Improving Endpoint Detection to Support Automated Systematic Reviews

Ana Lucic, MS[1], Catherine L. Blake, PhD[1]
[1]School of Information Sciences, University of Illinois, Champaign, IL

Abstract

Authors of biomedical articles use comparison sentences to communicate the findings of a study, and to compare the results of the current study with earlier studies. The Claim Framework defines a comparison claim as a sentence that includes at least two entities that are being compared, and an endpoint that captures the way in which the entities are compared. Although automated methods have been developed to identify comparison sentences from the text, identifying the role that a specific noun plays (i.e. entity or endpoint) is much more difficult. Automated methods have been successful at identifying the second entity, but classification models were unable to clearly differentiate between the first entity and the endpoint. We show empirically that establishing if head noun is an amount or measure provides a statistically significant improvement that increases the endpoint precision from 0.42 to 0.56 on longer and from 0.51 to 0.58 on shorter sentences and recall from 0.64 to 0.71 on longer and from 0.69 to 0.74 on shorter sentences. The differences were not statistically significant for the second compared entity.

Introduction

Scientific literature in the field of biomedicine continues to grow at a staggering rate. The number of abstracts in PubMed already exceeds 24 million and every week, the National Library of Medicine adds thousands of new abstracts. Although systematic reviews can help both frontline health care professionals and researchers by accurately synthesizing high quality evidence, the manual processes used to conduct a systematic review are time consuming. A systematic review, which is the cornerstone of Evidence Based Medicine (EBM), can take 5-6 people more than 1000 hours to complete[1], so help is urgently needed to reduce the time between the publication of new results and their integration into practice. Figure 1 shows clearly that the number of meta-analyses (a systematic review that integrates results using quantitative methods) on diabetes have increased from 40 to 974 since 2000.

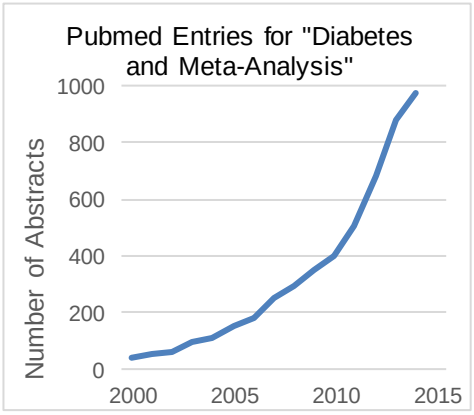

Figure 1. Number of Meta-analysis on Diabetes in PubMed since 2000.

Automating the systematic review process was first introduced almost a decade ago[2] and since then there have been several efforts to that focus on the information retrieval[3-8], and information extraction stages of the process[9] have been developed. In addition to automated strategies, manual efforts are underway to capture data required in a systematic review and tools are available to help with writing the manuscript. Our goal is to support the systematic review process by automatically identifying results from full-text articles.

In this paper we focus on comparison sentences. Within the context of biomedical collection of articles, comparison sentences frequently communicate the results of a study and include the information about the entities that were compared and the basis on which they were compared, which we call an endpoint. Although not a frequent structure in scholarly articles, comparison sentences contain a wealth of information that, when viewed in aggregate, can assist policy makers, health care providers, patients and general consumers of health information with insights about the entities of interest and their comparative characteristics. An analysis of clinical questions in the National Library of Health (NLH) Question Answering Service (http://www.clinicalanswers.nhs.uk) revealed that 16% of the 4,580 questions referred to direct comparisons of different drugs, treatment methods and intervention[11]. Although comparisons have been identified as an information need, current systems do not allow the extraction and synthesis of comparative data from scholarly articles. More broadly, the structure of comparison sentences and the methods that allow parsing of the comparison structure in an automated way can be seen as particularly relevant to Comparative Effectiveness Research whose goal is to provide evidence on the effectiveness, benefits, and drawbacks of different treatment options. Identifying comparison facets in an automated way can assist the process of generating a comparative summary and thus highlight the areas where comparative work has or has not been done. More recently, there has been a shift towards the identification of indirect comparisons in scholarly articles[12-

[13]. High quality evidence consisting of systematic review of randomized clinical trials that provide direct (head-to-head) comparison of two interventions are commonly rare, sometimes non-existent or inconclusive; occasionally, indirect comparisons can be more reliable than direct evidence due to methodological inadequacies of trials [14].

In this study we pay particular attention to direct comparisons (direct mention of at least two compared entities and the endpoint in the comparison sentence) and to the expression of endpoints in comparison sentences. In the following example, *fast-track* and *slow-track patients* represent the compared entities whereas *HbA1c*, *blood pressure* and *serum creatinine levels* represent the basis on which slow-track and fast-track patients were compared. This sentence is considered a direct comparison sentence:

(1) HbA1c [Endpoint_1], blood pressure [Endpoint_2], and serum creatinine levels [Endpoint_3] were significantly higher in fast-track [Entity 1] than in slow-track patients [Entity 2]. 12453917

The following is an example of a comparison sentence that features *hypoglycemia* as the endpoint modifier:

(2) Glucagon levels [Endpoint_1] were significantly lower (P < 0.0001) during hypoglycemia [Endpoint_1_modifier] with tolbutamide [Entity 1] than without tolbutamide [Entity 2]. 11916913

In the above example, *hypoglycemia* is used to modify the main endpoint, *glucagon levels*, and to compare it with relation to drug tolbutamide.

