Evaluating Phrasal Semantics: 
Figurative vs. Literal

Prashant Jalan (11523), Mohit Garg (10409) 
Department of Computer Science & Engineering 
Indian Institute of Technology Kanpur

November 19, 2013
Contents

1 Introduction .................................................... 1
   1.1 The Task ................................................. 1
   1.2 Dataset .................................................. 1

2 Previous Work .................................................. 2

3 Our Approach ................................................... 3

4 Results .......................................................... 4

5 Conclusion & Future Work .................................... 5
   5.1 Conclusion ............................................... 5
   5.2 Future Work .............................................. 5
Acknowledgements

We would like to thank Dr. Amitabha Mukerjee under whose supervision we did this work. If not for his guidance, motivation and support, this work would not have been possible. We are thankful that he gave us an opportunity to work on the project.

We are thankful to Mr. M.S. Ram and Mr. Ravi Teja who were always there to help us throughout the project.
Abstract

Our project is based on SemEval Task 5b Evaluating Phrasal Semantics. We have adopted a token based approach to determine whether a given phrase has been used literally or figuratively in a given context. Inspired from IIRG submission, which provided the best results for the given task, we found that word overlapping method significantly improves after taking both noun and verb tokens into account as against taking noun tokens only in the original submission. We get the overall accuracy of 68%.
Chapter 1

Introduction

1.1 The Task

The task at our disposal is: Given a target phrase and its context (sentence/paragraph in which it has been used), we need to make a binary choice whether the given phrase has been used in its figurative sense or literal sense. The problem has always been a very challenging task in the history of Natural Language Processing. Such phrases occur frequently and often exhibit irregular behaviour by not adhering to grammatical constraints. For example, consider the sentence, Climbing Mt. Everest is not everybody’s cup of tea. In this sentence, the phrase cup of tea does not fit into existing grammatical structure of the language.

Further, having a language independent approach would also help us in various several other tasks such as machine translation and information retrieval, we are not particularly very keen to include a technique which is based on grammatical structure of the language.

However, this NLP problem has a very unique property associated with it and most of the previous approaches (including the one which we followed) have been largely based on it. This property is the property of cohesiveness. UNAL (2013) states the definition of cohesiveness as follows. **The cohesiveness of a phrase and its context is the degree of relatedness between the two.** It suggests that the cohesiveness between a phrase and its context can be measured aggregating the relatedness of the context words against the target phrase. Cohesiveness should be high for phrases used literally. Conversely, figurative usages can occur in a large variety of contexts implying low cohesiveness.

In the next section, we discuss some previous work in the field in brief which would give us the underlying idea behind the technique which has been implemented, followed by our approach, followed by the results and conclusion.

1.2 Dataset

The data provided for this task consists of two data sets LexSample and AllWords, which are divided into development, training and test sets. We considered a single training set aggregating the development and training parts from both data sets for a total of 3,230 instances for 28 phrases.
Chapter 2

Previous Work

In 1999, Lin observed that the distributional characteristics of the literal and figurative usage are different. Katz and Giesbrecht (2006) showed that the similarities among contexts are correlated with their literal or figurative usage. Then in 2006, Birke and Sarkar clustered literal and figurative contexts using a word sense disambiguation approach. Fazly et al. (2009) showed that literal and figurative usages are related to particular syntactical forms.

Sporleder and Li (2010) showed that for a particular phrase the contexts of its literal usages are more cohesive than those of its figurative usages. They present a system involving identifying the global and local contexts of a phrase. Global context was determined by looking for occurrences of semantically related words in a given passage. Local context focuses on the words immediately preceding and following the phrase.

The key idea that emerges out from the above discussion is that one can use a set of cohesive and syntactic features to train a machine learning classifier to solve the above task.

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Run</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IIRG</td>
<td>3</td>
<td>.779</td>
</tr>
<tr>
<td>2</td>
<td>UNAL</td>
<td>2</td>
<td>.754</td>
</tr>
<tr>
<td>3</td>
<td>UNAL</td>
<td>1</td>
<td>.722</td>
</tr>
<tr>
<td>5</td>
<td>IIRG</td>
<td>1</td>
<td>.530</td>
</tr>
<tr>
<td>4</td>
<td>Baseline MFC</td>
<td>-</td>
<td>.503</td>
</tr>
<tr>
<td>6</td>
<td>IIRG</td>
<td>2</td>
<td>.502</td>
</tr>
</tbody>
</table>

Figure 2.1: Results of previous work.

