$SL - FII$: Syntactic and Lexical Constraints with Frequency based Iterative Improvement for Disease Mention Recognition in News Headlines

Sidak Pal Singh, Sopan Khosla, Sajal Rustagi, Manisha Patel
Graduate student
IIT Roorkee
(sidakuec,sajalume,manipubt)@iitr.ac.in,khoslasopan@gmail.com

Dhaval Patel
Faculty
IIT Roorkee
patelfec@iitr.ac.in

Abstract

News headlines are a vital source of information for the masses. Identifying diseases that are being spread or discovered is important to take necessary steps for their prevention and cure. Our system uses a syntactic and lexical constraint-based approach which then goes through a frequency analysis phase to extract meaningful disease names. In the task of top-150 (unique) disease mention recognition on the 2015 news headlines dataset, our approach shows **40%** gain in accuracy in comparison to other baseline approaches, illustrating the benefit of our approach.

1 Introduction

Disease Mention Recognition involves extraction of disease name from a given text. News provides us access to current events and up-to-date information regarding varied fields. Rather than analyzing the entire news text across different sources, headlines are a quick and viable option to extract useful knowledge.

Disease names found in news headlines can inform us about the kind of diseases which are getting spread or are prevalent in different regions at various points of time. This has several advantages: taking adequate measures for the prevention and control of diseases, investing in research and development for their cures, predicting future epidemic outbreaks, etc. Correctly recognizing a disease mention is vital for improvement of disease-centric knowledge extraction tasks, like drug discovery [Agarwal and Searls, 2008].

We aim to discover patterns in the news headlines that contain diseases and use them to generalize over diseases that haven't been seen. We identify a set of significantly covering word roots that signal disease mentions and then extract the sentence structure using rule-based inference techniques. In other words, we use syntactic and lexical (SL) constraints to extract the disease names from headlines in an initial pass. The highlight of our approach is the *Frequency-based Iterative Improvement* (FII) that leads to more accurate results by weeding out the false positives.

We experimented on a total of *664163* headlines for year 2014 and *408052* headlines for year 2015, collected using the iMM (Indian news Media Monitoring) system described in [Mazumder *et al.*, 2014]. Our $SL - FII$ [1] system was able to extract a total of **3157** and **5058** correct occurrences of disease names for 2014 and 2015 respectively. In order to compare the performance of our method with baseline approaches, we use both manual analysis as well as an external knowledge source. Our system gives *40%* gain in accuracy in comparison to other baseline approaches in the task of *top-150* (unique) disease mention recognition on the 2015 news headlines dataset.

2 Related Work

One of the essential requirements for a text mining application is the ability to identify relevant entities. There has been an increasing amount of research in Biomedical Named Entity Recognition (BNER), which is the task of locating boundaries of the biomedical entities in the given corpus and tagging them with corresponding semantic type (e.g. proteins, vitamins, viruses etc.). With the various events, i2b2 [Uzuner *et al.*, 2011] and scientific challenges [Kim *et al.*, 2009], BNER has seen huge development in recognizing the mention of genes [Lu *et al.*, 2011], [Torii *et al.*, 2009], organisms [Naderi *et al.*, 2011] and diseases [Dogan and Lu, 2012].

Most of the research related to Biomedical Named Entity Recognition has been focused on clinical texts [Pradhan *et al.*, 2014], medical records and PubMed queries [Névéol *et al.*, 2009], [Doan *et al.*, 2014]. But as far as our knowledge is concerned, extracting disease mentions *from news headlines* hasn't been significantly explored in the literature.

Most of the techniques that have been used in these tasks are based on machine learning approaches such as Support Vector Machines (SVM) and Conditional Random Fields (CRF). Disease mention recognition by [Jonnagaddala *et al.*, 2015] was performed using CRF approach on PubMed Abstracts. BANNER, developed by [Leaman *et al.*, 2008] is also based on CRF approach using syntactic, lexical and orthographic features extracted to recognize disease mentions. This work was further extended in the context of biomedical texts by [Chowdhury *et al.*, 2010] by use of contextual features in addition to features extracted by Banner.

In all these approaches, the used corpus (for example NCBI, [Doan *et al.*, 2014]) has detailed annotations at both

[1] Code and data for our system can be found at `https://github.com/sidak/Disease_Mention_Recognition`

mention and concept level. But, our headline dataset **does not** have any form of annotations. Such a scenario makes it difficult to apply supervised machine learning techniques.

