A 5W1H Based Annotation Scheme for Semantic Role Labeling of English Tweets

Kunal Chakma¹, Amitava Das²

¹ National Institute of Technology Agartala, India
² Indian Insitute of Information Technology, Sri City, India

Abstract. Semantic Role Labeling (SRL) is a well researched area of Natural Language Processing. State-of-the-art lexical resources have been developed for SRL on formal texts that involve a tedious annotation scheme and require linguistic expertise. The difficulties increase manifold when such complex annotation scheme is applied on tweets for identifying predicates and role arguments. In this paper, we present a simplified approach for annotation of English tweets for identification of predicates and corresponding semantic roles. For annotation purpose, we adopted the 5W1H (Who, What, When, Where, Why and How) concept which is widely used in journalism. The 5W1H task seeks to extract the semantic information in a natural language sentence by distilling it into the answers to the 5W1H questions: Who, What, When, Where, Why and How. The 5W1H approach is comparatively simple and convenient with respect to the ProbBank Semantic Role Labeling task. We report an the performance of our annotation scheme for SRL on tweets and show that non-expert annotators can produce quality SRL data for tweets. This paper also reports the difficulties and challenges involved with semantic role labeling on twitter data and propose solutions to them.

Keywords. 5W1H, semantic role labeling, twitter.

1 Introduction

Recently, research and development in natural language understanding on social media texts has significantly taken a momentum. Semantic role labeling(SRL) is one such natural language understanding task that involves shallow semantic parsing and has wide applications in other natural language processing areas such as question and answering(QA), information extraction(IE), machine translation, event tracking and so on. For understanding an event from a given sentence means being able to answer “Who did what to whom when where why and how”. To answer such questions of “who,what”etc., it is important to identify each syntactic constituents of a sentence such as predicates, subjects, objects etc. In SRL, the task is to assign syntactic constituents called arguments with semantic roles of predicates (mostly verbs) at sentence level.

The relationship that a syntactic constituent has with a predicate is considered as a semantic role. For instance, for a given sentence, the SRL task consists of analyzing the propositions expressed by some target verbs of the sentence. Particularly, the task is to recognize the semantic role of all the constituents of a sentence for each target verb. Typical semantic arguments include Agent, Patient, Instrument, etc. and also adjuncts such as Locative, Temporal, Manner, Cause, etc.

The study of semantic roles was first proposed by the Indian grammarian Panini [2] in his “Karaka” theory. Karaka theory assigns generic semantic roles to words in a natural language sentence. The relationship of the arguments with the verb is described using relations called Karaka relations. Karaka relations describe the
way in which arguments participate in the action
described by the verb.

[18] describes a syntactic annotation scheme
for English based on Panini's concept of Karakas.
Subsequent work [6] revived Panini's Karaka
theory and developed state-of-the-art SRL system.

There are several lexical resources available
for SRL such as PropBank [13], FrameNet [1],
VerbNet [16]. In this paper, we discuss about an
annotation scheme for SRL on Social Media Texts,
particularly Twitter texts (also called tweet) which
are informal in nature. Twitter is a micro-blogging
site which allows users to post a message(tweet)
within the limitation of 140 characters. However,
in late 2017, Twitter announced to increase the
limit up to 280 characters. At the time when we
collected the tweets, we did not get tweets with
280 characters. Due to the restriction of maximum
140 characters, use of abbreviations, word play,
phonetic typing, emoticons are often found in
tweets. For example, in the tweet given below
Tweet(1):

— You made me ROFL ... pls dooooon't do it again

“ROFL” is the abbreviated form of rolling on the
floor laughing, “pls” for please and “dooooon’t” (for
don’t) is a word play. It is evident that tweets
are free form and may not contain grammatically
correct phrases. Therefore, performing SRL on such texts is a difficult task and state-of-the-art
SRL systems do not perform well on them. The
concept of “5W1H” (Who, What, When, Where,
Why, How) adopted in this paper, aims at
developing an annotation scheme for semantic role
labeling of English tweets. The major contributions
of our work are:

— Prepare a corpus for SRL on tweets.
— Propose a simple annotation scheme for SRL
  on tweets based on 5W1H concept.

