IIRG: A Naïve Approach to Evaluating Phrasal Semantics

Lorna Byrne, Caroline Fenlon, John Dunnion
School of Computer Science and Informatics
University College Dublin
Ireland
{lorna.byrne@ucd.ie,caroline.fenlon@ucdconnect.ie,john.dunnion@ucd.ie}

Abstract

This paper describes the IIRG [1] system entered in SemEval-2013, the 7th International Workshop on Semantic Evaluation. We participated in Task 5 Evaluating Phrasal Semantics. We have adopted a token-based approach to solve this task using 1) Naïve Bayes methods and 2) Word Overlap methods, both of which rely on the extraction of syntactic features. We found that the word overlap method significantly out-performs the Naïve Bayes methods, achieving our highest overall score with an accuracy of approximately 78%.

1 Introduction

The Phrasal Semantics task consists of two related subtasks. Task 5A requires systems to evaluate the semantic similarity of words and compositional phrases. Task 5B requires systems to evaluate the compositionality of phrases in context. We participated in Task 5B and submitted three runs for evaluation, two runs using the Naïve Bayes Machine Learning Algorithm and a Word Overlap run using a simple bag-of-words approach.

Identifying non-literal expressions poses a major challenge in NLP because they occur frequently and often exhibit irregular behavior by not adhering to grammatical constraints. Previous research in the area of identifying literal/non-literal use of expressions includes generating a wide range of different features for use with a machine learning prediction algorithm. (Li and Sporleder, 2010) 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, while local context focuses on the words immediately preceding and following the phrase. Windows of five words at each side of the target were taken as features. More syntactic features were also used, including details of nodes from the dependency tree of each example. The system produced approximately 90% accuracy when tested, for both idiom-specific and generic models. It was found that the statistical features (global and local contexts) performed well, even on unseen phrases. (Katz and Giesbrecht, 2006) found that similarities between words in the expression and its context indicate literal usage. This is comparable to (Sporleder and Li, 2009), which used cohesion-based classifiers based on lexical chains and graphs. Unsupervised approaches to classifying idiomatic use include clustering (Fazly et al., 2009), which classified data based on semantic analyzability (whether the meaning of the expression is similar to the meanings of its parts) and lexical and syntactic flexibility (measurements of how much variation exists within the expression).

2 Task 5B

In Task 5B, participants were required to make a binary decision as to whether a target phrase is used figuratively or literally within a given context. The phrase "drop the ball" can be used figuratively, for example in the sentence
We get paid for completing work, so we've designed a detailed workflow process to make sure we don't

[1] Intelligent Information Retrieval Group

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 103–107, Atlanta, Georgia, June 14-15, 2013.

drop the ball.
and literally, for example in the sentence
In the end, the Referee drops the ball with the attacking player nearby.

In order to train systems, participants were given training data consisting of approximately 1400 text snippets (one or more sentences) containing 10 target phrases, together with real usage examples sampled from the WaCky (Baroni et al., 2009) corpora. The number of examples and distribution of figurative and literal instances varied for each phrase.

Participants were allowed to submit three runs for evaluation purposes.

2.1 Approach

The main assumption for our approach is that tokens preceding and succeeding the target phrase might indicate the usage of the target phrase, i.e. whether the target phrase is being used in a literal or figurative context. Firstly, each text snippet was processed using the Stanford Suite of Core NLP Tools[2] to tokenise the snippet and produce part-of-speech tags and lemmas for each token.

During the training phase, we identified and extracted a target phrase boundary for each of the target phrases. A target phrase boundary consists of a window of tokens immediately before and after the target phrase. The phrase boundaries identified for the first two runs were restricted to windows of one, i.e. the token immediately before and after the target phrase were extracted, tokens were also restricted to the canonical form.

