Entity Centric Opinion Mining from Blogs

Akshat Bakliwal, *Piyush Arora*, *Vasudeva Varma*
Search and Information Extraction Lab,
LTRC, International Institute of Information Technology, Hyderabad
akshat.bakliwal@research.iiit.ac.in, piyush.arora@research.iiit.ac.in,
vv@iiit.ac.in

ABSTRACT
With the growth of web 2.0, people are using it as a medium to express their opinion and thoughts. With the explosion of blogs, journal like user-generated content on the web, companies, celebrities and politicians are concerned about mining and analyzing the discussions about them or their products. In this paper, we present a method to perform opinion mining and summarize opinions at entity level for English blogs. We first identify various objects (named entities) which are talked about by the blogger, then we identify the modifiers which modify the orientation towards these objects. Finally, we generate object centric opinionated summary from blogs. We perform experiments like named entity identification, entity-modifier relationship extraction and modifier orientation estimation. Experiments and Results presented in this paper are cross verified with the judgment of human annotators.

KEYWORDS: Sentiment Analysis, Opinion Mining, English Blog, Object Identification, Opinion Summary.

Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), pages 53–64,
COLING 2012, Mumbai, December 2012.

1 Introduction

A Blog is a web page where an individual or group of users record opinions, information, etc. on a regular basis. Blogs are written on many diverse topics like politics, sports, travel and even products. However, the quality of the text generated from these sources is generally poor and noisy. These texts are informally written and suffer from spelling mistakes, grammatical errors, random/irrational capitalization (Dey and Haque, 2008).

Opinion Mining from blogs aims at identifying the viewpoint of the author about the objects[1]. Summarizing these expressed viewpoints can be useful for many business and organizations where they analyze the sentiment of the people on a product, or for an individual(s) who are curious to know opinions of other people. Current approaches on opinion identification divide the larger problem (document) into sub-problems (sentences) and approach each sub-problem separately. These approaches have a drawback that they cannot capture the context flow and opinion towards multiple objects within the blog.

Blog summarization task is considered as normal text summarization, without giving significance to the nature and structure of the blog. Current state of art summarization systems perform candidate sentences selection from the content and generate the summary.

In this paper[2], we present a new picture to blog opinion mining, an entity perspective blog opinion mining and summarization. Here, we identify the objects which the blogger has mentioned in the blog along with his view points on these objects. In this work, named entities are potential objects for opinion mining. We perform opinion mining for each of these objects by linking modifiers to each of these objects and deciding the orientation of these modifiers using a pre-constructed subjective lexicon. And finally, we generate two different concept summaries: an object wise opinionated summary of the document and opinionated summary of the object across the dataset.

2 Related Work

The research we propose here is a combination of Opinion Mining and Summarization. (Pang et al., 2002; Turney, 2002) started the work in the direction of document level sentiment analysis. Major work in phrase level sentiment analysis was initially performed in (Agarwal et al., 2009; Wilson, 2005). (Hu and Liu, 2004; Liu and Hu, 2004; Popescu and Etzioni, 2005) concentrated on feature level product review mining. They extracted features from product reviews and generated a feature wise opinionated summary. Blog sentiment classification is primarily performed at document and sentence level. (Ku et al., 2006) used TREC and NTCIR blogs for opinion extraction. (Chesley, 2006) performed topic and genre independent blog classification, making novel use of linguistic features. (Zhang and et al., 2007) divided the document into sentences and used Pang (Pang et al., 2002) hypothesis to decide opinion features. (He et al., 2008) proposed dictionary based statistical approach which automatically derives the evidence for subjectivity from blogs. (Melville et al., 2009) and (Draya et al., 2009) are among few other works on blog sentiment analysis.

MEAD (Radev et al., 2004) is among the first few and most widely used summarizing systems. (Arora and Ravindran, 2008) perform multi-document summarization. In (Zhou and Hovy, 2005), authors try to summarize technical chats and email discussions.

[1] In this article, we shall refer to named entity(ies) as object(s).