Further, due to differences in structural patterns of news corpus and biomedical texts, aforementioned approaches cannot be effectively used for disease mention recognition from news corpus. Apart from this, an advantage of our approach is that it can generalize for entities belonging to domains as different as cricket, politics etc. We illustrate this with an example in Section 3.3.

3 Proposed Solution

Our solution involves four basic stages: Pre-processing, Relational Extraction, Frequency-based Iterative Improvement and Post-processing. The architecture of our $SL - FII$ system is shown in Figure 1.

3.1 Dataset

We use the iMM system [Mazumder *et al.*, 2014] to collect news headlines for year 2014 and 2015. We make use of a manually prepared list of 95 diseases to simulate annotations (see Section 4) and then use it to extract word roots that cover a significant portion of disease containing headlines. Some of the headlines collected by the iMM system that would be further used in this paper for the explanation of our technique are mentioned in Table 1.

Sample Headlines
Healthcare worker in Scotland diagnosed with Ebola
Beyonce Reaches Out to Grieving Family of Teen Who Died of Cancer
Bird flu outbreak in Kottayam, Alappuzha
More details on Dan Uggla's concussion symptoms
IMF rules out special treatment for Greece

Table 1: Sample headlines collected by iMM system

3.2 Pre-processing

The first step of pre-processing is to remove apostrophe inconsistencies from the corpus (headlines). After this we use $NLTK$ [2] to tokenize the headlines and then tag the produced tokens with parts of speech (i.e. PoS tagging). News headlines may also contain grammatical inconsistencies. For example, capitalizing the first letter of every word, punctuation mistakes, or missing articles etc. In such cases, PoS taggers might incorrectly tag certain words as explained in the following example. Consider the headline,

"India Seeks Revenge From Australia"

```
POS tags: [('India', 'NNP'), ('Seeks',
'NNP'), ('Revenge', 'NNP'), ('From', 'NNP'),
('Australia', 'NNP')]
```

[2] http://www.nltk.org/ Natural Language Toolkit (NLTK v3.1)

(The symbols for the PoS tags and their corresponding description is shown in Table 2.)

To handle such inconsistencies we use the following approach:

1. Convert headlines to lower-case and then compare the respective PoS tags of tokens with that of the original sentence.
2. If PoS tag differs, use lower-case form. Otherwise, use the original one.

Thus, the example headline is compared to *"india seeks revenge from australia"*

```
PoS tags: [('india', 'NN'), ('seeks',
'VBZ'), ('revenge', 'NN'), ('from', 'IN'),
('australia', 'NN')]
```

On converting to lower-case, PoS tag of *'India'* still semantically portrays a noun, whereas the PoS tag of *'Seeks'* changes from noun to verb (NNP to VBZ). So the example headline finally gets converted to *"India seeks Revenge from Australia"*

```
POS tagged: [('India', 'NNP'), ('seeks',
'VBZ'), ('Revenge', 'NNP'), ('from', 'IN'),
('Australia', 'NNP')]
```

3.3 Relational Extraction

In this section, we introduce two types of constraints, namely lexical and syntactic. These constraints help us to discover and extract *disease name - word root* relations. We also discuss how to generalize this idea for entities pertaining to different domains.

Symbol	**Description**
DT	determiner
JJ	adjective
NN	noun, singular
NNS	noun plural
NNP	proper noun, singular
$NNPS$	proper noun, plural
RB	adverb
VB	verb, base form
VBZ	verb, 3rd person
CD	cardinal number

Table 2: Notation used for PoS tags.

Lexical Constraints

News headlines contain multiple entities. Our task is to identify the correct set of entities that correspond to disease names. In other words, we need to formulate a context that signals the occurrence of disease names.

