Association for Information Systems
AIS Electronic Library (AISeL)

CONF-IRM 2014 Proceedings

International Conference on Information Resources Management (CONF-IRM)

2014

An Algorithm to Extract Jamaican Geographic Locations from News Articles – Using NLP Techniques

Jean-Mark Wright
The University of the West Indies, Mona Campus, Jamaica, jeanmark.wright@gmail.com

Gunjan Mansingh
The University of the West Indies, Mona Campus, Jamaica, gunjan.mansingh@uwimona.edu.jm

5P. An Algorithm to Extract Jamaican Geographic Locations from News Articles – Using NLP Techniques

Jean-Mark Wright
The University of the West Indies, Mona Campus, Jamaica
jeanmark.wright@gmail.com

Gunjan Mansingh
The University of the West Indies, Mona Campus, Jamaica
gunjan.mansingh@uwimona.edu.jm

Abstract
Natural Language Processing (NLP) has long been used to extract information from large bodies of text. NLP is often used to intelligently parse large volumes of data where the manual alternative may be infeasible. Named Entity Recognition (NER) is used to extract named entities such as people, places or organizations from text written in natural language. Using NER, NLP algorithms can be created to extract the mentions of geographic locations of different types from current and archived news articles. This information can be used to add a spatial window into previously flat datasets, allowing users to access information by filtering location information. Information that is derived can be used to support intelligent decision making and influence expert systems. This paper describes the development of an algorithm that uses the principles of both NLP and NER to extract references to geographic locations within news articles. The algorithm has been developed using the NLTK and Pattern Web Toolkit for Python and performs with a precision and accuracy above eighty (80) percent.

Keywords
Natural Language Programming, Named Entity Recognition, NLTK, Gazetteer

1. Introduction
Newspaper articles often contain several references to names of places. If these references can be mapped and displayed on a map, a spatial perspective into the data is created. This perspective allows the user to better appreciate the scope and concentration of the references; the user is able to clearly see where the references are, the distance between each, and also visualize concentration or clusters of the references. Using NLP technologies, algorithms can examine the sentences containing these locations to extract themes of interest and begin to track trends.

Several sectors can benefit from the information a spatial perspective brings. Security personnel can adopt this spatial perspective to understand the concentration of crime and derive relationships between clusters of incidents. Similarly, accidents at locations can be mapped to aid in the identification of crash hot spots and tracking of the frequency. Any sector can track a particular theme of interest by using a spatial perspective to visualize the distribution and concentration of references to place names related to that theme.

In order to identify references to place names, we need to adapt techniques that will help us to search for these references. References are often very ambiguous and need to be verified; names of persons may resemble names of places or there may be clashes of location names within different geographic regions. Natural Language Processing (NLP) helps us to understand what is being said in a body of text and also provides techniques for disambiguation. The aim of this research is to use Natural Language Processing (NLP) techniques to develop an algorithm that extracts the names of Jamaican geographic features from news articles. Once these references are extracted, they can be plotted on a map.

2. Background Literature

2.1 Natural Language Processing

NLP is the processing of human language to understand or make derivations about what is being said. NLP has long been used to extract information from large bodies of text (Berger, Pietra & Pietra 1996). NLP can eliminate the need for manual perusing of such sources (Pouliquen et al. 2006), which quickly becomes infeasible as data grows to volumes of gigabytes and terabytes. When algorithms are able to understand written text, they can be used to process very large volumes of data and make derivations. Many NLP systems are able to perform unsupervised parsing of natural language, i.e. the system is able to parse natural language without human aid.

NLP systems generally use training sets to refine the system's ability to parse natural language. Using training sets, computers can be "taught" and can arguably be more consistent, focused and unbiased than their human counterparts. NLP has also been used to perform document level sentiment summarization (Godbole, Srinivasaiah & Skiena 2007), automatic encoding of documents (Friedman, Shagina, Lussier & Hripcsak 2004) and also producing more granular derivations based on sentences or phrases written in natural language. Such algorithms have been used to power commercial grade systems with very high levels of accuracy and precision (Yi, Nasukawa, Bunescu & Niblack 2003). Information that is derived can be used confidently to support intelligent decision making.