For example, the target phrase boundary identified for the snippet: *"The returning team will drop the ball and give you a chance to recover ."* is as follows:

```
before:will
after:and
```

and the target phrase boundary identified for the snippet: *"Meanwhile , costs are going through the roof ."* is as follows:

```
before:go
after:.
```

[2]http://nlp.stanford.edu/software

IIRG Training Runs	
RunID	Accuracy (%)
Run0	85.29
Run1	81.84
Run2	95.92

Table 1: Results of IIRG Training Runs

We then trained multiple Naïve Bayes classifiers on these extracted phrase boundaries. The first classifier was trained on the set of target phrase boundaries extracted from the entire training set of target phrases and usage examples (Run0); the second classifier was trained on the set of target phrase boundaries extracted from the entire training set of target phrases and usage examples including the phrase itself as a predictor variable (Run1); and a set of target-phrase classifiers, one per target phrase, were trained on the set of target phrase boundaries extracted from each individual target phrase (Run2).

The results of the initial training runs can be seen in Table 1. Although Run0 yielded very high accuracy scores on the training data, outperforming Run1, in practice this approach performed poorly on unseen data and was biased towards a figurative classification. We thus opted not to implement this run in the testing phase and instead concentrated on Run1 and Run2.

For our third submitted run, we adopted a word overlap method which implemented a simple bag-of-words approach. For each target phrase we created a bag-of-words by selecting the canonical form of all of the noun tokens in each corresponding training usage example. The frequency of occurrence of each token within a given context was recorded and each token was labeled as *figurative* or *literal* depending on its frequency of occurrence within a given context. The frequency of occurrence of each token was also recorded in order to adjust the threshold of token occurrences for subsequent runs. For this run, Run3, the token frequency threshold was set to 2, so that a given token must occur two or more times in a given context to be added to the bag-of-words.

3 Results

System performance is measured in terms of accuracy. The results of the submitted runs can be seen in Table 2.

Of the submitted runs, the Word Overlap method (Run3) performed best overall. This approach was also consistently good across all phrases, with scores ranging from 70% to 80%, as seen in Table 3.

The classifiers trained on the canonical phrase boundaries (Run1 and Run2) performed poorly on unseen data. They were also biased towards a figurative prediction. For several phrases they incorrectly classified all literal expressions as figurative. They were not effective at processing all of the phrases: in Run1, some phrases had very high scores relative to the overall score (e.g. "break a leg"), while others scored very poorly (e.g. "through the roof"). In Run2, a similar effect was found. Interestingly, even though separate classifiers were trained for each phrase, the accuracy was lower than that of Run1 in several cases (e.g. "through the roof"). This may be a relic of the small, literally-skewed, training data for some of the phrases, or may suggest that this approach is not suitable for those expressions. The very high accuracy of the classifiers tested on a subset of the training data may be attributed to overfitting. The approach used in Run1 and Run2 is unlikely to yield very accurate results for the classification of general data, due to the potential for many unseen canonical forms of word boundaries.

3.1 Additional Runs

After the submission deadline, we completed some additional runs, the results of which can be seen in Table 4.

These runs were similar to Run1 and Run2, where we used Naïve Bayes Classifiers to train on extracted target phrase boundaries. However, for Run4 and Run5 we restricted the phrase boundaries to the canonical form of the nearest verb (Run4) or nearest noun (Run5) that was present in a bag-of-words.

We used the same bag-of-words created for Run3 for the noun-based bag-of-words, and this same approach was used to create the (canonical form) verb-based bag-of-words. If there were no such verbs or nouns present then the label NULL was applied. If a phrase occurred at the start or end of a text snippet this information was also captured. The Naïve Bayes classifiers were then trained using labels from the following set of input labels: FIGURATIVE, LITERAL, START, END or NULL, which indicate the target phrase boundaries of the target phrases.

For example, the target phrase boundaries identified for the snippet: *"Meanwhile , costs are going through the roof ."* for Run4 and Run5, respectively, are as follows:

```
before:FIGURATIVE
after:END
```

where the FIGURATIVE label is the classification of the token *'going'* as indicated in the verb-based bag-of-words, and

```
before:FIGURATIVE
after:END
```

where the FIGURATIVE label is the classification of the token *'costs'* as indicated in the noun-based bag-of-words.

