Finding ARG1 of Partitive Nouns in NomBank with DistilBERT

Ziyang Zeng
Dept. of Computer Science
New York University
251 Mercer Street, New York, NY
zz2960@nyu.edu

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

NomBank (Meyers et al., 2004c) is a databank project at New York University that adds annotation layer of argument structure for instances in the Penn Treebank II corpus. Apart from the more commonly researched nominalizations of verbs and nominalizations of adjectives, it also covers relational nouns, partitive nouns and several other types of argument-taking nouns.

More recently, the NLP community has seen huge popularity in heavily adopting pretraining-based langauge models. BERT (Devlin et al., 2018) is a large-size language representation model trained on a large corpus of English text that can achieve state-of-the-art results on a variety of natural language processing tasks. BERT came a decade later than the release of NomBank and there's been little previous work using it to conduct NomBank-based tasks and related experiments. However, as powerful as BERT is, its large parameters size makes it significantly more computationally expensive to train and use. On the other hand, DistilBERT is a distilled student model of BERT that retains 97% of BERT's language understanding capabilities while being 40% smaller and 60% faster. This makes it a great candidate to train when facing with limited resources.

This paper presents a token classification approach using pre-trained DistilBERT to find ARG1 of partitive nouns in NomBank. It focuses on partitive nouns (nouns that are used to describe a part or quantity of something) and the partitive task. The task is described as finding one argument (ARG1) of % (the percent sign), or in other words, finding the none group that is being sub-divided or quantified over. For example, for the sentence "Output in the energy sector rose 3.8%.", the ARG1 to be found is "Output" since it is the partitive noun that "3.8%" is referring to. This task can be translated into a binary classification problem on each token in the sentence where the model predicts whether the token is an ARG1 or not.

In addition to the token classification approach above, this paper also experiments with question-answer task, transforming the original ARG1 finding task into the task of make the model answer the question "What is the ARG1 of this sentence?". The advantage of this adaptation is that the model would stably output exactly one ARG1 for each sentence whereas token classification could give each sentence 0 to N ARG1s.

2 Related Work

NomBank (Meyers et al., 2004c), as a databank extending the frames of NOMLEX and PropBank, annotates argument structures of common nouns similar to how PropBank annotates predicating verbs. The development of the NomBank corpus has made several other work on argument structure extraction more accessible. Jiang and Ng (2006) attempts the first NomBank-based automatic semantic role labeling system after the databank's release. Ping (2006) adapts a PropBank-based SRL system to the SRL task of NomBank, achieving an overall F1 score of 72.73 on section 23 of the NomBank corpus.

BERT, or Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), has shown state-of-the-art results on many NLP tasks. There have been many attempts to compress BERT into a smaller model, among which DistilBERT (Sanh et al., 2019) successfully reduces the size of a BERT

model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster, leveraging knowledge distillation (Bucila et al., 2006, Hinton et al., 2015) during the pre-training phase.

More recent development on semantic role labeling (SRL) has shifted from feature engineering to architecture-based modeling leveraging deep neural networks (Collobert et al., 2011). Several recent notable approaches suggest only using raw tokenized text as input and let the model learn the features from the text itself (He et al., 2017; Sahin and Steedman, 2018; Marcheggiani et al., 2017; Strubell et al., 2018). This end-to-end approach allows the model to easily generalize from one SRL task to another without very task-specific feature engineering effort.

3 ARG1 Finding Pipelines

3.1 Maximum Entropy Baseline Pipeline

The baseline we use is a maximum entropy machine learning method in a token-by-token classification manner. Maximum entropy is a general technique for estimating probability distributions from data. Labeled training data is used to derive a set of constraints for the model that characterize the class-specific expectations for the distribution. To perform well, the model relies on the good representation of the features we feed it. In this paper, we select features including the word stem, the neighboring 2 words, POS tags, NG-BIO tags, whether capitalized and the position of the word in the sentence.

3.2 Plain Raw Input Pipeline

The simplest pipeline for finding ARG1 of partitive nouns in NomBank assumes that the input text is raw and unprocessed. The pipeline first tokenizes the text using the tokenizer provided by DistilBERT. Then, the model predicts whether each token is an ARG1 or not. The pipeline is illustrated in the following figure 1.