The rest of the paper is organized as follows:
Section 2 discusses related work. In section
3, the corpus collection and annotation process
is discussed. Analysis of the annotation task is
discussed in section 4. Section 5 discusses the
ambiguities followed by the concluding remarks
and future work in section 6.

2 Related Work

We categorized the related work into two types:
first, we discuss about previous work on 5W1H and
second, SRL on tweets. Previous work on 5W1H
such as [20] describe a verb-driven approach
to extract 5W1H event semantic information from
Chinese online news. [5] present different
methodologies to extract semantic role labels of
Bengali nouns using 5W distillation process. They
used lexico-syntactic features such as POS and
morphological features such as root word, gender,
case and modality for identification of the 5W1H.
[3, 4] describe a 5W1H based visualization system
that facilitates users to generate sentiment tracking
with textual summary and sentiment polarity. [14]
describe a 5W1H based Cross-Lingual Machine
Translation system from Chinese to English. [21]
propose an algorithm named 5WTAG for detecting
microblog topics based on the model of five Ws.

[10] are the first to study SRL on tweets. They
considered only those tweets that reported
news events. They mapped predicate-argument
structures from news sentences to news tweets to
get training data, based on which a tweet specific
system is trained. Using hierarchical agglomerative
clustering algorithm [17], news excerpts were
divided into groups in terms of content similarity
and predicate argument structures by removing
all meta data from the tweets except the main
text. [11] is an extension of [10] where similar
tweets are grouped by clustering. Then for each
cluster a two-stage SRL labeling is conducted.
[12] describe a system for emotion detection from
tweets. Their work mainly focuses on identification
of roles for Experiencer, State and Stimulus of an
emotion.

1www.twitter.com

2https://en.wikipedia.org/wiki/Bengali_language
Table 1. Corpus statistics

<table>
<thead>
<tr>
<th>Type of tweet</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain tweets</td>
<td>48</td>
</tr>
<tr>
<td>Tweets with only @mentions</td>
<td>28</td>
</tr>
<tr>
<td>Tweets with only # tags</td>
<td>2297</td>
</tr>
<tr>
<td>Tweets with both @mentions and # tags</td>
<td>627</td>
</tr>
<tr>
<td>Total</td>
<td>3000</td>
</tr>
</tbody>
</table>

Our work presented in this paper, reports on the 5W1H based annotation process of English tweets for SRL task. Given a tweet, the objective is to identify the predicate \( p \) first and then extract the corresponding role arguments. The arguments of a predicate is the answer to the 5W1H entity <Who, What, When, Where, Why, How>. All the 5W1H may not be present in a tweet.

3 Data Collection and Annotation

3.1 Data Collection

For collection of tweets, we crawled Twitter data related to the US elections held in 2016 using Twitter API \(^3\). Hash tags such as #USElections, #USElections2016, #USElectionsUpdate, #ElectionNight, #HillaryClinton, #DonaldTrump, #hillary, #trump and #DonaldTrumpWins were used as query for fetching the tweets. Apart from hashtags, we also crawled tweets using terms such as “Donald Trump”, “Donald”, “Trump”, “Hillary Clinton”, “Hillary”, “Clinton”.

We crawled a total of 38,984 tweets which are further reduced to 24,679 tweets after manually removing the Non-English tweets as well as re-tweets (tweets with RT). We randomly sampled 3000 tweets and tokenized them with CMU tokenizer \(^7\).

We further manually segregated the 3000 tweets based on whether a tweet contains @user mentions or hash tags or both. The corpus distribution is shown in Table 1.

3.2 Annotation based on PropBank

We deployed five annotators for identifying predicates and PropBank role arguments from the tweets. The annotators are not linguists but well conversant in English. Our annotation task involve the following steps:

Step 1: Automatic Predicate Identification and Argument Prediction:

We use a SRL system \(^{15}\) for automatically identifying the predicates and labeling the semantic roles. Since the SRL system of \(^{15}\) is not designed for tweets, a high accuracy is not desired. Therefore, the output of such a system requires manual evaluation.