In order to define this context or neighbourhood, we extract certain word roots that indicate the presence of disease

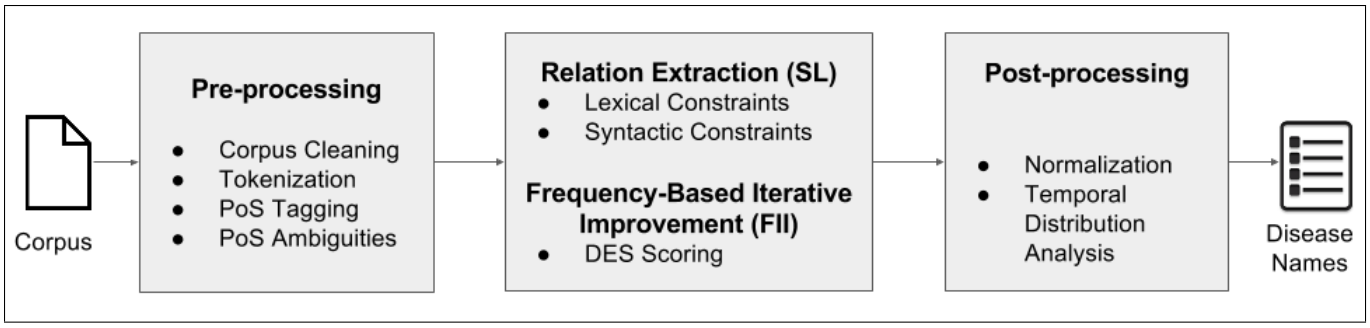

Figure 1: The architecture of our $SL - FII$ system.

names in headlines (with high confidence). The word roots are obtained by analyzing the headline data and the initial list of disease names. Based on how well they cover the disease containing headlines, both quantitatively and qualitatively, we select a certain subset of these word roots. From our experiments, we find that a specific set of 10 word roots covers a significant portion of disease containing headlines and are listed as follows:

- 'diagnos' derivatives
- 'outbreak' derivatives
- 'cur' derivatives
- 'vaccin' derivatives
- 'die' derivatives
- 'battling' derivatives
- 'symptom' derivatives
- 'treatment' derivatives
- 'virus' derivatives
- 'hospital' derivatives

Note that, derivatives (over here) implies the inflected forms of the keywords mentioned along with certain prepositions. For example, consider the derivatives of *'diagnos'*

[mis]diagnos(e | es | ed | is | tic) [with | for | of | by]

Apart from the above list, word roots like *'drug'*, *'patient'*, *'therapy'* and more are also identified. Consideration of these word roots leads to a marginal improvement in the number of identified disease mentions. An intuitive justification for the above is that quite often such word roots are used along with entities other than disease names. In several cases, they tend to identify more false positives than true positives. Thus choosing this small list doesn't lead to any significant loss, and at the same time speeds up the system.

Syntactic Constraints

The headlines containing inflected forms of different word roots are extracted using lexical constraints. Using the PoS tags of obtained headlines, we develop syntactic constraints/rules to capture the position of occurrence of disease names, in relation to word roots identified above.

Besides eliminating incoherent extractions, syntactic constraints also reduce uninformative extractions by capturing relation phrases that are expressed by only certain combinations.

Figure 2 shows the syntactic constraints developed for the inflected form *'diagnoses'* of word root *'diagnos'* and is described in detail below.

- In the headline, *"Stool test diagnoses **bowel disease**"*, inflected form *'diagnoses'* of word root *'diagnos'* is used as 3rd person verb (VBZ). The disease mention, *'bowel disease'* extracted as pair of singular nouns (NN NN), occurs to the right of *'diagnoses'*.
- In another headline, *"**Autism** diagnoses surge by 30 percent in kids"*, inflected form *'diagnoses'* of word root *'diagnos'* is used as plural noun (NNS). The disease mention , *'Autism'* extracted as noun (NN), occurs to the left of *'diagnoses'*.

.

- If the PoS tag for 'diagnoses' is <VBZ>:
  - Disease name is towards right
  - The structure is (J J)* (NN | NNP | NNS | NNPS | CD) +
    - Skip any article <DT> before this structure
- Else if the PoS tag is <NNS> or <NNP>:
  - Disease name is towards left
  - The structure is (J J)* (NN | NNP | NNS | NNPS | CD) +

Figure 2: Rules for the inflected form 'diagnoses' of word root 'diagnos'.

Syntactic constraints for other word roots are developed in a similar manner. Disease mentions in news headlines generally occur around the word roots as the regular expression given below:

$$E = [DT](JJ)^*(NN|NNP|NNS|NNPS|CD)^+$$

Or in other words, disease mentions are phrases that contain an optional determiner or article (e.g. a, an) followed by multiple optional adjectives (e.g. fractured) and at least one noun (e.g. elbow, non-Hodgskins lymphoma) with a maximum length of four.