There were two submission which were made to SemEval 2013- UNAL and IISG. Both of these are language independent models. As Figure 1 shows, and because it is comparatively easier to implement and understand we preferred IIRG to UNAL to design our system, details of which have been provided in the sections that follow.
In first stage, we first extract the nouns and then in second stage both noun and verb tokens and take their canonical form. We use Stanford NLP suite for this purpose. But there is nothing sacrosanct about the tool and any package which does the same can be used for this purpose. For each training instance of the target phrase, we choose the threshold value to be 2, which means that we add only those words to our bag-of-words which appear at least twice in the context (instance). The choice of this threshold depends on the length of each training instance. In our case, it is 2-3 sentences in each case and hence 2 is the optimal choice. A very low value of threshold would result in adding even those words which are not cohesive. Similarly, a very high value of threshold would result in missing of some of the cohesive tokens.

We also store the frequency of occurrence of each token within a given context. Each token is labeled as figurative or literal depending on its frequency of occurrence within a given context.

To conclude, the above exercise gives us a list of tokens, their type of usage and frequencies for each target phrase. Now we train Nave Bayes Classifier using these tokens as attributes. Hence, for a test instance, this classifier gives us the output by looking at the tokens which overlap with tokens in the list and their corresponding frequencies. Hence this method is also called as word overlapping method.

The performance of this algorithm has been discussed in next section.
Chapter 4

Results

System performance is measured in terms of overall accuracy, precision and recall.

In Figure 4.1, we present our result when we used only noun tokens and then in Figure 4.2, our results are shown after taking both noun and verb tokens into account. In Figure 4.3, we show the performance of our classifier for some selected phrases.

<table>
<thead>
<tr>
<th>Overall Accuracy</th>
<th>63%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (Figurative)</td>
<td>65%</td>
</tr>
<tr>
<td>Recall (Figurative)</td>
<td>59%</td>
</tr>
<tr>
<td>Precision (Literal)</td>
<td>62%</td>
</tr>
<tr>
<td>Recall (Literal)</td>
<td>68%</td>
</tr>
</tbody>
</table>

Figure 4.1: With only nouns as context words.

<table>
<thead>
<tr>
<th>Overall Accuracy</th>
<th>68%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (Figurative)</td>
<td>71%</td>
</tr>
<tr>
<td>Recall (Figurative)</td>
<td>63%</td>
</tr>
<tr>
<td>Precision (Literal)</td>
<td>67%</td>
</tr>
<tr>
<td>Recall (Literal)</td>
<td>72%</td>
</tr>
</tbody>
</table>

Figure 4.2: With nouns and verbs as context words.

| Bread and butter | 71% |
| Break a leg | 62% |
| Drop the ball | 69% |
| Through the roof | 56% |
| Under the microscope | 73% |

Figure 4.3: Per phrase accuracy for selected phrases.
Chapter 5

Conclusion & Future Work

5.1 Conclusion

Our approach gives an overall high value of precision and recall. SemEval 2013 has not revealed much information about their model of testing. Therefore, despite being largely based on IIRG, our approach gives accuracy less than reported accuracy of IIRG. In contrast to the findings of IIRG that only nouns describe the context, we found out that verb tokens also described the context. The technique works very well on seen phrases and unseen contexts. Since its a largely supervised model largely based on cohesiveness of phrases with context words, the approach does not do well on unseen phrases. The phrases having words describing more specific objects have better accuracy than others.

5.2 Future Work

We will try to examine whether implementing more sophisticated strategies for selecting tokens for the bags of words improves the effectiveness of the Word Overlap methods. We can also try experimenting with different classifiers to see if we can get better results. Furthermore, given proper annotated data or accumulating data through crowd-sourcing we may try to run this method for other languages such as Hindi.
Bibliography


