Identifying Comparative Claim Sentences in Full-Text Scientific Articles

Dae Hoon Park[a]
[a]Department of Computer Science

University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
dpark34@illinois.edu

Catherine Blake[a,b]
[b]Center for Informatics Research in Science and Scholarship at the Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820-6211, USA
clblake@illinois.edu

Abstract

Comparisons play a critical role in scientific communication by allowing an author to situate their work in the context of earlier research problems, experimental approaches, and results. Our goal is to identify comparison claims automatically from full-text scientific articles. In this paper, we introduce a set of semantic and syntactic features that characterize a sentence and then demonstrate how those features can be used in three different classifiers: Naïve Bayes (NB), a Support Vector Machine (SVM) and a Bayesian network (BN). Experiments were conducted on 122 full-text toxicology articles containing 14,157 sentences, of which 1,735 (12.25%) were comparisons. Experiments show an F1 score of 0.71, 0.69, and 0.74 on the development set and 0.76, 0.65, and 0.74 on a validation set for the NB, SVM and BN, respectively.

1 Introduction

Comparisons provide a fundamental building block in human communication. We continually compare products, strategies, and political candidates in our daily life, but comparisons also play a central role in scientific discourse and it is not a surprise that comparisons appear in several models of scientific rhetoric. The Create a Research Space (CARS) model includes counter-claiming and establishing a gap during the 'establishing a niche' phase (Swales, 1990), and the Rhetorical Structure Theory includes a contrast schema and antithesis relation that is used between different nucleus and satellite clauses (Mann & Thompson, 1988). However, neither of these models identify where scientists make these comparisons. In contrast, Kircz's (1991) study of physics articles only mentions comparisons with respect to the use of data to compare with other experimental results (sections 4.3 and 8.1, respectively) with earlier work. Similarly, Teufel and Moen's contrast category (which includes the action lexicon s better_solution, comparison and contrast) is also restricted to contrasts with other work (Teufel & Moens, 2002). Lastly the Claim Framework (CF) includes a comparison category, but in contrast to the earlier comparisons that reflect how science is situated within earlier work, the CF captures comparisons between entities (Blake, 2010).

Identifying comparisons automatically is difficult from a computational perspective (Friedman, 1989). For example, the following sentence is not a comparison even though it contains two words (more than) which are indicative of comparisons. *More than five methods were used*. Bresnan claimed that 'comparative clause construction in English is almost notorious for its syntactic complexity' (Bresnan, 1973), p275. Perhaps due to this complexity, several instructional books have been written to teach such constructs to non-native speakers.

Our goal in this paper is to automatically identify comparison claims from full-text scientific articles, which were first defined in Blake's Claim Framework (Blake, 2010). Comparisons capture a binary relationship between two concepts within a sentence and the aspect on which the comparison is made. For example, 'patients with AML' (a type of

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 1–9,
Jeju, Republic of Korea, 12 July 2012.

leukemia) and 'normal controls' are being compared in the following sentence, and the aspect on which the comparison is made is 'the plasma concentration of nm23-H1'. *The plasma concentration of nm23-H1 was higher in patients with AML than in normal controls (P = .0001)*. In this paper, we focus on identifying comparison sentences and leave extraction of the two concepts and the aspect on which the comparison is made as future work. Similar to earlier comparison sentences in biomedicine, we consider the sentence as the unit of analysis (Fiszman, et al, 2007).

To achieve this goal, we cast the problem as a classification activity and defined both semantic and syntactic features that are indicative of comparisons based on comparison sentences that were kindly provided by Fiszman (2007) and Blake (2010). With the features in place, we conducted experiments using the Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers, which both work well on text. We then introduce a Bayesian Network (BN) that removes some of the independence assumptions made in NB model. The subsequent evaluation considers more than 1,735 comparison claim sentences that were identified in 122 full text toxicology articles.