Sentence (3) features *body weight*, as an endpoint:

(3) Body weight [Endpoint] of the high-fat-fed C57BL/6J mice [Entity 1] was 28% higher at 3 months (P < 0.001) and 69% higher at 15 months (P < 0.001) compared with the normal diet-fed C57BL/6J mice [Entity 2]. 12941783

Sentence (4) features *fasting plasma glucose* (FPG) as an endpoint that indicates the difference that was observed in ZDF rats at 12 weeks of age versus 6 weeks of age:

(4) The fasting plasma glucose level [Endpoint] of ZDF rats [Entity 1] was significantly elevated at 12 weeks of age [Entity 1 modifier] compared with the level observed at 6 weeks of age [Entity 2_modifier]. 11679425

These examples demonstrate that comparison sentences represent a convenient medium for examining and identifying endpoints reported in biomedical literature. Furthermore, this paper will demonstrate how by focusing on comparison sentences in biomedical articles we can facilitate a better identification, retrieval, aggregation as well as examination of endpoints. Once entities and endpoints are identified information from comparison sentences can be organized into a tabular summary that shows a detailed summary of which comparisons have already been made and which comparisons are currently missing from the current biomedical literature (see ref 18 for details). Such a summary can be used when writing a systematic review to establish areas where there is enough literature and to identify areas where the results between studies differ.

Several automated methods have been developed to identify comparison *sentences* from text [15-17] and we extend that work by identifying the specific noun phrase within a sentence that fulfills each of the entity and endpoint roles. Previous work[18] achieved good results with respect to Entity 2 where 0.74, 0.80, and 0.91 were reported for precision, recall and accuracy; however, differentiating between Entity 1 and the endpoint is challenging because they are used in similar contexts and with similar grammatical structures. This paper introduces a new method to improve the predictive accuracy of Entity 1 and the endpoint. The approach employs a set of heuristics that capture measurements, and leverages a multi-class classifier instead of a binary classifier. The hypothesis of this study is that endpoints frequently, although not exclusively, represent dependent entities that lend themselves to measurement. Also, we hypothesize that this property can be useful for separating endpoints from other entities in the sentence and for improving the precision and recall for this comparison facet.

Materials and Methods

A set of 100 comparison sentences from the journals *Diabetes*, *Carcinogenesis*, and *Endocrinology* (TREC Genomics collection) that comprise 641 noun phrases used in an earlier study[18] were enriched with information about whether the head noun of the candidate noun is likely to be categorized as an Amount or whether it is more likely to be as a (population) group. This set of 100 comparison sentences was used for training the models (training set). A locally created dictionary consisting of 91 terms was used for the purpose of enriching the feature set with the information on whether the head noun of the candidate noun phrase in a comparison sentence was likely to be an

Amount or a population group. 71 unique terms such as *level, concentration, rate, mass, proportion,* and *degree* were categorized as an Amount and 20 terms were used to indicate a population group such as, *control, arm, trial, treatment*. Drugs were also identified as a group because drugs, within the context of comparison sentences extracted from biomedical scholarly articles, are frequently used as a population group that is compared to another group of drugs. Drugs were identified using the UMLS Pharmacologic Substance semantic class. A number of candidate nouns that occur in comparison sentence will not be identified with either Amount or Group semantic class and was assigned a Null value.

Once the models were built (Table 1), they were then applied to the entire collection and evaluated on a test set of 132 sentences with <=40 words that were drawn at random from the collection. The test set comprised 939 noun phrases. Results are also shown for a sub-sample of the test set consisting of 66 short sentences (<=30 words) comprising 385 noun phrases[18] to explore the impact of sentence length on system performance. Our research question is to determine if the additional amount and group information improves classification performance for the three crucial facets of a comparison sentence – the two compared entities and the endpoint on which they are compared (Entity 1, Entity 2, and Endpoint).

Table 1. Description of the six experiments. Support Vector Machine classification algorithm and linear and Gaussian kernel were used for all experiments.

Experiment title	Description
BC (Binary classifier)	Support vector machine binary classification method, one versus all (OVA).
MC_4 (Multi-class classifier, 4 classes)	Support vector machine multi-class classification with no additional features (4 classes, Entity 1, Endpoint, Entity 2 and the nouns that do not belong to either of these classes)
MC_3 (Multi-class classifier, 3 classes)	Support vector machine multi-class classification with no additional features (3 classes, Entity 1, Endpoint, and the rest of the nouns that do not belong to either of these classes)
BC + A & G (Binary classifier with Amount and Group added features)	Support vector machine binary classification method, one versus all + additional features, Amount and Group (OVA + Amount and Group).
MC_4 + A & G (Multi-class classifier with Amount and Group added features)	Support vector machine multi-class classification + additional features, Amount and Group (4 classes, Entity 1, Endpoint, Entity 2and the nouns that do not belong to either of these classes)
MC_3 + A & G (Multi-class classifier with Amount and Group added features)	Support vector machine multi-class classification + additional features, Amount and Group (3 classes, Entity 1, Endpoint, and the noun that do not belong to either of these classes)

Earlier work reported results for Support Vector Machine algorithm and linear kernel[18]. We were interested in contrasting linear to Gaussian kernel as well as binary to multi-class classification method. Each of the experiments in Table 1 was run with linear and then contrasted with Gaussian kernel method. Of particular interest is whether the non-linear separator and transforming data into an n-dimensional space may provide better results than the linear equation on the attributes in the data set. Two types of multi-class classification experiments were conducted: 1) prediction of 2 compared entities, endpoint and all the rest of the nouns that did not belong to either of these categories (4 classes) and 2) prediction of Entity 1, Endpoint and all the rest of the nouns that did not belong to either of these categories (3 classes). Oracle Data Miner, version 3.2 was used as the platform.

