**Introduction**

Named entity recognition is the process of finding and determining the types of named entities in a text for example people, locations and organizations and it is an important step in computers understanding written text. When named entity recognition works the computer can determine that the White House is the official residence of the president of the United States instead of a just a regular house that has been painted white, an important distinction to make if one wants to understand a text. However, to truly understand the importance of named entity recognition one needs to understand its place in the overarching field of statistical natural language programing which has a goal of allowing computers to understand all types of speech in a meaningful way. Furthermore, inside of the field of natural language programing named entity recognition falls into a category called information extraction which is in general trying to understand the underlying meaning of a text.

The term named entity recognition comes from the Sixth Message Understanding Conference when researchers were determining the sub tasks involved in information extraction from unstructured text (R. Grishman & Sundheim 1996). The task was separated out from general information extraction due to its difficulty and importance. In order that the task be completed two different approaches have been created, the traditional rule based approach, meaning that the type of an entity is determined based off of a set of hardcoded rules and the classification approach that is based off of machine learning after training data has been used to teach the machine. However, the results discussed in this paper have been found by using a combination of both approaches.

This paper uses both approaches in order to get the best of both and to decrease the chances of errors occurring. The rule-based approach is the now falling out of favor because of the increased capabilities of machine learning however it still has many strengths such as its lack of ambiguity. Rule based systems are very good at categorizing Named Entities that fall easily into predefined categories but do less well with entities that appear to fit in few different categories. Machine learning on the other hand handles the more ambiguous entities well and while it also does a good job with the less ambiguous, which is why it is rising in popularity, it sometimes makes a mistake that a rule based system would not. Furthermore, gazetteers will be used to make a first pass at incoming data for a few different reasons. First of all, using a gazetteer is not difficult and can add a lot to named entity recognition especially for things like cities and more prominent organizations. Secondly, even if a gazetteer finds that an entity can be found on multiple lists, meaning that it could for example be both a location and an organization that at least tells one that it is probably not a persons name. Lastly, one loses nothing by using a gazetteer since even if it cannot find an entity on any of its lists one is exactly where one would have been if it had not been used with nothing lost.

For the purpose of this paper the different categories that the named entities will be sorted into are person, organization, location and misc. Other papers have used more such art or event or product but the over specification involved in that can also lead to problems and a much larger amount of ambiguity. Also the more specific categories are not universally agreed upon both what exactly should fall into them but also whether or not they should be used and so to make this paper more generally applicable only those three will be used.

**Literature Survey**

There have been many papers written on the topic of named entity recognition since the term was invented but in this section three will be surveyed for their contributions to the field by this paper to better understand the work done by researchers in the past. The first paper to be examined is a Nadeau and Sekine (2007), which is an overview of the research done in the field. This is an important paper to survey due to its examination of the past research done in the field of named entity recognition over a fifteen-year period. It explores how the field has changed from one based solely off of rule based systems to the increasing use of machine learning today showing how our research which uses a combination of the two is bringing past and current methods together to increase accuracy.

The second paper to be examined is Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections (Hooland, S. Van, M. De Wilde, R. Verborgh, T. Steiner, and R. Van De Walle 2013), which explores the use of, named entity recognitions and term extraction in the field of digital humanities. The end result of the paper was that both forms of recognitions are useful and important however, one must be careful when using it to get meaningful results. Similar to our paper Hooland and his co-authors used a combination approach focusing on both qualitative and quantitative methods. However, this paper focuses on a much more specific type of text while ours is more general in its uses.

The last paper is on using multiple different named entity recognition systems together to create a better system (Illig, Jens, Charles La Rosse, Iliana Simova, Gu ̈nter Neumann, and Bogdan Sacaleanu 2011). The systems work together through a voting system that tries to cover the individual weaknesses of the systems and therefore have fewer errors. This paper also relates to our combination approach since like them we are using multiple approaches to try and cover for individual weaknesses. In conclusion this paper is built upon the work of the many researchers that have come before it.

**References**

Hooland, S. Van, M. De Wilde, R. Verborgh, T. Steiner, and R. Van De Walle. "Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections." *Digital Scholarship in the Humanities* 30.2 (2013): 262-79. Web. 15 Dec. 2016.

Illig, Jens, Charles La Rosse, Iliana Simova, Gu ̈nter Neumann, and Bogdan Sacaleanu. "A Platform for Named Entity Processing." (2011): n. pag. Print.

Nadeau, David, and Satoshi Sekine. "A Survey of Named Entity Recognition and Classification." *David Nadeau and Satoshi Sekine [li.30.1.03nad]*. National Research Council Canada, 10 Aug. 2007. Web. 15 Dec. 2016.

Grishman, Ralph; Sundheim, B. 1996. Message Understanding Conference - 6: A Brief History. In *Proc. International Conference on Computational Linguistics*.