Although automatically detecting comparison sentences in full-text articles is challenging, we believe that the information conveyed from such sentences will provide a powerful new way to organize scientific findings. For example, a student or researcher could enter a concept of interest and the system would provide all the comparisons that had been made. Such a system would advance our general knowledge of information organization by revealing what concepts *can* be compared. Such a strategy could also be used for query expansion in information retrieval, and comparisons have already been used for question answering (Ballard, 1989).

2 Related Work

Comparisons play an important role in models of scientific discourse (see Introduction), because authors can compare research hypotheses, data collection methods, subject groups, and findings. Comparisons are similar to the antithesis in the CARS model (Swales, 1990), the contrast schema in RST (Mann & Thompson, 1988) and in (Teufel & Moens, 2002) and the comparisons category of the CF model (Blake, 2010).

From a computational linguistic perspective, Bresnan (1973) described the comparative clause construction in English as 'almost notorious for its syntactic complexity'. Friedman (1989) also pointed out that comparative structure is very difficult to process by computer since comparison can occur in a variety of forms pervasively throughout the grammar and can occur almost anywhere in a sentence. In contrast to the syntax description of comparison sentences, Staab and Hahn (1997) provided a description logic representation of comparative sentences. Each of these linguists studied the construction of comparative sentence, but did not distinguish comparatives from non-comparative sentences.

Beyond the linguistic community, Jindal and Liu (2006) have explored comparisons between products and proposed a comparative sentence mining method based on sequential rule mining with words and the neighboring words' Part-of-Speech tags. The sequential rules are then used as features in machine learning algorithms. They report that their method achieved a precision of 79% and a recall of 81% on their data set. We too frame the problem as a classification activity, but Jindal and Liu use Part-of-Speech tags and indicator words as features while we use a dependency tree representation to capture sentence features. We also constructed a Bayesian Network to remove the independence assumption of Naïve Bayes classifier. The comparison definition used here also reflects the work of Jindal and Liu (2006).

The work on product review comparisons was subsequently extended to identify the preferred product; for example, camera X would be extracted from the sentence "*the picture quality of Camera X is better than that of Camera Y.*" (Ganapathibhotla and Liu, 2008). Features used for this subsequent work included a comparative word, compared features, compared entities, and a comparison type. Most recently, Xu et al. (2011) explored comparative opinion mining using Conditional Random Fields (CRF) to identify different types of comparison relations where two product names must be present in a sentence. They report that their approach achieved a higher F1 score than the Jindal and Liu's method on mobile phone review data.

Yang and Ko (2011) used maximum entropy method and Support Vector Machines (SVM) to identify comparison sentences from the web based on keywords and Part-of-Speech tags of their neighboring words. They achieved an F1-score of 90% on a data set written in Korean.

The experiments reported here consider articles in biomedicine and toxicology which are similar to those used by Fiszman et al. who identified comparisons between drugs reported in published clinical trial abstracts (Fiszman et al., 2007). However, their definition of comparative sentence is narrower than ours in that non-gradable comparative sentences are not considered. Also, the goal is to classify type of comparative sentences which is different from identifying comparative sentences from a full-text article that contains non-comparative sentences as well.

From a methodological standpoint, Naïve Bayes (NB), Support Vector Machines (SVM), and Bayesian Network (BN) have been explored for variety of text classification problems (Sebastiani, 2002). However, we are not aware of any studies that have explored these methods to identify comparison sentences in full-text scientific articles.

3 Method

Our goal is to automatically identify comparison sentences from full text articles, which can be framed as a classification problem. This section provides the definitions used in this paper, a description of the semantic and syntactic features, and the classifiers used to achieve the goal. Stated formally: Let $S = \{S_1, S_2, \ldots, S_N\}$ be a set of sentences in a collection D. The features extracted automatically from those sentences will be $X = \{X_1, X_2, \ldots, X_M\}$. Each feature X_j is a discrete random variable and has a value X_{ij} for each sentence S_i. Let C_i be a class variable that indicates whether a sentence S_i is a comparative. Thus, the classifier will predict C_i based on the feature values $X_{i1}, X_{i2}, \ldots, X_{iM}$ of S_i.

