208	230
Groups	363	168
Outcomes	1131	494

Table 1: Numbers of entities of each type in the corpus.

2.1 Corpus Characteristics

The corpus comprises the abstracts of the first 100 randomized controlled trials. published online in 2005 and 2006 (publication dates range from 2005 to 2007). The corpus contains a total 1344 sentences which and consists of 31,324 tokens. All sentences from each abstract are included in the corpus regardless of whether or not they contain entities. All mentions of treatments, groups, and outcomes in the corpus were annotated by the first author and the annotations were then reviewed and verified by the third author.

The following example contains all three types of entities:

> [Mortality]$_{outcome}$ was higher in [the [quinine]$_{treatment}$ group]$_{group}$ than in [the [artemether]$_{treatment}$ group]$_{group}$

The total numbers of entity occurrences and distinct entity phrases (*unique entities*) of each kind are given in Table 1. Note that a given article may discuss (e.g.) two different treatments, but each may be referred to multiple times in the same abstract. Since multiple references to the same unique entity may vary slightly in wording, each occurrence is annotated with an ID that indicates the unique entity to which it belongs, such that all entity occurrences in a given abstract that refer to the same unique entity in that abstract have the same ID.

2.2 Entity Characteristics

One challenge to recognizing treatments, groups, and outcomes is that they can vary greatly in terms of length. Table 2 contains statistics characterizing the average number of tokens in an entity and how much this can vary. In this corpus, treatments may vary in length from consist of just one token (e.g. *paroxetine*), or as many as 15 tokens (e.g. *remain within a department for the care of elderly people in a district general hospital*). Likewise, outcomes can vary from *death* to the more lengthy *the proportion of participants developing a new pressure ulcer of grade 2 or worse*.

	Treatment	Group	Outcome
Avg. length	3.1	3.0	3.6
Std. dev.	2.3	1.5	2.2
Min. length	1	1	1
Max length	15	12	14

Table 2: Entity length statistics (in terms of number of tokens) for each entity type in corpus.

Compared with treatments and outcomes, most group mentions are relatively predictable and usually consist of a treatment mention followed by the words "group" or "arm" (e.g. *the acupuncture group*). However, when results are reported in an abstract, a group may simply be referred to by the treatment assigned to it as in *five had fatty liver alone (1 [tamoxifen], 4 [placebo])*. In this case, *tamoxifen* and *placebo* are group mentions in addition to being treatment mentions. Finally, although less common than with treatments and outcomes, group mentions may occasionally consist of long descriptive phrases (e.g. *patients who underwent conventional coronary artery bypass grafting surgery using cardiopulmonary bypass*).

3 Methods

To find these three types of entities in medical abstracts, we apply a named-entity recognizer that uses a Conditional Random Field (CRF) classifier. Conditional Random Fields (Lafferty et al., 2001) were designed for segmenting and labeling sequential data (e.g. labeling words in a sentence) and have been successfully applied to the task of named-entity recognition (McCallum and Li, 2003).

Entities are recognized by applying a trained CRF classifier to each word in every sentence of an abstract. The classifiers label each word (or token) as an entity word (i.e. part of a treatment, group, or outcome), or not an entity word (i.e. not part of any entity). The classifiers are trained and tested separately on treatments, groups, and outcomes so it is possible for the same word to be part of more than one type of entity. For instance, a word may be labeled as part of a treatment and it may be separately labeled as part of a group. Consecutive words with the same label (e.g. treatment) are grouped together and are considered a *detected* entity. To evaluate the accuracy of the classifiers, the detected entities are matched with the *annotated* entities using the matching criteria described in Section 3.3.

3.1 Features

Features used to label each word include: the word itself, its part of speech (POS), its Medical Subject Heading[2] (MeSH) Id, implemented as described in (Rosario and Hearst, 2004), its semantic tag(s), if any, the title of the section in the abstract where the word occurs, and a set of four context words to the left and right of the word along with their POS and semantic tags.