[2] Due to space constraints, we have eliminated some parts and discussions. Extended version of this paper is available at http://akshatbakliwal.in

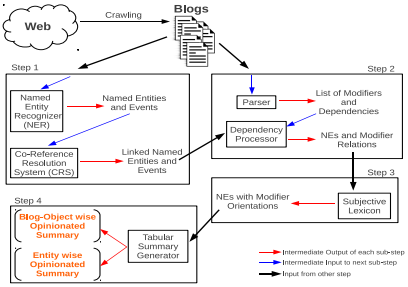

Figure 1: Algorithmic Flow of the proposed approach

Our work derives its motivation from (Hu and Liu, 2004; Liu and Hu, 2004). They identified product features and generated an opinionated summary from product reviews. In our task, the data is more formal (structured with less grammatical and spelling errors) and we generate opinionated summaries of objects (person, organizations, etc). Our summary differs from a conventional summary because we don't pick candidate sentences directly from the text, we pick only entities and opinion words towards those entities.

3 Proposed Approach

In this research, we present a new and different approach towards blog opinion analysis. Apart from the traditional approaches of classifying a blog at the sentence and document level, we describe an approach which uses the connectivity and contextual information mined using parsing. The approach proposed here depends on two lexical resources, Stanford CoreNLP[3] Tool and SentiWordNet (Baccianella et al., 2010). Stanford CoreNLP tool includes Parser, Part-of-Speech Tagger (PoS), Named Entity Recognizer (NER), Co-reference Resolution System (CRS). Our approach can be viewed as comprising of four major steps. *Figure 1* represents the architecture of the approach proposed highlighting all the sub-modules and intermediate inputs and outputs.

1. Object Identification: What is an Object? An Object is the entity the blogger is talking and expressing his views about. A blog can have multiple objects which are discussed at varied level of depths. In this step we extract the objects from the input blog. There are various techniques that can be used for performing object identification, using Named Entity Recognizer and Noun Phrase patterns (Hu and Liu, 2004; Liu and Hu, 2004).

2. Modifier Identification: What are object modifiers? An object modifier is usually an adjective, an adverb or a verb which modifies the object directly or indirectly. In this step we extract the modifiers from the blog and map them to the objects identified in Step 1. In this step, we find all the modifiers and link them to corresponding objects. In the subsequent steps we use only those modifiers which are linked to any object. We focus primarily on adjective modifier (amod), adverb modifier (advmod), nominal subject (nsubj), etc types of dependencies from Stanford parser to link modifiers to their

[3] http://nlp.stanford.edu/software/corenlp.shtml

Source	Telegraph (telegraph.co.uk)
Domain	Sports
Number of Blogs	100
Number of Unique Objects	1984
Number of Modifiers	4540

Table 1: Summary of Blog Dataset

objects. For ex. 'Ram Lal is a good boy.', for this example following are the collapsed dependencies: nn(Lal-2,Ram-1); nsubj(boy-6,Lal-2); cop(boy-6,is-3); det(boy-6,a-4); amod(boy-6,good-5). Using the nn tag 'Ram' and 'Lal' are combined a single entity [Ram Lal], amod tag maps 'good' (adjective) to 'boy' (noun) and nsubj tag maps 'boy' (noun) to 'Lal' [Ram Lal].

3. Modifier Orientation: What really is Modifier Orientation? Every adjective, adverb and verb have some implicit polarity (positive, negative or neutral) associated with them. With this polarity they modify the orientation of the objects. Once we get the objects being talked about in the blog and also have the modifiers with respect to each of these objects, the next important task is to assign subjectivity scores to these modifiers. We use SentiWordNet(Baccianella et al., 2010) to determine the polarity of each modifier.

4. Summary Generation: We generate an entity wise opinionated summary in the last step. We generate two different kinds of summaries, blog level and entity wise.