Step 2: Manual Argument and Predicate Identification:

Annotators are trained on PropBank role labels and asked to curate the output of Step 1. It took approximately three months (due to irregularity of annotators) to train the annotators to get acquainted with the PropBank argument role set. We call them the “Experienced Annotators (EA)”. In this step, we ask annotators to either accept or reject or correct the predicates identified and arguments predicted in step 1. Each predicate identified in step 1 is manually checked in the PropBank database for the correct arguments. On an average, it took 6 minutes to annotate one tweet. As an illustration, in the below given tweet Tweet(2):

Tweet(2):
— Hillary Clinton lost because of being Hillary Clinton!

the SRL system in step 1 predicted predicate lose as lose.02. As per PropBank, the predicate lose.02 means “no longer have” with arguments ARG0:entity losing something, ARG1:thing lost and ARG2:beneactive or entity gaining thing lost. This suggests that in the example tweet, “Hillary Clinton” is ARG0 and “because of being Hillary Clinton” is ARG-CAU. In this example, argument ARG1 is missing.

Step 3: Identify Missing Arguments and Predicates:

3http://twitter4j.org

ISSN 2007-9737
A 5W1H Based Annotation Scheme for Semantic Role Labeling of English Tweets 749
doi: 10.13053/CyS-22-3-3016

In this step, annotators are asked to identify the missing arguments and predicates. As an example, the SRL system in step 1 could not identify the predicate *provide.01* for the below given tweet **Tweet(3)**:

— *Trump’s Pyrrhic Victory Provides a BIG Silver Lining for Democrats*

As per PropBank, predicate *provide.01* means “to give” with arguments ARG0: *provider*, ARG1: *thing provided* and ARG2: *entity provided for (benefactive)*.

An annotation is considered “accept” only if all the five annotators agree on the annotation. The steps are illustrated in Fig.1. The annotation agreement is reported in Table 3.

### 3.3 5W1H Annotation

The concept of 5W1H (Who, What, When, Where, Why and How) was first introduced by [8] and widely used in journalism. In journalism, a news article or a story is considered to be complete and correct only when the 5W1H are present. The 5W1H provide the facts about a news article or a story being written such as:

— **Who?**: Who was involved?
— **What?**: What happened?
— **When?**: When did it happen?
— **Where?**: Where did it happen?
— **Why?**: Why did it happen?
— **How?**: How did it happen?

For 5W1H annotation, we adopted a Question and Answer (QA) based approach similar to [9] and [19] to extract the answers to the 5W1H questions. The following steps explain our approach.

**Step 1: Predicate Identification:**

The first task in the annotation process is to identify the predicates. For this task, we deployed three new annotators without any training on PropBank argument role set. We call them the **EA** Auto identify Predicate: *P* and Argument: *arg* with Roth’s system [15]

**Step 2: Semantic Role Identification with QA:**

We prepared QA pairs with the help of two Post Graduate Scholars in English Language. For every predicate identified in the previous step, QA pairs
3.4 Handling @user mentions

The major difference between a formal sentence and a sentence in a tweet is the presence of @user mentions. A username on Twitter is also known as "handle" which is considered as a user’s identity. Twitter usernames typically appear with an at sign (@) before the name. A username could be an individual’s name or name of an organization. In a tweet, one twitter user sometime may prefer to mention another twitter user’s name either to emphasize on an opinion expressed or for some other reasons. Twitter has a restriction on the length of usernames to 15 characters. The presence of @user mentions creates difficulty in identifying semantic role arguments. Let us consider the following tweet Tweet(4):

— @abc @xyz @pqr4 Y’all should chill. I wanted Hillary, too. But she lost. Move on ...

Tweet(4) has three @user mentions. For predicate chill.02, the 5W1H is extracted as; **Question:** Who should chill? **Answer:** Y’all. However, in this sentence, “Y’all” refers to the three usernames. So, for such cases, we adopted a simple approach of extending the span of the 5W1H to @user mentions. Therefore, in this case, the answer would be all the three usernames. But this approach is not uniform for every occurrence of @user mentions. Let us consider another example Tweet(5):

— Thousands across the USA protest Trump victory https://t.co/nsS5k4MoTV via @uvwxyz5

Tweet(5) is a news feed. The information delivered in this case is from external sources (https://t.co/nsS5k4MoTV and @uvwxyz). Moreover, the username “@uvwxyz”, is not an argument of the predicate protest.01. Therefore, in this case, @user mention is ignored.

