

Insert here your thesis' task.



**FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE**

Master's thesis

Domain-specific Name Entity Recognition

Bc. Bogoljub Jakovcheski

Department of theoretical computer science

Supervisor: Ing. Milan Dojchinovski, Ph.D.

May 14, 2018

Acknowledgements

I would like to thank my family and friends for support during writing this thesis.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Article 46(6) of the Act, I hereby grant a nonexclusive authorization (license) to utilize this thesis, including any and all computer programs incorporated therein or attached thereto and all corresponding documentation (hereinafter collectively referred to as the “Work”), to any and all persons that wish to utilize the Work. Such persons are entitled to use the Work in any way (including for-profit purposes) that does not detract from its value. This authorization is not limited in terms of time, location and quantity. However, all persons that makes use of the above license shall be obliged to grant a license at least in the same scope as defined above with respect to each and every work that is created (wholly or in part) based on the Work, by modifying the Work, by combining the Work with another work, by including the Work in a collection of works or by adapting the Work (including translation), and at the same time make available the source code of such work at least in a way and scope that are comparable to the way and scope in which the source code of the Work is made available.

In Prague on May 14, 2018

.....

Czech Technical University in Prague

Faculty of Information Technology

© 2018 Bogoljub Jakovcheski. All rights reserved.

This thesis is school work as defined by Copyright Act of the Czech Republic. It has been submitted at Czech Technical University in Prague, Faculty of Information Technology. The thesis is protected by the Copyright Act and its usage without author's permission is prohibited (with exceptions defined by the Copyright Act).

Citation of this thesis

Jakovcheski, Bogoljub. *Domain-specific Name Entity Recognition*. Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2018.

Abstrakt

Klíčová slova

Abstract

Keywords

Contents

Introduction	1
Motivation	1
Goals of the thesis	1
Thesis outline	1
1 Background and related work	3
1.1 Information extraction	3
1.2 NER	3
1.3 RDF/NIF	3
1.4 DBpedia ontology	3
1.5 Evaluation/training datasets	3
1.6 Domain specific NER	3
1.7 Stanford NER	4
2 Domain specific NER	5
2.1 Data pre-processing	5
2.2 Types retrieval	5
2.3 Domain specification	5
2.4 Data transformation/output for Stanford	6
3 Experiments	7
3.1 Goals of the experiments	7
3.2 List of experiments	7
3.3 Summary of results	44
Conclusion	47
A Contents of CD	49

List of Figures

3.1	All 3 Domains Coarse BBC Coarse	7
3.2	All 3 Domains Coarse Top 10 Links	8
3.3	All 3 Domains Coarse Top 10 Links Politics	8
3.4	All 3 Domains Coarse Top 10 Links Sport	8
3.5	All 3 Domains Coarse Top 10 Links Transportation	9
3.6	All 3 Domains Fine Top 10 Links	9
3.7	All 3 Domains Fine Top 10 Links Politics	9
3.8	All 3 Domains Fine Top 10 Links Sport	9
3.9	All 3 Domains Fine Top 10 Links Transportation	9
3.10	Politics Coarse Top 10 Links Politics	10
3.11	Politics Fine Top 10 Links Politics	10
3.12	Politics Coarse Top 10 Links Politics	10
3.13	Sport Fine Top 10 Links Sport	10
3.14	All 3 Domains Coarse Top 20 Links	11
3.15	All 3 Domains Coarse Top 20 Links Politics	11
3.16	All 3 Domains Coarse Top 20 Links Sport	11
3.17	All 3 Domains Coarse Top 20 Links Transportation	11
3.18	All 3 Domains Fine Top 20 Links	11
3.19	All 3 Domains Fine Top 20 Links Politics	12
3.20	All 3 Domains Fine Top 20 Links Sport	12
3.21	All 3 Domains Fine Top 20 Links Transportation	12
3.22	Politics Coarse Top 20 Links Politics	12
3.23	Politics Fine Top 20 Links Politics	13
3.24	Politics Coarse Top 20 Links Politics	13
3.25	Sport Fine Top 20 Links Sport	13
3.26	All 3 Domains Coarse Top 40 Links	13
3.27	All 3 Domains Coarse Top 40 Links Politics	13
3.28	All 3 Domains Coarse Top 40 Links Sport	13
3.29	All 3 Domains Coarse Top 40 Links Transportation	14
3.30	All 3 Domains Fine Top 40 Links	14

3.31	All 3 Domains Fine Top 40 Links Politics	14
3.32	All 3 Domains Fine Top 40 Links Sport	14
3.33	All 3 Domains Fine Top 40 Links Transportation	15
3.34	Politics Coarse Top 40 Links Politics	15
3.35	Politics Fine Top 40 Links Politics	15
3.36	Politics Coarse Top 40 Links Politics	15
3.37	Sport Fine Top 40 Links Sport	16
3.38	All 3 Domains Coarse Top 100 Links	16
3.39	All 3 Domains Coarse Top 100 Links Politics	16
3.40	All 3 Domains Coarse Top 100 Links Sport	16
3.41	All 3 Domains Coarse Top 100 Links Transportation	16
3.42	All 3 Domains Fine Top 100 Links	17
3.43	All 3 Domains Fine Top 100 Links Politics	17
3.44	All 3 Domains Fine Top 100 Links Sport	17
3.45	All 3 Domains Fine Top 100 Links Transportation	18
3.46	Politics Coarse Top 100 Links Politics	18
3.47	Politics Fine Top 100 Links Politics	18
3.48	Politics Coarse Top 100 Links Politics	18
3.49	Sport Fine Top 100 Links Sport	18
3.50	All 3 Domains Coarse Top 300 Links	19
3.51	All 3 Domains Coarse Top 300 Links Politics	19
3.52	All 3 Domains Coarse Top 300 Links Sport	19
3.53	All 3 Domains Coarse Top 300 Links Transportation	19
3.54	All 3 Domains Fine Top 300 Links	19
3.55	All 3 Domains Fine Top 300 Links Politics	20
3.56	All 3 Domains Fine Top 300 Links Sport	20
3.57	All 3 Domains Fine Top 300 Links Transportation	20
3.58	Politics Coarse Top 300 Links Politics	20
3.59	Politics Fine Top 300 Links Politics	21
3.60	Politics Coarse Top 300 Links Politics	21
3.61	Sport Fine Top 300 Links Sport	21
3.62	All 3 Domains Coarse Top 400 Links	21
3.63	All 3 Domains Coarse Top 400 Links Politics	22
3.64	All 3 Domains Coarse Top 400 Links Sport	22
3.65	All 3 Domains Coarse Top 400 Links Transportation	22
3.66	All 3 Domains Fine Top 400 Links	22
3.67	All 3 Domains Fine Top 400 Links Politics	23
3.68	All 3 Domains Fine Top 400 Links Sport	23
3.69	All 3 Domains Fine Top 400 Links Transportation	23
3.70	Politics Coarse Top 400 Links Politics	24
3.71	Politics Fine Top 400 Links Politics	24
3.72	Politics Coarse Top 400 Links Politics	24
3.73	Sport Fine Top 400 Links Sport	24
3.74	All 3 Domains Coarse Top 500 Links	24

3.75	All 3 Domains Coarse Top 500 Links Politics	24
3.76	All 3 Domains Coarse Top 500 Links Sport	25
3.77	All 3 Domains Coarse Top 500 Links Transportation	25
3.78	All 3 Domains Fine Top 500 Links	25
3.79	All 3 Domains Fine Top 500 Links Politics	26
3.80	All 3 Domains Fine Top 500 Links Sport	26
3.81	All 3 Domains Fine Top 500 Links Transportation	26
3.82	Politics Coarse Top 500 Links Politics	26
3.83	Politics Fine Top 500 Links Politics	27
3.84	Politics Coarse Top 500 Links Politics	27
3.85	Sport Fine Top 500 Links Sport	27
3.86	All 3 Domains Fine Grained Top 300 With All 3 Domains Fine Grained Top 500 Links And All 3 Domains Fine Grained Top 500 Links With Lower PageRank	27
3.87	All 3 Domains Fine Grained Top 500 Links With All 3 Domains Fine Grained Top 500 Links And All 3 Domains Fine Grained Top 500 Links With Lower Page Rank	28
3.88	All 3 Domains Fine Grained Top 500 Links With All 3 Domains Fine Grained Top 500 Links With Lower PageRank	28
3.89	Transportation Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 300 Links And Transportation Fine Grained Top 300 Links	29
3.90	All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 100 Links	29
3.91	All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 300 Links	30
3.92	All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 500 Links	30
3.93	All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 500 Links With Lower PageRank	31
3.94	All 3 Domains Fine Grained Top 100 Links Runned With All 3 Domains Fine Grained Top 500 Links	31
3.95	All 3 Domains Fine Grained Top 100 Links Runned With All 3 Domains Fine Grained Top 500 Links With Lower PageRank	32
3.96	All 3 Domains Fine Grained Top 100 Links Runned With Politics Fine Grained Top 500 Links	32
3.97	All 3 Domains Fine Grained Top 100 Links Runned With Politics Fine Grained Top 500 Links With Lower PageRank	32
3.98	All 3 Domains Fine Grained Top 100 Links Runned With Sport Fine Grained Top 500 Links	33
3.99	All 3 Domains Fine Grained Top 100 Links Runned With Sport Fine Grained Top 500 Links With Lower PageRank	33
3.100	All 3 Domains Fine Grained Top 100 Links Runned With Trans- portation Fine Grained Top 500 Links	33

3.101	All 3 Domains Fine Grained Top 100 Links Runned With Transportation Fine Grained Top 500 Links With Lower PageRank . . .	34
3.102	All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 500 Links	34
3.103	All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 500 Links With Lower PageRank . . .	34
3.104	All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 500 Links	35
3.105	All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 500 Links With Lower PageRank	35
3.106	All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 500 Links	35
3.107	All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 500 Links With Lower PageRank	36
3.108	All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 500 Links	36
3.109	All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 500 Links With Lower PageRank . . .	36
3.110	All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 10 Links	37
3.111	All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 100 Links	37
3.112	All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 10 Links	37
3.113	All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 100 Links	37
3.114	All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 10 Links	38
3.115	All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 100 Links	38
3.116	All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 10 Links	38
3.117	All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 100 Links	38
3.118	All 3 Domains Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 10 Links	39
3.119	All 3 Domains Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 100 Links	39
3.120	All 3 Domains Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 300 Links	39
3.121	All 3 Domains Fine Grained Top 500 Links Runned With Politics Fine Grained Top 10 Links	40
3.122	All 3 Domains Fine Grained Top 500 Links Runned With Politics Fine Grained Top 100 Links	40

3.123All 3 Domains Fine Grained Top 500 Links Runned With Politics	
Fine Grained Top 300 Links	40
3.124All 3 Domains Fine Grained Top 500 Links Runned With Sport	
Fine Grained Top 10 Links	40
3.125All 3 Domains Fine Grained Top 500 Links Runned With Sport	
Fine Grained Top 100 Links	41
3.126All 3 Domains Fine Grained Top 500 Links Runned With Sport	
Fine Grained Top 300 Links	41
3.127All 3 Domains Fine Grained Top 500 Links Runned With Trans-	
portation Fine Grained Top 10 Links	41
3.128All 3 Domains Fine Grained Top 500 Links Runned With Trans-	
portation Fine Grained Top 100 Links	41
3.129All 3 Domains Fine Grained Top 500 Links Runned With Trans-	
portation Fine Grained Top 300 Links	42
3.130Politics Fine Grained Top 300 Links Runned With All 3 Domains	
Fine Grained Top 100 Links	42
3.131Politics Fine Grained Top 500 Links Runned With All 3 Domains	
Fine Grained Top 100 Links	42
3.132Politics Fine Grained Top 500 Links Runned With All 3 Domains	
Fine Grained Top 300 Links	43
3.133Sport Fine Grained Top 300 Links Runned With All 3 Domains	
Fine Grained Top 100 Links	43
3.134Top 500 Links F1	44
3.135All 3 Domains Coarse Grained Runned With All 3 Domains Coarse	
Grained-tsv F1	45

Introduction

Motivation

Goals of the thesis

The main goal of the thesis is to show (test) does extracting information from a global domain gives us better results than extracting information from a specific domain, for example politics domain or any other domain.

Thesis outline

Background and related work

1.1 Information extraction

information extraction, that is, automatically building a relational database from information contained in unstructured text. Unlike linear-chain models, general CRFs can capture long distance dependencies between labels.

1.2 NER

NER is the problem of identifying and classifying proper names in text, including locations, such as China; people, such as George Bush; and organizations, such as the United Nations. The named-entity recognition task is, given a sentence, first to segment which words are part of entities, and then to classify each entity by type (person, organization, location, and so on). The challenge of this problem is that many named entities are too rare to appear even in a large training set, and therefore the system must identify them based only on context

1.3 RDF/NIF

1.4 DBpedia ontology

<http://wiki.dbpedia.org/services-resources/ontology> <http://mappings.dbpedia.org/server/ontology/class>

1.5 Evaluation/training datasets

1.6 Domain specific NER

<https://arxiv.org/pdf/1612.00148.pdf>

1.7 Stanford NER

Domain specific NER

There are parameters of the computer used for tests shown in Table ??.