3.1 Definitions

A ***comparative sentence*** describes at least one similarity or difference relation between two entities. The definition is similar to that in (Jindal & Liu, 2006). A sentence may include more than one comparison relation and may also include an aspect on which the comparison is made. We require that the entities participating in the comparison relation should be non-numeric and exist in the same sentence.

A ***comparison word*** expresses comparative relation between entities. Common comparison words include 'similar', 'different', and adjectives with an '-er' suffix. A ***compared entity*** is an object in a sentence that is being compared with another object. Objects are typically noun phrases, such as a chemical name or biological entity. Other than being non-numeric, no other constraints apply to the compared entities. A ***compared aspect*** captures the aspect on which two comparison entities are compared. The definition is similar to a *feature* in (Jindal & Liu, 2006). For example: *the level of significance differed greatly between the first and second studies*. A compared aspect is optional in comparative sentence.

There are two comparative relation types: *gradable* and *non-gradable* (Jindal & Liu, 2006), and we further partition the latter into non-gradable similarity comparison and non-gradable difference comparison. Also, we consider equative comparison (Jindal & Liu, 2006) as non-gradable. ***Gradable comparisons*** express an ordering of entities with regard to a certain aspect. For example, sentences with phrases such as 'greater than', 'decreased compared with', or 'shorter length than' are typically categorized into this type. The sentence "*The number of deaths was higher for rats treated with the Emulphor vehicle than with corn oil and increased with dose for both vehicles"* is a gradable difference comparison where *'higher'* is a comparison word, *'rats treated with the Emulphor vehicle'* and *'rats treated with corn oil'* are compared entities, and *'the number of deaths'* is a compared aspect.

Non-gradable similarity comparisons state the similarity between entities. Due to nature of similarity, it has a non-gradable property. Phrases such as 'similar to', 'the same as', 'as ~ as', and 'similarly' can indicate similarity comparison in the sentence. The sentence "*Mean maternal body weight was similar between controls and treated groups just prior to the beginning of dosing."* is an example of similarity comparison where *'similar'* is a comparison word, *'controls'* and *'treated*

groups' are compared entities, and *'Mean maternal body weight'* is a compared aspect.

Non-gradable difference comparisons express the difference between entities without stating the order of the entities. For example, comparison phrases such as 'different from' and 'difference between' are present in non-gradable difference comparison sentences. In the sentence "*Body weight gain and food consumption were not significantly different between groups*" there is a single term entity 'groups', and a comparison word 'different'. With the entity and comparison word, this sentence has two comparative relations: one with a compared aspect 'body weight gain' and another with 'food consumption'.

3.2 Feature representations

Feature selection can have significant impact on classification performance (Mitchell, 1997). We identified candidate features in a pilot study that considered 274 comparison sentences in abstracts (Fiszman et al., 2007) and 164 comparison claim sentences in full text articles (Blake, 2010). Thirty-five features were developed that reflect both lexical and syntactic characteristics of a sentence. Lexical features explored in these experiments include:

L1: The first lexical feature uses terms from the SPECIALIST lexicon (Browne, McCray, & Srinivasan, 2000), a component of the Unified Medical Language System (UMLS[1], 2011AB) and is set to true when the sentence contains any inflections that are marked as comparisons. We modified the lexicon by adding terms in {'better', 'more', 'less', 'worse', 'fewer', 'lesser'} and removing terms in {'few', 'good', 'ill', 'later', 'long-term', 'low-dose', 'number', 'well', 'well-defined'}, resulting in 968 terms in total.

L2: The second lexical feature captures direction. A lexicon of 104 words was created using 82 of 174 direction verbs in (Blake, 2010) and an additional 22 manually compiled words. Selections of direction words were based on how well the individual word predicted a comparison sentence in the development set. This feature is set to true when a sentence contains any words in the lexicon.

