Technology Review: Stanford Named Entity Recognition
October 27, 2022

Adam Michalsky
adamwm3@illinois.edu
CS 410 - Text Information Systems

In the digital-age, technology has been so embedded in everyday life that it is often difficult to know exactly what happens when one posts a tweet, clicks a link, or even searches for an answer to a question. For an end-user, submitting a question to a search engine is a straight-forward task, but knowing how many tasks that search engine completes to find an answer is the true marvel. It has been said that mankind has built computers in their own image. While this is a valid statement in some respects, the irony is that computers cannot speak human language. Computer communication was built for efficiency between computers and in that scenario, an understanding of human language is not required.

Today, data is stored in many formats that can be classified as structured, semi-structured, and unstructured data. A good rule of thumb is the more structure a data set has, the easier it is to catalog and subsequently search. It is the responsibility of a search engine to be able to catalog or index this information from all sources including unstructured sources such as text on a webpage. In order for a computer to understand text on any webpage, a computer or an application must be able to understand the rules of a language.

Natural language processing (NLP) refers to the branch of artificial intelligence concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. [1]. NLP is composed of several areas of research or tasks such as speech recognition, part-of-speech (POS) tagging, and name entity recognition (NER)[1] to name a few. NER is a task under the NLP umbrella that labels sequences of words in a text which are names of things[2]. These things include but are not limited to a person's name, company name, or even gene and protein names.

Similiar to any implementation of an NLP task, there is not a one size fits all approach. Some NER implementations are designed to perform tasks in different languages and some of them could be better at labeling a person's name compared to a name of a place. In order to measure the efficacy of models, named corpas have been created and made available to train NER models based on their precision, recall, and F1 measure.

These corpas consist of text from various sources such as the American and British newswires and are classified based on language support and entity diversity. For instance CoNLL-2002 and CoNLL-2003 are composed of sources on the British Newswire, and it has training data for German, Spanish, Dutch, and English with a focus on the person, location, organization, and miscellaneous entities. In contrast to the CoNLL corpora, MOC-6 and MOC-7 are based on the American newswire, and they have training data for English with a focus on person, location, organization, time, date, percent, and money entities[3].

In 2006, a group at Stanford University released their own implementation of a name entity recognizer trained on different compositions of CoNLL, MUC-6, MUC-7, and ACE named corpa [2]. The Stanford named entity recognizer (Stanford NER) has models based on a mixture of the aforementioned named corpora as well as models trained on a single named corpa. Since the initial release, language support has expanded from just English to include Chinese and Spanish modules. Some models support German as well but are somewhat dated. The diversity in Stanford's training approach has contributed to very robust models that can label person, organization, and locations entities with precision and recall as high as 93.28% and 92.71% respectively [4]. Additional details on the Stanford NER model performance by training set can be found in the table below.

Stanford Name Entity Recognizer Performance Measures

Corpus		**# Word Tokens**		**# Entities**		**Exact Match Score (conlleval)**		
Name	**Language**	**Train**	**Test**	**Types**	**Instances**	**Precision**	**Recall**	**F_1**
CoNLL 2003	English	219553	51578	4	5942	91.37%	91.22%	91.29%
CoNLL 2003	English	219554	51578	4	5942	92.15%	92.39%	92.27%
CoNLL 2003	English	219553	46666	4	5648	85.65%	85.41%	85.53%
CoNLL 2003	English	219554	46666	4	5648	86.12%	86.49%	86.31%
CoNLL 2003	English	219553	51578	4	5942	91.64%	90.93%	91.28%
CoNLL 2003	English	219553	51578	4	5942	93.28%	92.71%	92.99%
CoNLL 2003	English	219553	46666	4	5648	88.21%	87.68%	87.94%

The table above is a subset of results for different implementations of NER and more performance results on different languages can be found on their website [4].

Stanford's approach to NER attempts to blend the approach of using maximum entropy Markov models (MEMM) or conditional Markov models (CMM) with hidden Markov models (HMM). This blend of approaches has led to a solution that contains the best features of MEMM and HMM based approaches. The approach, conditional random field (CRF), is discriminative and does not assume features or words are independent when labeling [3]. In other words, CRF based models are able to consider the context of words better than models based on MEMM, CMM, or HMM based models.

Between the initial release and today others have made contributions beyond just improving the feature and functionality of the tool. Stanford NER has features such as being able to run in server mode and documented APIs so that it can be used outside the Java environment with popular tools in the data science community. These integrations have been developed for Python, Perl, Ruby, .NET, and C# [2]. A full list of integrations can be found on Stanford NER's documentation.

The integrations for Stanford's NER are key to making it accessible to the masses. It can be used in a variety of ways, including web applications via JavaScript or in a data science writeup utilizing a Jupyter notebook with Python. Stanford's NER is a great choice for those just getting started with NLP and are looking for a robust tool that performs the name entity recognition task. There are variants of models trained on specific datasets like CoNLL-2003 [2], and variants of models that leverage a pure CMM based approach in lieu of the CRF approach [4].

This diversity and active support has made the Stanford NER a great choice for experienced and new NLP researchers alike. It is amazing that something as simple as asking a question to a browser can be so involved and complex. NER is just one computing technique of many that occurs when someone uses a search engine.. The abstraction of advanced statistical modeling enables larger numbers of people to start their own research into NLP by making it simple to work with. One could argue that it makes it so simple that it is almost as simple as answering a question on Google, but not quite

References:

1. (2020-07-02) - IBM Cloud Education - *What is Natural Language Processing?* https://www.ibm.com/cloud/learn/natural-language-processing
2. (2020) - Stanford NLP Group - *Software* > *Stanford Named Entity Recognizer (NER)* https://nlp.stanford.edu/software/CRF-NER.shtml
3. (2007-03-09) - Jenny Finkel - *Named Entity Recognition and Stanford NER Software* https://downloads.cs.stanford.edu/nlp/software/jenny-ner-2007.pdf
4. (2009) Stanford NLP Group - *Stanford NLP Named Entity Recoginition Results* *https://nlp.stanford.edu/projects/project-ner.shtml*