Named Entity Recognition (NER), a branch of NLP is used to identify named entities mentioned in a body of text. In general, the three entity types recognized are names of persons, organizations and geographic locations. When attempting to recognize locations, a gazetteer is often used as part of the training set to aid the system's recognition. A gazetteer is a dictionary of geographic locations where each location has a coordinate attached to it. NER can also be used for text extraction, data mining, document indexing and has also been used in other sciences such as Genetics and Biomedical Sciences (Bast 2011). The evidence based on literature reviewed presents a very strong case for the relevance and impact of NLP technologies.

2.2 NLP Systems

We examined literature to determine how NLP and NER technologies have been used by various systems. The Perseus Digital Library (Smith & Crane 2001) system, a vast Historical Digital Library contains geocoded locations i.e. locations for which geographic coordinates were obtained. This system relied on the use of a gazetteer, linguistic information and bibliographic

information. Nominator (Ravin & Wacholder 1997), a system developed at IBM was built to recognize and classify names of organizations and persons. Nominator used a highly statistical model whose heuristics were based on assumptions of how names and organizations are typically used in the English language. In Geocoding Multilingual Texts (Pouliquen et al. 2006), a gazetteer was used for identification of named entities. Consequently, weightings, geoparsing and the distance between references were also considered to aid the disambiguation process. Yi et al. (2003), in their work on sentiment analysis demonstrated how direct methods of derivation could be used to extract meaning from human written reviews about cameras. We also reviewed literature on combined classifiers where separate systems were combined to perform NER and make advanced derivations from the context (Florian, Ittycheriah, Jing & Zhang 2003). Finally, we considered a paper where very little emphasis was placed on the use of gazetteers (Mikheev, Moens & Grover 1999). This paper explored the role, importance and impact of gazetteers.

The literature revealed that statistical (indirect), direct methods and hybrid combinations have all resulted in very high rates of accuracy and precision. IBM's Nominator (Ravin & Wacholder 1997), a model based mainly on statistics and aggregation obtained scores of precision and accuracy above eighty (80) percent. The research reviewed in the area of Sentiment Extraction also showed performance on par with their indirect counterparts (Yi et al. 2003).

As it pertains to training sets and the use of gazetteers; these were identified as being particularly useful for identification of locations, and less critical for identifying names (Smith & Crane 2001). This holds true despite it being proved that NLP systems can function without heavy reliance on a gazetteer (Mikheev et al. 1999). It was generally observed that most researches used some level of heuristics from the English Language to supplement research techniques. For example, the literature reviewed on Classifier Combination (Florian et al. 2003) employed extensive use of grammar rules. Finally, another technique employed was to use the output of an external, well-trained NER system to aid in the process of disambiguation.

3. Implementation

In preparation to extract Jamaican locations from news articles, we used the steps outlined below.

- Build a gazetteer of Jamaican locations. The literature reviewed (Pouliquen et al. 2006) supported and encouraged the use of gazetteers for the extraction of named entities for geographic locations. This technique has been widely used and also helped to confine the scope of the research, i.e. only locations in the gazetteer would be extracted and used to test the system, since no gazetteer exists that can confirm whether a location is indeed a valid location in Jamaica. A local corpus is needed so that oddities and differences for the Jamaican context can be properly understood and integrated into the algorithm for recognizing places. The gazetteer was populated with Jamaica's fourteen (14) parishes and five hundred (500) communities.
- Build a repository of downloaded articles consisting of manually confirmed references to locations in Jamaica. These references were used to build a set of articles for training the system and a separate set for evaluating the system. The literature reviewed underlined the importance of training the system properly so that context rules may be derived and hence

increase the relevance within the domain of text it will process. In particular David Smith (Smith 2001) notes that proper training is necessary for the derivation of context rules. A total of 4839 articles were downloaded from the Gleaner website. These articles were saved in a database (Step 2, **Figure 1**) where the articles and their references could be easily searched and referenced. After the articles were downloaded using a Python script, each of the articles were manually geo-tagged (Step 3, **Figure 1**). For the training set 562 references were manually confirmed (310 articles). The Pattern toolkit was used to download articles from the Jamaica Gleaner website (http://www.jamaica-gleaner.com). Consequently, the references were separated into training and evaluation sets (see steps 4 and 5 **Figure 1**).