L3: Set to true when a sentence includes any of the following words: *from, over* or *above*.

L4: Set to true when the sentence includes either *versus* or *vs*.

L5: Set to true when the sentence includes the phrase *twice the*.

L6: Set to true when the sentence includes any of the following phrases *times that of, half that of, third that of, fourth that of*

The 27 syntactic features use a combination of semantics (words) and syntax. Figure 1 shows a dependency tree that was generated using the Stanford Parser (version 1.6.9) (Klein & Manning, 2003). The tree shown in Figure 1 would be represented as:

ROOT [*root* orders [*nsubj* DBP, *cop* is, *amod* several, *prep* of [*pobj* magnitude [*amod* mutagenic/carcinogenic [*advmod* more], *prep* than [*pobj* BP]], *punct* .]]

where dependencies are shown in italics and the tree hierarchy is captured using []. The word ROOT depicts the parent node of the tree.

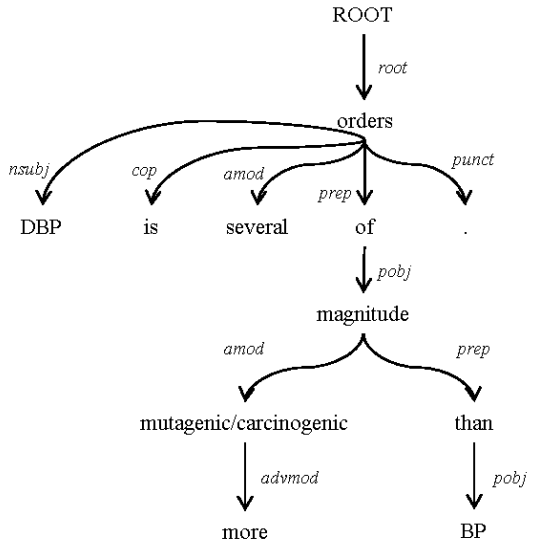

Figure 1. Dependency tree for the sentence "*DBP is several orders of magnitude more mutagenic/carcinogenic than BP.*"

We compiled a similarity and difference lexicon (*SIMDIF*), which includes 31 words such as *similar*, *different*, and *same*. Words were selected in the same way as the direction words (see L2). Each term in the *SIMDIF* lexicon has a corresponding set of prepositions that were

[1] http://www.nlm.nih.gov/research/umls/quickstart.html

collected from dictionaries. For example, the word *different* in the *SIMDIF* lexicon has two corresponding prepositions: 'from' and 'than'.

The first four syntactic rules capture comparisons containing words in *SIMDIF,* and rules 5 through 24 capture comparisons related to the features L1, L2, or both. Each of the rules 25 and 26 consists of a comparative phrase and its syntactic dependency. Each rule is reflected as a Boolean feature that is set to true when the rule applies and false otherwise. For example, rule S1 would be true for the sentence "X is similar to Y".

Subscripts in the templates below depict the word identifier and constraints applied to a word. For example W_{2_than} means that word 2 is drawn from the domain of (than), where numeric values such as 2 are used to distinguish between words. Similarly, W_{4_SIMDIF} means that the word 4 is drawn from terms in the *SIMDIF* lexicon. The symbols |, ¬, ?, and * depict disjunctions, negations, optional, and wildcard operators respectively.

S1: [*root* W_{1_SIMDIF} [*nsubj*|*cop* W_2, (*prep* W_3)?]]
S2: [¬*root* W_{1_SIMDIF} [*nsubj*|*cop* W_2, (*prep* W_3)?]]

Syntactic rules 3 and 4 capture other forms of non-gradable comparisons with connected prepositions.

S3: [(prep W_1)?, (* W_2)? [(*prep* W_3)?, (*acomp*|*nsubjpass*|*nsubj*|*dobj*|*conj*) W_{4_SIMDIF} [(*prep* W_5)?]]]
S4: [(prep W_1)?, (* W_2)? [(*prep* W_3)?, ¬(*acomp*|*nsubjpass*|*nsubj*|*dobj*|*conj*) W_{4_SIMDIF} [(*prep* W_5)?]]]

