Using Subjectivity Analysis to Improve the Precision of Information Extraction System – A Review

Ms. Heena J Sharma #1, Mr. M. B. Chandak #2, Ms. A. M. Karandikar #3

M.Tech Scholar, CSE Department, SRCOEM, Nagpur, India #1
Professor and Head, CSE Department, SRCOEM, Nagpur, India #2
Assistant Professor, CSE Department, SRCOEM, Nagpur, India #3

Abstract—“Information Extraction” systems, face a crucial problem of false hits because of a number of reasons, observations show that a large number of these false hits are due to the presence of subjectivity in the text. This observation motivates the exploration of the notion of using subjectivity analysis to decrease the number of false hits by the Information Extraction systems. “Subjectivity Analysis” system is the one which identifies and extracts information related to opinions, attitudes and sentiments from the text. This paper presents a review on the method of subjectivity analysis, and then using its results to improve precision of information extraction system.

Keywords—Subjectivity Analysis, Information Extraction, Natural Language Processing.

I. INTRODUCTION

Information extraction system is the one which automatically identifies and extracts factual information relating to events of interest. The Information Extraction Systems is usually eager to extract information that appears relevant to its domain of interest. They search exhaustively through the available text for required information. As a result, Information Extraction systems generate a number of false hits, many of which are because of the subjective sentences present in text.

Sentences containing the colorful language of sentiments, using metaphor or hyperbola, can easily misguide an Information Extraction system, consider for example an information extraction system searching for information related to bombing is applied to a text containing sentences like: (1) “The comedian bombed last night”, (2) “The dancers exploded the stage”. In the sentence (1) the system may consider there was a bombing last night and the comedian was responsible for it. This is incorrect because here the word “bombed” is used metaphorically. Similarly the sentence (2) may be misinterpreted as the word “exploded” is used with a metaphorical sense. Other sources of false hits for IE systems are sentences with opinions, allegations, conjectures, and speculations, all of which are subjective sentences. This clearly shows that many false hits can be prevented by identifying the sentences with subjectivity and filtering them before running an information extraction system.

The main objective of this paper is to increase the precision of the information extraction system.

This paper is organized as follows: Section 2 comprises related work done; section 3 comprises of concept overview and Conclusion is in section 4.

II. RELATED WORK DONE

This section gives review of the existing approaches for exploiting subjective expressions to improve the results of information extraction systems. The authors in [1] have explored ways in which these seemingly disparate areas (subjectivity analysis and information extraction) can benefit one another. The first research direction investigates several ways that information extraction techniques can be used to learn and recognize subjective language. The approach in [1] uses weakly supervised IE learning method to automatically generate lists of subjective terms and expressions from unannotated texts. This work focuses on two types of subjective language: nouns that have a subjective meaning or connotation, and multword expressions that capture subjectivity, which have been extracted using bootstrapping algorithms (Basilisk [5] and Meta-bootstrapping [6]) and AutoSlog-TS [7] respectively. The authors in [1] and [3] have demonstrated that high-precision subjectivity classification can be used to generate a large amount of labeled training data for subsequent learning algorithms to exploit. Second, they showed that an extraction pattern learning technique can learn subjective expressions that are linguistically richer than individual words or fixed phrases. Third, they augmented their original high-precision subjective classifier with these newly learned extraction patterns. This bootstrapping process resulted in a higher recall with a minimal loss in precision.

In the second research direction in [1] they have explored the idea of using a subjective sentence classifier, on data which have not been annotated for subjectivity, to proactively identify and filter subjective sentences before extracting information from them. A series of experiments, have been carried out, exploring several strategies, including an aggressive strategy that discards all extractions in subjective sentences, and more complex ones that are more selective. The paper [1] also provides evidence that topic-based text filtering and subjectivity filtering are complementary ways to increase information extraction performance.

The authors [1] and [2] have checked the performance of these different filtering techniques on the MUC-4 terrorism data set and found that indiscriminately filtering extractions
from subjective sentences was highly aggressive, and more selective filtering strategies improved IE precision with a minimal loss of recall. It also shows that topic-based classification and subjectivity filtering are complementary methods for improving performance.

The authors in [4] have described a hybrid approach to the problem of extracting sources of opinions in text. They cast this problem as an information extraction task, using both CRFs (Conditional Random Fields) and extraction patterns. Their research is the first to identify both direct and indirect sources for all types of opinions, speculations and sentiments.