The baseline approach used the same set of lexico-syntactic features was used as in earlier study (See ref 18) although a different version of the parser was used: Stanford dependency parser, version 3.5.1. Multi-class classification methods used 26 features from the previous study for the baseline model whereas binary classification method used 26 features for Entity 1 and 21 for Endpoint and Entity 2[18]. The features used in the earlier and this study rely on the syntactic parse of the sentence provided through Stanford dependency parser and measure the syntactic but also raw distance of each of the candidate nouns from each comparison, evidence, and change anchor in the sentence. Comparison anchors represent the phrases such as *compared with* and *similar to, different from.* A set of 65 comparison anchors is used. Evidence anchor represent verbs that indicate a finding such as *demonstrate, explain, transform* etc. The system uses a set of 432 evidence verbs. Change anchors represent verbs that demonstrate change, such as *increase, decrease,* and *accelerate.* The system uses a lexicon of 770 change verbs. Syntactic paths that connect the syntactic root of the sentence as well as comparison anchor to each candidate noun

are also included as the features in the model. Finally, the classifier is provided with the information on whether the candidate noun appears as a terminal leaf in the syntactic parse of the sentence. The only new features that are added in this experiment is the information on whether the head noun of the candidate noun phrase is more likely to be identified as an Amount or as a population group.

Results

Entity 1 Prediction

For prediction of Entity 1, Table 2 indicates that the results do not improve when multi-class classifier was used and no additional features were added to the model. When four classes were predicted and no additional information was added, precision dropped 0.01 point and recall 0.30 points. When we reduced the number of classes to three (focus on Entity 1 and Endpoint only), the precision increased 0.01 point but recall dropped 0.14 points. However, adding information about whether the head noun of the candidate noun phrase is likely to be categorized as Amount or Group improved precision and recall with binary classifier and linear kernel (BC + A & G). Compared to baseline (BC), precision increased 0.05 points and recall 0.02. The combination of a multi-class classifier and Gaussian kernel plus additional features (MC_4 + Amount and Group) also improved the results. Precision increased from 0.39 to 0.53 (0.14 increase) and recall from 0.47 to 0.57 (0.10 increase). However, given that baseline recall was 0.58 (BC, linear) this actually represents a drop in recall of 0.01 point.

Table 2. Entity 1 results for all test set sentences (95% confidence intervals are shown in parenthesis).

Entity 1 (939 noun phrases, 132 sentences)							**Entity 1 (939 noun phrases, 132 sentences)**					
Linear kernel							**Gaussian kernel**					
	BC	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**	**BC**	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**
Precision	0.38 (0.35, 0.41)	0.37 (0.34, 0.40)	0.39 (0.36, 0.42)	0.43 (0.40, 0.46)	0.49 (0.46, 0.52)	0.50 (0.47, 0.53)	0.39 (0.36, 0.42)	0.48 (0.45, 0.51)	0.38 (0.35, 0.41)	0.49 (0.46, 0.52)	0.53 (0.50, 0.56)	0.56 (0.53, 0.59)
Recall	0.58 (0.55, 0.61)	0.28 (0.25, 0.31)	0.44 (0.41, 0.47)	0.60 (0.57, 0.63)	0.47 (0.44, 0.50)	0.52 (0.49, 0.55)	0.47 (0.44, 0.50)	0.33 (0.30, 0.36)	0.41 (0.38, 044)	0.56 (0.53, 0.59)	0.57 (0.54, 0.60)	0.54 (0.51, 0.57)
F_1	0.46 (0.43, 0.49)	0.32 (0.29, 0.35)	0.41 (0.38, 0.44)	0.50 (0.47, 0.53)	0.48 (0.45, 0.51)	0.51 (0.48, 0.54)	0.43 (0.40, 0.46)	0.39 (0.36, 0.42)	0.39 (0.36, 0.42)	0.52 (0.49, 0.55)	0.55 (0.52, 0.58)	0.55 (0.52, 0.58)
Accuracy	0.73 (0.70, 0.76)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.76 (0.73, 0.79)	0.79 (0.76, 0.82)	0.81 (0.78, 0.84)	0.75 (0.72, 0.78)	0.80 (0.77, 0.83)	0.76 (0.73, 0.79)	0.80 (0.77, 0.83)	0.82 (0.80, 0.84)	0.83 (0.81, 0.85)

With shorter sentences that are not longer than 30 words and that typically have fewer candidate nouns, precision improved from 0.46 using binary linear kernel classifier (BC, linear) to 0.68 using multi-class classifier and Gaussian kernel (0.22 increase) (MC_4 + A & G) while the recall dropped 0.01 point from 0.63 to 0.62 (see Table 3).