The following syntactic rules capture other non-gradable comparisons and gradable comparisons. For example, the comparative sentence example in Figure 1 has the component [*prep* than], which is satisfied by rule S5. One additional rule (rule **S27**) uses a construct of 'as … as', but it's not included here due to space limitations.

S5: [*prep* W_{1_than}]
S6: [*advmod* W_{1_than}]
S7: [*quantmod*|*mwe* W_{1_than}]
S8: [*mark* W_{1_than}]
S9: [*dep* W_{1_than}]
S10: [¬(*prep*|*advmod*|*quantmod*|*mwe*|*mark*|*dep)* W_{1_than}]
S11: [*advcl*|*prep* $W_{1\_compared}$]
S12: [*dep* $W_{1\_compared}$]
S13: [¬ *(advcl*|*prep*|*dep)* $W_{1\_compared}$]
S14: [*advcl* $W_{1\_comparing}$]
S15: [*partmod*|*xcomp* $W_{1\_comparing}$]
S16: [*pcomp* $W_{1\_comparing}$]
S17: [*nsubj* $W_{1\_comparison}$]
S18: [*pobj* $W_{1\_comparison}$]
S19: [¬ *(nsubj*|*pobj)* $W_{1\_comparison}$]
S20: [*dep* $W_{1\_contrast}$]
S21: [*pobj* $W_{1\_contrast}$]
S22: [*advmod* $W_{1\_relative}$]
S23: [*amod* $W_{1\_relative}$]
S24: [¬*(advmod*|*amod)* $W_{1\_relative}$]
S25: $W_{1\_compare}$ [*advmod* $W_{2\_(well|favorably)}$]
S26: $W_{1\_\%}$ [*nsubj* W_2 [*prep* W_{3_of}]]

Two additional general features were used. The preposition feature (*PREP*) captures the most indicative preposition among connected prepositions in the rules 1 through 4. It is a nominal variable with six possible values, and the value assignment is shown in Table 1. When more than two values are satisfied, the lowest value is assigned. The plural feature (*PLURAL*) for the rules 1 through 4 is set to true when the subject of a comparison is in the plural form and false otherwise. These two features provide information on if the sentence contains compared entities which are required in a comparison sentence.

Value	Preposition connected to *SIMDIF* word
1	*between*, *among*, or *across*
2	proper preposition provided in *SIMDIF*
3	*between*, *among*, or *across*, but may not be connected to *SIMDIF* word
4	*in* or *for*
5	any other prepositions or no preposition
6	no *SIMDIF* word is found

Table 1: *PREP* value assignment

3.3 Classifiers

The Naïve Bayes (NB), Support Vector Machine (SVM) and Bayesian Network (BN) classifiers were used in these experiments because they work well with text (Sebastiani, 2002).

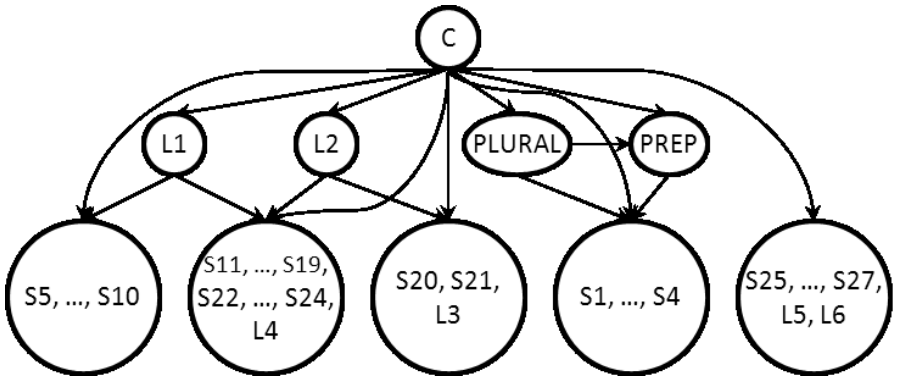

Figure 2: Bayesian Network for comparative sentences. Multiple features having the same connections are placed in a big circle node for the purpose of simple representation. C is a class variable (comparative).