2.1 Data pre-processing

â€” Coarse grained â€” Fine grained

2.2 Types retrieval

At the main file where we write text we retrieve all types that are available for that text. After that, that text is processed with coarse and fine grained types. In coarse grained we have politics, sport and transportation specific domains. In fine grained in politics domain we retrieve in total 26 types (write them all), which we sort in 11 more specific types like for example Ambassador, Mayor, Deputy etc. are fused in one specific domain Politician. We do the same for sport domain where we retrieve in total 171 types, so those types, same as politics domain, are specified in 8 types, like SportClub, SportsLeague, SportsTeam, Athlete, Coach, OrganizationMember, SportsManager, SportsEvent. We also do the same for transportation domain, where we retrieve in total 78 types and minimized in 14 more specific types like Aircraft, Automobile, Locomotive, MilitaryVehicle, Motorcycle, On-SiteTransportation, Rocket, Ship, SpaceShuttle, SpaceStation, Spacecraft, Train, PublicTransitSystem and Infrastructure.

2.3 Domain specification

â€” How we choose those domains (Politics, Sport, Transportation)?

2.4 Data transformation/output for Stanford

â€” How we process data to be ready to use on Stanford NER system?

Experiments

We have provide various types of experiments. In next sections we will discuss more about every provided experiment.

3.1 Goals of the experiments

We set a few goals of the experiments. First of all we waned to test does we will get better results if we run the model of all domains in coarse grained, against the model of all domains in fine grained. In this test we run the models also with all domains texts. Then we get those models and we run it with specific domain texts, in both fine and coarse grained. Also we make experiments with specific domain model runned with domain specific texts, for example, politics domain model in coarse grained is runned with politics domain text also annotated in coarse grained, politics domain model in fine grained is runned with politics domain text also annotated in fine grained, and the same for sport and transportation domains.

3.2 List of experiments

Make it like experiment - discussion

```
CRFClassifier tagged 4307 words in 1 documents at 5867,85 words per second.
  Entity P      R      F1      TP      FP      FN
POLITICS 1,0000  0,0299  0,0580  2       0       65
SPORT    0,5000  0,0127  0,0247  1       1       78
TRANSPORTATION 1,0000  0,2857  0,4444  4       0       10
Totals   0,8750  0,0438  0,0833  7       1      153

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksCoarseGrained.ser.gz -testFile texts\All3DomainsBBCCoar
seGrainedSpecified.tsv
```

Figure 3.1: All 3 Domains Coarse BBC Coarse

3. EXPERIMENTS

```
CRFClassifier tagged 10336 words in 1 documents at 6764,40 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9655  0,9333  0,9492  28      1      2
SPORT    1,0000  1,0000  1,0000  15      0      0
TRANSPORTATION 1,0000  1,0000  1,0000  9       0      0
Totals   0,9811  0,9630  0,9720  52      1      2

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op10Links\All3DomainsTop10LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.2: All 3 Domains Coarse Top 10 Links

```
CRFClassifier tagged 4504 words in 1 documents at 5118,18 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9655  0,3636  0,5283  28      1     49
Totals   0,9655  0,3636  0,5283  28      1     49

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op10Links\PoliticsCoarseGrainedTop10Links.tsv
```

Figure 3.3: All 3 Domains Coarse Top 10 Links Politics

```
CRFClassifier tagged 2854 words in 1 documents at 4298,19 words per second.
Entity P      R      F1      TP      FP      FN
SPORT    1,0000  1,0000  1,0000  15      0      0
Totals   1,0000  1,0000  1,0000  15      0      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op10Links\SportCoarseGrainedTop10Links.tsv
```

Figure 3.4: All 3 Domains Coarse Top 10 Links Sport

3.2.1 Top 10 Links

In this experiment we get the model that is trained with all 3 domains together with coarse grained and we make an experiment with the text that is trained, that means text from all 3 domains annotated with coarse grained. Here we have for every domain 10 abstracts, so in total 30 annotated abstracts. This is our smallest model. How we can see from the picture in politics domain F1 value is lower than in sport and transportation. This results with lower F1 value in total, but still that is very satisfying value, because we have only 1 false positive and 2 false negative words.

In this experiment we get the previous trained model and test with politics coarse grained text. Result is not satisfying, because we have a very low recall, so that results with bad F1 value and very big false negative words. We can clearly say that this is not the model that we want to use in annotating. With those results we can say that all 3 domains together with all text perform better results than all 3 domain model tested with specified domain.

Then we get the same model, and like previous, we tested with sport coarse grained text. So how we can see from the picture, we have perfect result, there is no false positive or false negative words, there are only true positive which results with F1 value of 1.000 (no lost words).

TEXT
TEXT

3.2. List of experiments

```
CRFClassifier tagged 2978 words in 1 documents at 4532,72 words per second.
Entity P      R      F1      TP      FP      FN
TRANSPORTATION 1,0000  1,0000  1,0000  9       0       0
Totals 1,0000  1,0000  1,0000  9       0       0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op10Links\TransportationCoarseGrainedTop10Links.tsv
```

Figure 3.5: All 3 Domains Coarse Top 10 Links Transportation

```
CRFClassifier tagged 10336 words in 1 documents at 3395,53 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  1,0000  1,0000  1       0       0
Athlete 1,0000  1,0000  1,0000  1       0       0
Coach 1,0000  1,0000  1,0000  1       0       0
PoliticalParty 0,9600  0,9231  0,9412  24      1       2
Politician 1,0000  1,0000  1,0000  4       0       0
PublicTransitSystem 1,0000  1,0000  1,0000  7       0       0
Ship 1,0000  1,0000  1,0000  1       0       0
SportsClub 1,0000  1,0000  1,0000  7       0       0
SportsEvent 1,0000  1,0000  1,0000  1       0       0
SportsLeague 1,0000  1,0000  1,0000  4       0       0
SportsTeam 1,0000  1,0000  1,0000  1       0       0
Totals 0,9811  0,9630  0,9720  52      1       2

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10
Links\All3DomainsTop10LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.6: All 3 Domains Fine Top 10 Links

```
CRFClassifier tagged 4504 words in 1 documents at 2861,50 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000  0,0000  0,0000  0       0       15
PoliticalParty 0,9600  0,9231  0,9412  24      1       2
Politician 1,0000  0,1111  0,2000  4       0       32
Totals 0,9655  0,3636  0,5283  28      1       49

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10
Links\PoliticsFineGrainedTop10Links.tsv
```

Figure 3.7: All 3 Domains Fine Top 10 Links Politics

TEXT
TEXT
TEXT
TEXT
TEXT
TEXT
TEXT

```
CRFClassifier tagged 2854 words in 1 documents at 2548,21 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000  1,0000  1,0000  1       0       0
Coach 1,0000  1,0000  1,0000  1       0       0
SportsClub 1,0000  1,0000  1,0000  7       0       0
SportsEvent 1,0000  1,0000  1,0000  1       0       0
SportsLeague 1,0000  1,0000  1,0000  4       0       0
SportsTeam 1,0000  1,0000  1,0000  1       0       0
Totals 1,0000  1,0000  1,0000  15      0       0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10
Links\SportFineGrainedTop10Links.tsv
```

Figure 3.8: All 3 Domains Fine Top 10 Links Sport

3. EXPERIMENTS

```
CRFClassifier tagged 2978 words in 1 documents at 2538,79 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 1,0000 1,0000 1        0        0
PublicTransitSystem 1,0000 1,0000 1,0000 7        0        0
Ship 1,0000 1,0000 1,0000 1        0        0
Totals 1,0000 1,0000 1,0000 9        0        0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10
Links\TransportationFineGrainedTop10Links.tsv
```

Figure 3.9: All 3 Domains Fine Top 10 Links Transportation

```
CRFClassifier tagged 4504 words in 1 documents at 5941,95 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9737 0,9610 0,9673 74       2        3
Totals 0,9737 0,9610 0,9673 74       2        3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-PoliticsTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top10
Links\PoliticsCoarseGrainedTop10Links.tsv
```

Figure 3.10: Politics Coarse Top 10 Links Politics

```
CRFClassifier tagged 4504 words in 1 documents at 3568,94 words per second.
Entity P      R      F1      TP      FP      FN
Election 1,0000 0,9333 0,9655 14       0        1
PoliticalParty 0,9600 0,9231 0,9412 24       1        2
Politician 1,0000 1,0000 1,0000 36       0        0
Totals 0,9867 0,9610 0,9737 74       1        3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-PoliticsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Li
ks\PoliticsFineGrainedTop10Links.tsv
```

Figure 3.11: Politics Fine Top 10 Links Politics

```
CRFClassifier tagged 2854 words in 1 documents at 4920,69 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 1,0000 1,0000 1,0000 15       0        0
Totals 1,0000 1,0000 1,0000 15       0        0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-SportTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Li
nks\SportCoarseGrainedTop10Links.tsv
```

Figure 3.12: Politics Coarse Top 10 Links Politics

```
CRFClassifier tagged 2854 words in 1 documents at 2525,66 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 1,0000 1,0000 1        0        0
Coach 1,0000 1,0000 1,0000 1        0        0
SportsClub 1,0000 1,0000 1,0000 7        0        0
SportsEvent 1,0000 1,0000 1,0000 1        0        0
SportsLeague 1,0000 1,0000 1,0000 4        0        0
SportsTeam 1,0000 1,0000 1,0000 1        0        0
Totals 1,0000 1,0000 1,0000 15       0        0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top10Links\ner-SportTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Li
ks\SportFineGrainedTop10Links.tsv
```

Figure 3.13: Sport Fine Top 10 Links Sport

3.2. List of experiments

```
CRFClassifier tagged 18733 words in 1 documents at 6802,11 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 1,0000 0,9615 0,9804 50      0      2
SPORT 1,0000 1,0000 1,0000 27      0      0
TRANSPORTATION 1,0000 1,0000 1,0000 12      0      0
Totals 1,0000 0,9780 0,9889 89      0      2

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op20Links\All3DomainsTop20LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.14: All 3 Domains Coarse Top 20 Links

```
CRFClassifier tagged 7218 words in 1 documents at 5710,44 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 1,0000 0,3906 0,5618 50      0      78
Totals 1,0000 0,3906 0,5618 50      0      78

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op20Links\PoliticsCoarseGrainedTop20Links.tsv
```

Figure 3.15: All 3 Domains Coarse Top 20 Links Politics

```
CRFClassifier tagged 6810 words in 1 documents at 5198,47 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 1,0000 1,0000 1,0000 26      0      0
Totals 1,0000 1,0000 1,0000 26      0      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op20Links\SportCoarseGrainedTop20Links.tsv
```

Figure 3.16: All 3 Domains Coarse Top 20 Links Sport

3.2.2 Top 20 Links

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

```
CRFClassifier tagged 4705 words in 1 documents at 5535,29 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,0000 1,0000 0,0000 0      1      0
TRANSPORTATION 1,0000 1,0000 1,0000 12      0      0
Totals 0,9231 1,0000 0,9600 12      1      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op20Links\TransportationCoarseGrainedTop20Links.tsv
```

Figure 3.17: All 3 Domains Coarse Top 20 Links Transportation

3. EXPERIMENTS

```
CRFClassifier tagged 18733 words in 1 documents at 2912,47 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  0,5000  0,6667  1       0       1
Athlete 1,0000  1,0000  1,0000  1       0       0
Coach 1,0000  1,0000  1,0000  1       0       0
Infrastructure 1,0000  0,5000  0,6667  1       0       1
PoliticalParty 1,0000  0,9512  0,9750  39      0       2
Politician 1,0000  1,0000  1,0000  11      0       0
Ship 1,0000  1,0000  1,0000  1       0       0
SpaceShuttle 1,0000  1,0000  1,0000  6       0       0
SpaceStation 1,0000  1,0000  1,0000  1       0       0
SportsClub 1,0000  1,0000  1,0000  10      0       0
SportsEvent 1,0000  1,0000  1,0000  1       0       0
SportsLeague 1,0000  1,0000  1,0000  12      0       0
SportsTeam 1,0000  1,0000  1,0000  2       0       0
Totals 1,0000  0,9560  0,9775  87      0       4

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top20
Links\All3DomainsTop20LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.18: All 3 Domains Fine Top 20 Links

```
CRFClassifier tagged 7218 words in 1 documents at 2440,99 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000  0,0000  0,0000  0       0       32
PoliticalParty 1,0000  0,9512  0,9750  39      0       2
Politician 1,0000  0,2000  0,3333  11      0       44
Totals 1,0000  0,3906  0,5618  50      0       78

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top20
Links\PoliticsFineGrainedTop20Links.tsv
```

Figure 3.19: All 3 Domains Fine Top 20 Links Politics

```
CRFClassifier tagged 6810 words in 1 documents at 2508,29 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000  1,0000  1,0000  1       0       0
Coach 1,0000  1,0000  1,0000  1       0       0
SportsClub 1,0000  1,0000  1,0000  10      0       0
SportsEvent 1,0000  1,0000  1,0000  1       0       0
SportsLeague 1,0000  1,0000  1,0000  12      0       0
SportsTeam 1,0000  1,0000  1,0000  1       0       0
Totals 1,0000  1,0000  1,0000  26      0       0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top20
Links\SportFineGrainedTop20Links.tsv
```

Figure 3.20: All 3 Domains Fine Top 20 Links Sport

```
CRFClassifier tagged 4705 words in 1 documents at 2180,26 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  0,5000  0,6667  1       0       1
Infrastructure 1,0000  0,5000  0,6667  1       0       1
Ship 1,0000  1,0000  1,0000  1       0       0
SpaceShuttle 1,0000  1,0000  1,0000  6       0       0
SpaceStation 1,0000  1,0000  1,0000  1       0       0
SportsTeam 0,0000  1,0000  0,0000  0       1       0
Totals 0,9091  0,8333  0,8696  10      1       2

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-All3DomainsTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top20
Links\TransportationFineGrainedTop20Links.tsv
```