Given that a series of 12 classification tasks was conducted and given that six of them did not include Amount and Group information and six did include Amount and Group information the question of interest was whether these apparent differences in the results can be attributed to chance. To establish whether adding the Amount and Group information boosts the performance by chance, a series of matched t-test on the two contrasted groups (BC, MC4, MC3—Linear and BC, MC4, MC4—Gaussian) versus (BC + A & G, MC4 + A & G, MC3 + A & G—Linear and BC + A & G, MC4 + A & G, MC3, A & G—Gaussian kernel) for each of the reported metrics was conducted. For Entity 1 prediction, the differences for individual metric—precision, recall, F1, and accuracy—on the entire test set (939 noun phrases, 132 sentences) were statistically significant ($P < .05$).

Table 3. Entity 1 results for short sentences (95% confidence intervals are shown in parenthesis).

Entity 1 (385 noun phrases, 66 sentences)							Entity 1 (385 noun phrases, 66 sentences)					
Linear kernel							Gaussian kernel					
	BC	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**	**BC**	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**
Precision	0.46 (0.35, 041)	0.45 (0.34, 0.40)	0.45 (0.42, 0.48)	0.55 (0.52, 0.58)	0.63 (0.60, 0.66)	0.65 (0.62, 0.68)	0.44 (0.41, 0.47)	0.62 (0.59, 0.65)	0.50 (0.47, 0.53)	0.64 (0.61, 0.67)	0.68 (0.65, 0.71)	0.72 (0.69, 0.75)
Recall	0.63 (0.60, 0.66)	0.32 (0.29, 0.35)	0.46 (0.43, 0.49)	0.60 (0.57, 0.63)	0.50 (0.47, 0.53)	0.51 (0.48, 0.54)	0.53 (0.50, 0.56)	0.35 (0.32, 0.38)	0.45 (0.42, 0.48)	0.57 (0.54, 0.60)	0.62 (0.59, 0.65)	0.57 (0.54, 0.60)
F_1	0.53 (0.50, 0.56)	0.37 (0.34, 0.40)	0.46 (0.43, 0.49)	0.58 (0.55, 0.61)	0.56 (0.53, 0.59)	0.57 (0.54, 0.60)	0.48 (0.45, 0.51)	0.45 (0.42, 0.48)	0.48 (0.45, 0.51)	0.6 (0.57, 0.63)	0.65 (0.62, 0.68)	0.64 (0.61, 0.67)
Accuracy	0.75 (0.72, 0.78)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.80 (0.77, 0.83)	0.83 (0.81, 0.85)	0.83 (0.81, 0.85)	0.74 (0.71, 0.77)	0.81 (0.78, 0.84)	0.78 (0.75, 0.81)	0.83 (0.81, 0.85)	0.85 (0.83, 0.87)	0.86 (0.84, 0.88)

In conclusion, associating the head noun of a candidate compound noun with categories such as Amount and Group improved the precision of identifying Entity 1. Compared to the baseline method, recall did not increase and typically dropped 0.01 or 0.02 points except with binary classifier, linear kernel when additional features were used (BC + A & G). Generally, multi-class classifier with additional features (regardless of the number of classes predicted) raised precision of the classifier while the recall dropped minimally.

Endpoint Prediction

Table 4 indicates the results for endpoint prediction. Similar to Entity 1 classification, setting the problem as a binary or multi-class classifier does not make a difference until information about the type of head noun is added. Such an addition boosts performance with both binary and multi-class classification methods. More particularly, using multi-class classifier, linear kernel, and additional features (MC_3 + A & G) on all 132 sentences improves the precision from 0.42 to 0.56 (0.14 points) and recall from 0.64 to 0.71 (0.07 improvement). Consequently, F_1 measure improves to 0.62 (0.09 improvement) and accuracy to 0.79 (0.06 improvement) (Table 4).

Table 4. Endpoint results for all test sentences (95% confidence interval in parenthesis).

Endpoint (939 noun phrases, 132 sentences)							Endpoint (939 noun phrases, 132 sentences)					
Linear kernel							Gaussian kernel					
	BC	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**	**BC**	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**
Precision	0.42 (0.39, 0.45)	0.42 (0.39, 0.45)	0.37 (0.34, 0.40)	0.45 (0.42, 0.48)	0.47 (0.44, 0.50)	0.56 (0.53, 0.59)	0.39 (0.36, 0.42)	0.42 (0.39, 0.45)	0.40 (0.37, 0.43)	0.50 (0.47, 0.53)	0.47 (0.44, 0.50)	0.45 (0.42, 0.48)
Recall	0.64 (0.61, 0.67)	0.57 (0.54, 0.60)	0.61 (0.58, 0.64)	0.67 (0.64, 0.70)	0.65 (0.62, 0.68)	0.71 (0.68, 0.74)	0.58 (0.55, 0.61)	0.62 (0.59, 0.65)	0.54 (0.51, 0.57)	0.56 (0.53, 0.59)	0.64 (0.61, 0.67)	0.68 (0.65, 0.71)
F_1	0.51 (0.48, 0.54)	0.48 (0.45, 0.51)	0.46 (0.43, 0.49)	0.54 (0.51, 0.57)	0.55 (0.52, 0.58)	0.62 (0.59, 0.65)	0.47 (0.44, 0.50)	0.50 (0.47, 0.53)	0.46 (0.43, 0.49)	0.53 (0.50, 0.56)	0.55 (0.52, 0.58)	0.55 (0.52, 0.58)
Accuracy	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.69 (0.66, 0.72)	0.74 (0.71, 0.77)	0.76 (0.73, 0.79)	0.79 (0.76, 0.82)	0.71 (0.68, 0.74)	0.73 (0.70, 0.76)	0.71 (0.68, 0.74)	0.78 (0.75, 0.81)	0.76 (0.73, 0.79)	0.75 (0.72, 0.78)