The Bayesian Network model was developed to remove the independence assumption in the NB model. BN is a directed acyclic graph that can compactly represent a probability distribution because only the conditional probabilities (rather than the joint probabilities) need to be maintained. Each node in the BN represents a random variable X_i and each directed edge reflects influence from the parent node to the child node.

In order to improve Naïve Bayes classifier, we designed our Bayesian Network model by capturing proper conditional dependencies among features. Figure 2 shows the BN model used in our experiments. The relationships between features in BN were determined heuristically. Based on our observation, most gradable comparisons contain both comparison words and corresponding prepositions, so we connected such pairs. Also, most non-gradable comparisons contained comparison words and different kinds of prepositions depending on syntactic structure and plurality of subjects, and these relations are captured in the network. For example, features S5 through S10 depend on L1 because a preposition 'than' can be a good indicative word only if there is a comparison word of L1 in the same sentence. Parameters for the BN were estimated using maximum likelihood estimation (MLE) with additive smoothing. Exact inference is feasible because all nodes except for the class node are observed.

4 Results and Discussion

A pilot study was conducted using 297 and 165 sentences provided by (Fiszman et al., 2007) and (Blake, 2010) respectively to identify an initial set of features. Features were then refined based on the development set described below (section 3 reports the revised features). The BN model was also created based on results in the development set.

Sentence Type	Develop-ment	Valid-ation
Comparative Sentences	1659 (12.15%)	76 (15.2%)
Non-comparative sentences	11998 (87.85%)	424 (84.8%)
Total	13657 (100%)	500 (100%)

Table 2: Distribution of comparative and non-comparative sentences.

Experiments reported in this paper consider 122 full text articles on toxicology. Figures, tables, citations, and references were removed from the corpus, and a development set comprising 83 articles were drawn at random which included 13,657 headings and sentences (the *development set)*. Articles in the development set were manually inspected by three annotators to identify comparison claim sentences. Annotators met weekly to discuss problematic sentences and all comparison sentences were subsequently reviewed by the first author and updated where required to ensure consistency. Once the feature refinements and BN were complete, a random

sample of 500 sentences was drawn from the remaining 39 articles (the *validation set*) which were then annotated by the first author. Table 2 shows that the number of comparison and non-comparison sentences are similar between the development and validation sets.

The NB, SVM (LIBSVM package), and BN implementations from WEKA were used with their default settings (Hall et al., 2009; Chang and Lin, 2011). Classifier performance was measured using stratified 10-fold cross validation and a paired t-test was performed (using two-tail p-values 0.05 and 0.01) to determine if the performance of the BN model was significantly different from the NB and SVM.

We measured accuracy, the proportion of correct predictions, and the area under a ROC curve (ROC AUC), which is a plot of true positive rate vs. false positive rate. Given the skewed dataset (only 12% of the development sentences are comparisons), we recorded precision, recall, and F1 score of each class, where F1 score is a harmonic mean of precision and recall.

	NB	SVM	BN
Accuracy	0.923	0.933	$\mathbf{0.940}^{++}{}_{++}$
ROC AUC	0.928	0.904	$\mathbf{0.933}^{++}{}_{++}$
Comp. Precision	0.653	0.780	$\mathbf{0.782}^{++}$
Comp. Recall	**0.778**	0.621	$0.706^{--}{}_{++}$
Comp. F1 score	0.710	0.691	$\mathbf{0.742}^{++}{}_{++}$
Non-comp. Precision	**0.968**	0.949	$0.960^{--}{}_{++}$
Non-comp. Recall	0.943	**0.976**	$0.973^{++}{}_{-}$
Non-comp. F1 score	0.955	0.962	$\mathbf{0.966}^{++}{}_{++}$

Table 3: Development set results. Superscripts and subscripts depict statistical significance for BN vs. NB and BN vs. SVM respectively. +/- is significant at p=0.05 and ++/-- is significant at p=0.01. Bold depicts the best performance for each metric.