Since disease names basically represent a kind of entities, the syntactic constraints extract the potential disease name phrases from news headlines. For obvious reasons, we omit the determiner or article (represented by the PoS tag: DT), in order to get the disease names. Below, is an example headline, depicting the application of these constraints, with extracted disease mention highlighted in bold.

Former Butler forward Andrew Smith diagnosed with ***non-Hodgskins lymphoma*** .

Generalization

Since a basic entity is represented by the regular expression discussed earlier, our $SL - FII$ approach can be generalized for entities of other domains. This can be possible via a simple modification of lexical constraints as per the context requirements. For example in order to identify the names of cricketers, we can use word roots like:

- ' wicket haul' derivatives
- 'ton' derivatives

In the headlines below, we can successfully extract the cricketer's name using our $SL - FII$ model.

Gayle's 47-ball ***ton*** *wipes out England*
Ashwin's 5 ***wicket haul*** *takes India to the semis*

Generalization of this approach to other languages mainly depends on the quality of the standard NLP pre-processing tools like PoS tagger and stemmer, that are available for the required language, while the approach will remain pretty much the same.

3.4 Frequency-based Iterative Improvement (FII)

The main motivation is to use the previous experience gained in recognizing disease mentions with different word roots, to weed out incorrect diseases names. The constraints formulated above are used to extract potential disease names from the corpus in an initial pass, which are then passed through the FII phase. In order to filter out false positives, we assign Disease Expectancy Score (DES) values (based on the probability of a word being disease name) to outputs of the initial pass. The idea behind our FII phase can be easily understood with following examples and equations 1 to 4.

- Mentions identified along with word root 'battling' can come as 'battling insecurity','battling cancer','battling lust', but for word root 'treatment' most of the occurrences are of the form, 'Ebola treatment', 'treatment of cancer'. So if disease mention is extracted using word root 'treatment', it is more probable to be an instance of disease name as compared to mention extracted using word root 'battling'.
- Disease mention such as 'Cancer' which can be used along with multiple word roots 'treatment','battling' and 'cure', are more probable to be instances of disease names as compared to mentions like 'seasonal' that are only extracted using a single word root 'outbreak'.

1. Probability of a phrase(p) being an instance of disease name(D) depends on the probability of phrase being used with different lexical rules ($p \cap rule_i$) as defined in Equation 1.

$$P(p \in D) = \sum_{rules} P(p \in D | p \cap rule_i) \times P(p \cap rule_i) \quad (1)$$

2. Based on the training set (2014 headlines), a weight is assigned each to lexical rule ($W[rule_i]$) which corresponds to the probability that the phrase extracted using that lexical rule($p \cap rule_i$) is a valid disease-name, as shown by Equation 2.

$$P(p \in D | p \cap rule_i) = W[rule_i] \quad (2)$$

   For example, rules involving 'diagnosed with' give better results than rules involving 'outbreak' (which also gives false positives with natural disasters). Thus much higher weight is associated with phrases output via 'diagnosed with'.

   Thus, the weight assigned to each rule depends on the number of correctly recognized disease mentions and the total number of disease mentions extracted using it.

3. Probability of a phrase occurring with the lexical rule in consideration ($p \cap rule_i$) depends on the frequency of phrase occurring with lexical rule and size of our entire corpus($size$), as shown by Equation 3.

$$P(p \cap rule_i) = \frac{F[rule_i][p]}{size} \quad (3)$$

4. Final score (also termed as DES in Equation 4) is calculated taking into account the above probabilities for each rule.

   The score is equivalent to finding the probability that the phrase detected after the initial pass is a disease.

$$DES[p] = \sum_{rules} W[rule_i] \times F[rule_i][p] \quad (4)$$

FII increases accuracy and reduces false positives based on probabilistic measures. Words with higher DES have higher chances of being a valid disease-name. Phrases which occur with more rules are more reliable as a disease-name. e.g. Both Ebola and Tornado occur with 'outbreak', but Ebola can also occur with 'diagnosed with', 'died of' and several other rules whereas Tornado cannot. Thus, FII increases belief in Ebola as a valid disease-name.