   At blog level, after assigning orientation to each of the modifier for a particular entity, we generate a tabular summary of the whole blog which has two main parts, different entities and events being talked about and opinion orientation with respect to each entity. Template of the summary is as follows: <Entity, Opinion Words, Opinion Orientation>

   In entity wise summary, we collect all the opinions expressed across all the blogs on that entity. After collecting all the opinions we generate a summary for each of the entities. Template of the summary is as follows: <Entity, Opinion Words, Opinion Orientation,Number of Blogs, List of Blog titles>

4 Dataset

We have collected 100 blogs from Telegraph[4] in sports domain. These blogs are from various categories like London Olympics, Cricket, Boxing, etc. While working on this data, we observed that blogs are usually comparison between two or more objects like "X is better than Y under some circumstances". Hence, we have found many objects in the blog but we find very few opinion words for various entities. Refer *Table 1* for the dataset used in this research.

We decided to go with a more formal dataset because of two reasons. Firstly, there was no dataset available aprior which was annotated in the required format. Secondly, to avoid any form of biasness while collecting the dataset, we simply crawled top 100 blogs from Telegraph sports section. Although this dataset is free from most of the anomalies present in general blog data, but helps us to present the essence of our proposed approach very clearly. A few observations we made while working on blog dataset were: Much of the information present in

[4]telegraph.co.uk

the blog(s) were factual, most of the opinions expressed were either in comparison format or negatively orientated.

4.1 Evaluation Setup

In this subsection, we explain the method used for evaluating our approach. We hired three human annotators for this task and calculation of their mutual agreement is done using Cohen's Kappa measurement[5]. Validation task was divided into three basic steps

1. Object Identification: Each human annotator was asked to identify all the named entities (person, organizations, location, etc) from the text. This process is similar to step 1 of our proposed approach. *Table 2* gives the agreement of human annotators for object identification.

	Total Unique Objects Identified	Total Unique Modifiers Identified
Annotator 1	1984	3690
Annotator 2	1698	3740
Annotator 3	1820	3721
Average κ Score	**0.856**	**0.818**

Table 2: Manual agreement scores for Object and Modifier Identification

2. Modifier Identification: After step 1, they were asked to mark and decide the orientation (positive or negative) for all the modifier words (adjectives, adverbs and verbs) from the text. This step involved a good understanding of English language and word usage. This corresponds to step 2 of our approach. *Table 2* gives the agreement of human annotators for modifier identification.

3. Object-Modifier Relation: Here, they were asked to assign/link modifiers to named entities i.e. to determine the opinion of the blogger towards the objects. This step was the most tricky step as it requires a clear understanding of language construct(s). This corresponds to step 3 in our approach where we use dependency processor to handle this. Dependency Processor is a module which reads the typed dependencies retrieved from stanford parser and relates the attributes of these dependency tags with each other.

In the end, for the cases where the annotators failed to achieve an agreement, first and second authors of this paper performed the task of annotation to resolve the disagreement. *Table 3* gives the kappa (κ) statistics of human agreement for each of these tasks.

One striking observation we made was that majority of the modifiers were negatively orientated i.e. blogs are frequently written to express negative sentiments (or disagreement) about the object.

5 Experiments and Results

We divide the experiment into four steps as discussed in *Section 3* (Approach). In this section, we describe the experiment using a small running example[6] from the corpus. We illustrate the

[5] http://en.wikipedia.org/wiki/Cohen's_kappa

[6] Title: Channel 4 unveils Paralympic Games broadcast team - Clare Balding, Jon Snow leading lights

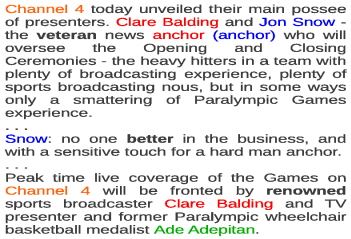

Figure 2: Image highlights named entities in the piece of text. Words in bold highlight the modifiers and Words coloured using same colour highlight same entities.