- Create heuristics to disambiguate locations. After the gazetteer is used to find potential references, the heuristics must be applied to validate the reference. A mere search for named entities is not sufficient to allow a system to score high marks for precision and accuracy. Heuristics are critical for boosting the precision and accuracy of the system. Of the literature reviewed, all employed direct or indirect techniques were employed for disambiguation.

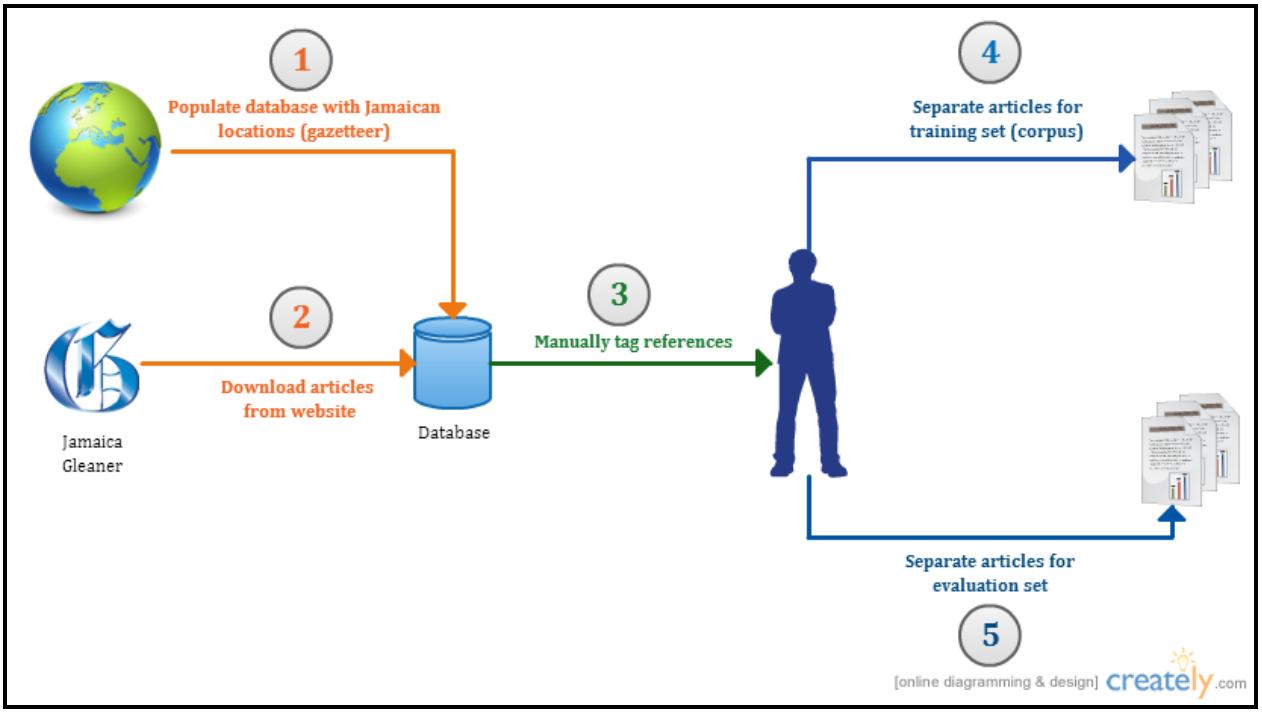

Figure 1: Building a training set

3.1 Tackling ambiguity

In this research, we were able to define several individual heuristics that played different roles in the process of disambiguation. Below are the set of heuristics used to perform the process of disambiguation that would be executed after identifying locations.