3.5 Post-processing

The results of FII are sent for post-processing. Firstly we normalize the results and then analyze them based on its temporal distribution. Valid disease-names show distinct peaks whereas non-disease entities show uniform distribution over the specified time-interval.

4 Experimental Results

Across all 664163 headlines for the year 2014 and 408052 headlines for the year 2015, *none of them is annotated* in any manner. To resolve this problem, we manually prepare

a list of 95 disease names. Then any headline which contains atleast one of these diseases is taken to be a disease containing headline. In this manner, we obtain 1562 and **1884** headlines for the year 2014 and 2015 respectively, that are considered to have a correct disease mention. Figure 3 shows the number of occurrences for the most frequently occurring diseases, in these headlines for the year 2014.

Few of these headlines may actually contain diseases names (from the prepared list) used in different contexts, for example, 'fever'. Despite this fact, our system is able to handle such slight inconsistencies, because of the FII phase.

The system is trained on headlines from 2014 to extract suitable SL constraints and learn the weights of lexical rules for the FII phase. When tested on all the headlines from 2015, the system was able to extract a total of **5058** correct disease mentions. There is more than **2.5x** gain in the number of disease containing headlines for 2015.

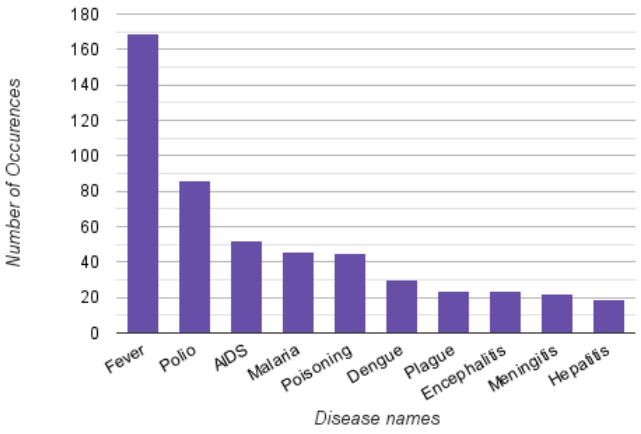

Figure 3: Number of occurrences of Disease names in 2014

Besides this quantitative improvement, we also obtain many diseases like *'Ebola', 'Cancer', 'Concussion'* etc., which were not there in our input set of diseases, as shown in Table 3.

Disease Name	# instances in 2014	# instances in 2015
Cancer	1072	494
Ebola	182	990
Concussion	190	79
Autism	82	35
Listeria	18	13

Table 3: Example of new extracted disease-names

Table 4 shows a sample of potential disease names extracted from headlines using syntactic and lexical (SL) constraints. The 5th headline is a sample false positive, where SL constraints incorrectly label *'Greece'* as a disease. But this will be handled once we pass it through FII phase, as illustrated in Table 5. The DES value of Greece is 0.32 which is much less than the DES values for other entities and can be filtered out as per the threshold accepted DES value. This threshold can be decided on the basis of accuracy requirements.

Headline	Lexical rule	Disease name extracted
Healthcare worker in Scotland diagnosed with **Ebola**	diagnosed with	Ebola
Beyonce Reaches Out to Grieving Family of Teen Who died of **Cancer**	died of	Cancer
Bird flu outbreak in Kottayam, Alappuzha	outbreak	Bird flu
More details on Dan Uggla's **concussion** symptoms	symptoms	Concussion
IMF rules out special treatment for **Greece**	treatment for	Greece

Table 4: Diseases extracted from headlines using SL rules

Disease	DES
Ebola	80.13
Cancer	14.66
Bird Flu	12.38
Concussion	1.86
Greece	0.32

Table 5: Disease Expectancy Scores after FII

Extracted disease names are sorted in descending order of their DES values. Due to our post-processing step i.e. *Temporal Distribution Analysis*, the accuracy of our $SL - FII$ system gets improved by a great extent since uniformly occurring words like *'fight', 'water', 'state'* etc. are filtered out. The difference in accuracy (% of correct disease mentions in the top K unique predictions) on test data (headlines of 2015) before and after the Post-processing step can be observed in Figure 4.

Now, we compare our $SL - FII$ approach to other techniques that can be used for extraction of disease names.