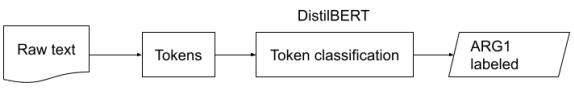

Figure 1: The simplest pipeline for finding ARG1 of partitive nouns in NomBank

3.3 Feature Enhanced Pipeline

NomBank provides a number of features for each token in the databank, including POS tags and NG-BIO tags. They could be helpful for the task of finding ARG1 of partitive nouns. The pipeline below incorporate these features to the raw text input in 3.2 to find ARG1 of partitive nouns. The pipeline is illustrated in the figure 2.

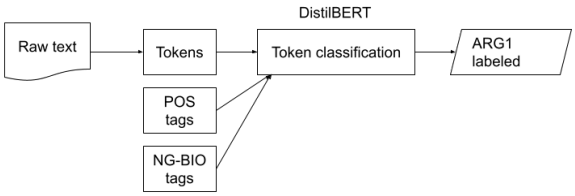

Figure 2: The feature-enhanced pipeline for finding ARG1 of partitive nouns in NomBank

Because the original pre-trained DistilBERT model does not take extra features including POS and NG-BIO tags as input, we have to modified the token classification model and alter the last linear layer, concatenating the original BERT hidden layer output with the extra features, then doing token classification through the linear layer.

3.4 Question-Answer Pipeline

Apart from token classification, we also explored question-answer approach which rephrase the original problem into a QA problem. Question-answering, just as the plain token classification approach, assumes tokenized input and no extra features. Additionally, question-answering tasks also requires a question prompt and answer location for each sample while training. The pipeline is illustrated in the figure 3.

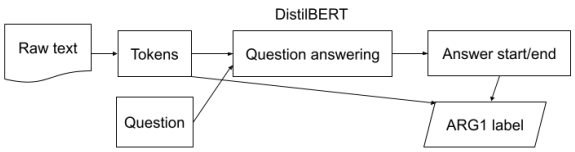

Figure 3: The question-answering pipeline for finding ARG1 of partitive nouns in NomBank

The output of the question-answering model is the location (start index and end index) of the predicted answer. By locating the token on the original input token list with the start index and end index, we can reconstruct the ARG1 label of the sentence.

4 Experiments

4.1 Dataset

The dataset we use in this task is part of the NomBank that specifically focus on partitive nouns. In the dataset is split into training, dev and test set. The training set has 2367 sentences or 66186 words, the dev set has 83 sentences or 2225 words, and the test set has 150 sentences or 4276 words. Each token in a sentence takes a line in the data file, the line containing the token itself, the POS tag, the NG-BIO tag, the index of the sentence and the index of the token in the sentence. The semantic role label is put at the end of each line, if the label is not one of "ARG1", "PRED" and "SUPPORT", then it's omitted. Figure 4 is one sample sentence from the training dataset.

```
But CC  O   0   0
about   IN  B-NP    1   0
25  CD  I-NP    2   0
%   NN  I-NP    3   0   PRED
of  IN  B-PP    4   0
the DT  B-NP    5   0
insiders    NNS I-NP    6   0   ARG1
COMMA   COMMA   O   7   0
according   VBG B-PP    8   0
to  TO  B-PP    9   0
SEC NNP B-NP    10  0
figures NNS I-NP    11  0
COMMA   COMMA   O   12  0
file    VBP B-VP    13  0
their   PRP$    B-NP    14  0
reports NNS I-NP    15  0
late    RB  B-ADVP  16  0
.   .   O   17  0
```

Figure 4: One sample sentence in the training dataset

In a sentence in the % dataset, the PRED labels % symbol, the ARG1 is the partitive noun that % symbol refers to, and the SUPPORT marks the word connecting the ARG1 and the PRED. It is guaranteed that there is one and only one ARG1 in each sentence of the dataset.

Acknowledgements

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

A Example Appendix

This is an appendix.