Semantic tags were defined for *anatomy*, *time*, *disease*, *symptom*, *drug*, *procedure* and *measurement* terms; each word appearing in one of the lists was assigned a corresponding semantic tag. The lists of words for disease, symptom and procedure were obtained from MedicineNet[3]. The list of drug words was obtained from RxList[4]. A list of anatomy words was obtained from Wikipedia[5]. Words were removed from the lists if they also appeared in a list of common words[6]. The lists for time and measurement were created by hand and contain common units of time and measurement (length, volume, weight and mass) and their abbreviations.

3.2 Entity recognition

For a Conditional Random Field (CRF) classifier we used the MALLET SimpleTagger (McCallum, 2002).

For a baseline comparison we also applied the general purpose CRF-based Stanford Named Entity Recognizer (Finkel et al., 2005). The features used by this system: the word itself, character n-grams within the word of lengths 2 to 6, the identities of the words within four words on either side of the word in question, and several binary "word shape" features such as "contains digit" or "all uppercase". Parameter settings were those mentioned in the example in the online documentation[7] except that we also included character n-grams in the middle of a word.

Both approaches used the OpenNLP[8] tokenizer to segment sentences into words. In addition, the Mallet approach used the OpenNLP part of speech tagger to generate POS tags for each word. Since the corpus consists of medical abstracts, we used models trained on the PennBioIE biomedical corpus[9], obtained from the JULIE Lab[10].

3.3 Matching criteria

The difficulty with analyzing the accuracy of an entity recognition system is that entity boundaries are often ambiguous. Consider the following phrase.

Trimethoprim 300 mg daily for three days

Is the treatment *Trimethoprim*, or *Trimethoprim 300 mg daily*, or the entire phrase? In many cases, a system that is able to recognize at least one of the three options is acceptable. One method for handling this situation is to annotate all acceptable versions of an entity (e.g. annotate all three possibilities in this case). However, annotating every possibility can greatly add to the complexity of the annotation process and the annotator may miss some acceptable versions of the entity. An alternative approach is to relax the criteria for determining when an entity recognized by the system matches an annotated entity in the corpus. It is this approach that is the one used in this paper.

When determining if a detected entity occurrence matches an annotated entity occurrence, we use various matching criteria described in (Tsai et al., 2006). The criteria are: Exact (both entities are identical), Left (first token is the same for both entities), Right (last token is the same for both entities), Left/Right (either the first or last tokens match) and Partial (at least one token in the detected entity matches a token in the annotated entity). These criteria range from very strict (Exact match) to very loose (Partial match). These two extremes provide an upper and lower bound on the performance of the recognition system, while the more moderate criteria (Left,

[2] http://www.nlm.nih.gov/mesh/

[3] http://www.medicinenet.com/

[4] http://www.rxlist.com

[5] http://en.wikipedia.org/wiki/List_of_anatomical_topics

[6] http://www.cse.unsw.edu.au/~dons/code/fps/tests/Usr.Dict.Words

[7] http://nlp.stanford.edu/software/crf-faq.shtml

[8] http://opennlp.sourceforge.net/

[9] http://bioie.ldc.upenn.edu/

[10] http://www.julielab.de/

Right, Left/Right) provide an intermediate assessment of the system's performance.

When determining if the unique entities in an abstract have been correctly recognized, two different matching criteria (Exact and Partial) are used. With Exact match, all detected entities in an abstract comprising the same set of tokens (ignoring their order) are considered to refer to the same unique entity. If the set of tokens for a detected unique entity is identical to the set of tokens in at least one version of an annotated unique entity, the detected unique entity is considered to match the annotated unique entity. The Partial match criteria is similar, except that matching criteria for building detected unique entities is less strict. A detected entity occurrence is assigned to the detected unique entity with which it shares the most words, provided at least two words match. Function words are ignored except for negation words (i.e., *no, not, neither, never, none, without*).

4 Results and discussion

Table 3 shows results for the entity recognition system using the features described in Section 3.1. It also contains the results achieved when each feature was individually removed from the feature set. Finally, it also contains results obtained with the Stanford Named Entity Recognizer. 10-fold cross-validation was performed over abstracts. Recall, precision, and F-score (the harmonic mean of precision and recall) were computed for token classification as well as for entity recognition using the matching criteria described above. Table 4 shows results for recognizing unique entities in each abstract. In each table, the highest score is in boldface.