A. Summarization of Literature Reviewed
A sentence is said to be subjective if it contains subjective expressions. Subjective expressions are either words or group of words (phrases) being used to express opinions, emotions, etc. Following are some examples, with subjective expressions with single word showing subjectivity:

- Noise drives me up the wall

Thus we need to consider both kinds of expressions while dealing with subjectivity.

Once subjectivity analysis is done we need to decide the strategy of “How and how much to discard” based on the domain knowledge.

The results can then be used for information Extraction and the observations prove that it increases the precision of the extraction system.

III. CONCEPT OVERVIEW: SUBJECTIVITY ANALYSIS AND INFORMATION EXTRACTION
The concept, of using Subjectivity Analysis to improve the precision of Information Extraction system, consists of two basic steps:

1. Subjectivity Analysis – to filter the subjective sentences from the text.
2. Information Extraction – to extract the required information from the text resulting from previous step.

For the first step the work focuses on two types of subjective language: nouns that have a subjective meaning or connotation, and multiword expressions that capture subjectivity.

In order to learn subjective nouns, IE bootstrapping techniques can be explored, (say for example Basilisk [5] and Meta-Bootstrapping [6]), which were originally designed for semantic lexicon induction. These bootstrapping algorithms begin with a few “seed nouns” that are used to identify additional nouns that occur in the same extraction pattern contexts.

To learn multiword subjective expressions, the use of extraction pattern learning techniques can be investigated and the AutoSlog- TS learner [7] can be used to automatically identify extraction patterns that are correlated with subjective text. AutoSlog-TS does not require annotated training data, but it does require relevant and irrelevant text samples for training (in this case, subjective and objective text).

First an existing subjectivity lexicon, needs to be exploited, to construct domain-independent rule-based classifiers that identify subjective and objective sentences with high precision (but low recall) as shown in fig 1. Then, one needs to run the rule-based classifiers over a huge text (unannotated) corpus, which automatically generates a small but high-quality collection of subjective and objective sentences. These sentences are given to AutoSlog-TS as training data to learn IE patterns associated with subjectivity.

The AutoSlog-TS learning process has two steps. First, the syntactic templates are applied to the training corpus in an exhaustive fashion so that extraction patterns are generated for (literally) every possible instantiation of the templates that appears in the corpus. The second step of AutoSlog-TS’s learning process applies all of the learned extraction patterns to the training corpus and gathers statistics for how often each pattern occurs in subjective versus objective sentences. The Sundance system [13] can be used to parse the documents and apply the extraction patterns. AutoSlog-TS then ranks the extraction patterns using a conditional probability measure: the probability that a sentence is subjective given that it contains a specific extraction pattern. The exact formula is

\[
Pr(\text{subjective} | \text{pattern}) = \frac{\text{subjfreq(pattern)}}{\text{freq(pattern)}}
\]

where subjfreq(pattern) is the frequency of pattern in subjective training sentences, and freq(pattern) is the
frequency of pattern, in all training sentences. (This may also be viewed as the subjective precision of the pattern on the training data.) Finally, two thresholds are used to select extraction patterns that are strongly associated with subjectivity in the training data. In [1] the authors choose extraction patterns for which \( \text{freq}(\text{pattern}) \geq \theta_1 = 5 \) and \( \text{Pr} (\text{subjective} | \text{pattern}) \geq \theta_2 = 0.95 \).

The labeled sentences identified by the rule-based classifiers provide with the opportunity to apply supervised learning algorithms to the sentence classification task. Previous work [9], [10], [11] has found that Naive Bayes [12] performs well for subjectivity recognition, and it is a simple and efficient algorithm to work with. Thus, the use of Naive Bayes as our learning algorithm is advantageous. The high accuracy of Support vector machine (SVM) comes with a trade off of more processing time but can be taken as an option for future task.

The overall process used to create the classifier is depicted in Fig.2

![Fig. 2 The self-training process](image)

Then various different methods are used to remove the subjective sentences from the corpus including aggressive strategies that discard all subjective sentences and a few more complex strategies that select certain subjective sentences to discard based on different statistics.

For the second step that is Information Extraction, Information is extracted from the corpus concerning some area of interest using different strategies for discarding subjectivity before information extraction. Then results for all strategies are compared with that without discarding subjectivity and the best one is selected.

IV. CONCLUSIONS

Subjectivity Analysis systems and Information extraction systems, remain critically important in this day and age of increasingly vast amounts of text and opinions available online, but Information extraction systems suffers from a problem of false hits and can be benefited from subjectivity analysis systems by the process stated above and efforts can be made to increase precision to a higher level by using some other classifying algorithms (like SVM or Decision tree) instead of Naive Bayes.

REFERENCES