With respect to short sentences (<=30 words), the multi-class classifier with the Gaussian kernel and additional features improved the results from 0.51 (BC) to 0.58 (MC_3 + A & G) and 0.59 (MC_4 + A & G). Similarly, recall improved from 0.69 (BC) to 0.74 (MC_3 + A & G). Consequently, the F_1 measure and accuracy increased to 0.63 and 0.78 (MC_3 + A & G) and 0.65 and 0.78 (MC_4 + A & G) (see Table 5).

With endpoint prediction, both precision and recall increase when multi-class classifier and additional information are used and we do not see the precision-recall trade-off as with Entity 1. Both types of multi-class classifiers (MC3 and MC4) and both kernel methods, linear and Gaussian, benefit from the addition of Amount and Group features. To illustrate, compared with baseline (BC) precision of 0.42, multi-class Support Vector Machine, linear kernel classifier (3 classes) achieved precision of 0.56 (0.14 increase) while recall went from 0.64 to 0.71. With shorter sentences, multi-class (3 classes) and Gaussian kernel achieved precision of 0.58 compared to 0.51 earlier best result (0.07 increase) and recall of 0.74 compared to earlier 0.64 (0.10 increase).

Table 5. Endpoint results for short sentences (95% confidence intervals are shown in parenthesis).

Endpoint 385 noun phrases, 66 sentences)							**Endpoint (385 noun phrases, 66 sentences)**					
Linear kernel							**Gaussian kernel**					
	BC	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**	**BC**	**MC_4**	**MC_3**	**BC + A & G**	**MC_4 + A & G**	**MC_3 + A & G**
Precision	0.51 (0.48, 0.54)	0.51 (0.48, 0.54)	0.45 (0.42, 0.48)	0.53 (0.50, 0.56)	0.58 (0.55, 0.61)	0.52 (0.49, 0.55)	0.48 (0.45, 0.51)	0.51 (0.48, 0.54)	0.49 (0.46, 0.52)	0.63 (0.60, 0.66)	0.59 (0.56, 0.62)	0.58 (0.55, 0.61)
Recall	0.69 (0.66, 0.72)	0.59 (0.56, 0.62)	0.66 (0.63, 0.69)	0.67 (0.64, 0.70)	0.67 (0.64, 0.70)	0.75 (0.72, 0.78)	0.60 (0.57, 0.63)	0.66 (0.63, 0.69)	0.59 (0.56, 0.62)	0.58 (0.55, 0.61)	0.67 (0.64, 0.70)	0.74 (0.71, 0.77)
F_1	0.59 (0.56, 0.62)	0.55 (0.52, 0.58)	0.54 (0.51, 0.57)	0.59 (0.56, 0.62)	0.62 (0.59, 0.65)	0.61 (0.58, 0.64)	0.54 (0.51, 0.57)	0.58 (0.55, 0.61)	0.54 (0.51, 0.57)	0.60 (0.57, 0.63)	0.63 (0.60, 0.66)	0.65 (0.62, 0.68)
Accuracy	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.69 (0.66, 0.72)	0.75 (0.72, 0.78)	0.78 (0.75, 0.81)	0.74 (0.71, 0.77)	0.71 (0.68, 0.74)	0.74 (0.71, 0.77)	0.72 (0.69, 0.75)	0.79 (0.76, 0.82)	0.78 (0.75, 0.81)	0.78 (0.75, 0.81)

In conclusion, the endpoint prediction was similar to Entity 1, where the differences between results achieved with or without Amount and Group information could not be attributed to chance. The differences were statistically significant for each individual metric (precision, recall, F_1, accuracy) ($P < .05$).

Entity 2 Prediction

Interestingly, the identification of Entity 2 does not benefit from additional information. As Table 6 indicates, the performance of the linear kernel classifier dropped after the additional information was added with both binary and multi-class classifiers.

Earlier work18 reported the closeness of Entity 2 to comparison anchor terms such as compared with, similar to, and different from comprise some of the best indicators for the location of Entity 2. The only improvement in this instance was with respect to recall using binary classifier, Gaussian kernel on all sentences without additional information. In this instance, recall increased to 0.83 from the earlier value of 0.80. This increase, however, was accompanied with a drop in precision: 0.66 compared to 0.74. In conclusion, the addition of the new features boosted the performance for Entity 1 and Endpoint but not for Entity 2 that already boasts a high level of precision and recall. (0.74 precision and 0.80 recall on longer sentences). The implication is that shattering of the search space (multi-class classifier) for Entity 2 was not helpful and did not result in better prediction results.

With short sentences, recall also increased using Gaussian kernel and binary classifier with no additional features (BC) (0.87 compared to 0.83). However, this was accompanied with a drop in precision from 0.77 to 0.71 (Table 7). In conclusion, Entity 2 prediction does not benefit from additional information and setting up the problem as a multi-class classification did not bring any improvement over binary classifier (BC). A spike in recall was recorded with Gaussian kernel and binary classification method and no additional features.