An analysis of the features used reveals that the usefulness of each feature depends on the type of entity that the system is trying to recognize and the type of matching criteria used for analysis. In general, for treatments and outcomes, a token's context features (features from neighboring tokens), its POS, and its section label appear to be the most helpful. With groups, the word itself and its context features appear to be most useful. In some cases, certain features appear to be ineffective and do not improve recognition. POS and the section label features do not appear to aid in the recognition of groups. Semantic tag and MeSH Id features do not appear to improve the recognition of treatments or outcomes. This result is particularly surprising given that these features would seem to provide useful medical domain knowledge to the classifier. One explanation for this result is that many treatments and outcomes do not consist of medical terms (e.g. *preventive training* and *numbers of falls*). Also the presence of certain medical terms do not guarantee that the detected entity is of the same type that the system is looking for.

An examination of recognition errors made by the system reveals potential areas for improvement. Groups of the form *<TREATMENT> group* are rarely missed. However, the system often has trouble when the words "group" or "arm" are not present (e.g. *participants referred to the clinic*). When looking for treatments or outcomes the system will sometimes identify an entity of a completely different type. For instance, the system identified *an abdominal aortic aneurysm* as a treatment when it is really a disease and it recognized *good quality motivational interviewing* as an outcome when it is actually a treatment. Sometimes the entity that is detected does not exactly match what is annotated in the corpus. However, what is detected usually is enough to capture the "essential" part of the entity. For instance, the system identified *anaemia determined through passive case detection* and *haemoglobin concentration* as outcomes. However, the annotated entities were *anaemia* and *the decrease in haemoglobin concentration*. Discrepancies like these are not much of a concern and justify the use of looser matching criteria than Exact match. In general, the most common error made by the system is that of missing entity occurrences all together. This error is much more common than the recognition of invalid entities. This can be seen from the fact that the system usually has much higher precision than recall.

Overall, recognizing treatments and outcomes appears to be more difficult than recognizing groups. This is expected since groups are often short noun phrases that end with the word "group", whereas treatments and outcomes are more varied and may be long descriptions of therapies or conditions. Given this variability, our corpus of 1344 sentences is relatively small, and so we expect our approach to benefit from more training data.