---
4For privacy related issues, the @user mentions have been replaced with abc, xyz and pqr
5Actual @user mention has been replaced with uvwxyz
Table 2. Comparison of PropBank role set and 5W1H

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Predicate</th>
<th>PropBank Argument</th>
<th>5W1H Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Trump’s Pyrrhic Victory provides a BIG Silver Lining for Democrats</td>
<td>provide.01</td>
<td>A0: Trump’s Pyrrhic Victory</td>
<td>Who is the provider?</td>
<td>Trump’s Pyrrhic Victory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1: a BIG Silver Lining for Democrats</td>
<td>What is being provided?</td>
<td>a BIG Silver Lining for Democrats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2: for Democrats</td>
<td>Who is being provided?</td>
<td>for Democrats</td>
</tr>
<tr>
<td>(2) Watch President Obama Address Nation Following Trump’s Election Victory</td>
<td>watch.01</td>
<td>A0: not identified</td>
<td>Who is the watcher?</td>
<td>Viewers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1: President Obama Address Nation Following Trump’s Election Victory</td>
<td>What is being watched?</td>
<td>President Obama Address Nation Following Trump’s Election Victory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>When something is watched?</td>
<td>Following Trump’s Election Victory President Obama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A0: President Obama A1: Nation</td>
<td>Who is the addresser?</td>
<td>Not defined Nation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>What is the address about?</td>
<td>Nation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Who is being addressed?</td>
<td>Following Trump’s Election Victory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>When is it addressed?</td>
<td>Following Trump’s Election Victory</td>
</tr>
</tbody>
</table>

3.5 Handling hashtags (#)

A Twitter hash tag is simply a keyword phrase, spelled out without spaces, with a pound sign (#) in front of it. For example, #DonaldTrumpWins and #ILoveMusic are both hashtags. A Twitter hash tag ties the conversations of different users into one stream so that un-linked Twitter users could discuss on the same topic. Hash tags could occur anywhere in a tweet (beginning, in between words, end). In our corpus, we found 2297 tweets with hash tags. Handling hash tags is difficult when extracting 5W1H. Some hash tags are simple Named Entities such as #DonaldTrump, #HillaryClinton whereas, some are phrases such as #DonaldTrumpWins. The position and type of a hash tag is important while extracting the 5W1H.

An example explains our approach for handling hash tags Tweet(6):

— Will the GOP find a reason to impeach #Trump & usher in Pence ? #p2 #topprog

For the predicate impeach.01, 5W1H question is “Who is the impeacher?” “Who is being impeached?”. Here hash tag “#Trump” is the one being impeached. Therefore, we consider “#Trump” as the answer. The other two hash tags (#p2 and #topprog) do not play a significant role here. But this approach is not applicable for all the hash tags. The following example tweet explains the problem.

Also consider Tweet(7):
Table 3. Annotation agreement ratio of EA and IA annotators for identification of PropBank role set vs. 5W1H extraction

<table>
<thead>
<tr>
<th>Agreement of EA on PropBank task</th>
<th>#Tasks</th>
<th>#Correct</th>
<th>#Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>all 5 EA agree on answer</td>
<td>8375</td>
<td>6198</td>
<td>2177</td>
<td>0.74</td>
</tr>
<tr>
<td>4 out of 5 agree</td>
<td>2512</td>
<td>1733</td>
<td>779</td>
<td>0.69</td>
</tr>
<tr>
<td>3 out of 5 agree</td>
<td>1025</td>
<td>666</td>
<td>359</td>
<td>0.65</td>
</tr>
<tr>
<td>no agreement</td>
<td>52</td>
<td>0</td>
<td>52</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>11964</td>
<td>8597</td>
<td>3367</td>
<td>0.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agreement of IA on 5W1H task</th>
<th>#Tasks</th>
<th>#Correct</th>
<th>#Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>all 3 IA agree on answer</td>
<td>9368</td>
<td>8618</td>
<td>750</td>
<td>0.92</td>
</tr>
<tr>
<td>2 out of 3 agree</td>
<td>1405</td>
<td>1166</td>
<td>239</td>
<td>0.83</td>
</tr>
<tr>
<td>1 out of 3 agree</td>
<td>1172</td>
<td>833</td>
<td>339</td>
<td>0.71</td>
</tr>
<tr>
<td>no agreement</td>
<td>19</td>
<td>0</td>
<td>19</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>11964</td>
<td>10617</td>
<td>1347</td>
<td>0.89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agreement of EA on 5W1H task</th>
<th>#Tasks</th>
<th>#Correct</th>
<th>#Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>all 5 EA agree on answer</td>
<td>9368</td>
<td>8900</td>
<td>468</td>
<td>0.95</td>
</tr>
<tr>
<td>4 out of 5 agree</td>
<td>1502</td>
<td>1307</td>
<td>195</td>
<td>0.87</td>
</tr>
<tr>
<td>3 out of 5 agree</td>
<td>1087</td>
<td>848</td>
<td>239</td>
<td>0.78</td>
</tr>
<tr>
<td>no agreement</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>11964</td>
<td>11055</td>
<td>909</td>
<td>0.92</td>
</tr>
</tbody>
</table>

