Phrasal Semantics - Figurative Vs Literal

Prashant Jalan, Mohit Garg

October 22, 2013

1 The Problem

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.

In SemEval 2013 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 weve designed a detailed workflow process to make sure we dont drop the ball*" and literally, for example in the sentence "*In the end, the Referee drops the ball with the attacking player nearby.*"

Being able to detect whether a phrase is used literally or figuratively is especially important for information retrieval, where figuratively used words should be treated separately to avoid false positives. For example, the example sentence "*He will go down in history as one of the old school, a true gentlemen.*" should probably not be retrieved for the query '*school*'. Rather, the insights generated from subtask a could be utilized to retrieve sentences using a similar phrase such as '*gentleman-like behaviour*'.

2 Previous Work

In *SemEval 2013*, there were primarily two approached discussed, UNAL and IIRG. For known phrases the accuracy of IIRG was reported to be 78% and that of UNAL 75.4%[1]. For unseen phrases the accuracy was 66.8% and 65.4% respectively[1].

In IIRG[2], tha authors adopted a token-based approach to solve the task using 1) Naive Bayes methods and 2) Word Overlap methods, both of which rely on the extraction of syntactic features. 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. In the word overlap method,

they create a bag-of-words for each target phrase by selecting the canonical form of all of the noun tokens. The final run employing the word overlap method yielded the highest accuracy.

In UNAL[3], the authors proposed a set or features based on cohesiveness, syntax and stylometry, which are used to train a machine learning classifier. The cohesiveness between a phrase an its context can be measured aggregating the relatedness of the context words against the target phrase using measures like pointwise mutual information(PMI). Stylistic trend says that figurative usage tends to occur later in the document in comparison with the literal usage.

There are several other approaches proposed, using LSA (latent semantic analysis)[5], using lype and token identification[4], unsupervised approach discourse connectivity/cohesion[6] and lexical cohesion[7].

3 Our approach

We plan to start with the two approached followed in *SemEval 2013*, employing both simultaneously in a hope to achieve better results. Furhtermore, if time permits, we plan to employ more complex methods such as latent semantic analysis, lexical cohesion or discourse cohesion.

We will be using the annoted dataset provided during *SemEval 2013* and will be analysing the problem only for the **English** language.

References

[1] Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*

[2] Lorna Byrne, Caroline Fenlon and John Dunnion. 2013. IIRG: A Naive Approach to Evaluating Phrasal Semantics. *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*

[3] Sergio Jimenez, Claudia Becerra and Alexander Gelbukh. 2013. UNAL: Discriminating between Literal and Figurative Phrasal Usage Using Distributional Statistics and POS tags. *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*

[4] Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61103, March.

[5] 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 1219.

[6] 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 683691.

[7] 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)*