	Token			**Exact**			**Left**			**Right**			**Left/Right**			**Partial**		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
Treatments:																		
All features	**.51**	**.75**	**.60**	**.40**	.60	**.48**	**.47**	.70	**.56**	**.49**	.74	**.59**	**.55**	**.82**	**.66**	.56	.84	**.67**
Without word	.49	.73	.59	.39	.57	.46	**.47**	.67	.55	**.49**	.71	.58	**.55**	.79	.65	**.57**	.82	**.67**
No context	.45	.74	.56	.33	.58	.42	.39	.68	.50	.43	**.75**	.54	.47	**.82**	.60	.49	.85	.62
No POS	.45	.74	.56	.34	.59	.43	.40	.68	.50	.43	**.75**	.55	.48	**.82**	.61	.50	**.86**	.63
No section	.45	.73	.56	.39	.60	.47	.44	.69	.54	.47	.73	.57	.51	.80	.62	.53	.83	.65
No semantic tags	.50	**.75**	**.60**	.39	.61	**.48**	.45	.71	.55	.47	.74	.58	.52	**.82**	.64	.54	.85	.66
No MeSH Id	50	.74	**.60**	**.40**	.61	**.48**	.46	.69	.55	**.49**	.74	**.59**	.54	.81	.64	.55	.83	.66
Stanford NER	.43	.74	.54	.39	**.63**	**.48**	.45	**.72**	.55	.44	.71	.54	.49	.79	.61	.50	.81	.62
Groups:																		
All features	.70	.93	.80	.67	.91	.77	**.69**	.94	.79	.67	.92	.78	.69	.94	.80	.69	.94	.80
Without word	.50	.74	.60	.40	.61	.48	.46	.69	.55	.49	.74	.59	.54	.81	.64	.55	.83	.66
No context	.70	.91	.79	.62	.88	.73	.66	.93	.77	.64	.91	.75	.67	.94	.78	.67	.95	.79
No POS	**.72**	**.95**	**.82**	.67	.92	**.78**	**.69**	**.95**	**.80**	**.68**	.93	.78	**.70**	**.96**	**.81**	**.70**	**.97**	**.81**
No section	.71	.94	.81	**.68**	**.93**	**.78**	**.69**	.94	**.80**	**.68**	**.94**	**.79**	.69	.95	.80	.69	.95	.80
No semantic tags	.70	.92	.79	.66	.91	.76	.68	.94	.79	.66	.92	.77	.68	.94	.79	.68	.94	.79
No MeSH Id	.70	.93	.80	.66	.91	.77	.67	.93	.78	.67	.92	.77	.68	.94	.79	.68	.94	.79
Stanford NER	.71	**.95**	.81	.67	.89	.76	**.69**	.92	.79	**.68**	.90	.78	**.70**	.93	.80	**.70**	.93	.80
Outcomes:																		
All features	**.62**	.75	**.68**	.45	.58	**.51**	**.54**	**.69**	**.60**	**.57**	.73	**.64**	**.63**	.80	**.71**	**.64**	.82	**.72**
Without word	.59	**.76**	.66	.42	.58	.49	.50	.68	.57	.54	.73	.62	.59	**.81**	.68	.61	.83	.70
No context	.50	.73	.59	.32	.56	.41	.37	.65	.47	.43	**.74**	.55	.47	**.81**	.59	.49	.84	.62
No POS	.58	.75	.66	.39	.57	.46	.46	.68	.55	.51	**.74**	.60	.56	**.81**	.66	.58	**.85**	.69
No section	.55	.71	.62	.41	.57	.48	.47	.66	.55	.51	.71	.59	.55	.77	.64	.57	.79	.66
No semantic tags	**.62**	.75	**.68**	**.46**	**.59**	**.51**	.53	**.69**	**.60**	**.57**	.73	**.64**	**.63**	**.81**	**.71**	**.64**	.82	**.72**
No MeSH Id	.61	.74	.67	.45	.57	**.51**	.53	.68	**.60**	.56	.71	.63	.62	.79	.69	.63	.81	.71
Stanford NER	.52	.71	.60	.42	**.59**	.49	.48	.66	.55	.50	.70	.58	.54	.75	.63	.56	.78	.65

Table 3: Annotated entity recognition results: The first column contains results for recognizing tokens of annotated entities, and the others contain results recognizing annotated entities using different matching criteria.

	Exact			**Partial**		
	R	P	F	R	P	F
Treatments:						
All features	.78	.34	.47	**.73**	.54	**.62**
Without word	.77	.31	.45	.70	.48	.57
No context	.67	**.36**	.47	.62	**.58**	.60
No POS	.73	**.36**	.48	.68	.55	.61
No section	.75	.35	.48	.70	.53	.61
No semantic tags	**.79**	.35	**.49**	.72	.54	**.62**
No MeSH Id	.78	.33	.47	**.73**	.53	.61
Stanford NER	.67	.34	.46	.62	.51	.56
Groups:						
All features	**.80**	.83	.81	**.77**	.87	.82
Without word	.78	.82	.80	.76	.86	.80
No context	.75	.77	.76	.73	.86	.79
No POS	.77	**.84**	.81	.76	**.89**	.82
No section	.79	.83	.81	.76	.88	.82
No semantic tags	**.80**	**.84**	**.82**	**.77**	.88	**.83**
No MeSH Id	.79	.83	.80	.76	.88	.82
Stanford NER	.77	.82	.80	.74	.88	.81
Outcomes:						
All features	.69	.43	.53	.61	.61	.61
Without word	.65	.43	.52	.57	.61	.59
No context	.53	**.45**	.49	.48	**.67**	.56
No POS	.63	**.45**	.52	.56	.64	.60
No section	.64	.43	.51	.55	.60	.57
No semantic tags	**.70**	.44	**.54**	**.62**	.62	**.62**
No MeSH Id	.68	.42	.52	.60	.59	.60
Stanford NER	.61	.42	.50	.55	.60	.57

Table 4: Unique entity recognition results using both exact and partial match criteria.