- NLTK Named Entity Tagger - The NLTK toolkit is an open source Natural Language toolkit that provides among other functionalities, a rigid named entity tagger. The named entity tagger is able to tag named entities that are found in a sentence. The use of a trained NER system has proved to be useful in disambiguation (Florian et al. 2003).

- Consideration of English grammar rules and surrounding words to better determine if a word is indeed a location (Florian et al. 2003).
- The distance of references relative to each other. Smith et al. (2001) utilized geographic distances between references in their process of disambiguation (Smith & Crane 2001).
- Use of weightings applied to locations based on their administrative boundary level (parish or community). Smith et al. (2001) assigned weights by their boundary level.

In order to make derivations from the training set, the references in the training set were analyzed to obtain statistics and aid in the decision for allotment of points. The following statistics were extracted from the 310 articles in the training set:

Statistics of the training set	
Average references per article	1.81
Maximum references in an article	19
Number of references tagged by NLTK as being named entities	495 (98%)
Number of references with parent boundary region mentioned in the same article	92 (16.37%)
Average standard deviation of references	1.23 km
Average number of articles not surrounded by proper nouns	483 (86%)
Number of locations clashing with a common English word	2 (0.36%)
Most common words preceding a location	the, and, of, downtown, from, west, western, a, are, for, include, to, at, comprises, covers, 's, frequent, near, new, avoid, frequents, including, several, touch, windalco, writer, affects, alleged, along, association, avenue, bartons, bay, between, close, damaged, hit, into, its, malvern, on, popular, portmore, prominent, reconstruct, runaway, s, says, seek, sent, some, towards, town, two, way

Table 1: Statistics of the training set

3.2 The Algorithm

An algorithm was developed in Python using the aid of the Natural Language toolkit (NLTK) for Python and the Pattern toolkit. NLTK is a well-documented module that houses many techniques for doing Natural Language analysis. The Pattern toolkit for Python (Smedt & Daelemans 2012) was developed at MIT and contains functionality for crawling web pages, parsing results from search engines, HTML parsing and text analysis.

The algorithm was built with the ability to recognize as many references as exist in a given article. The algorithm (illustrated below in Figure 2) shows the individual heuristics we have designed. The algorithm starts by using a case-insensitive search to identify references in a given

article for all locations in the gazetteer. This initial search is naive and is used to build a list of potential references.

Each reference that was previously identified using the naïve search was either marked as being a valid or invalid reference. In manually tagging references, the following decisions were made to control the scope of references that would be tested.

- Only title or upper cased references would be deemed valid references. Given that the domain is journalism, we expect proper use of the English language and respect of the rule that proper nouns begin with an upper case letter (title cased).
- A proper noun followed by other non-reference proper nouns is deemed as a reference to a place rather than to a parish or community. For example, in the sentence "The Kingston Police Division is on the attack!", "Kingston Police Division" is considered a reference to a place or organization, and not a reference to Kingston, the parish. However, in the sentence "The Kingston police are on the attack!", "Kinston" is marked as a reference to Kingston.