Figure 3.21: All 3 Domains Fine Top 20 Links Transportation

```
CRFClassifier tagged 7218 words in 1 documents at 4903,53 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9921  0,9766  0,9843  125     1       3
Totals 0,9921  0,9766  0,9843  125     1       3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-PoliticsTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top20
Links\PoliticsCoarseGrainedTop20Links.tsv
```

Figure 3.22: Politics Coarse Top 20 Links Politics

3.2. List of experiments

```
CRFClassifier tagged 7218 words in 1 documents at 6153,45 words per second.
Entity P      R      F1      TP      FP      FN
Election 1,0000 0,9688 0,9841 31      0      1
PoliticalParty 1,0000 0,9512 0,9750 39      0      2
Politician 1,0000 1,0000 1,0000 55      0      0
Totals 1,0000 0,9766 0,9881 125     0      3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-PoliticsTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top20Lin
ks\PoliticsFineGrainedTop20Links.tsv
```

Figure 3.23: Politics Fine Top 20 Links Politics

```
CRFClassifier tagged 6810 words in 1 documents at 6418,47 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 1,0000 1,0000 1,0000 26      0      0
Totals 1,0000 1,0000 1,0000 26      0      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-SportTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top20Li
nks\SportCoarseGrainedTop20Links.tsv
```

Figure 3.24: Politics Coarse Top 20 Links Politics

```
CRFClassifier tagged 6810 words in 1 documents at 4518,91 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 1,0000 1,0000 1      0      0
Coach 1,0000 1,0000 1,0000 1      0      0
SportsClub 1,0000 1,0000 1,0000 10     0      0
SportsEvent 1,0000 1,0000 1,0000 1      0      0
SportsLeague 1,0000 1,0000 1,0000 12     0      0
SportsTeam 1,0000 1,0000 1,0000 1      0      0
Totals 1,0000 1,0000 1,0000 26      0      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top20Links\ner-SportTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top20Lin
ks\SportFineGrainedTop20Links.tsv
```

Figure 3.25: Sport Fine Top 20 Links Sport

3.2.3 Top 40 Links

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

```
CRFClassifier tagged 42657 words in 1 documents at 7408,30 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9890 0,9375 0,9626 90      1      6
SPORT 1,0000 1,0000 1,0000 93      0      0
TRANSPORTATION 1,0000 0,9846 0,9922 64     0      1
Totals 0,9960 0,9724 0,9841 247     1      7

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op40Links\All3DomainsTop40LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.26: All 3 Domains Coarse Top 40 Links

3. EXPERIMENTS

```
CRFClassifier tagged 17912 words in 1 documents at 6553,97 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9890  0,3529  0,5202  90      1      165
Totals 0,9890  0,3529  0,5202  90      1      165

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op40Links\PoliticsCoarseGrainedTop40Links.tsv
```

Figure 3.27: All 3 Domains Coarse Top 40 Links Politics

```
CRFClassifier tagged 13802 words in 1 documents at 6470,70 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 1,0000  1,0000  1,0000  91      0      0
Totals 1,0000  1,0000  1,0000  91      0      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op40Links\SportCoarseGrainedTop40Links.tsv
```

Figure 3.28: All 3 Domains Coarse Top 40 Links Sport

```
CRFClassifier tagged 10943 words in 1 documents at 6560,55 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,0000  1,0000  0,0000  0      2      0
TRANSPORTATION 1,0000  0,9846  0,9922  64      0      1
Totals 0,9697  0,9846  0,9771  64      2      1

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\T
op40Links\TransportationCoarseGrainedTop40Links.tsv
```

Figure 3.29: All 3 Domains Coarse Top 40 Links Transportation

```
CRFClassifier tagged 42657 words in 1 documents at 4228,07 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  1,0000  1,0000  16      0      0
Athlete 1,0000  1,0000  1,0000  8      0      0
Coach 1,0000  1,0000  1,0000  1      0      0
Infrastructure 1,0000  1,0000  1,0000  14      0      0
PoliticalParty 0,9863  0,9730  0,9796  72      1      2
Politician 1,0000  0,8182  0,9000  18      0      4
PublicTransitSystem 1,0000  1,0000  1,0000  27      0      0
Ship 1,0000  1,0000  1,0000  1      0      0
SpaceShuttle 1,0000  1,0000  1,0000  6      0      0
SpaceStation 1,0000  1,0000  1,0000  1      0      0
SportsClub 1,0000  1,0000  1,0000  20      0      0
SportsEvent 1,0000  1,0000  1,0000  29      0      0
SportsLeague 1,0000  1,0000  1,0000  23      0      0
SportsTeam 1,0000  1,0000  1,0000  12      0      0
Totals 0,9960  0,9764  0,9861  248      1      6

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top40
Links\All3DomainsTop40LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.30: All 3 Domains Fine Top 40 Links

```
CRFClassifier tagged 17912 words in 1 documents at 3556,79 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000  0,0000  0,0000  0      0      66
PoliticalParty 0,9863  0,9730  0,9796  72      1      2
Politician 1,0000  0,1565  0,2707  18      0      97
Totals 0,9890  0,3529  0,5202  90      1      165

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top40
Links\PoliticsFineGrainedTop40Links.tsv
```

Figure 3.31: All 3 Domains Fine Top 40 Links Politics

3.2. List of experiments

```
CRFClassifier tagged 13802 words in 1 documents at 3215,75 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 1,0000 1,0000 8        0        0
Coach 1,0000 1,0000 1,0000 1        0        0
SportsClub 1,0000 1,0000 1,0000 19       0        0
SportsEvent 1,0000 1,0000 1,0000 29       0        0
SportsLeague 1,0000 1,0000 1,0000 23       0        0
SportsTeam 1,0000 1,0000 1,0000 11       0        0
Totals 1,0000 1,0000 1,0000 91        0        0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top40
Links\SportFineGrainedTop40Links.tsv
```

Figure 3.32: All 3 Domains Fine Top 40 Links Sport

```
CRFClassifier tagged 10943 words in 1 documents at 3228,02 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 1,0000 1,0000 16       0        0
Infrastructure 1,0000 1,0000 1,0000 14       0        0
PublicTransitSystem 1,0000 1,0000 1,0000 27       0        0
Ship 1,0000 1,0000 1,0000 1        0        0
SpaceShuttle 1,0000 1,0000 1,0000 6        0        0
SpaceStation 1,0000 1,0000 1,0000 1        0        0
SportsClub 0,0000 1,0000 0,0000 0        1        0
SportsTeam 0,0000 1,0000 0,0000 0        1        0
Totals 0,9701 1,0000 0,9848 65       2        0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-All3DomainsTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top40
Links\TransportationFineGrainedTop40Links.tsv
```

Figure 3.33: All 3 Domains Fine Top 40 Links Transportation

```
CRFClassifier tagged 17912 words in 1 documents at 6014,78 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9921 0,9804 0,9862 250      2        5
Totals 0,9921 0,9804 0,9862 250      2        5

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-PoliticsTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top4
0Links\PoliticsCoarseGrainedTop40Links.tsv
```

Figure 3.34: Politics Coarse Top 40 Links Politics

```
CRFClassifier tagged 17912 words in 1 documents at 6014,78 words per second.
Entity P      R      F1      TP      FP      FN
Election 1,0000 0,9848 0,9924 65       0        1
PoliticalParty 0,9863 0,9730 0,9796 72       1        2
Politician 1,0000 1,0000 1,0000 115      0        0
Totals 0,9960 0,9882 0,9921 252      1        3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-PoliticsTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top40Li
nks\PoliticsFineGrainedTop40Links.tsv
```

Figure 3.35: Politics Fine Top 40 Links Politics

```
CRFClassifier tagged 13802 words in 1 documents at 5808,92 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 1,0000 1,0000 1,0000 91        0        0
Totals 1,0000 1,0000 1,0000 91        0        0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-SportTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top40Li
nks\SportCoarseGrainedTop40Links.tsv
```

Figure 3.36: Politics Coarse Top 40 Links Politics

3. EXPERIMENTS

```
CRFClassifier tagged 13802 words in 1 documents at 4370,49 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 1,0000 1,0000 8       0       0
Coach 1,0000 1,0000 1,0000 1       0       0
SportsClub 1,0000 0,9474 0,9730 18      0       1
SportsEvent 1,0000 1,0000 1,0000 29      0       0
SportsLeague 1,0000 1,0000 1,0000 23      0       0
SportsTeam 1,0000 1,0000 1,0000 11      0       0
Totals 1,0000 0,9890 0,9945 90      0       1

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top40Links\ner-SportTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top40Links\
SportFineGrainedTop40Links.tsv
```

Figure 3.37: Sport Fine Top 40 Links Sport

```
CRFClassifier tagged 97089 words in 1 documents at 6866,27 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9920 0,9612 0,9764 248     2      10
SPORT 0,9963 0,9926 0,9944 268     1       2
TRANSPORTATION 1,0000 0,9735 0,9865 110     0       3
Totals 0,9952 0,9766 0,9858 626     3      15

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.38: All 3 Domains Coarse Top 100 Links

```
CRFClassifier tagged 40673 words in 1 documents at 8047,69 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9920 0,3615 0,5299 248     2     438
Totals 0,9920 0,3615 0,5299 248     2     438

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top100Links\PoliticsCoarseGrainedTop100Links.tsv
```

Figure 3.39: All 3 Domains Coarse Top 100 Links Politics

TEXT

3.2.4 Top 100 Links

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

```
CRFClassifier tagged 31872 words in 1 documents at 7887,16 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,9962 0,9888 0,9925 265     1       3
Totals 0,9962 0,9888 0,9925 265     1       3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top100Links\SportCoarseGrainedTop100Links.tsv
```

Figure 3.40: All 3 Domains Coarse Top 100 Links Sport

3.2. List of experiments

```
CRFClassifier tagged 24544 words in 1 documents at 7745,03 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,0000  1,0000  0,0000  0       2       0
TRANSPORTATION 1,0000  0,9735  0,9865  110     0       3
Totals 0,9821  0,9735  0,9778  110     2       3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\
Top100Links\TransportationCoarseGrainedTop100Links.tsv
```

Figure 3.41: All 3 Domains Coarse Top 100 Links Transportation

```
CRFClassifier tagged 97089 words in 1 documents at 3538,10 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  0,6957  0,8205  16      0       7
Athlete 1,0000  0,4167  0,5882  15      0      21
Automobile 1,0000  1,0000  1,0000  3       0       0
Coach 1,0000  0,6667  0,8000  2       0       1
Infrastructure 1,0000  1,0000  1,0000  22      0       0
PoliticalParty 0,8774  0,6700  0,7598  136     19      67
Politician 1,0000  0,7455  0,8542  41      0      14
PublicTransitSystem 0,9744  0,7308  0,8352  38      1      14
Ship 1,0000  0,6000  0,7500  3       0       2
SpaceShuttle 1,0000  1,0000  1,0000  6       0       0
SpaceStation 1,0000  1,0000  1,0000  1       0       0
SportsClub 0,9512  0,9398  0,9455  78      4       5
SportsEvent 0,9737  0,8605  0,9136  37      1       6
SportsLeague 0,9500  0,8636  0,9048  57      3       9
SportsManager 1,0000  1,0000  1,0000  6       0       0
SportsTeam 1,0000  0,6364  0,7778  21      0      12
Train 1,0000  1,0000  1,0000  1       0       0
Totals 0,9452  0,7535  0,8385  483     28     158

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.42: All 3 Domains Fine Top 100 Links

```
CRFClassifier tagged 40673 words in 1 documents at 4051,50 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000  0,0000  0,0000  0       0      164
PoliticalParty 0,8774  0,6700  0,7598  136     19      67
Politician 1,0000  0,1285  0,2278  41      0     278
SportsEvent 0,0000  1,0000  0,0000  0       1       0
Totals 0,8985  0,2580  0,4009  177     20     509

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
100Links\PoliticsFineGrainedTop100Links.tsv
```

Figure 3.43: All 3 Domains Fine Top 100 Links Politics

```
CRFClassifier tagged 31872 words in 1 documents at 3906,84 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000  0,4167  0,5882  15      0      21
Coach 1,0000  0,6667  0,8000  2       0       1
SportsClub 0,9506  0,9390  0,9448  77      4       5
SportsEvent 1,0000  0,8605  0,9250  37      0       6
SportsLeague 0,9500  0,8636  0,9048  57      3       9
SportsManager 1,0000  1,0000  1,0000  6       0       0
SportsTeam 1,0000  0,6250  0,7692  20      0      12
Totals 0,9683  0,7985  0,8753  214     7       54

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
100Links\SportFineGrainedTop100Links.tsv
```

Figure 3.44: All 3 Domains Fine Top 100 Links Sport

3. EXPERIMENTS

```
CRFClassifier tagged 24544 words in 1 documents at 3679,76 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 0,6957 0,8205 16      0      7
Automobile 1,0000 1,0000 1,0000 3      0      0
Infrastructure 1,0000 1,0000 1,0000 22     0      0
PublicTransitSystem 0,9744 0,7308 0,8352 38     1      14
Ship 1,0000 0,6000 0,7500 3      0      2
SpaceShuttle 1,0000 1,0000 1,0000 6      0      0
SpaceStation 1,0000 1,0000 1,0000 1      0      0
SportsClub 0,0000 1,0000 0,0000 0      1      0
SportsTeam 0,0000 1,0000 0,0000 0      1      0
Train 1,0000 1,0000 1,0000 1      0      0
Totals 0,9677 0,7965 0,8738 90     3      23

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
100Links\TransportationFineGrainedTop100Links.tsv
```