Kappa (κ) score between annotator 'i' and annotator 'j' (κ_{ij})	
κ_{12}	**0.875**
κ_{13}	0.827
κ_{23}	0.842
Average κ Score	**0.848**

Table 3: Kappa Scores for Manual Agreement

tools we have used for each step with a small description. We also highlight the task done in each step for snippet example in *Figure 2*.

- Step 1, we identify the objects using NER (Stanford NER). We use NER over noun phrase patterns because noun phrase patterns tend to introduce more noise. There can be many noun phrases which have no named entities. And also, we have to use some method (like association rule mining) to discard non relevant noun phrases. After performing named entity identification, we then perform co-reference resolution to link all the instances of these entities together, using CRS (Stanford Co-reference Resolution). Stanford NER and Stanford CRS tools were available in Stanford CoreNLP toolkit.

  Using Stanford NER, our system discovered a total 1756 unique named entities from 1984 unique named entities tagged by human annotators. In the sample snippet shown in *Figure 2*, we have 4 named entities "Channel 4", "Clare Balding", "Jon Snow" and "Ade Adepitan".

- In Step 2, we identify adjectives, adverbs and verbs which modify the named entities identified in step 1. We link the named entities and modifiers using the dependencies (like amod, advmod, nsubj, etc) given by Stanford parser. We perform dependency association to a level of depth 2. We also discard all the modifiers which are not mapped to any named entity as they are of no use to our system later. Stanford parser and Stanford part-of-speech used in this step is also available in Stanford CoreNLP toolkit.

  Using Stanford parser and part-of-speech tagger, our system discovered 3755 correct modifiers (adjectives + adverbs + verbs). This is 82.7% of what human annotators identified (4540). For the sample snippet in *Figure 2* mappings (modifier to entity) are shown in *Table 4*.

- Using SentiWordNet, we identified the orientation of these modifiers in Step 3. We use the most common used sense of each word for scoring in order to handle multiple senses of each word. While deciding the orientation of a modifier, we perform negation handling and take in account for negative words (like not, no , never, *n't, etc.) preceding it within a window of 3 words to the left..

Modifier	Object	Parser Dependency
<>	Channel 4	<>
Veteran	Clare Balding	amod(veteran, anchor); dep(anchor, Balding)
Veteran	John Snow	amod(veteran, anchor); dep(anchor, Balding); conj_and(Balding, Snow)
Better	John Snow	advmod(better, one); dep(one, Snow)
Renowned	Clare Balding	amod(renowned, Balding)
<>	Ade Adepitan	<>

Table 4: Object to Modifier Mapping steps

Object	Modifier(s)	Orientation
Channel 4	< >	Neutral
Clare Balding	<veteran, renowned>	Positive
Jon Snow	<veteran, better>	Positive
Ade Adepitan	<medalist>	Positive

Table 5: Blog-Object Summary for the example

- In Step 4, we create a tabular summary of objects and their respective modifiers (Refer *Table 5*). This summary belongs to type 1 : Blog level summary. Using this kind of summary, we can draw a picture of user's mind and how he/she thinks about various entities. The second type of summary generated can be used to compare two different entities.

Table 6 reports the accordance of our proposed algorithm with human annotators. Opinion orientation agreement is calculated as an aggregate opinion towards an entity.

One plausible reason for decent agreement of our system with manual annotation is that, most of the external tools (Stanford CoreNLP, Parser, PoS tagger) we used in this research are trained on Wall Street Journal News wire data. Our dataset is also taken from a news website, and is written by professional content writers.

Unique objects identified by human annotators	1984
Unique objects identified by our system	1919
Correct Unique objects identified by our system	1756
Object identification coverage	**88.5%**
Total modifiers tagged by human annotators	4540
Total modifiers tagged by our system	4690
Correct Total modifiers tagged by our system	3755
Modifier identification coverage	**82.7%**
Opinion orientation agreement (aggregate)	**81.4%**

Table 6: Results of the system proposed and developed in this research

6 Discussion

In this section, we discuss some of the challenges we faced while working with the tools used in the above approaches. We try to illustrate the drawbacks of the tools with help of examples, specific for each step.