— #DonaldTrumpWins I think ppl r fed up of traditional way of politics and governance. They r expecting radical changes, aggressive leadership.

For phrase based hash tags, we simply segmented them into their semantic constituents. Therefore, #DonaldTrumpWins is expanded to Donald Trump wins. On expanding the hash tag, we get win.01 as the predicate with “Donald Trump” as the argument.

This further helps in finding the context for the argument of predicate think.01 and the answer to the 5W1H question of “Why one thinks?”.

4 Analysis

On performing the three sets of annotation tasks, we observe that the agreement on the correct answers increases when more annotators agree. From Table 3, we observe that the overall accuracy of EA is only 72% for PropBank role identification task, whereas it is 89% for IA for the 5W1H task. This suggests that even without prior training, IAs could easily identify the presence of 5W1H. A comparison of the IA and the EA shows that when all three IA agreed for an answer, they identified more tasks while extracting the 5W1H.

EA identified only 8375 tasks while identifying PropBank arguments. When all EA agreed for an answer, there is a significant increase in the Accuracy from 92% to 95% against the agreement when all IA agreed. Finally, we get an accuracy of 95% for EA which is a significant improvement over the previous annotation tasks. This suggests that our approach is comparatively easier to annotate with respect to ProbBank argument identification.

5 Discussion on Ambiguities

In this section, we discuss the ambiguous cases where it is difficult to come to an agreement. While curating our corpus, we observed that certain tweets which are direct news feeds, mostly do not explicitly mention the AGENT of the predicate. As an example, let us consider the following tweet Tweet(8):

— Watch President Obama Adress Nation Following Trump’s Election Victory

For the predicate watch.01, the AGENT or the ARG0 is explicitly not mentioned. However, there is an implicit AGENT or ARG0 present in the above tweet which semantically refers to the “viewers” or “readers” of the news feed. For such cases, it is
difficult to extract an answer to a 5W1H question and difficult to come to an agreement for the annotators. The absence of proper punctuation is also a great concern while annotating the tweets. Some tweets do not have proper punctuation for marking the boundary of an utterance.

For instance, Tweet(9):

— #DonaldTrump is a #racists liar & a #fascist do u really wanna vote for that America #USElections2016

In Tweet(9), there are two possible utterances, one being “#DonaldTrump is a #racists liar & a #fascist” and the other being “do u really wanna vote for that America #USElections2016”. There are two possible annotations, one without breaking the utterances and the other after breaking the utterances. In all such cases, we instructed the EA and IA to treat them as two utterances. Detecting the boundary of utterances is itself a difficult task and currently outside the scope of our work.

6 Conclusion and Future Work

In this paper, we described an annotation scheme to assign semantic roles on tweets by 5W1H extraction. Initially, we did not get satisfactory inter-annotator agreement for the PropBank predicate and argument identification task. The 5W1H based approach reported better annotator agreement without any expert level knowledge about the task as compared to the argument identification task based on PropBank. This suggests that our approach is simpler and convenient for identification of semantic roles. There is no single universal set of semantic roles that can be applied across all domains. The PropBank semantic role labels are too specific and complex in nature. Assigning such complex semantic role labels on tweets are ambiguous in certain cases. The simple and convenient approach for annotation discussed in the paper for SRL can be useful to some NLP application areas such as opinion mining, textual entailment and event detection. In the near future, we intend to incorporate a system for utterance boundary detection and evaluate how SRL could be done.

References


Article received on 10/01/2018; accepted on 05/03/2018. Corresponding author is Kunal Chakma.