Figure 3.45: All 3 Domains Fine Top 100 Links Transportation

```
CRFClassifier tagged 40673 words in 1 documents at 7738,39 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9956 0,9898 0,9927 679     3      7
Totals 0,9956 0,9898 0,9927 679     3      7

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-PoliticsTop100LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\To
p100Links\PoliticsCoarseGrainedTop100Links.tsv
```

Figure 3.46: Politics Coarse Top 100 Links Politics

```
CRFClassifier tagged 40673 words in 1 documents at 7513,95 words per second.
Entity P      R      F1      TP      FP      FN
Election 1,0000 0,9878 0,9939 162     0      2
PoliticalParty 0,9950 0,9852 0,9901 200     1      3
Politician 0,9937 0,9906 0,9922 316     2      3
Totals 0,9956 0,9883 0,9920 678     3      8

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-PoliticsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100
Links\PoliticsFineGrainedTop100Links.tsv
```

Figure 3.47: Politics Fine Top 100 Links Politics

TEXT
TEXT

3.2.5 Top 300 Links

TEXT
TEXT
TEXT
TEXT
TEXT
TEXT

```
CRFClassifier tagged 31872 words in 1 documents at 8157,67 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,9963 0,9963 0,9963 267     1      1
Totals 0,9963 0,9963 0,9963 267     1      1

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-SportTop100LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top100
Links\SportCoarseGrainedTop100Links.tsv
```

Figure 3.48: Politics Coarse Top 100 Links Politics

3.2. List of experiments

```
CRFClassifier tagged 31872 words in 1 documents at 5172,35 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000  0,9722  0,9859  35      0      1
Coach 1,0000  1,0000  1,0000  3      0      0
SportsClub 1,0000  0,9878  0,9939  81      0      1
SportsEvent 1,0000  0,9767  0,9882  42      0      1
SportsLeague 1,0000  1,0000  1,0000  66      0      0
SportsManager 1,0000  1,0000  1,0000  6      0      0
SportsTeam 1,0000  1,0000  1,0000  32      0      0
Totals 1,0000  0,9888  0,9944  265     0      3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top100Links\ner-SportTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Lin
ks\SportFineGrainedTop100Links.tsv
```

Figure 3.49: Sport Fine Top 100 Links Sport

```
CRFClassifier tagged 270361 words in 1 documents at 4557,90 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9872  0,9462  0,9662  615     8      35
SPORT 0,9846  0,9629  0,9736  1090    17     42
TRANSPORTATION 0,9940  0,9823  0,9881  500     3      9
Totals 0,9875  0,9625  0,9748  2205    28     86

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top300Links\All3DomainsTop300LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.50: All 3 Domains Coarse Top 300 Links

```
CRFClassifier tagged 107491 words in 1 documents at 5438,73 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9839  0,4025  0,5713  611     10     907
TRANSPORTATION 0,0000  1,0000  0,0000  0      3      0
Totals 0,9792  0,4025  0,5705  611     13     907

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top300Links\PoliticsCoarseGrainedTop300Links.tsv
```

Figure 3.51: All 3 Domains Coarse Top 300 Links Politics

```
CRFClassifier tagged 89224 words in 1 documents at 4834,68 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,0000  1,0000  0,0000  0      2      0
SPORT 0,9846  0,9628  0,9736  1086    17     42
TRANSPORTATION 0,0000  1,0000  0,0000  0      1      0
Totals 0,9819  0,9628  0,9722  1086    20     42

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top300Links\SportCoarseGrainedTop300Links.tsv
```

Figure 3.52: All 3 Domains Coarse Top 300 Links Sport

```
CRFClassifier tagged 73646 words in 1 documents at 4706,42 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,0000  1,0000  0,0000  0      4      0
TRANSPORTATION 0,9940  0,9822  0,9880  496     3      9
Totals 0,9861  0,9822  0,9841  496     7      9

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation
\Top300Links\TransportationCoarseGrainedTop300Links.tsv
```

Figure 3.53: All 3 Domains Coarse Top 300 Links Transportation

3. EXPERIMENTS

```
CRFClassifier tagged 270361 words in 1 documents at 1875,59 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 1,0000 1,0000 102     0      0
Athlete 1,0000 0,9802 0,9900 198     0      4
Automobile 1,0000 1,0000 1,0000 20      0      0
Coach 1,0000 1,0000 1,0000 4        0      0
Infrastructure 1,0000 0,9820 0,9909 100     0      2
PoliticalParty 0,9860 0,9628 0,9743 492     7      19
Politician 1,0000 0,9353 0,9665 130     0      9
PublicTransitSystem 0,9919 0,9839 0,9879 244     2      4
Ship 1,0000 1,0000 1,0000 15      0      0
SpaceShuttle 1,0000 1,0000 1,0000 6        0      0
SpaceStation 1,0000 1,0000 1,0000 1        0      0
SportsClub 0,9796 0,9683 0,9739 336     7      11
SportsEvent 1,0000 0,9242 0,9606 61      0      5
SportsLeague 0,9647 0,9805 0,9725 301     11     6
SportsManager 1,0000 0,9423 0,9703 49      0      3
SportsTeam 1,0000 0,9805 0,9902 151     0      3
Train 1,0000 1,0000 1,0000 6        0      0
Totals 0,9880 0,9712 0,9795 2225     27     66

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
300Links\All3DomainsTop300LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.54: All 3 Domains Fine Top 300 Links

```
CRFClassifier tagged 107491 words in 1 documents at 3895,87 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000 0,0000 0,0000 0        0      322
PoliticalParty 0,9860 0,9628 0,9743 492     7      19
Politician 1,0000 0,1849 0,3120 127     0      560
PublicTransitSystem 0,0000 1,0000 0,0000 0        2      0
Ship 0,0000 1,0000 0,0000 0        1      0
SportsLeague 0,0000 1,0000 0,0000 0        1      0
Totals 0,9825 0,4072 0,5758 619     11     901

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
300Links\PoliticsFineGrainedTop300Links.tsv
```

Figure 3.55: All 3 Domains Fine Top 300 Links Politics

```
CRFClassifier tagged 89224 words in 1 documents at 4185,19 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 0,9802 0,9900 198     0      4
Coach 1,0000 1,0000 1,0000 4        0      0
Politician 0,0000 1,0000 0,0000 0        2      0
SportsClub 0,9794 0,9680 0,9737 333     7      11
SportsEvent 1,0000 0,9242 0,9606 61      0      5
SportsLeague 0,9678 0,9805 0,9741 301     10     6
SportsManager 1,0000 0,9423 0,9703 49      0      3
SportsTeam 1,0000 0,9804 0,9901 150     0      3
Train 1,0000 1,0000 0,0000 0        1      0
Totals 0,9821 0,9716 0,9768 1096     20     32

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
300Links\SportFineGrainedTop300Links.tsv
```

Figure 3.56: All 3 Domains Fine Top 300 Links Sport

3.2. List of experiments

```
CRFClassifier tagged 73646 words in 1 documents at 3681,01 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 1,0000 1,0000 102     0       0
Automobile 1,0000 1,0000 1,0000 20      0       0
Infrastructure 1,0000 0,9820 0,9909 109     0       2
Politician 0,0000 1,0000 0,0000 0        1       0
PublicTransitSystem 0,9918 0,9837 0,9878 242     2       4
Ship 1,0000 1,0000 1,0000 14      0       0
SpaceShuttle 1,0000 1,0000 1,0000 6        0       0
SpaceStation 1,0000 1,0000 1,0000 1        0       0
SportsClub 0,0000 1,0000 0,0000 0        3       0
SportsTeam 0,0000 1,0000 0,0000 0        1       0
Train 1,0000 1,0000 1,0000 5        0       0
Totals 0,9862 0,9881 0,9871 499     7       6

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top
300Links\TransportationFineGrainedTop300Links.tsv
```

Figure 3.57: All 3 Domains Fine Top 300 Links Transportation

```
CRFClassifier tagged 107491 words in 1 documents at 9027,55 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,8039 0,6779 0,7355 1029    251     489
Totals 0,8039 0,6779 0,7355 1029    251     489

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-PoliticsTop300LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\To
p300Links\PoliticsCoarseGrainedTop300Links.tsv
```

Figure 3.58: Politics Coarse Top 300 Links Politics

```
CRFClassifier tagged 107491 words in 1 documents at 6511,05 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,8240 0,6398 0,7203 206     44     116
PoliticalParty 0,8100 0,7006 0,7513 358     84     153
Politician 0,8599 0,7234 0,7858 497     81     190
Totals 0,8354 0,6980 0,7606 1061    209     459

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-PoliticsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300
Links\PoliticsFineGrainedTop300Links.tsv
```

Figure 3.59: Politics Fine Top 300 Links Politics

TEXT
TEXT
TEXT
TEXT
TEXT

3.2.6 Top 400 Links

TEXT
TEXT

```
CRFClassifier tagged 89224 words in 1 documents at 9812,38 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,9432 0,8839 0,9126 997     60     131
Totals 0,9432 0,8839 0,9126 997     60     131

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier
modelsWithOneAnnotation\Top300Links\ner-SportTop300LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top300
Links\SportCoarseGrainedTop300Links.tsv
```

Figure 3.60: Politics Coarse Top 300 Links Politics

3. EXPERIMENTS

```
CRFClassifier tagged 89224 words in 1 documents at 5142,89 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 0,9713  0,8366  0,8989  169     5       33
Coach   1,0000  0,7500  0,8571  3       0       1
SportsClub 0,9453  0,9041  0,9242  311    18      33
SportsEvent 1,0000  0,7879  0,8814  52     0       14
SportsLeague 0,9418  0,8958  0,9182  275    17      32
SportsManager 1,0000  0,9615  0,9804  50     0       2
SportsTeam 0,9845  0,8301  0,9007  127    2       26
Totals 0,9592  0,8750  0,9152  987    42      141

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-SportTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\SportFineGrainedTop300Links.tsv
```

Figure 3.61: Sport Fine Top 300 Links Sport

```
CRFClassifier tagged 351466 words in 1 documents at 3777,09 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9804  0,9434  0,9615  800    16      48
SPORT    0,9832  0,9590  0,9709  1519   26      65
TRANSPORTATION 0,9941  0,9754  0,9847  674    4       17
Totals 0,9849  0,9584  0,9714  2993   46     130

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\All3DomainsTop400LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.62: All 3 Domains Coarse Top 400 Links

```
CRFClassifier tagged 137620 words in 1 documents at 5615,08 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9754  0,4082  0,5756  792    20     1148
SPORT    0,0000  1,0000  0,0000  0       4       0
TRANSPORTATION 0,0000  1,0000  0,0000  0      15       0
Totals 0,9531  0,4082  0,5716  792    39     1148

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\PoliticsCoarseGrainedTop400Links.tsv
```

Figure 3.63: All 3 Domains Coarse Top 400 Links Politics

TEXT
TEXT
TEXT
TEXT
TEXT
TEXT
TEXT
TEXT
TEXT

```
CRFClassifier tagged 116878 words in 1 documents at 4957,08 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,0000  1,0000  0,0000  0       3       0
SPORT    0,9837  0,9588  0,9711  1511   25      65
TRANSPORTATION 0,0000  1,0000  0,0000  0       2       0
Totals 0,9805  0,9588  0,9695  1511   30      65

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\SportCoarseGrainedTop400Links.tsv
```

Figure 3.64: All 3 Domains Coarse Top 400 Links Sport

3.2. List of experiments

```
CRFClassifier tagged 96968 words in 1 documents at 6509,67 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,0000  1,0000  0,0000  0       1       0
SPORT 0,0000  1,0000  0,0000  0       5       0
TRANSPORTATION 0,9939  0,9762  0,9850  657     4       16
Totals 0,9850  0,9762  0,9806  657     10      16

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\TransportationCoarseGrainedTop400Links.tsv
```

Figure 3.65: All 3 Domains Coarse Top 400 Links Transportation

```
CRFClassifier tagged 351466 words in 1 documents at 2788,70 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  1,0000  1,0000  123     0       0
Athlete 1,0000  0,9899  0,9949  294     0       3
Automobile 1,0000  1,0000  1,0000  24      0       0
Coach 1,0000  1,0000  1,0000  4       0       0
Infrastructure 1,0000  0,9896  0,9948  190     0       2
PoliticalParty 0,9766  0,9486  0,9624  627     15      34
Politician 1,0000  0,9893  0,9946  185     0       2
PublicTransitSystem 0,9935  0,9776  0,9855  305     2       7
Ship 1,0000  0,9231  0,9600  24      0       2
SpaceShuttle 1,0000  1,0000  1,0000  6       0       0
SpaceStation 1,0000  1,0000  1,0000  2       0       0
SportsClub 0,9796  0,9658  0,9726  480     10      17
SportsEvent 1,0000  0,8636  0,9268  76      0       12
SportsLeague 0,9698  0,9835  0,9766  417     13       7
SportsManager 1,0000  0,9726  0,9861  71      0       2
SportsTeam 1,0000  0,9851  0,9925  198     0       3
Train 1,0000  1,0000  1,0000  6       0       0
Totals 0,9870  0,9709  0,9789  3032    40      91

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\All3DomainsTop400LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.66: All 3 Domains Fine Top 400 Links

```
CRFClassifier tagged 137620 words in 1 documents at 2890,33 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000  1,0000  0,0000  0       1       0
Election 0,0000  0,0000  0,0000  0       0       415
PoliticalParty 0,9766  0,9484  0,9623  625     15      34
Politician 1,0000  0,2092  0,3460  182     0       688
PublicTransitSystem 0,0000  1,0000  0,0000  0       4       0
Ship 0,0000  1,0000  0,0000  0       9       0
SportsClub 0,0000  1,0000  0,0000  0       1       0
SportsLeague 0,0000  1,0000  0,0000  0       2       0
Totals 0,9619  0,4151  0,5799  807     32     1137

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\PoliticsFineGrainedTop400Links.tsv
```