As in Run1 and Run2, an entire-set classifier and individual target-phrase classifiers were trained for both runs. These additional runs performed well, yielding high accuracy results and significantly outperforming Run1 and Run2.

The Run4 classifiers did not perform comparatively well across all phrases. In particular, the target phrase "break a leg", had very low accuracy scores, possibly because the training data for the phrase was small and contained mostly literal examples. The ranges of phrase scores for the noun classification runs (Run5) were similar to those of the Word Overlap runs. The results across each phrase were also consistent, with no scores significantly lower than the overall accuracy. Using target phrase boundaries based on noun classifications may prove to yield reasonable results when extended to more phrases, as opposed to the erratic results found when using verb classifications.

In both Run4 and Run5, very similar overall results were produced from both the entire-set and target-phrase classifiers. In most cases, the run performed poorly on the same phrases in both instances, indicating that the approach may not be appropriate for the particular phrase. For example, the verb classifications runs scored low accuracy for "drop the ball", while the noun classifications run was approximately 80% accurate for the same phrase using both

IIRG Submitted Runs (%)					
RunID	Overall Accuracy	Precision (Figurative)	Recall (Figurative)	Precision (Literal)	Recall (Literal)
Run1	53.03	52.03	89.97	60.25	15.65
Run2	50.17	50.81	41.81	54.06	58.84
Run3	77.95	79.65	75.92	76.62	80.27

Table 2: Results of Runs Submitted to Sem-Eval 2013

IIRG Submitted Runs - Per Phrase Accuracy (%)										
RunID	At the end of the day	Bread and butter	Break a leg	Drop the ball	In the bag	In the fast lane	Play ball	Rub it in	Through the roof	Under the micro-scope
Run1	68.92	57.89	40.00	40.82	43.42	67.86	52.63	66.67	64.94	33.33
Run2	45.95	38.16	83.33	57.14	48.68	75.00	46.05	56.67	29.87	62.82
Run3	75.68	82.89	73.33	83.67	72.37	75.00	78.95	60.00	80.52	83.33

Table 3: Results of Runs Submitted to Sem-Eval 2013 (per phrase)

IIRG Additional Runs - Accuracy (%)		
RunID	Entire-Set Classifier	Target-Phrase Classifier
Run4	64.81	65.99
Run5	75.25	76.60

Table 4: Accuracy of Additional Unsubmitted Runs

an entire-set and target-phrase classifier.

4 Conclusion

This is the first year we have taken part in the Semantic Evaluation Exercises, participating in Task 5b, Evaluating Phrasal Semantics. Task 5B requires systems to evaluate the compositionality of phrases in context. We have adopted a token-based approach to solve this task using 1) Naïve Bayes methods whereby target phrase boundaries were identified and extracted in order to train multiple classifiers; and 2) Word Overlap methods, whereby a simple bag-of-words was created for each target phrase. We submitted three runs for evaluation purposes, two runs using Naïve Bayes methods (Run1 and Run2) and one run based on a Word Overlap approach (Run3). The Word Overlap approach, which limited each bag-of-words to using the canonical form of the nouns in the text snippets, yielded the highest accuracy scores of all submitted runs, at approximately 78% accurate. An additional run (Run5), also using the canonical form of the nouns in the usage examples but implementing a Naïve Bayes approach, yielded similar results, almost 77% accuracy. The approaches which were restricted to using the nouns in the text snippets yielded the highest accuracy results, thus indicating that nouns provide important contextual information for distinguishing literal and figurative usage.

In future work, we will explore whether we can improve the performance of the target phrase boundaries by experimenting with the local context window sizes. Another potential improvement might be to examine whether implementing more sophisticated strategies for selecting tokens for the bags-of-words improves the effectiveness of the Word Overlap methods.

References

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43*, 3(3):209–226.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103, March.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19.

Linlin Li and Caroline Sporleder. 2010. Linguistic Cues for Distinguishing Literal and Non-Literal Usages. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, pages 683–691.

Caroline Sporleder and Linlin Li. 2009. Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*.