Baseline I : Machine learning approach developed by [Chowdhury *et al.*, 2010] for Disease mention recognition in biomedical texts. This technique makes use of decision trees and extracts orthographic and linguistic features such as PoS tag, suffixes and prefixes, word beginning with upper-case letter, checking if all letters are in the upper-case and more. 15,000 instances were extracted from news corpus in order to train our classifier. 10-fold cross validation accuracy of 74.89% was obtained using decision trees for classification.

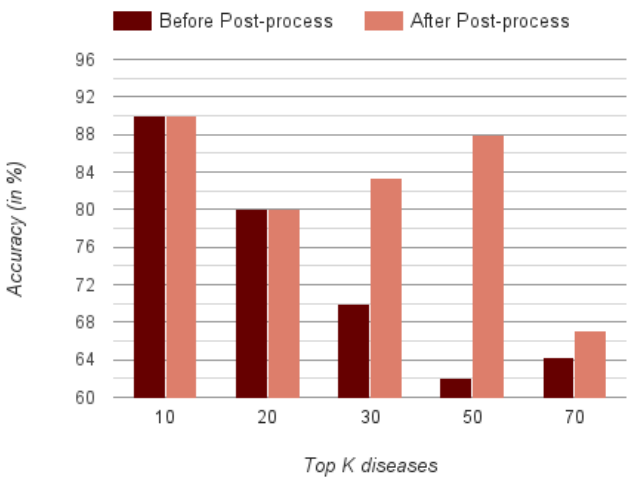

Figure 4: Accuracy obtained for top-K unique instances

Other classifiers such as SVM (72.3%.) gave lower accuracy in comparison to decision trees over our data set.

Baseline II : Sequence approach which is based on the hypothesis that context for disease names can be defined using prefixes and suffixes used around them. This approach extracts all the unigrams and bigrams that occur around disease names i.e. all the one/two word prefixes and suffixes. These prefix and suffix are assigned weights on the basis of the probability of their occurrence with disease names. Some of the prefixes extracted by this approach were 'cured of','deaths from' and 'deadly'. Similarly 'outbreak rises','cases reported', 'vaccine' were some of the suffixes extracted by this approach. Only mentions that have a score above a particular threshold value are considered as disease mentions. Similarly probability of misclassification by each prefix/suffix with which disease mention occurs is used to calculate the probability of error for each disease mention. If this error probability is below 5%, disease mention is considered to be correct by the approach.

In our comparison, all the approaches are trained using headlines from 2014 and are tested on headlines from 2015. Disease names extracted using all three approaches are then sorted in order of confidence-values given by the approaches. Disease instances/names extracted using each approach are further analyzed manually to estimate the top-K (unique) instance accuracies, as shown in Figure 5. Our method achieves around **88%** accuracy in the top-50 instances.

To automatically compare our results with other approaches, we used the list of diseases provided by the Centers for Disease Control and Prevention (CDC) [3], USA. CDC identifies around 841 diseases, along with their conditions and variants. Figure 6 shows the automatic comparison of accuracies for top-K (unique) disease instances extracted by each approach using disease names provided by CDC.

Using these experiments, we observe that our $SL-FII$ technique performs considerably better than other approaches. In the top-150 (unique) disease instances, our approach *outperforms* the best baseline by **40%** and similarly in the top-50 by **22%**.