- In step 1, we proposed the use of NER. We covered $\sim$ 82.7% of the named entities tagged by human annotators. Stanford NER failed to detect multi word organization names at many places. For example "Great Britain basketball squad" in this example Stanford NER tagged "Great Britain" as a location and didn't tag basketball squad (tagged as 'Other'). But in actual, this whole should be tagged as an organization. For another example "GB Performance Director Chris Spice", in this example all the words in this phrase were tagged as an organization but here "Chris Spice" should have been tagged as person.
- In step 2, we use the parser and part-of-speech information derived using Stanford Parser and PoS tagger. These tools also produced some errors, for example in one of the blogs, "match-winning" is tagged as an adjective and in another example, "fine-innings" is tagged as an adjective.
- We use SentiWordNet to decide the polarity of each modifier in step 3. This is a general lexicon built for large purpose sentiment analysis and doesn't cover various words which are specific to sports domain. Words which are specific to sports domain like "medalist", "winner" are not present in such lexicons, and thus we need to build a domain specific lexicon.

We perform entity centric opinion mining on blogs because blogs are document like big collection of text. In blogs, context flows within sentences, across sentences and across multiple paragraphs. It is very hard to perform sentiment analysis at document and sentence level because there are multiple objects being talked about and also at varied level of depths. Calculating the overall opinion is difficult and also it will not present the correct picture. Thus, in this research, we first identify the objects (entities) and perform sentiment analysis and opinion mining across the entities. For example, *"England batted poorly, but credit to Saeed Ajmal for a quite superb performance, ending up with career-best figures of 7-55."* in this sentence, there are multiple opinions. We discuss this example in more detail in *Appendix A*.

Traditional N-Gram based approaches have following limitations: limited training data, diverse topics, context dependency and vocabulary mismatch. Problem of limited training data, context dependency (partial) and vocabulary mismatch are addressed by far using our approach. Our proposed approach is not completely hassle free. Our approach has these limitations: no prior established annotated dataset, determining modifier orientation and poor performance on complex dependency relations.

Conclusion

Blog opinion identification and summarization is an interesting task which will be very useful for businesses to analyze users' opinion at a fine grained feature level, for governments to understand the fall backs in the policies introduced. We described a method to generate opinionated summary of various entities within the blog and also across the corpus in an automated manner. We achieved $\sim$86% agreement in object identification, $\sim$83% accordance in modifier orientation and $\sim$81% agreement in opinion orientation identification.

References

Agarwal, A., Biadsy, F., and Mckeown, K. R. (2009). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*.

Arora, R. and Ravindran, B. (2008). Latent dirichlet allocation based multi-document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, AND '08.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Chesley, P. (2006). Using verbs and adjectives to automatically classify blog sentiment. In *In Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*.

Dey, L. and Haque, S. K. M. (2008). Opinion mining from noisy text data. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, AND '08.

Draya, G., Plantié, M., Harb, A., Poncelet, P., Roche, M., and Trousset, F. (2009). Opinion mining from blogs. In *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*.

He, B., Macdonald, C., He, J., and Ounis, I. (2008). An effective statistical approach to blog post opinion retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04)*.

Ku, L.-W., Liang, Y.-T., and Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Liu, B. and Hu, M. (2004). Mining opinion features in customer reviews. In *Proceedings of American Associatoin for Artifical Intelligence (AAAI'04)*.

Melville, P., Gryc, W., and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*.

Radev, D., Allison, T., Blair-Goldensohn, S., and Blitzer (2004). Mead- a platform for multidocument multilingual text summarization. In *Conference on Language Resources and Evaluation (LREC)*.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Wilson, T. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *In Proceedings of HLT-EMNLP*.

Zhang, W. and et al. (2007). Opinion retrieval from blogs. In *In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (2007)*.