Table 3 shows the development set results. The accuracy and area under the ROC curve was significantly higher in BN compared to the NB and SVM models. For comparative sentences, recall was the highest with NB, but F1 score was significantly higher with BN. Although the difference was small, the F1 score for non-comparative sentences was significantly highest in the BN model.

Table 4 shows the validation set results, which are similar to the development set in that the BN model also achieved the highest accuracy and area under the ROC curve. The BN model had the highest non-comparative F1 score, but NB had a higher F1 score on comparatives.

	NB	SVM	BN
Accuracy	0.924	0.916	**0.932**
ROC AUC	0.948	0.883	**0.958**
Comp. Precision	0.726	**0.886**	0.875
Comp. Recall	**0.803**	0.513	0.645
Comp. F1 score	**0.763**	0.650	0.742
Non-comp. Precision	**0.964**	0.919	0.939
Non-comp. Recall	0.946	**0.988**	0.983
Non-comp. F1 score	0.955	0.952	**0.961**

Table 4: Validation set results.

The results suggest that capturing dependencies between features helped to improve the BN performance in some cases. For example, unlike the BN, the NB and SVM models incorrectly classified the following sentence as comparative: "*The method of forward difference was selected for calculation of sensitivity coefficients."* The words 'forward' and 'difference' would activate features L2 and S4, respectively, and 5 would be assigned for PREP. Since the BN model captures dependencies between L and S features and between S and the PREP feature, the probability in the BN model would not increase as much as in the NB model. To better understand the features, we conducted an error analysis of the BN classifier on validation set (see Table 5).

		Predicted	
	Class	0	1
Actual	Non-comparative (0)	417	7
	Comparative (1)	27	49

Table 5. Validation confusion matrix for BN.

We conducted a closer inspection of the seven false positives (i.e. the non-comparative sentences that were predicted comparative). In four cases, sentences were predicted as comparative because two or more independent

weak features were true. For example, in the sentence below, the features related to 'compared' (rule S11) and 'different' (rule S4) were true and produced an incorrect classification. "*Although these data cannot be compared directly to those in the current study because they are in a different strain of rat (Charles River CD), they clearly illustrate the variability in the incidence of glial cell tumors in rats."* This sentence is not comparative for *compared* since there is no comparison word between *these data* and *current study*. Similarly, this sentence is not comparative for *different* since only one *compared entity* is present for it.

Two of the remaining false positive sentences were misclassified because the sentence had a comparison word and comparison entities, but the sentence was not a *claim*. The last incorrect sentence included a comparison with a numeric value.

Reason of misclassification	# errors
Probability is estimated poorly	10
Comparison is partially covered by dependency features	7
Comparison word is not in lexicon	7
Dependency parse error	3
Total	27

Table 6. Summary of false negative errors.

We also investigated false negatives (i.e. comparative sentences that were predicted as non-comparative by the BN). The reasons of errors are summarized in Table 6. Out of 27 errors, poor estimation was responsible for ten errors. These errors mostly come from the sparse feature space. For example, in the sentence below, the features related to 'increased' (rule L2) and 'comparison' (rule S18) were active, but the probability of comparison is 0.424 since the feature space of 'comparison' feature is sparse, and the feature is not indicative enough. "*Mesotheliomas of the testicular tunic were statistically (p < 0.001) increased in the high-dose male group in comparison to the combined control groups."*

Seven of the false negative errors were caused by poor dependency features. In this case, the comparison was covered by either the parent or the child feature node, not by both. Other seven errors were caused by missing terms in the lexicons, and the last three were caused by a dependency parse error.