Overall, it appears that our CRF classifier performs slightly better than the existing CRF-based NER system. The primary difference between the two approaches is the feature sets. Both share some features (word, context window of 4 words on each side). However, our approach uses additional features based on outside information about a word (its POS tag, semantic class, MeSH Id, and its section location in the abstract), whereas the NER system has additional features based on word shape (character n-grams and various binary word shape features). While word shape features can be useful for recognizing some types of entities, it appears that features based on information about a word are more useful for recognizing the types of entities we are concerned with in this paper. This difference is most apparent when recognizing outcomes, where our approach consistently outperformed the NER system. Furthermore, we found that adding the character n-grams and binary word shape features did not improve performance.

For comparison purposes, we also evaluated the classifiers on the slightly different task of finding treatments in the corpus used by Rosario and Hearst (2004). Results were similar to those on our corpus, though they are not comparable to Rosario and Hearst's, as they only evaluated over both treatments and diseases together using a 75/25 train/test split instead of 10-fold cross-validation.

5 Related Work

While there has been much research on entity recognition of various sorts, little has been done on finding the sorts of entities we seek here. We describe here the most relevant related work on extracting such entities from medical abstracts.

Rosario and Hearst (2004) describe use of a hidden Markov model for identifying treatments and diseases in sentences from medical texts and classifying their relationships. The features used by their system, for a given word, were: the word itself, its part of speech, the phrase constituent it belongs to, its Medical Subject Headings (MeSH) id, various orthographic features and whether the MeSH subheirarchy of the word is usually corresponds to treatments, diseases or neither. They found that the most important features for deciding if a word was part of a treatment or disease were: the word itself, its MeSH id and part of speech.

Paek et al. (2006) used shallow semantic parsing to identify agent, patient and effect (i.e. treatment, group, and outcome) entities in sentences containing one of five predicates ("reduce", "improve", "suggest", "increase", and "use"). Sentences were parsed into their constituents and a classifier was used to identify the constituents that were arguments for the predicate in the sentence.

Dawes et al. (2007) investigated the feasibility of identifying patient/population/problem, exposure/intervention, comparison, outcome, duration and results entities in a set of 20 abstracts from clinical studies. They compiled a list of terms that often indicate their key entities.

Leaman and Gonzalez (2008) described a CRF-based biomedical named entity recognition (NER) system. They applied their system to various publicly available biomedical data sets including (Rosario and Hearst, 2004) and achieved good results compared with other extant NER systems.

6 Conclusion

We have presented here an initial approach to the novel task of recognizing treatments, groups, and outcomes in medical abstracts. Our results suggest that features that include information about a word are more useful for recognizing these entities than features based on word shape. This result is especially true for recognizing outcome entities. Future work will include adding features based on syntactic relations, as well as using syntactic analysis of entity occurrences to account for ellipsis and other variability in entity mentions.

References

Martin Dawes, Pierre Pluye, Laura Shea, Roland Grad, Arlene Greenberg, and Jian-Yun Nie. 2007. The identification of clinically important elements within medical journal abstracts: Patient-population-problem, exposure-intervention, comparison, outcome, duration and results (pecodr). *Information in Primary Care*, 15:9–16.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

Rober Leaman and Graciela Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, 13:652–663.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Hyung Paek, Yacov Kogan, Prem Thomas, Seymour Codish, and Michael Krauthammer. 2006. Shallow semantic parsing of randomized controlled trial reports. In *AMIA Annual Symp Proc. 2006*, pages 604–608.

Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, page 430.

Alan Schwartz. 2006. Evidence based medicine (ebm) and decision tools. *MedEdPORTAL*. Available from: `http://www.aamc.org/mededportal`, ID = 209.

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Jieh Hsiang Yu-Chun Lin, Ding He, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7:92.