Zhou, L. and Hovy, E. (2005). Fine-grained clustering for summarizing chat logs. In *Proceedings of the Workshop on Beyond Threaded Conversation, held at the Computer-Human Interaction conference (CHI2005)*.

Appendix A

Word	**Lemma**	**PoS**	**NER**
England	England	NNP	Location
batted	bat	VBD	O
poorly	poorly	RB	O
,	,	,	O
but	but	CC	O
credit	credit	NN	O
to	to	TO	O
Saeed	Saeed	NNP	Person
Ajmal	Ajmal	NNP	Person
for	for	IN	O
a	a	DT	O
quite	quite	RB	O
superb	superb	JJ	O
performance	performance	NN	O
,	,	,	O
ending	end	VBG	O
up	up	RP	O
with	with	IN	O
career-best	career-best	JJ	O
figures	figure	NNS	O
of	of	IN	O
7-55	7-55	CD	Number
.	.	.	O

Table 7: Results of Stanford CoreNLP on the sample sentence

Here, we provide the output of Stanford CoreNLP for the sentence *"England batted poorly, but credit to Saeed Ajmal for a quite superb performance, ending up with career-best figures of 7-55."* in *Table 7*. *Figure 3* shows the output of Part-of-speech tagger. *Figure 4* shows the output of

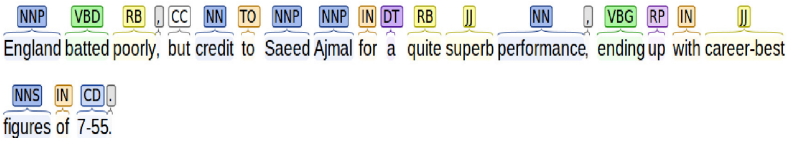

Figure 3: Figure shows the output of Part-of-speech tagger for the sample sentence

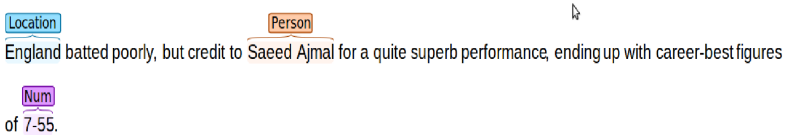

Figure 4: Figure shows the output of Named Entity Recognizer for the sample sentence

named entity recognizer and in *Figure 5* we show the collapsed dependencies for the sample sentence. We have taken *Figure 3, 4, 5* from here[7]. If we consider the sentence as a whole unit, we cannot predict concretely whether it is positive, negative or neutral sentence but if we look entity wise, we can construct a clear representation for it. We have mainly 2 entities "England" and "Saeed Ajmal" which are referred in this sentence. The entity wise analysis will yield us the following results:
– England: batted poorly. (Negative). Negative sentiment is imparted by adverb "poor". "poor" is connected with "batted" using an adverb modifier (advmod) tag and "batted" is connected to "England" by nominal subject (nsubj) tag. *Figure 5* highlight all these connections.
– Saeed Ajmal: quite superb performance. (Positive). Positive adjective ("superb") is linked to "performance" using adjective modifier (amod) relation and "performance" is linked to "Ajmal" via preposition for (prep_for). In this way, we get the positive sentiment towards *Saeed Ajmal*. This entity wise representation provides us with a clear picture of the sentence and overcomes the limitation of sentence level sentiment classification.

In this example, we can also see that "England" is identified incorrectly by Stanford NER. Here "England" should have been identified as an organization (England Cricket Team) rather than as location.

[7] http://nlp.stanford.edu:8080/corenlp/

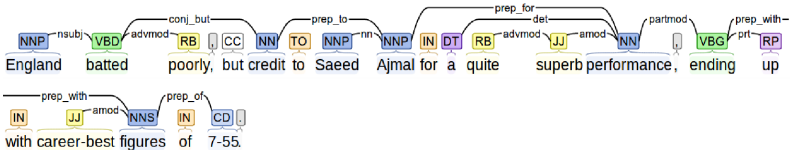

Figure 5: Figure shows the output of Parser for the sample sentence