Table 6. Entity 2 results for all test sentences (95% confidence intervals shown in parenthesis).

Entity 2 (939 noun phrases, 132 sentences)					Entity 2 (939 noun phrases, 132 sentences)			
Linear kernel					Gaussian kernel			
	BC	**MC_4**	**BC + A & G**	**MC_4 + A & G**	**BC**	**MC_4**	**BC + A & G**	**MC_4 + A & G**
Precision	0.74 (0.71, 0.77)	0.69 (0.66, 0.72)	0.72 (0.69, 0.75)	0.71 (0.68, 0.74)	0.66 (0.63, 0.69)	0.67 (0.64, 0.70)	0.67 (0.64, 0.70)	0.69 (0.66, 0.72)
Recall	0.80 (0.77, 0.83)	0.70 (0.67, 0.73)	0.79 (0.76, 0.82)	0.69 (0.66, 0.72)	0.83 (0.81, 0.85)	0.78 (0.75, 0.81)	0.80 (0.77, 0.83)	0.74 (0.71, 0.77)
F_1	0.77 (0.74, 0.80)	0.69 (0.66, 0.72)	0.75 (0.72, 0.78)	0.70 (0.67, 0.73)	0.73 (0.70, 0.76)	0.72 (0.69, 0.75)	0.73 (0.70, 0.76)	0.72 (0.69, 0.75)
Accuracy	0.91 (0.89, 0.93)	0.89 (0.87, 0.91)	0.91 (0.89, 0.93)	0.90 (0.88, 0.92)	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.90 (0.88, 0.920

Table 7. Entity 2 results for short sentences (95% confidence intervals are shown in parenthesis).

Entity 2 (385 noun phrases, 66 sentences)					Entity 2 (385 noun phrases, 66 sentences)			
Linear kernel					Gaussian kernel			
	BC	**MC_4**	**BC + A & G**	**MC_4 + A & G**	**BC**	**MC_4**	**BC + A & G**	**MC_4 + A & G**
Precision	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.72 (0.69, 0.75)	0.75 (0.72, 0.78)	0.71 (0.68, 0.74)	0.69 (0.66, 0.72)	0.69 (0.66, 0.72)	0.75 (0.72, 0.78)
Recall	0.83 (0.81, 0.85)	0.78 (0.75, 0.81)	0.82 (0.80, 0.84)	0.78 (0.75, 0.81)	0.87 (0.85, 0.89)	0.83 (0.81, 0.85)	0.82 (0.80, 0.84)	0.82 (0.80, 0.84)
F_1	0.80 (0.77, 0.83)	0.77 (0.74, 0.80)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.78 (0.75, 0.81)	0.75 (0.72, 0.78)	0.75 (0.72, 0.78)	0.78 (0.75, 0.81)
Accuracy	0.92 (0.90, 0.94)	0.91 (0.89, 0.93)	0.90 (0.88, 0.92)	0.90 (0.88, 0.92)	0.90 (0.88, 0.92)	0.89 (0.87, 0.91)	0.90 (0.88, 0.92)	0.91 (0.89, 0.93)

When the results of the experiments that use the additional information were compared to the results that do not use additional information were contrasted it was not clear that these individual differences cannot be attributed to chance only—the difference between these two groups for each individual metric (precision, recall, F_1, accuracy) was *not* statistically significant ($P > .05$).

Discussion

The concept of mechanism in the sciences[19] is useful to understand the nature of comparison sentences. Mechanisms consist of entities and activities: entities are the things that engage in activities and activities are producers of change[19]. Entities, within the context of a comparison sentence, represent things that are compared whereas activities represent the change that has occurred between the entities. The endpoint can be seen as part of the activity process, an entity, a dependent more likely than a continuant[20], whose role is to communicate the change that has occurred between the entities. Seen from this perspective, compared entities and endpoints would likely belong to different semantic classes. We contend that it is the addition of a semantic class to the feature set that would help separate the first compared entity from endpoint. The question, however, is how to obtain the information about the semantic

class for compared entities and endpoints? From an ontological point of view, entities that occur in comparison sentences can be matched to an ontological class such as species or population group. For example, the Unified Medical Language System Metathesaurus semantic class Population group can be seen as helpful for their identification. Consider the following sentence:

(5) The plasma insulin concentration [Endpoint_1] at 8 weeks of age [Modifier] and the pancreatic insulin content [Endpoint_2] and the beta-cell mass [Endpoint_3] on day 8 and 8 weeks of age [modifier] in STZ-treated rats [Entity1] were severely reduced compared with those of normal rats [Entity 2] ($P < 0.001$). 14988244