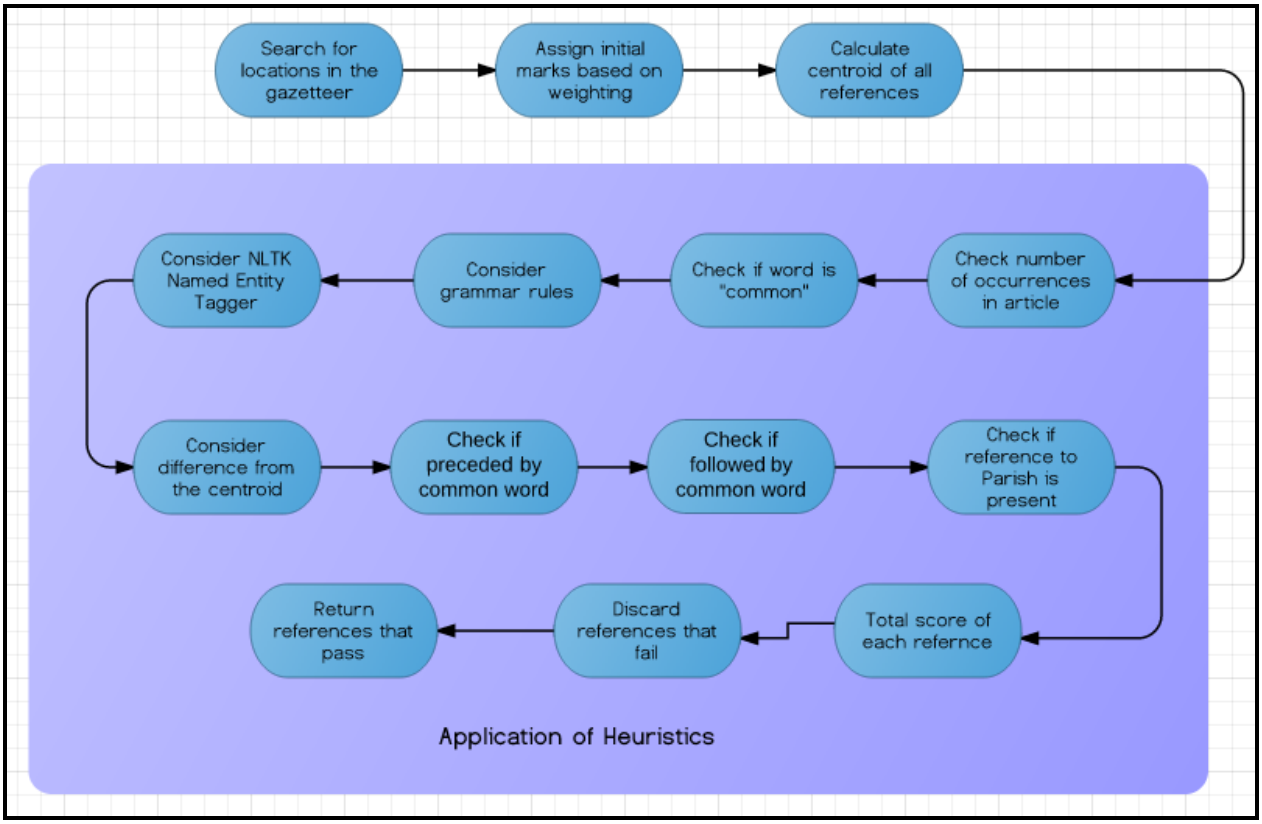

Figure 2: Algorithm Overview

The steps that follow are used to refine the choices before the algorithm returns its list of references. After the potential references are identified, the centroid of all the references are calculated and stored. The "Application of Heuristics" (see Figure 2) represents several tests that each reference would undergo. Each test has a mark that is assigned if the reference passes the test. The following tests were designed for use in evaluation of grammar rules test:

Description of Tests	
Number of occurrences in the same article (MO)	If the reference in mentioned multiple times in the same article a mark is assigned.
Common Word Test (CW)	A negative weighting is assigned to those references to a community whose name appears very commonly in the English language. One such example is the community Industry, in Portland. A list of 5000 common English words was used to determine whether the name of the reference was common. This file was read in at the start of the algorithm and used to compare each reference.
Consider grammar rules	General information known about how place names tend to appear in written text is considered. Primarily, references to parishes or communities in news articles are expected to be title cased or upper cased (this was observed in a surprisingly large number of articles). References to such locations are also not to be followed by other title cased words as this generally indicates reference to a name of organization or other place. For example, “May Pen Cemetery” is understood to be a reference to a cemetery and not to May Pen, the community. This caution is necessary since the existence of a parish or community reference in a name does not necessarily indicate that the place being referenced is in the parish or community. The May Pen Cemetery is actually located in Kingston, Jamaica as opposed to May Pen, Clarendon. Not Lower Cased Test (NLC) This test gives a positive score to references whose name is not lower cased. The majority of valid references are either title cased or upper cased. No Near Pronouns Test (NNP) This test is used to examine the context in which the reference is used with special consideration of the proximity of other pronouns or title cased words. Specifically, it considers the following: If a reference is surrounded by title cased words that are not references, they are regarded as invalid. For example, the sentence “The shooting could be heard for miles outside the May Pen Cemetery” has a reference to May Pen but is suffixed by the title-cased word “Cemetery” which is not a reference. Therefore, the system ignores this reference as it will be considered as a reference to a place and not the community May Pen. Similarly, the sentence “The police have been turning up the heat in Kingston and St. Andrew.” has valid references to Kingston and St. Andrew. However, the sentence “The Kingston Parish Council need the support of the police” is marked as having no valid references since “Parish Council” follows Kingston. Possessives are assessed to determine to prevent matching incomplete references. In the sentence “… many events are now happening in St. Ann’s Bay”, a reference to parish St. Ann would be incorrect since St. Ann’s Bay is the name of a community. The surrounding words often aid in understanding the context in which