Figure 3.67: All 3 Domains Fine Top 400 Links Politics

```
CRFClassifier tagged 116878 words in 1 documents at 3423,29 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000  1,0000  0,0000  0       1       0
Athlete 1,0000  0,9899  0,9949  294     0       3
Coach 1,0000  1,0000  1,0000  4       0       0
PoliticalParty 0,0000  1,0000  0,0000  0       1       0
Politician 0,0000  1,0000  0,0000  0       2       0
SportsClub 0,9794  0,9654  0,9724  475     10      17
SportsEvent 1,0000  0,8636  0,9268  76      0       12
SportsLeague 0,9696  0,9834  0,9765  415     13       7
SportsManager 1,0000  0,9726  0,9861  71      0       2
SportsTeam 1,0000  0,9850  0,9924  197     0       3
Train 0,0000  1,0000  0,0000  0       1       0
Totals 0,9821  0,9721  0,9770  1532    28      44

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\SportFineGrainedTop400Links.tsv
```

Figure 3.68: All 3 Domains Fine Top 400 Links Sport

3. EXPERIMENTS

```
CRFClassifier tagged 96968 words in 1 documents at 3314,01 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 1,0000 1,0000 121     0       0
Automobile 1,0000 1,0000 1,0000 24      0       0
Infrastructure 1,0000 0,9896 0,9948 190     0       2
PoliticalParty 0,0000 1,0000 0,0000 0        1       0
Politician 0,0000 1,0000 0,0000 0        1       0
PublicTransitSystem 0,9934 0,9773 0,9853 301     2       7
Ship 1,0000 1,0000 1,0000 15      0       0
SpaceShuttle 1,0000 1,0000 1,0000 6        0       0
SpaceStation 1,0000 1,0000 1,0000 2        0       0
SportsClub 0,0000 1,0000 0,0000 0        4       0
SportsTeam 0,0000 1,0000 0,0000 0        1       0
Train 1,0000 1,0000 1,0000 5        0       0
Totals 0,9866 0,9866 0,9866 664     9       9

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-All3DomainsTop400LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\TransportationFineGrainedTop400Links.tsv
```

Figure 3.69: All 3 Domains Fine Top 400 Links Transportation

```
CRFClassifier tagged 137620 words in 1 documents at 6639,01 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9866 0,9479 0,9669 1839    25     101
Totals 0,9866 0,9479 0,9669 1839    25     101

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-PoliticsTop400LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\PoliticsCoarseGrainedTop400Links.tsv
```

Figure 3.70: Politics Coarse Top 400 Links Politics

```
CRFClassifier tagged 137620 words in 1 documents at 6109,93 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,9975 0,9590 0,9779 398     1      17
PoliticalParty 0,9767 0,9530 0,9647 628     15     31
Politician 0,9977 0,9920 0,9948 863     2       7
Totals 0,9906 0,9717 0,9810 1889    18     55

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-PoliticsTop400LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\PoliticsFineGrainedTop400Links.tsv
```

Figure 3.71: Politics Fine Top 400 Links Politics

```
CRFClassifier tagged 116878 words in 1 documents at 7155,50 words per second.
Entity P      R      F1      TP      FP      FN
SPORT 0,9858 0,9676 0,9766 1525    22     51
Totals 0,9858 0,9676 0,9766 1525    22     51

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-SportTop400LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\SportCoarseGrainedTop400Links.tsv
```

Figure 3.72: Politics Coarse Top 400 Links Politics

```
CRFClassifier tagged 116878 words in 1 documents at 4135,52 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 0,9731 0,9863 289     0       8
Coach 1,0000 1,0000 1,0000 4        0       0
SportsClub 0,9815 0,9715 0,9765 478     9      14
SportsEvent 1,0000 0,9091 0,9524 80      0       8
SportsLeague 0,9718 0,9787 0,9752 413     12     9
SportsManager 1,0000 0,9726 0,9861 71      0       2
SportsTeam 1,0000 0,9900 0,9950 198     0       2
Totals 0,9865 0,9727 0,9796 1533    21     43

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top400Links\ner-SportTop400LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top400Links\SportFineGrainedTop400Links.tsv
```

Figure 3.73: Sport Fine Top 400 Links Sport

3.2. List of experiments

```
CRFClassifier tagged 423454 words in 1 documents at 4057,94 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9788 0,9444 0,9613 969    21     57
SPORT    0,9850 0,9596 0,9721 1899   29     80
TRANSPORTATION 0,9962 0,9750 0,9855 779    3     20
Totals   0,9857 0,9587 0,9720 3647   53     157

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractCoarseGrainedSpecified.tsv
```

Figure 3.74: All 3 Domains Coarse Top 500 Links

```
CRFClassifier tagged 164060 words in 1 documents at 7056,04 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,9734 0,4095 0,5765 953    26    1374
SPORT    0,0000 1,0000 0,0000 0       4       0
TRANSPORTATION 0,0000 1,0000 0,0000 0      15       0
Totals   0,9549 0,4095 0,5732 953    45    1374

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsCoarseGrainedTop500Links.tsv
```

Figure 3.75: All 3 Domains Coarse Top 500 Links Politics

```
CRFClassifier tagged 142322 words in 1 documents at 8310,77 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,0000 1,0000 0,0000 0       9       0
SPORT    0,9849 0,9594 0,9720 1890   29     80
TRANSPORTATION 0,0000 1,0000 0,0000 0       3       0
Totals   0,9788 0,9594 0,9690 1890   41     80

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportCoarseGrainedTop500Links.tsv
```

Figure 3.76: All 3 Domains Coarse Top 500 Links Sport

3.2.7 Top 500 Links

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

TEXT

```
CRFClassifier tagged 117072 words in 1 documents at 6921,19 words per second.
Entity P      R      F1      TP      FP      FN
POLITICS 0,0000 1,0000 0,0000 0       2       0
SPORT    0,0000 1,0000 0,0000 0       5       0
TRANSPORTATION 0,9961 0,9769 0,9864 761    3     18
Totals   0,9870 0,9769 0,9819 761    10     18

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\TransportationCoarseGrainedTop500Links.tsv
```

Figure 3.77: All 3 Domains Coarse Top 500 Links Transportation

3. EXPERIMENTS

```
CRFClassifier tagged 423454 words in 1 documents at 2276,66 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 0,9929 0,9964 139    0      1
Athlete 1,0000 0,9896 0,9948 379    0      4
Automobile 1,0000 1,0000 1,0000 25     0      0
Coach 1,0000 1,0000 1,0000 6      0      0
Infrastructure 1,0000 0,9783 0,9890 225    0      5
PoliticalParty 0,9775 0,9403 0,9585 740    17     47
Politician 1,0000 0,9874 0,9937 236    0      3
PublicTransitSystem 0,9944 0,9807 0,9875 355    2      7
Ship 1,0000 0,9259 0,9615 25     0      2
SpaceShuttle 1,0000 1,0000 1,0000 7      0      0
SpaceStation 1,0000 1,0000 1,0000 2      0      0
SportsClub 0,9756 0,9553 0,9654 599    15     28
SportsEvent 1,0000 0,8796 0,9360 95     0      13
SportsLeague 0,9700 0,9810 0,9755 517    16     10
SportsManager 1,0000 0,9780 0,9889 89     0      2
SportsTeam 1,0000 0,9831 0,9915 233    0      4
Train 1,0000 1,0000 1,0000 6      0      0
Totals 0,9866 0,9669 0,9766 3678 50     126

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.78: All 3 Domains Fine Top 500 Links

```
CRFClassifier tagged 164060 words in 1 documents at 3562,26 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 1,0000 0,0000 0      1      0
Election 0,0000 0,0000 0,0000 0      0     494
PoliticalParty 0,9774 0,9400 0,9583 736    17     47
Politician 1,0000 0,2171 0,3567 229    0     826
PublicTransitSystem 0,0000 1,0000 0,0000 0      4      0
Ship 0,0000 1,0000 0,0000 0      9      0
SpaceShuttle 0,0000 1,0000 0,0000 0      1      0
SportsClub 0,0000 1,0000 0,0000 0      2      0
SportsLeague 0,0000 1,0000 0,0000 0      3      0
Totals 0,9631 0,4138 0,5789 965 37 1367

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsFineGrainedTop500Links.tsv
```

Figure 3.79: All 3 Domains Fine Top 500 Links Politics

```
CRFClassifier tagged 142322 words in 1 documents at 3839,48 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 1,0000 0,0000 0      2      0
Athlete 1,0000 0,9896 0,9948 379    0      4
Coach 1,0000 1,0000 1,0000 6      0      0
PoliticalParty 0,0000 1,0000 0,0000 0      3      0
Politician 0,0000 1,0000 0,0000 0      6      0
SportsClub 0,9753 0,9549 0,9650 593    15     28
SportsEvent 1,0000 0,8796 0,9360 95     0      13
SportsLeague 0,9717 0,9810 0,9763 515    15     10
SportsManager 1,0000 0,9780 0,9889 89     0      2
SportsTeam 1,0000 0,9831 0,9915 232    0      4
Train 0,0000 1,0000 0,0000 0      1      0
Totals 0,9785 0,9690 0,9737 1909 42 61

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportFineGrainedTop500Links.tsv
```

Figure 3.80: All 3 Domains Fine Top 500 Links Sport

3.2. List of experiments

```
CRFClassifier tagged 117072 words in 1 documents at 3763,28 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 0,9927 0,9963 136 0 1
Automobile 1,0000 1,0000 1,0000 25 0 0
Infrastructure 1,0000 0,9783 0,9890 225 0 5
PoliticalParty 0,0000 1,0000 0,0000 0 1 0
Politician 0,0000 1,0000 0,0000 0 1 0
PublicTransitSystem 0,9943 0,9804 0,9873 351 2 7
Ship 1,0000 1,0000 1,0000 16 0 0
SpaceShuttle 1,0000 1,0000 1,0000 6 0 0
SpaceStation 1,0000 1,0000 1,0000 2 0 0
SportsClub 0,0000 1,0000 0,0000 0 4 0
SportsTeam 0,0000 1,0000 0,0000 0 1 0
Train 1,0000 1,0000 1,0000 5 0 0
Totals 0,9884 0,9833 0,9858 766 9 13

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\TransportationFineGrainedTop500Links.tsv
```

Figure 3.81: All 3 Domains Fine Top 500 Links Transportation

```
CRFClassifier tagged 164060 words in 1 documents at 7597,48 words per second.
Entity P R F1 TP FP FN
POLITICS 0,9808 0,9450 0,9626 2199 43 128
Totals 0,9808 0,9450 0,9626 2199 43 128

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-PoliticsTop500LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsCoarseGrainedTop500Links.tsv
```

Figure 3.82: Politics Coarse Top 500 Links Politics

```
CRFClassifier tagged 164060 words in 1 documents at 5517,59 words per second.
Entity P R F1 TP FP FN
Election 0,9915 0,9393 0,9647 464 4 30
PoliticalParty 0,9777 0,9502 0,9637 744 17 39
Politician 0,9962 0,9877 0,9919 1042 4 13
Totals 0,9890 0,9648 0,9768 2250 25 82

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-PoliticsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsFineGrainedTop500Links.tsv
```

Figure 3.83: Politics Fine Top 500 Links Politics

```
CRFClassifier tagged 142322 words in 1 documents at 8207,73 words per second.
Entity P R F1 TP FP FN
SPORT 0,9856 0,9706 0,9780 1912 28 58
Totals 0,9856 0,9706 0,9780 1912 28 58

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-SportTop500LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportCoarseGrainedTop500Links.tsv
```

Figure 3.84: Politics Coarse Top 500 Links Politics

```
CRFClassifier tagged 142322 words in 1 documents at 5008,69 words per second.
Entity P R F1 TP FP FN
Athlete 1,0000 0,9791 0,9894 375 0 8
Coach 1,0000 1,0000 1,0000 6 0 0
SportsClub 0,9771 0,9630 0,9700 598 14 23
SportsEvent 1,0000 0,9074 0,9515 98 0 10
SportsLeague 0,9755 0,9848 0,9801 517 13 8
SportsManager 1,0000 0,9780 0,9889 89 0 2
SportsTeam 1,0000 0,9915 0,9957 234 0 2
Totals 0,9861 0,9731 0,9796 1917 27 53

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-SportTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportFineGrainedTop500Links.tsv
```

Figure 3.85: Sport Fine Top 500 Links Sport

3. EXPERIMENTS

```
CRFClassifier tagged 760067 words in 2 documents at 3091,92 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,9242  0,5755  0,7093  122    10     90
Athlete  0,8182  0,3778  0,5169  252    56     415
Automobile 0,9565  0,3607  0,5238  22     1     39
Coach    1,0000  0,2000  0,3333  4       0     16
Infrastructure 0,8457  0,2937  0,4359  148    27    356
Locomotive 0,0000  0,0000  0,0000  0       0     2
Motorcycle 0,0000  0,0000  0,0000  0       0     1
OrganisationMember 0,0000  0,0000  0,0000  0       0     1
PoliticalParty 0,7656  0,5819  0,6613  735    225    528
Politician 0,8925  0,4099  0,5618  166    20     239
PublicTransitSystem 0,8291  0,6178  0,7080  388    80     240
Rocket    0,0000  0,0000  0,0000  0       0     5
Ship      0,9375  0,3409  0,5000  15     1     29
SpaceShuttle 1,0000  0,3750  0,5455  6       0     10
SpaceStation 1,0000  0,3333  0,5000  1       0     2
SportsClub 0,8009  0,4276  0,5575  543    135    727
SportsEvent 0,9559  0,3171  0,4762  65     3     140
SportsLeague 0,8071  0,5912  0,6824  548    131    379
SportsManager 0,9643  0,2903  0,4463  54     2     132
SportsTeam 0,8856  0,5838  0,7037  209    27     149
Train     1,0000  0,5455  0,7059  6       0     5
Totals    0,8206  0,4837  0,6087  3284   718    3505