In this sentence, *STZ-treated rats* and *normal rats* represent two entities that can be seen as two population groups that were compared. And yet matching these two concepts to their semantic classes using the UMLS Metathesaurus does not bring us to the Population but rather to the Animals class. Neither *STZ-treated rats* nor normal rats has its full match in the UMLS because they represent very specific groups of rats: rats treated with streptozotocin versus rats that were not treated. With both of these examples, however, it is the head noun—rats—that provides sufficient basis for inferring the semantic class for these two entities—Mammals—and then, through its parent relation, to Vertebrae and Animals higher up in the hierarchy. Animals, however, do not link directly to Population group in the UMLS. In the higher levels of semantic network, Animals is an Entity, the broad type used for grouping conceptual and physical entities. Population group in the UMLS Metathesaurus is defined as "an individual or individuals classified according to their sex, racial origin, religion, common place of living, financial or social status, or some other cultural or behavioral attribute" and as such does not extend to Animals. Similarly, matching endpoints to an ontological class is far from being a straightforward process. Endpoints represent the processes, mechanisms, activities that are happening on the molecular, cellular, tissue, organ, or body level and as such can span semantic classes or be comprised of several semantic classes. By their nature, endpoints represent very specific processes and are typically expressed as a compound noun. In the sentence above, *plasma insulin concentration*, *pancreatic insulin content* and *beta-cell mass* were identified as endpoints. Matching *plasma insulin concentration* to its semantic class would fall under the category of a complex match as *plasma insulin* would be matched to one concept and *concentration* to another. What complicates things further is the fact that *concentration* also represents the case of overmatching because it is identified as a Mental Concept but also as a Quantitative Concept. It is the surrounding context that can determine the concept that should be used for *concentration* which in this case is a Quantitative Concept. *Pancreatic insulin* matches to Neoplastic Process which is not an ideal match for *pancreatic insulin content*. Ideally, we would have *pancreatic insulin content* matched to one semantic class that would be identified as the measurement of insulin in the pancreas. It is the head noun *content* in this compound noun that adds this quantitative quality to *pancreatic insulin* and steers the meaning of the noun in the direction of measurement. The situation is somewhat better with *beta-cell mass*, an example of a complex match. *Beta cell* matches to Cell semantic type and *mass* to Quantitative Concept semantic class, both of which identify the parts correctly.

Most often, endpoints are specific phrases and terms that sometimes indicate the outcome measure and sometimes the property of a compared entity that experienced a change. The following sentence is used to demonstrate the level of endpoint specificity:

(6) Nonfasting plasma glucose levels [Endpoint_1] and the overall glycemic excursion (area under the curve) to a glucose load [Endpoint_2] were significantly reduced (1.6-fold; $P < 0.05$) in (Pro_3) GIP-treated mice [Entity 1] compared with controls [Entity 2]. 16046312

The endpoints in this sentence are *nonfasting plasma glucose levels* and *glycemic excursion to a glucose load*. While *plasma glucose* gets matched to the semantic type Laboratory procedure, *nonfasting* is matched to semantic class Finding which in this case is not ideal. Within the context of a comparison sentence and the information it conveys, the modifier, *nonfasting* provides a very important nuance for the meaning of the entire sentence and it should be retrieved as part of the endpoint.

Previous research[18] reported a number of endpoints related to metformin drug comparison to other interventions. That study reported the following endpoints that relate to insulin: *proinsulin concentrations*, *insulin*, *insulin action*, *insulin concentrations*, *serum insulin concentrations*, *insulin sensitivity*, *% suppression by insulin*. The following endpoints relate to glucose: *glucose*, *fasting glucose*, *hepatic glucose production during hyperinsulinemia*, *glucose disposal*, *glucose disposal rate*, *serum fructosamine*, *glycated hemoglobin (HbA1c)*. These endpoints were grouped based on the main substance they were measuring, insulin or glucose, whereas in this study we grouped the endpoints based on the property that they frequently share: measurement characteristic. The question remains what kind of grouping or matching system is better for the particular and specific nature of endpoints and at what modifier

and what level we can start to draw the line. These questions require a medical specialist to intervene and assist with the process of endpoint categorization.

The problems described above fall under the categories of complex, partial matches, gapped partial matches and overmatching. Most typically, endpoints are very specific phrases and it is this level of specificity that prevents them from being matched to an ontology effectively. This study showed that the fact that many endpoints lend themselves to measurement was the feature that improved their overall identification and retrieval. And yet, not all endpoints lend themselves to measurement. Consider the following sentence:

(7) There is evidence to suggest that the somatic mutational pathway [Endpoint_1] may differ between invasive [Entity 1] and LMP ovarian tumours [Entity 2] and invasive tumours [Entity 1] are more likely than LMP [Entity 2] to exhibit p53 overexpression [Endpoint_2]. 11159743

Somatic mutational pathway is identified as one of the endpoints in this sentence. It is not clear that the concept of mechanism can extend to pathways[22] but even if this is the case, this type of mechanism and the change that is indicated in the above sentence does not involve measurement of any kind, only the statement that the pathway was different. Clearly, in this case, the endpoint does not lend itself to measurement in the same way as the endpoints that comprise the head noun, such as *concentration*, *level*, *degree*, or *mass*. The second endpoint, however, *p53 overexpression*, can be measured. This sentence provides an example where one of the endpoints lends itself to measurement and the other does not indicating that the endpoints, even within the context of the same sentence, do not need to share the same characteristics.