	the word is used. In the sentence the "The University of Technology is now accepting submissions", a reference to the community University, St. Andrew would be incorrect even though the word "University" appears as part of the name of the organization.
NLTK Named Entity Tag (NNE)	The Python NLTK toolkit has a named entity tagger that has been properly trained to recognize named entities. Each reference is passed through the NTLK name entity tagger. If NLTK tags the reference as being a named entity, the reference passes the test.
Within One Standard Deviation (WOS)	When references are found, its centroid is calculated. Then the centroid of all the references (article centroid) and standard deviation from the centroid is calculated. All references within one standard deviation of the article centroid pass this test.
Preceded by a common word (PBCW)	A reference passes this test if it is prefixed by a word that is in the top thirty (30) words that commonly precede valid references.
Followed by common word (FBCW)	A reference passes this test if it is followed by a word that is in the top thirty (30) words that commonly follow valid references.
Parish Present (PP)	A community reference passes this test if a reference is made to the community's parish.

Table 2: Description of tests

Our system keeps a tally of all the marks obtained for each test then compares the total score to the system pass mark. If the reference's total points fall below the pass mark, the reference is discarded. The references that remain are returned as the references found by the algorithm.

4. Evaluation

For evaluation, the articles previously downloaded for the evaluation set were used. The algorithm was evaluated by using it to recognize and classify 100 references of geographic names. The desired result was to achieve an F-measure above 0.8. Since these articles were previously manually tagged, the list of locations returned from the system can be compared with the list that was previously confirmed. By doing this comparison, our system is able to calculate its precision and accuracy.

5. Results

After the algorithm was created, different weightings for each heuristics were tested in an effort to achieve a precision and accuracy above 0.8. The evaluation set was used for testing the system. The evaluation set contained a total of 100 articles containing 189 valid references. As discussed earlier, parishes were given an initial weight of 20, while communities were given a weight of 10.

The configuration of the dataset that achieved an F-measure above 0.8 is presented below in Table 3.

The pass mark used for each reference for this test was 35 marks. All references achieving 35 marks or higher would be submitted as a valid reference for the article being considered. The parishes were given an initial weighting of 20 marks. Therefore, a parish reference would only

need to pass the No Near Proper Nouns (NNP) test to be marked as a valid reference. If the parish test failed the NNP test, it is difficult for the parish reference to pass. A parish reference that fails the NNP test is also likely to fail the NLTK NE Tag Weighting (NNE) test. A community was started with an initial weighting of 10 marks. A community reference passing the NNP and NLC test is sufficient to place the reference at the pass mark. If the community reference fails the NNP test it is significantly harder for the reference to score above the pass mark. The observation was made that only 16.37% of community references in the training set had a reference to their parish in the article, therefore suggesting a fairly low probability that a reference will also gain marks based on the appearance of the parish. Additionally, failing the NNP test also predisposes the reference to failing NNE test. In the training set, it was observed that 85% of references that passed the NNP test also passed the NNE test. This test resulted in a precision of 0.84 and accuracy of 0.81, giving an F-measure of 0.82.