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFiles domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv,domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecifiedWithLowerPageRank.tsv
```

Figure 3.86: All 3 Domains Fine Grained Top 300 With All 3 Domains Fine Grained Top 500 Links And All 3 Domains Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 760067 words in 2 documents at 3060,01 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,9735  0,6934  0,8099  147     4     65
Athlete  0,9101  0,6222  0,7391  415    41    252
Automobile 1,0000  0,4098  0,5814  25     0     36
Coach    1,0000  0,3000  0,4615  6       0     14
Infrastructure 0,8885  0,5218  0,6575  263    33    241
Locomotive 0,0000  0,0000  0,0000  0       0     2
Motorcycle 0,0000  0,0000  0,0000  0       0     1
OrganisationMember 0,0000  0,0000  0,0000  0       0     1
PoliticalParty 0,8393  0,7403  0,7867  935    179    328
Politician 0,9271  0,6593  0,7706  267    21     138
PublicTransitSystem 0,9027  0,7389  0,8126  464    50     164
Rocket    0,0000  0,0000  0,0000  0       0     5
Ship      0,9615  0,5682  0,7143  25     1     19
SpaceShuttle 1,0000  0,4375  0,6087  7       0     9
SpaceStation 1,0000  0,6667  0,8000  2       0     1
SportsClub 0,8722  0,6071  0,7159  771    113    499
SportsEvent 0,9000  0,4829  0,6286  99     11     106
SportsLeague 0,8622  0,7357  0,7939  682    109    245
SportsManager 0,9787  0,4946  0,6571  92     2     94
SportsTeam 0,9276  0,7514  0,8302  269    21     89
Train     1,0000  0,5455  0,7059  6       0     5
Totals    0,8844  0,6592  0,7553  4475   585    2314

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFiles domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv,domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecifiedWithLowerPageRank.tsv
```

Figure 3.87: All 3 Domains Fine Grained Top 500 Links With All 3 Domains Fine Grained Top 500 Links And All 3 Domains Fine Grained Top 500 Links With Lower Page Rank

3.2.8 MIXED

3.2.9 With lower links on model

3.2.10 With higher links on model

3.2. List of experiments

```
CRFClassifier tagged 336613 words in 1 documents at 2685,15 words per second.
Entity P R F1 TP FP FN
Aircraft 0,6667 0,1111 0,1905 8 4 64
Athlete 0,4675 0,1268 0,1994 36 41 248
Automobile 0,0000 0,0000 0,0000 0 0 36
Coach 0,0000 0,0000 0,0000 0 0 14
Infrastructure 0,5352 0,1387 0,2203 38 33 236
Locomotive 0,0000 0,0000 0,0000 0 0 2
Motorcycle 0,0000 0,0000 0,0000 0 0 1
OrganisationMember 0,0000 0,0000 0,0000 0 0 1
PoliticalParty 0,5462 0,4097 0,4682 195 162 281
Politician 0,5962 0,1867 0,2844 31 21 135
PublicTransitSystem 0,6943 0,4098 0,5154 109 48 157
Rocket 0,0000 0,0000 0,0000 0 0 5
Ship 0,0000 0,0000 0,0000 0 1 17
SpaceShuttle 0,0000 0,0000 0,0000 0 0 9
SpaceStation 0,0000 0,0000 0,0000 0 0 1
SportsClub 0,6370 0,2675 0,3768 172 98 471
SportsEvent 0,2667 0,0412 0,0714 4 11 93
SportsLeague 0,6395 0,4125 0,5015 165 93 235
SportsManager 0,6000 0,0316 0,0600 3 2 92
SportsTeam 0,6316 0,2975 0,4045 36 21 85
Train 0,0000 0,0000 0,0000 0 0 5
Totals 0,5983 0,2670 0,3692 797 535 2188

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecifiedWithLowerPageRank.tsv
```

Figure 3.88: All 3 Domains Fine Grained Top 500 Links With All 3 Domains Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 344007 words in 2 documents at 4801,01 words per second.
Entity P R F1 TP FP FN
Aircraft 0,9950 0,9706 0,9826 198 1 6
Athlete 0,0000 0,0000 0,0000 0 0 202
Automobile 1,0000 0,8500 0,9180 34 0 6
Coach 0,0000 0,0000 0,0000 0 0 4
Infrastructure 1,0000 0,9820 0,9909 218 0 4
PoliticalParty 0,0000 0,0000 0,0000 0 0 511
Politician 0,0000 0,0000 0,0000 0 0 139
PublicTransitSystem 0,9835 0,9676 0,9755 478 8 16
Ship 1,0000 0,9655 0,9825 28 0 1
SpaceShuttle 1,0000 0,6667 0,8000 8 0 4
SpaceStation 1,0000 1,0000 1,0000 2 0 0
SportsClub 0,0000 0,0000 0,0000 0 0 347
SportsEvent 0,0000 0,0000 0,0000 0 0 66
SportsLeague 0,0000 0,0000 0,0000 0 0 307
SportsManager 0,0000 0,0000 0,0000 0 0 52
SportsTeam 0,0000 0,0000 0,0000 0 0 154
Train 1,0000 0,9091 0,9524 10 0 1
Totals 0,9909 0,3491 0,5163 976 9 1820

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-TransportationTop500LinksFineGrained.ser.gz -testFiles domainsWithOneAnnotation\Top300Links\All3DomainsTop300LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv,domainsWithOneAnnotation\Top300Links\TransportationFineGrainedTop300Links.tsv
```

Figure 3.89: Transportation Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 300 Links And Transportation Fine Grained Top 300 Links

3. EXPERIMENTS

```
CRFClassifier tagged 97089 words in 1 documents at 2905,12 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 0,0435 0,0833 1       0       22
Athlete 1,0000 0,0278 0,0541 1       0       35
Automobile 0,0000 0,0000 0,0000 0       0       3
Coach 1,0000 0,3333 0,5000 1       0       2
Infrastructure 0,0000 0,0000 0,0000 0       0       22
PoliticalParty 0,6458 0,3054 0,4147 62      34      141
Politician 1,0000 0,0727 0,1356 4       0       51
PublicTransitSystem 0,8750 0,1346 0,2333 7       1       45
Ship 1,0000 0,2000 0,3333 1       0       4
SpaceShuttle 0,0000 0,0000 0,0000 0       0       6
SpaceStation 0,0000 0,0000 0,0000 0       0       1
SportsClub 0,9000 0,1084 0,1935 9       1       74
SportsEvent 1,0000 0,0233 0,0455 1       0       42
SportsLeague 0,6667 0,2121 0,3218 14      7       52
SportsManager 0,0000 0,0000 0,0000 0       0       6
SportsTeam 1,0000 0,0303 0,0588 1       0       32
Train 0,0000 0,0000 0,0000 0       0       1
Totals 0,7034 0,1591 0,2595 102      43      539

C:\Dev\stanford-ner-2017-06-09>java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.90: All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 100 Links

```
CRFClassifier tagged 270361 words in 1 documents at 3542,70 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 0,0000 0,0194 1       0       101
Athlete 1,0000 0,0050 0,0099 1       0       201
Automobile 0,0000 0,0000 0,0000 0       0       20
Coach 1,0000 0,2500 0,4000 1       0       3
Infrastructure 0,0000 0,0000 0,0000 0       0       111
PoliticalParty 0,5517 0,2192 0,3137 112     91      399
Politician 0,8000 0,0288 0,0556 4       1       135
PublicTransitSystem 0,6364 0,0282 0,0541 7       4       241
Ship 1,0000 0,0667 0,1250 1       0       14
SpaceShuttle 0,0000 0,0000 0,0000 0       0       6
SpaceStation 0,0000 0,0000 0,0000 0       0       1
SportsClub 0,8750 0,0403 0,0771 14      2       333
SportsEvent 1,0000 0,0152 0,0299 1       0       65
SportsLeague 0,6102 0,1173 0,1967 36      23      271
SportsManager 0,0000 0,0000 0,0000 0       0       52
SportsTeam 1,0000 0,0065 0,0129 1       0       153
Train 0,0000 0,0000 0,0000 0       0       6
Totals 0,5967 0,0781 0,1382 179      121     2112

C:\Dev\stanford-ner-2017-06-09>java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\All3DomainsTop300LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.91: All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 300 Links

```
CRFClassifier tagged 423454 words in 1 documents at 3197,54 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 0,0071 0,0142 1       0       139
Athlete 1,0000 0,0026 0,0052 1       0       382
Automobile 0,0000 0,0000 0,0000 0       0       25
Coach 1,0000 0,1667 0,2857 1       0       5
Infrastructure 0,0000 0,0000 0,0000 0       0       230
PoliticalParty 0,5385 0,1957 0,2870 154     132     633
Politician 0,8000 0,0167 0,0328 4       1       235
PublicTransitSystem 0,4118 0,0193 0,0369 7       10      355
Ship 1,0000 0,0370 0,0714 1       0       26
SpaceShuttle 0,0000 0,0000 0,0000 0       0       7
SpaceStation 0,0000 0,0000 0,0000 0       0       2
SportsClub 0,8148 0,0351 0,0673 22      5       605
SportsEvent 1,0000 0,0093 0,0183 1       0       107
SportsLeague 0,5556 0,0854 0,1480 45      36      482
SportsManager 0,0000 0,0000 0,0000 0       0       91
SportsTeam 1,0000 0,0295 0,0574 7       0       230
Train 0,0000 0,0000 0,0000 0       0       6
Totals 0,5701 0,0641 0,1153 244      184     3560

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top10Links\ner-All3DomainsTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.92: All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 500 Links

3.2. List of experiments

```
CRFClassifier tagged 336613 words in 1 documents at 2416,32 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 0,0000 0,0000 0      0      72
Athlete 0,0000 0,0000 0,0000 0      0      284
Automobile 0,0000 0,0000 0,0000 0      0      36
Coach 0,0000 0,0000 0,0000 0      0      14
Infrastructure 0,0000 0,0000 0,0000 0      0      274
Locomotive 0,0000 0,0000 0,0000 0      0      2
Motorcycle 0,0000 0,0000 0,0000 0      0      1
OrganisationMember 0,0000 0,0000 0,0000 0      0      1
PoliticalParty 0,3916 0,1176 0,1809 56      87      420
Politician 0,0000 0,0000 0,0000 0      0      166
PublicTransitSystem 0,0000 0,0000 0,0000 0      0      10
Rocket 0,0000 0,0000 0,0000 0      0      5
Ship 0,0000 0,0000 0,0000 0      0      17
SpaceShuttle 0,0000 0,0000 0,0000 0      0      9
SpaceStation 0,0000 0,0000 0,0000 0      0      1
SportsClub 0,5455 0,0187 0,0361 12      10      631
SportsEvent 0,0000 0,0000 0,0000 0      0      97
SportsLeague 0,4348 0,0500 0,0897 20      26      380
SportsManager 0,0000 0,0000 0,0000 0      0      95
SportsTeam 1,0000 0,0331 0,0640 4      0      117
Train 0,0000 0,0000 0,0000 0      0      5
Totals 0,4089 0,0308 0,0573 92      133      2893

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecifiedWithLowerPageRank.tsv
```

Figure 3.93: All 3 Domains Fine Grained Top 10 Links Runned With All 3 Domains Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 423454 words in 1 documents at 2649,15 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,9677 0,2143 0,3509 30      1      110
Athlete 1,0000 0,0392 0,0754 15      0      368
Automobile 1,0000 0,1600 0,2759 4      0      21
Coach 1,0000 0,3333 0,5000 2      0      4
Infrastructure 0,8065 0,1087 0,1916 25      6      205
PoliticalParty 0,7288 0,4473 0,5543 352      131      435
Politician 0,8851 0,3222 0,4724 77      10      162
PublicTransitSystem 0,8217 0,3564 0,4971 129      28      233
Ship 1,0000 0,1111 0,2000 3      0      24
SpaceShuttle 1,0000 0,8571 0,9231 6      0      1
SpaceStation 1,0000 0,5000 0,6667 1      0      1
SportsClub 0,8193 0,3254 0,4658 204      45      423
SportsEvent 0,9048 0,3519 0,5067 38      4      70
SportsLeague 0,8392 0,4953 0,6229 261      50      266
SportsManager 1,0000 0,0769 0,1429 7      0      84
SportsTeam 1,0000 0,1814 0,3071 43      0      194
Train 1,0000 0,5000 0,6667 3      0      3
Totals 0,8136 0,3155 0,4546 1200      275      2604

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.94: All 3 Domains Fine Grained Top 100 Links Runned With All 3 Domains Fine Grained Top 500 Links