5 Conclusion

Comparison sentences play a critical role in scientific discourse as they enable an author to fully engage the reader by relating work to earlier research hypotheses, data collection methods, subject groups, and findings. A review scientific discourse models reveals that comparisons have been reported as the thesis/antithesis in CARS (Swales, 1990), the contrast category in RST (Mann & Thompson, 1988) in Teufel & Moens (2002) and as a comparisons category in CF (Blake, 2010).

In this paper, we introduce 35 features that capture both semantic and syntactic characteristics of a sentence. We then use those features with three different classifiers, Naïve Bayes, Support Vector Machines, and Bayesian Networks to predict comparison sentences. Experiments consider 122 full text documents and 14,157 sentences, of which 1,735 express at least one comparison. To our knowledge, this is the largest experiment on comparison sentences expressed in full-text scientific articles.

Results show that the accuracy and F1 scores of the BN were statistically ($p<=0.05$) higher than those of both the NB and SVM classifiers. Results also suggest that scientists report claims using a comparison sentence in 12.24% of the full-text sentences, which is consistent with, but more prevalent than in an earlier Claim Framework study which reported a rate of 5.11%. Further work is required to understand the source of this variation and the degree to which the comparison features and classifiers used in this paper can also be used to capture comparisons of scientific papers in other domains.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. (1115774). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ballard, B.W. (1989). A General Computational Treatment of Comparatives for Natural Language Question Answering, Association of Computational Linguistics. Vancouver, British Columbia, Canada.

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. Journal of Biomedical Informatics, 43, 173-189.

Bresnan, J.W. (1973). Syntax of the Comparative Clause Construction in English. Linguistic Inquiry, 4(3), 275-343.

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

Browne, A.C., McCray, A.T., & Srinivasan, S. (2000). The SPECIALIST LEXICON. Bethesda, Maryland.

Fiszman, M., Demner-Fushman, D., Lang, F.M., Goetz, P., & Rindflesch, T.C. (2007). In Interpreting Comparative Constructions in Biomedical Text. (pp. 37-144).

Friedman, C. (1989). A General Computational Treatment Of The Comparative, Association of Computational Linguistics (pp. 161-168). Stroudsburg, PA.

Ganapathibhotla, M., & Liu, B. (2008). Mining Opinions in Comparative Sentences. International Conference on Computational Linguistics (Coling). Manchester, UK.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1).

Jindal, N., & Liu, B. (2006). Identifying Comparative Sentences in Text Documents, Special Interest Group in Information Retrieval (SIGIR) Seattle Washington USA, 244-251.

Jindal, N., & Liu, B. (2006). Mining Comparative Sentences and Relations, American Association for Artificial Intelligence Boston, MA.

Kircz, J.G. (1991). Rhetorical structure of scientific articles: the case for argumentation analysis in information retrieval. Journal of Documentation, 47(4), 354-372.

Klein, D., & Manning, C.D. (2003). In Fast Exact Inference with a Factored Model for Natural Language Parsing. Advances in Neural Information Processing Systems, 3-10.

Mann, W.C., & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, 8(3), 243-281.

Mitchell, T.M. (1997). Machine Learning: McGraw-Hill.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1 - 47.

Staab, S., & Hahn, U. (1997). Comparatives in Context. National Conference on AI. National Conference on Artificial Intelligence 616-621.

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings: Cambridge Applied Linguistics.

Teufel, S., & Moens, M. (2002). Summarizing Scientific Articles -- Experiments with Relevance and Rhetorical Status. Computational Linguistics, 28(4), 409-445.

Xu, K., Liao, S., Li, J., & Song, Y. (2011). Mining Comparative Opinions from Customer Reviews for Competitive Intelligence. Decision Support Systems, 50(4), 743-754.

Yang, S., & Ko, Y. (2011). Extracting Comparative Entities and Predicates from Texts Using Comparative Type Classification, Association of Computation Linguistics. Portland, OR.