

Part-Of-Speech tag Datasets for Patent Claims

Adaptive Signal Processing and Information Theory Research Group at Drexel University
aspitrg@gmail.com

March 27, 2016

1 Introduction

The set of datasets are collected by [Adaptive Signal Processing and Information Theory Research Group](#) at Drexel University. The datasets are based on the claims from 1491 patents that are randomly crawled by topics from [USPTO](#). The datasets provide the correct Part-of-Speech(POS) tags for the claims' words that are initially labeled as verbs by [Stanford NLP parser](#). The Release 1.0 provides 4 datasets that are based on 4 campaigns released on [Amazon Mechanical Turk platform](#). The prototype system using this dataset is described in [1].

2 Description

Each dataset from each campaign contains 4 files indicating the initial POS tags, the original text sentences, the answer from workers on AMT and the locations of the initially tagged verbs. The number of POS tags for the initially tagged verbs are listed in Table 1. Remark: the 4 datasets include the answers to the test questions in the AMT assignment.

Campaigns	Number of POS tags
Campaign 1	548
Campaign 2	7662
Campaign 3	45255
Campaign 4	124261

Table 1: Number of POS tags in each dataset: Answers to the test questions are included.

2.1 Campaign 1,2 and 3 Datasets

1. **InitPOS.txt**

As it is shown in Figure 1, all the sentences are labeled with Part-of-Speech tags by Stanford NLP parser. Each line contains one sentence. Within each sentence, the initially tagged verbs are bolded with the bold HTML tags which are used for identify the location of the initially labeled verbs. The sentences are repeated as many times as the number of initially tagged verbs within that sentence.

2. **OrigText.txt**

This file provides all the original sentences. Each line contains one sentence. The sentences within this file are corresponding to the sentences in the InitPOS.txt file. The example is shown in Figure 2 The sentences are repeated as many times as the number of initially tagged verbs within that sentence.

3. **TurkAnswer.txt**

This file contains the answers of the workers from Amazon Mechanical Turk as shown in Figure 3. The answers are only from the HITs that are approved by us. The answers provide the true label of the initially tagged verbs from workers. Each line contains one sentence. The

```

2 The/DT method/NN as/IN recited/JJ in/IN claim/NN 3/CD wherein/NNS
   <strong>said/VBD</strong> coefficient/JJ value/NN and/CC
   second/JJ coefficient/NN value/NN <strong>are/VBP</strong>
   different/JJ ./..
3 The/DT method/NN as/IN recited/JJ in/IN claim/NN 3/CD wherein/NNS
   <strong>said/VBD</strong> coefficient/JJ value/NN and/CC
   second/JJ coefficient/NN value/NN <strong>are/VBP</strong>
   different/JJ ./..

```

Figure 1: Campaign 1-3: InitPOS.txt file

```

2 The method as recited in claim 3 wherein said coefficient value
   and second coefficient value are different .
3 The method as recited in claim 3 wherein said coefficient value
   and second coefficient value are different .

```

Figure 2: Campaign 1-3: OrigText.txt file

answer-sentences are repeated as many times as the number of initially tagged verbs within that sentence.

4. VerbLocs.txt

The file provide the initially tagged verbs of each sentence shown in Figure4. The location of the initially tagged verbs are based on the term count. For example: Given the initially tagged sentence as follow shown in Figure 1:

There are 2 initially tagged verbs in this sentence. So you will find this sentence repeats 2 times(2 lines) in the files: **InitPOS.txt**, **OrigText.txt** and **TurkAnswer.txt**. The first initially tagged verb is the 9th word. The second initially tagged verb is at the 16th word. You will find 2 correspond lines in VerbLocs.txt where the first line is '9' and the second line is '16' indicating the location of the two initially tagged verbs.

2.2 Campaign 4 Dataset

The Campaign 4 dataset is different in InitPOS.txt and TurkAnswer.txt files due to a different AMT assignment design.

1. InitPOS.txt in Figure 5

Each line contains only one sentence. For each sentence, only the initially tagged verbs are bolded with the bold HTML tags. For other words in the sentence, the initial POS tags are not given. Sentences in the file are not repeated.

2. OrigText.txt

Each line contains only one sentence. The sentences within this file are corresponding to the sentences in the InitPOS.txt file. Sentences in the file are not repeated.

3. TurkAnswer.txt in Figure 6

Each line contains the true label of the initially tagged verb from the AMT worker. The lines are correspond to the sentences in InitPOS.txt.

4. VerbLocs.txt

Each line contains a number indicating the location of the initially tagged verb in the same line of InitPOS.txt.

2	The/DT method/NN as/IN recited/JJ in/IN claim/NN 3/CD wherein/NNS said/JJ(**) coefficient/JJ value/NN and/CC second/JJ coefficient/NN value/NN are/VBP different/JJ ./.
3	The/DT method/NN as/IN recited/JJ in/IN claim/NN 3/CD wherein/NNS said/JJ(**) coefficient/JJ value/NN and/CC second/JJ coefficient/NN value/NN are/VBP different/JJ ./.

Figure 3: Campaign 1-3: Answers from Amazon Mechanical Turk

2	9
3	16

Figure 4: Campaign 1-3: Locations of the initially tagged verbs

References

- [1] Mengke Hu, David Cinciruk, and John MacLaren Walsh. Improving Automated Patent Claim Parsing: Dataset, System, and Experiments. 2016.

3 The magnet valve as defined by claim 8 wherein the fixation means are **embodied** as a plastic injection-molded part .

Figure 5: Campaign 4: InitPOS.txt

3 **Verb**

Figure 6: Campaign 4: TurkAnswer