3. EXPERIMENTS

```
CRFClassifier tagged 336613 words in 1 documents at 3519,92 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 0,0000 0,0000 0      1      72
Athlete 0,0000 0,0000 0,0000 0      0      284
Automobile 0,0000 0,0000 0,0000 0      0      36
Coach 0,0000 0,0000 0,0000 0      0      14
Infrastructure 0,0000 0,0000 0,0000 0      7      274
Locomotive 0,0000 0,0000 0,0000 0      0      2
Motorcycle 0,0000 0,0000 0,0000 0      0      1
OrganisationMember 0,0000 0,0000 0,0000 0      0      1
PoliticalParty 0,5022 0,2374 0,3224 113    112    363
Politician 0,5517 0,0964 0,1641 16     13     150
PublicTransitSystem 0,7255 0,1391 0,2334 37     14     229
Rocket 0,0000 0,0000 0,0000 0      0      5
Ship 0,0000 0,0000 0,0000 0      0      17
SpaceShuttle 0,0000 0,0000 0,0000 0      0      9
SpaceStation 0,0000 0,0000 0,0000 0      0      1
SportsClub 0,6093 0,1431 0,2317 92     59     551
SportsEvent 0,1667 0,0103 0,0194 1      5      96
SportsLeague 0,6092 0,2650 0,3693 106    68     294
SportsManager 0,0000 0,0000 0,0000 0      1      95
SportsTeam 0,8571 0,0496 0,0937 6      1      115
Train 0,0000 0,0000 0,0000 0      0      5
Totals 0,5690 0,1243 0,2040 371    281    2614

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecifiedWithLowerPageRank.tsv
```

Figure 3.95: All 3 Domains Fine Grained Top 100 Links Runned With All 3 Domains Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 164060 words in 1 documents at 3254,58 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000 0,0000 0,0000 0      0      494
PoliticalParty 0,7282 0,4483 0,5549 351    131    432
Politician 0,9398 0,0739 0,1371 78     5      977
SportsEvent 0,0000 1,0000 0,0000 0      1      0
SportsLeague 0,0000 1,0000 0,0000 0      5      0
Totals 0,7513 0,1840 0,2956 429    142    1903

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsFineGrainedTop500Links.tsv
```

Figure 3.96: All 3 Domains Fine Grained Top 100 Links Runned With Politics Fine Grained Top 500 Links

```
CRFClassifier tagged 117053 words in 1 documents at 3034,66 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000 0,0000 0,0000 0      0      197
Infrastructure 0,0000 1,0000 0,0000 0      2      0
PoliticalParty 0,5045 0,2379 0,3233 113    111    362
Politician 0,6667 0,0245 0,0473 16     8      636
PublicTransitSystem 0,0000 1,0000 0,0000 0      3      0
SportsEvent 0,0000 1,0000 0,0000 0      2      0
SportsLeague 0,0000 1,0000 0,0000 0      3      0
Totals 0,5000 0,0974 0,1631 129    129    1195

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsFineGrainedTop500LinksWithLowerPageRank.tsv
```

Figure 3.97: All 3 Domains Fine Grained Top 100 Links Runned With Politics Fine Grained Top 500 Links With Lower PageRank

3.2. List of experiments

```
CRFClassifier tagged 142322 words in 1 documents at 3035,30 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 0,0392 0,0754 15      0      368
Coach 1,0000 0,3333 0,5000 2       0      4
Infrastructure 0,0000 1,0000 0,0000 0       1      0
PoliticalParty 0,0000 1,0000 0,0000 0       1      0
Politician 0,0000 1,0000 0,0000 0       1      0
SportsClub 0,8185 0,3269 0,4672 203     45     418
SportsEvent 0,9268 0,3519 0,5101 38      3      70
SportsLeague 0,8497 0,4952 0,6258 260     46     265
SportsManager 1,0000 0,0769 0,1429 7       0      84
SportsTeam 1,0000 0,1780 0,3022 42      0      194
Totals 0,8539 0,2878 0,4305 567     97     1403

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportFineGrainedTop500Links.tsv
```

Figure 3.98: All 3 Domains Fine Grained Top 100 Links Runned With Sport Fine Grained Top 500 Links

```
CRFClassifier tagged 118469 words in 1 documents at 3761,28 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 0,0000 0,0000 0,0000 0       0      271
Coach 0,0000 0,0000 0,0000 0       0      14
Infrastructure 0,0000 1,0000 0,0000 0       1      0
OrganisationMember 0,0000 0,0000 0,0000 0       0      1
Politician 0,0000 1,0000 0,0000 0       4      0
SportsClub 0,6216 0,1467 0,2374 92      56     535
SportsEvent 0,2500 0,0104 0,0200 1       3      95
SportsLeague 0,6082 0,2620 0,3662 104     67     293
SportsManager 0,0000 0,0000 0,0000 0       1      89
SportsTeam 0,8571 0,0504 0,0952 6       1      113
Totals 0,6042 0,1258 0,2082 203     133     1411

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportFineGrainedTop500LinksLowerPageRank.tsv
```

Figure 3.99: All 3 Domains Fine Grained Top 100 Links Runned With Sport Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 117072 words in 1 documents at 3791,68 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,9677 0,2190 0,3571 30      1      107
Automobile 1,0000 0,1600 0,2759 4       0      21
Infrastructure 0,8333 0,1087 0,1923 25      5      205
Politician 0,0000 1,0000 0,0000 0       3      0
PublicTransitSystem 0,8217 0,3603 0,5010 129     28     229
Ship 1,0000 0,1875 0,3158 3       0      13
SpaceShuttle 1,0000 1,0000 1,0000 6       0      0
SpaceStation 1,0000 0,5000 0,6667 1       0      1
SportsClub 0,0000 1,0000 0,0000 0       1      0
SportsTeam 0,0000 1,0000 0,0000 0       1      0
Train 1,0000 0,6000 0,7500 3       0      2
Totals 0,8375 0,2580 0,3945 201     39     578

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\TransportationFineGrainedTop500Links.tsv
```

Figure 3.100: All 3 Domains Fine Grained Top 100 Links Runned With Transportation Fine Grained Top 500 Links

3. EXPERIMENTS

```
CRFClassifier tagged 101091 words in 1 documents at 3831,82 words per second.
Entity P R F1 TP FP FN
Aircraft 0,0000 0,0000 0,0000 0 1 72
Automobile 0,0000 0,0000 0,0000 0 0 36
Infrastructure 0,0000 0,0000 0,0000 0 4 272
Locomotive 0,0000 0,0000 0,0000 0 0 2
Motorcycle 0,0000 0,0000 0,0000 0 0 1
PoliticalParty 0,0000 1,0000 0,0000 0 1 0
Politician 0,0000 1,0000 0,0000 0 1 0
PublicTransitSystem 0,7083 0,1399 0,2337 34 14 209
Rocket 0,0000 0,0000 0,0000 0 0 5
Ship 0,0000 0,0000 0,0000 0 0 12
SpaceShuttle 0,0000 0,0000 0,0000 0 0 9
SpaceStation 0,0000 0,0000 0,0000 0 0 1
SportsClub 0,0000 1,0000 0,0000 0 3 0
Train 0,0000 0,0000 0,0000 0 0 5
Totals 0,5862 0,0517 0,0950 34 24 624

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top100Links\ner-All3DomainsTop100LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\LowerPageRank\TransportationFineGrainedTop500Links\LowerPageRank.tsv
```

Figure 3.101: All 3 Domains Fine Grained Top 100 Links Runned With Transportation Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 423454 words in 1 documents at 2349,18 words per second.
Entity P R F1 TP FP FN
Aircraft 0,9573 0,8000 0,8716 112 5 28
Athlete 0,8889 0,5849 0,7055 224 28 159
Automobile 0,9565 0,8800 0,9167 22 1 3
Coach 1,0000 0,6667 0,8000 4 0 2
Infrastructure 0,9034 0,5696 0,6987 131 14 99
PoliticalParty 0,8854 0,7268 0,7983 572 74 215
Politician 0,9664 0,6025 0,7423 144 5 95
PublicTransitSystem 0,8966 0,7901 0,8399 286 33 76
Ship 1,0000 0,5556 0,7143 15 0 12
SpaceShuttle 1,0000 0,8571 0,9231 6 0 1
SpaceStation 1,0000 0,5000 0,6667 1 0 1
SportsClub 0,9018 0,6443 0,7516 404 44 223
SportsEvent 0,9688 0,5741 0,7209 62 2 46
SportsLeague 0,9079 0,7856 0,8423 414 42 113
SportsManager 0,9811 0,5714 0,7222 52 1 39
SportsTeam 0,9784 0,7637 0,8578 181 4 56
Train 1,0000 1,0000 1,0000 6 0 0
Totals 0,9124 0,6930 0,7877 2636 253 1168

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\All3DomainsTop500Links\TextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.102: All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 500 Links

```
CRFClassifier tagged 336613 words in 1 documents at 3545,42 words per second.
Entity P R F1 TP FP FN
Aircraft 0,6667 0,1389 0,2299 10 5 62
Athlete 0,5000 0,0986 0,1647 28 28 256
Automobile 0,0000 0,0000 0,0000 0 0 36
Coach 0,0000 0,0000 0,0000 0 0 14
Infrastructure 0,5667 0,0620 0,1118 17 13 257
Locomotive 0,0000 0,0000 0,0000 0 0 2
Motorcycle 0,0000 0,0000 0,0000 0 0 1
OrganisationMember 0,0000 0,0000 0,0000 0 0 1
PoliticalParty 0,5191 0,3424 0,4127 163 151 313
Politician 0,5946 0,1325 0,2167 22 15 144
PublicTransitSystem 0,6846 0,3835 0,4916 102 47 164
Rocket 0,0000 0,0000 0,0000 0 0 5
Ship 0,0000 0,0000 0,0000 0 1 17
SpaceShuttle 0,0000 0,0000 0,0000 0 0 9
SpaceStation 0,0000 0,0000 0,0000 0 0 1
SportsClub 0,6043 0,2162 0,3184 139 91 504
SportsEvent 0,7500 0,0309 0,0594 3 1 94
SportsLeague 0,6009 0,3350 0,4302 134 89 266
SportsManager 0,6667 0,0211 0,0408 2 1 93
SportsTeam 0,5490 0,2314 0,3256 28 23 93
Train 0,0000 0,0000 0,0000 0 0 5
Totals 0,5822 0,2171 0,3163 648 465 2337

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\LowerPageRank\All3DomainsTop500Links\TextWithSameProcessedLinksOnAbstractFineGrainedSpecified\LowerPageRank.tsv
```

Figure 3.103: All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 500 Links With Lower PageRank

3.2. List of experiments

```
CRFClassifier tagged 164060 words in 1 documents at 4066,23 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 0,0000 1,0000 0,0000 0      9      0
Election 0,0000 0,0000 0,0000 0      0      494
PoliticalParty 0,8853 0,7292 0,7997 571    74     212
Politician 0,9792 0,1336 0,2352 141    3      914
PublicTransitSystem 0,0000 1,0000 0,0000 0      2      0
Ship 0,0000 1,0000 0,0000 0      1      0
SportsClub 0,0000 1,0000 0,0000 0      1      0
SportsLeague 0,0000 1,0000 0,0000 0      2      0
SportsManager 0,0000 1,0000 0,0000 0      1      0
Totals 0,8845 0,3053 0,4539 712    93     1620

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\PoliticsFineGrainedTop500Links.tsv
```

Figure 3.104: All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 500 Links

```
CRFClassifier tagged 117053 words in 1 documents at 4211,90 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 0,0000 1,0000 0,0000 0      6      0
Election 0,0000 0,0000 0,0000 0      0      197
Infrastructure 0,0000 1,0000 0,0000 0      1      0
PoliticalParty 0,5209 0,3411 0,4122 162    149    313
Politician 0,5882 0,0307 0,0583 20     14     632
PublicTransitSystem 0,0000 1,0000 0,0000 0      9      0
SportsLeague 0,0000 1,0000 0,0000 0      2      0
SportsTeam 0,0000 1,0000 0,0000 0      1      0
Totals 0,5000 0,1375 0,2156 182    182    1142

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\WithLowerPageRank\PoliticsFineGrainedTop500Links\WithLowerPageRank.tsv
```

Figure 3.105: All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 142322 words in 1 documents at 4466,12 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 1,0000 0,0000 0      1      0
Athlete 0,9218 0,5849 0,7157 224    19     159
Automobile 0,0000 1,0000 0,0000 0      1      0
Coach 1,0000 0,6667 0,8000 4      0      2
PoliticalParty 0,0000 1,0000 0,0000 0      1      0
Politician 0,0000 1,0000 0,0000 0      3      0
SportsClub 0,9009 0,6441 0,7512 400    44     221
SportsEvent 0,9688 0,5741 0,7209 62     2      46
SportsLeague 0,9097 0,7867 0,8437 413    41     112
SportsManager 1,0000 0,5714 0,7273 52     0      39
SportsTeam 0,9783 0,7627 0,8571 180    4      56
Train 0,0000 1,0000 0,0000 0      1      0
Totals 0,9194 0,6777 0,7802 1335    117    635

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\SportFineGrainedTop500Links.tsv
```

Figure 3.106: All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 500 Links

3. EXPERIMENTS

```
CRFClassifier tagged 118469 words in 1 documents at 4461,27 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 0,5833  0,1033  0,1755  28      20      243
Coach 0,0000  0,0000  0,0000  0        0        14
OrganisationMember 0,0000  0,0000  0,0000  0        0        1
PoliticalParty 0,0000  1,0000  0,0000  0        3        0
Politician 0,0000  1,0000  0,0000  0        3        0
SportsClub 0,6096  0,2217  0,3251  139     89      488
SportsEvent 0,7500  0,0313  0,0600  3        1        93
SportsLeague 0,6000  0,3325  0,4279  132     88      265
SportsManager 0,6667  0,0225  0,0435  2        1        87
SportsTeam 0,5833  0,2353  0,3353  28      20      91
Totals 0,5961  0,2057  0,3058  332     225     1282

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\inksWithLowerPageRank\SportFineGrainedTop500Links\inksWithLowerPageRank.tsv
```

Figure 3.107: All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 500 Links With Lower PageRank