Evaluation dataset configuration	
Initial Parish Weighting	20
Initial Community Weighting	10
Parish Present Weighting (PP)	15
Is Common Weighting (CW)	-10
Not Lower Cased (NLC)	5
No Near Proper Nouns (NNP)	20
Within One Standard Deviation (WOS)	10
NLTK NE Tag Weighting (NNE)	10

Table 3: Evaluation dataset configuration

6. Conclusion

Using the heuristics developed in Python along with the aid of the Python NLTK and Pattern Web Toolkit, an algorithm was successfully developed that extracts Jamaican locations from newspaper articles with a precision of 0.84 and an accuracy of 0.81 (F-measure of 0.82). The major influencing factor in the system's accuracy and precision was the No Near Pronouns (NNP) test. Once a reference had been identified from the gazetteer, close consideration of the context was most important in disambiguating each reference. Derivations from the training set showed that the Parish Present (PP), Not Lower Cased (NLC) and NLTK NE Tag (NNE) tests all had relatively high rates of false positives. These tests (with the exception of NNE) did not consider the context or surrounding words and thus were not able to distinguish between a geographic location of interest (parish or community) and the name of a place. The NNP test however, was found to be the most informative of the list of heuristics, due to its consideration of the surrounding words.

The use of a gazetteer was also central to the research. The gazetteer allowed us to both confine the scope of the references and also start the process of identification of potential references. The combination of direct and indirect methods taken in this research proved to be successful and more than adequate to achieve the desired accuracy and precision of 0.8.

References

Bast, H. (2011). Efficient Natural Language Processing, Retrieved 24th March, 2013, from http://ad-wiki.informatik.uni-freiburg.de/teaching/EfficientNlpWS1112

Berger, A. L., V. J. D. Pietra, and S. a. D. Pietra (1996) "A maximum entropy approach to natural language processing", *Computational Linguistics*, (22)1, pp. 39-71.

Florian, R., A. Ittycheriah, H. Jing, and T. Zhang (2003) "Named entity recognition through classifier combination" in the proceedings of the *Seventh Conference on Natural language learning*, Edmonton, Canada, pp. 168-171.

Friedman, C., L. Shagina, Y. Lussier, and G. Hripcsak (2004) "Automated Encoding of Clinical Documents Based on Natural Language Processing", *Journal of the American Medical Informatics Association*, (11)5, pp. 392-402.

Godbole, N., M. Srinivasaiah, and S. Skiena (2007) "Large-Scale Sentiment Analysis for News and Blogs" in the proceedings of the *International conference of Weblogs and Social Media (ICWSM)*, Boulder, Colorado, USA.

Mikheev, A., M. Moens, and C. Grover (1999) "Named Entity recognition without gazetteers" in the proceedings of the *Ninth conference on European chapter of the Association for Computational Linguistics*, Bergen, Norway, pp. 1-8.

Pouliquen, B., M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, et al. (2006) "Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation" in the proceedings of the *5th International Conference on Language Resources and Evaluation (LREC' 2006)*, 24-26 May, Genoa, Italy, pp. 53-58.

Ravin, Y., and N. Wacholder (1997) "Extracting Names from Natural-Language Text", *Technical Report RC-20338*, Yorktown Heights, NY, USA: IBM Research Division, T. J. Watson.

Smedt, T. D., and W. Daelemans (2012) "Pattern for Python", *Journal of Machine Learning Research*, (13), pp. 2013-2035.

Smith, D. A., and G. Crane (2001) "Disambiguating Geographic Names in a Historical Digital Library". in P. Constantopoulos & I. T. Sølvberg (eds.), *Research and Advanced Technology for Digital Libraries, LNCS 2163*. Berline Heidelberg: Springer-Verlag, pp. 127-136

Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack (2003) "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques" in the proceedings of the *Third IEEE International Conference on Data Mining (ICDM)*, Washington D. C., USA, 19-22 November, pp. 427-434.