The Benefits of Crowdsourcing: Generating Better Estimates of Noun Phrases from Crowdsourced Data

Reno Kriz '16 and Professor Nancy Ide
Vassar College, 124 Raymond Avenue, Poughkeepsie, NY 12604

UNDERGRADUATE RESEARCH SUMMER INSTITUTE
Est. 1986 at Vassar College

Introduction

From the beginning of the computer age, humans have been fascinated with the concept of teaching a computer to communicate with humans. One of the most widely known examples of this is SIRI, a part of Apple's iPhone that attempts to answer questions posed by its human users.

In order for processes such as SIRI to continue to improve their performance, a computer has to be able to better recognize where parts of a sentence begin and end. Our project deals with generating more accurate labels of the boundaries of these sentence parts through crowdsourcing.

Figure 1. SIRI interacting with a user.

Background Information

Machine learning is a field that attempts to teach computers how to predict solutions to abstract problems. Natural language processing (NLP) is an application of machine learning that implements algorithms to help computers process and understand languages.

One way to help a computer learn is to have it associate words and sentences with certain labels, e.g. parts of speech. A computer can be trained with supervised data, a dataset of labeled sentences. Once this training is completed, the computer can be used to predict the labels of unsupervised data, sentences that are not labeled.

Previously, datasets were much smaller, because each word and sentence were manually annotated by a few people. Thus, to ensure the best accuracy, these people were highly trained, which was costly and time-consuming. Also, regardless of expertise, they still may have biases that lead them to incorrectly label certain phrases.

With today's Internet, crowdsourcing became an alternative to annotators for researchers. Crowdsourcing involves small contributions from many people to complete a task. In 2005, Amazon released Amazon Mechanical Turk (AMT), which allowed researchers to hire people around the world to perform various Human Intelligence Tasks.

Passonneau and Carpenter (2014) used AMT to generate annotations of word senses of several words from the Manually-Annotated Sub Corpus (MASC), a Vassar project through the American National Corpus (ANC). After determining the best labels from the crowdsourced data, the paper found that crowdsourced labels are more accurate and cheaper.

Methods

In our project, we extend the methods used by Passonneau and Carpenter. Specifically, we implement two algorithms to find the boundaries of noun chunks in the entire MASC corpus. A noun chunk is a noun, along with any adjectives and adverbs describing that noun, excluding any prepositions.

To find the boundaries, we use a BIO labeling system: B is the beginning of a noun chunk; I is the continuation of a noun chunk; and O is not part of a noun chunk. We also include the start and end values.

Token	Label	Start	End
Mr	B-NC	0	2
.	I-NC	2	3
Smith	I-NC	4	9
went	O	10	14
to	O	15	17
Washington	B-NC	18	28
.	O	28	29

Figure 2. In this example, we consider the sentence, "Mr. Smith went to Washington." We see that "Mr. Smith" and "Washington" are the noun chunks, while "went", "to", and "." are not noun chunks. Using the offset values, we know that "Mr. Smith" begins at 0 and ends at 9, while "Washington" starts at 8 and ends at 28.

To obtain our training data, we used several texts from MASC, which were automatically labeled using General Architecture for Text Engineering (GATE). To have variance in our training data, we generated annotations through three separate processes.

Passonneau and Carpenter (2014) used the Expectation Maximization (EM) algorithm to determine the maximum likelihood estimates of the labels, based on the crowdsourced annotations they obtained. In other words, the algorithm determines, using the annotations given, which label is most likely to be the gold-standard label, or the best prediction given the data. We also implemented a version of this algorithm, and applied it to our problem.

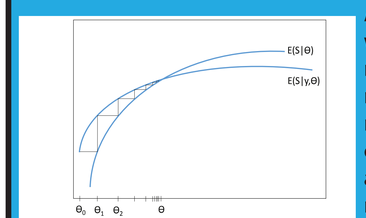

Figure 3. This figure is adapted from Navidi (1997); this shows the iterative process of the EM algorithm, in that each step is assured to be closer to a local maximum than the previous step.

As shown in Figure 3, the EM algorithm will always converge to a local maximum. However, that local maximum may not be the global maximum, and the EM algorithm cannot account for this. A different algorithm is the hierarchical full Bayes model. This model is more difficult to implement and work with; however, in theory it has a better chance of finding the global maximum, because it accounts for estimation uncertainty.

Future Work

Since the EM algorithm was implemented, the next step is to apply the Bayes algorithm. We will do this by using the Stan package in R. Once we have both implementations, we will run them on three automatically-generated noun chunk annotations of texts in MASC.

Then, we will compare the output of each implementation with the gold-standard noun chunk labels that are already included in MASC. Based on the percentage of noun chunks labeled correctly, we will choose the algorithm that is most accurate.

From there, we will finish collecting the crowdsourced annotations from the AMT, and run the chosen algorithm on these labels to determine the new gold-standard labels of the start and end of noun chunks in the sentences used from the ANC.

Figure 4. A cartoon of the capacity of a single expert, versus the capacity of many non-experts; this shows the potential power of crowdsourcing.

After successfully finding the gold-standard labels of the boundaries of noun chunks, we can move on to improving the current annotations of other parts of speech and word senses in MASC, such as verb chunks, prepositional phrases, among others.

Sources/Acknowledgements

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Navidi, W. (1997). A graphical illustration of the EM algorithm. American Statistician, 51(1), 29-31.

Passonneau, R., Carpenter,B. (2014). The benefits of a model of annotation. Transactions of the Association for Computational Linguistics, 2(0)

Yu, S., & Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden markov model. IEEE Signal Processing Letters, 10(1), 11-14.

We would like to thank Rebecca Passonneau, Bob Carpenter, and Ziheng Huang at Columbia University for allowing us to view and use parts of their source code for the Expectation Maximization algorithm; this project was funded and received support from the United States National Science Foundation.