[3] http://www.cdc.gov/diseasesconditions/

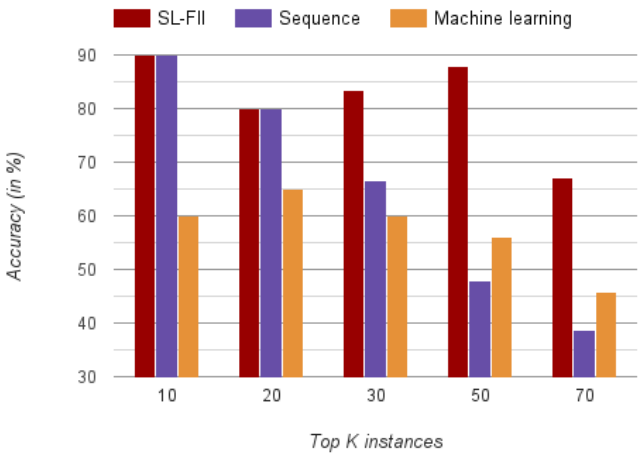

Figure 5: Comparison of Accuracies using Manual Approach

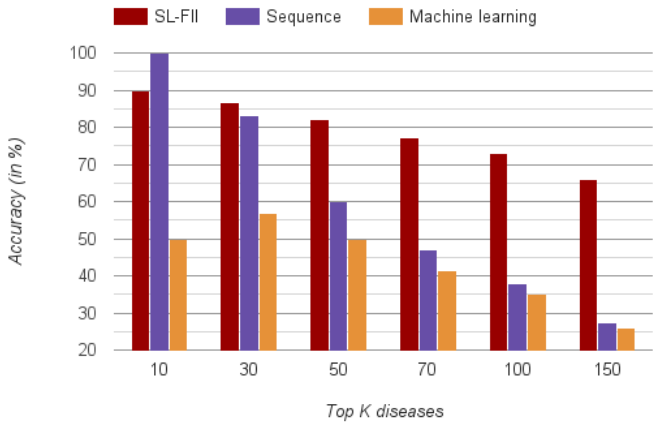

Figure 6: Comparison of Accuracies using CDC's Disease

5 Conclusion

Using syntactic and lexical constraints gives us potential disease names. We can then use the FII phase to weed out false diseases names based on the experience acquired in recognizing disease mentions with different word roots. Even in the absence of annotations on the corpus, the initial list of diseases helps us to simulate annotations. In future, we plan to find the minimum size of this initial list and the influence of removing some of the word roots on the accuracy.

6 Acknowledgments

We thank anonymous reviewers for their valuable comments and suggestions to improve this paper. We would also like to thank Shubham Kumar Pandey, Paarth Neekhara, Vikash Kumar, Baviskar Hrishikesh Hari, Chandan Singha and Kavin Motlani for setting up the baseline experiments.

References

[Agarwal and Searls, 2008] Pankaj Agarwal and David B. Searls. Literature mining in support of drug discovery. *Briefings in Bioinformatics*, 9(6):479–492, 2008.

[Chowdhury *et al.*, 2010] Mahbub Chowdhury, Md Faisal, et al. Disease mention recognition with specific features. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 83–90. Association for Computational Linguistics, 2010.

[Dogan and Lu, 2012] Rezarta Islamaj Dogan and Zhiyong Lu. An inference method for disease name normalization. In *Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012*, 2012.

[Doan *et al.*, 2014] Rezarta Islamaj Doan, Robert Leaman, and Zhiyong Lu. {NCBI} disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1 – 10, 2014.

[Jonnagaddala *et al.*, 2015] Jitendra Jonnagaddala, Nai-Wen Chang, Toni Rose Jue, and Hong-Jie Dai. Recognition and normalization of disease mentions in pubmed abstracts. 2015.

[Kim *et al.*, 2009] Jin-dong Kim, Tomoko Ohta, Sampo Pyysalo, and Yoshinobu Kano. Overview of bionlp09 shared task on event extraction. In *In Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*. Citeseer, 2009.

[Leaman *et al.*, 2008] Robert Leaman, Graciela Gonzalez, et al. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663. Citeseer, 2008.

[Lu *et al.*, 2011] Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hong-Jie Dai, Naoaki Okazaki, et al. The gene normalization task in biocreative iii. *BMC bioinformatics*, 12(8):1, 2011.

[Mazumder *et al.*, 2014] Sahisnu Mazumder, Bazir Bishnoi, and Dhaval Patel. News headlines: What they can tell us? In *Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE) on I-CARE 2014*, I-CARE 2014, pages 4:1–4:4, New York, NY, USA, 2014. ACM.

[Naderi *et al.*, 2011] Nona Naderi, Thomas Kappler, Christopher JO Baker, and René Witte. Organismtagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729, 2011.

[Névéol *et al.*, 2009] Aurélie Névéol, Won Kim, W. John Wilbur, and Zhiyong Lu. Exploring two biomedical text genres for disease recognition. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 144–152, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Pradhan *et al.*, 2014] Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. Semeval-2014 task 7: Analysis of clinical text. *SemEval*, 199(99):54, 2014.

[Torii *et al.*, 2009] Manabu Torii, Zhangzhi Hu, Cathy H. Wu, and Hongfang Liu. Biotagger-gm: A gene/protein name recognition system. *Journal of the American Medical Informatics Association*, 16(2):247–255, 2009.

[Uzuner *et al.*, 2011] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.