Word ambiguity, context of the article, precision of grammatical and semantic parsers are standing in the way of better alignment of the free form of textual information in scholarly articles and with entries in ontologies and their definitions. Commonly, a large number of pre-processing tasks is needed in order to convert the text of scholarly articles to a format in which it can be matched to an ontology to enable semantic processing of the text. This study demonstrated that the identification of crucial facets of comparison sentences has benefitted from additional information about the meaning of the candidate noun. The method outlined in this study requires testing on a larger dataset. Also, this work invites the examination of which semantic classes in the UMLS, or the subsets of them, can be used to indicate a Population Group effectively in the biomedical scholarly articles. Future work will also strive to examine the role of the Quantitative Concept UMLS semantic class in assisting with the process of identification, retrieval and definition of endpoints. Finally, given that not all endpoints lend themselves to measurement (for example, *pathway*) future work will need to establish other possible ways of modeling endpoints and establishing their significant properties that that can enable their more effective identification and retrieval.

Conclusion

The results from this study suggest that establishing if the head noun is an amount or measure enables the Support Vector Machine to differentiate nouns that play an endpoint role from other candidate noun phrases in a comparison sentence more effectively. Thus treating endpoints as an activity and seeing them through the lens of measurement provided a boost in performance with reference to their identification and retrieval. Classification performance on the test set improved for both entity 1 and the endpoint roles when amount and measure were provided as features and the improvement was statistically significant. The results were not statistically significant for Entity 2 prediction. This improved accuracy provides authors of a systematic review with more specific information about how treatments are compared. In addition, improved retrieval of endpoints will allows us to better examine and assess endpoints in aggregate and track changes over time.

References

1. Petrosino A. Lead Authors of Cochrane Reviews: Survey Results. Report to the Campbell Collaboration. Cambridge, MA: University of Pennsylvania, 1999 Dec 11,1999. Report No.
2. Blake C, Pratt W, Tengs T, editors. Automated Information Extraction and Analysis for Information Synthesis. American Medical Informatics Association Fall Symposium (AMIA); 2002; San Antonio, TX.
3. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006 Mar-Apr;13(2):206-19. PubMed PMID: 16357352. Pubmed Central PMCID: 1447545.
4. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. . J Am Med Inform Assoc. 2009;16:690-704.
5. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts

in performing systematic reviews. J Am Med Inform Assoc. 2010 Jul-Aug;17(4):446-53. PubMed PMID: 20595313. Pubmed Central PMCID: 2995653.
6. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. Artif Intell Med. 2012 Jul;55(3):197-207. PubMed PMID: 22677493. Pubmed Central PMCID: 3393813.
7. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):25-31. PubMed PMID: 18952929. Pubmed Central PMCID: 2605595.
8. Cohen AM, Demner-Fushman D, Iorio A, Sim I, Smalheiser NR. Tools for Identifying Reliable Evidence and Implementing it in Everyday Clinical Care. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science. 2013;2013:42-4. PubMed PMID: 24303233. Pubmed Central PMCID: 3845785.
9. Blake C, editor Information Synthesis: A New Approach to Explore Secondary Information in Scientific Literature. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries; 2005; Denver: ACM/IEEE.
10. Mishriky BM, Cummings DM, Tanenberg RJ. The efficacy and safety of DPP4 inhibitors compared to sulfonylureas as add-on therapy to metformin in patients with Type 2 diabetes: A systematic review and meta-analysis. Diabetes Res Clin Pract. 2015 Aug;109(2):378-88. PubMed PMID: 26059071.
11. Leonhard A. Towards retrieving relevant information for answering clinical comparison questions. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing; Boulder, Colorado. 1572386: Association for Computational Linguistics; 2009. p. 153-61.
12. Donegan S, Williamson P, Gamble C, Tudur-Smith C. Indirect Comparisons: A Review of Reporting and Methodological Quality. PLoS One. 2010 11/10 04/07/received 10/15/accepted;5(11):e11054. PubMed PMID: PMC2978085.
13. Guichard M, D'Andon A, Rumeau Pichon C, Borget I. PRM209 - The Use of Indirect Comparisons in Medicines Evaluation for their Access To Reimbursement by the Has. Value Health. 2015 11//;18(7):A719.
14. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol. 2008 May;61(5):455-63. PubMed PMID: 18394538.
15. Jindal N, Liu B. Identifying comparative sentences in text documents. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval; Seattle, Washington, USA. 1148215: ACM; 2006. p. 244-51.
16. Fiszman M, Demner-Fushman D, Lang FM, Goetz P, Rindflesch TC. Interpreting comparative constructions in biomedical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic. 1572417: Association for Computational Linguistics; 2007. p. 137-44.
17. Park DH, Blake C. Identifying comparative claim sentences in full-text scientific articles. Proceedings of the Workshop on Detecting Structure in Scholarly Discourse; Jeju, Republic of Korea. 2391173: Association for Computational Linguistics; 2012. p. 1-9.
18. Blake C, Lucic A. Automatic endpoint detection to support the systematic review process. Journal of biomedical informatics. 2015 Aug;56:42-56. PubMed PMID: 26003938.
19. Darden L, Craver C. Thinking about mechanisms. Philosophy of Science. 2000;67.
20. Smith B, Grenon P. The Cornucopia of Formal-Ontological Relations. Dialectica. 2004;58(3):279-96.
21. Pratt W, Yetisgen-Yildiz M. A Study of Biomedical Concept Identification: MetaMap vs. People. AMIA Annual Symposium Proceedings. 2003;2003:529-33. PubMed PMID: PMC1479976.
22. Röhl J. Mechanisms in biomedical ontology. Journal of Biomedical Semantics. 2012 09/21;3(Suppl 2):S9-S. PubMed PMID: PMC3448527.