```
CRFClassifier tagged 117072 words in 1 documents at 4190,42 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,9569  0,8102  0,8775  111     5        26
Automobile 1,0000  0,8800  0,9362  22      0        3
Infrastructure 0,9034  0,5696  0,6987  131     14      99
Politician 0,0000  1,0000  0,0000  0        2        0
PublicTransitSystem 0,8959  0,7933  0,8415  284     33      74
Ship 1,0000  0,8750  0,9333  14      0        2
SpaceShuttle 1,0000  1,0000  1,0000  6        0        0
SpaceStation 1,0000  0,5000  0,6667  1        0        1
SportsClub 0,0000  1,0000  0,0000  0        3        0
SportsTeam 0,0000  1,0000  0,0000  0        1        0
Train 1,0000  1,0000  1,0000  5        0        0
Totals 0,9082  0,7368  0,8136  574     58      205

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\TransportationFineGrainedTop500Links\inksWithLowerPageRank.tsv
```

Figure 3.108: All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 500 Links

```
CRFClassifier tagged 101091 words in 1 documents at 3283,24 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,6667  0,1380  0,2299  10      5        62
Athlete 0,0000  1,0000  0,0000  0        2        0
Automobile 0,0000  0,0000  0,0000  0        0        36
Infrastructure 0,5862  0,0625  0,1130  17      12      255
Locomotive 0,0000  0,0000  0,0000  0        0        2
Motorcycle 0,0000  0,0000  0,0000  0        0        1
PublicTransitSystem 0,6714  0,3868  0,4909  94      46      149
Rocket 0,0000  0,0000  0,0000  0        0        5
Ship 0,0000  0,0000  0,0000  0        1        12
SpaceShuttle 0,0000  0,0000  0,0000  0        0        9
SpaceStation 0,0000  0,0000  0,0000  0        0        1
SportsClub 0,0000  1,0000  0,0000  0        2        0
SportsLeague 0,0000  1,0000  0,0000  0        1        0
SportsTeam 0,0000  1,0000  0,0000  0        2        0
Train 0,0000  0,0000  0,0000  0        0        5
Totals 0,6302  0,1839  0,2847  121     71      537

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top500Links\inksWithLowerPageRank\TransportationFineGrainedTop500Links\inksWithLowerPageRank.tsv
```

Figure 3.109: All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 500 Links With Lower PageRank

3.2. List of experiments

```
CRFClassifier tagged 10336 words in 1 documents at 2532,71 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 1 0 0
Athlete 1,0000 1,0000 1,0000 1 0 0
Coach 1,0000 1,0000 1,0000 1 0 0
PoliticalParty 0,9545 0,8077 0,8750 21 1 5
Politician 1,0000 1,0000 1,0000 4 0 0
PublicTransitSystem 1,0000 0,8571 0,9231 6 0 1
Ship 1,0000 1,0000 1,0000 1 0 0
SportsClub 1,0000 1,0000 1,0000 7 0 0
SportsEvent 1,0000 1,0000 1,0000 1 0 0
SportsLeague 1,0000 0,7500 0,8571 3 0 1
SportsTeam 1,0000 1,0000 1,0000 1 0 0
Totals 0,9792 0,8704 0,9216 47 1 7

C:\Dev\stanford-nlp-2017-06-09>java -Xmx11g -cp stanford-nlp.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\All3DomainsTop10LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.110: All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 10 Links

```
CRFClassifier tagged 97089 words in 1 documents at 4145,56 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 23 0 0
Athlete 1,0000 1,0000 1,0000 36 0 0
Automobile 1,0000 1,0000 1,0000 3 0 0
Coach 1,0000 1,0000 1,0000 3 0 0
Infrastructure 1,0000 1,0000 1,0000 22 0 0
PoliticalParty 0,9848 0,9606 0,9726 195 3 8
Politician 1,0000 0,8909 0,9423 49 0 6
PublicTransitSystem 0,9800 0,9423 0,9608 49 1 3
Ship 1,0000 1,0000 1,0000 5 0 0
SpaceShuttle 1,0000 1,0000 1,0000 6 0 0
SpaceStation 1,0000 1,0000 1,0000 1 0 0
SportsClub 0,9639 0,9639 0,9639 80 3 3
SportsEvent 1,0000 0,9767 0,9882 42 0 1
SportsLeague 0,9403 0,9545 0,9474 63 4 3
SportsManager 1,0000 0,8333 0,9091 5 0 1
SportsTeam 1,0000 1,0000 1,0000 33 0 0
Train 1,0000 1,0000 1,0000 1 0 0
Totals 0,9825 0,9610 0,9716 616 11 25

C:\Dev\stanford-nlp-2017-06-09>java -Xmx11g -cp stanford-nlp.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.111: All 3 Domains Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 100 Links

```
CRFClassifier tagged 4504 words in 1 documents at 2969,02 words per second.
Entity P R F1 TP FP FN
Election 0,0000 0,0000 0,0000 0 0 15
PoliticalParty 0,9545 0,8077 0,8750 21 1 5
Politician 1,0000 0,1111 0,2000 4 0 32
Totals 0,9615 0,3247 0,4854 25 1 52

C:\Dev\stanford-nlp-2017-06-09>java -Xmx11g -cp stanford-nlp.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\PoliticsFineGrainedTop10Links.tsv
```

Figure 3.112: All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 10 Links

```
CRFClassifier tagged 40673 words in 1 documents at 4019,47 words per second.
Entity P R F1 TP FP FN
Election 0,0000 0,0000 0,0000 0 0 164
PoliticalParty 0,9848 0,9606 0,9726 195 3 8
Politician 1,0000 0,1536 0,2663 49 0 270
Totals 0,9879 0,3557 0,5230 244 3 442

C:\Dev\stanford-nlp-2017-06-09>java -Xmx11g -cp stanford-nlp.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\PoliticsFineGrainedTop100Links.tsv
```

Figure 3.113: All 3 Domains Fine Grained Top 300 Links Runned With Politics Fine Grained Top 100 Links

3. EXPERIMENTS

```
CRFClassifier tagged 2854 words in 1 documents at 2994,75 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000  1,0000  1,0000  1       0       0
Coach 1,0000  1,0000  1,0000  1       0       0
SportsClub 1,0000  1,0000  1,0000  7       0       0
SportsEvent 1,0000  1,0000  1,0000  1       0       0
SportsLeague 1,0000  0,7500  0,8571  3       0       1
SportsTeam 1,0000  1,0000  1,0000  1       0       0
Totals 1,0000  0,9333  0,9655  14      0       1

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\SportFineGrainedTop10Links.tsv
```

Figure 3.114: All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 10 Links

```
CRFClassifier tagged 31872 words in 1 documents at 3691,88 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000  1,0000  1,0000  36      0       0
Coach 1,0000  1,0000  1,0000  3       0       0
SportsClub 0,9634  0,9634  0,9634  79      3       3
SportsEvent 1,0000  0,9767  0,9882  42      0       1
SportsLeague 0,9403  0,9545  0,9474  63      4       3
SportsManager 1,0000  0,8333  0,9091  5       0       1
SportsTeam 1,0000  1,0000  1,0000  32      0       0
Totals 0,9738  0,9701  0,9720  260     7       8

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\SportFineGrainedTop100Links.tsv
```

Figure 3.115: All 3 Domains Fine Grained Top 300 Links Runned With Sport Fine Grained Top 100 Links

```
CRFClassifier tagged 2978 words in 1 documents at 2863,46 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  1,0000  1,0000  1       0       0
PublicTransitSystem 1,0000  0,8571  0,9231  6       0       1
Ship 1,0000  1,0000  1,0000  1       0       0
Totals 1,0000  0,8889  0,9412  8       0       1

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\TransportationFineGrainedTop10Links.tsv
```

Figure 3.116: All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 10 Links

```
CRFClassifier tagged 24544 words in 1 documents at 4180,55 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000  1,0000  1,0000  23      0       0
Automobile 1,0000  1,0000  1,0000  3       0       0
Infrastructure 1,0000  1,0000  1,0000  22      0       0
PublicTransitSystem 0,9800  0,9423  0,9608  49      1       3
Ship 1,0000  1,0000  1,0000  5       0       0
SpaceShuttle 1,0000  1,0000  1,0000  6       0       0
SpaceStation 1,0000  1,0000  1,0000  1       0       0
SportsClub 0,0000  1,0000  0,0000  0       1       0
SportsTeam 0,0000  1,0000  0,0000  0       1       0
Train 1,0000  1,0000  1,0000  1       0       0
Totals 0,9735  0,9735  0,9735  110     3       3

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-All3DomainsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\TransportationFineGrainedTop100Links.tsv
```

Figure 3.117: All 3 Domains Fine Grained Top 300 Links Runned With Transportation Fine Grained Top 100 Links

3.2. List of experiments

```
CRFClassifier tagged 10336 words in 1 documents at 3459,17 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 1 0 0
Athlete 1,0000 1,0000 1,0000 1 0 0
Coach 1,0000 1,0000 1,0000 1 0 0
PoliticalParty 0,9545 0,8077 0,8750 21 1 5
Politician 1,0000 1,0000 1,0000 4 0 0
PublicTransitSystem 0,8333 0,7143 0,7692 5 1 2
Ship 1,0000 1,0000 1,0000 1 0 0
SportsClub 0,8750 1,0000 0,9333 7 1 0
SportsEvent 1,0000 1,0000 1,0000 1 0 0
SportsLeague 1,0000 1,0000 1,0000 4 0 0
SportsTeam 1,0000 1,0000 1,0000 1 0 0
Totals 0,9400 0,8704 0,9038 47 3 7

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\All3DomainsTop10LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.118: All 3 Domains Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 10 Links

```
CRFClassifier tagged 97089 words in 1 documents at 3675,94 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 23 0 0
Athlete 1,0000 1,0000 1,0000 36 0 0
Automobile 1,0000 1,0000 1,0000 3 0 0
Coach 1,0000 1,0000 1,0000 3 0 0
Infrastructure 1,0000 0,9545 0,9767 21 0 1
PoliticalParty 0,9799 0,9606 0,9701 195 4 8
Politician 1,0000 1,0000 1,0000 55 0 0
PublicTransitSystem 0,9796 0,9231 0,9505 48 1 4
Ship 1,0000 1,0000 1,0000 5 0 0
SpaceShuttle 1,0000 1,0000 1,0000 6 0 0
SpaceStation 1,0000 1,0000 1,0000 1 0 0
SportsClub 0,9529 0,9759 0,9643 81 4 2
SportsEvent 1,0000 0,8837 0,9383 38 0 5
SportsLeague 0,9275 0,9697 0,9481 64 5 2
SportsManager 1,0000 1,0000 1,0000 6 0 0
SportsTeam 1,0000 1,0000 1,0000 33 0 0
Train 1,0000 1,0000 1,0000 1 0 0
Totals 0,9779 0,9657 0,9717 619 14 22

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.119: All 3 Domains Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 100 Links

```
CRFClassifier tagged 270361 words in 1 documents at 2605,86 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 102 0 0
Athlete 1,0000 0,9802 0,9900 198 0 4
Automobile 1,0000 1,0000 1,0000 20 0 0
Coach 1,0000 1,0000 1,0000 4 0 0
Infrastructure 1,0000 0,9640 0,9817 107 0 4
PoliticalParty 0,9718 0,9432 0,9573 482 14 29
Politician 1,0000 1,0000 1,0000 139 0 0
PublicTransitSystem 0,9918 0,9718 0,9817 241 2 7
Ship 1,0000 1,0000 1,0000 15 0 0
SpaceShuttle 1,0000 1,0000 1,0000 6 0 0
SpaceStation 1,0000 1,0000 1,0000 1 0 0
SportsClub 0,9709 0,9625 0,9667 334 10 13
SportsEvent 1,0000 0,8333 0,9091 55 0 11
SportsLeague 0,9551 0,9707 0,9628 298 14 9
SportsManager 1,0000 0,9615 0,9804 50 0 2
SportsTeam 1,0000 0,9805 0,9902 151 0 3
Train 1,0000 1,0000 1,0000 6 0 0
Totals 0,9822 0,9642 0,9731 2209 40 82

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\All3DomainsTop300LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.120: All 3 Domains Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 300 Links

3. EXPERIMENTS

```
CRFClassifier tagged 4504 words in 1 documents at 2429,34 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000 0,0000 0,0000 0      0      15
PoliticalParty 0,9545 0,8077 0,8750 21     1      5
Politician 1,0000 0,1111 0,2000 4      0      32
Totals 0,9615 0,3247 0,4854 25     1      52

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\PoliticsFineGrainedTop10Links.tsv
```

Figure 3.121: All 3 Domains Fine Grained Top 500 Links Runned With Politics Fine Grained Top 10 Links

```
CRFClassifier tagged 40673 words in 1 documents at 4212,64 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000 0,0000 0,0000 0      0      164
PoliticalParty 0,9799 0,9606 0,9701 195    4      8
Politician 1,0000 0,1724 0,2941 55     0      264
Totals 0,9843 0,3644 0,5319 250    4      436

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\PoliticsFineGrainedTop100Links.tsv
```

Figure 3.122: All 3 Domains Fine Grained Top 500 Links Runned With Politics Fine Grained Top 100 Links

```
CRFClassifier tagged 107491 words in 1 documents at 2506,26 words per second.
Entity P      R      F1      TP      FP      FN
Election 0,0000 0,0000 0,0000 0      0      322
PoliticalParty 0,9718 0,9432 0,9573 482    14     29
Politician 1,0000 0,1980 0,3305 136    0      551
PublicTransitSystem 0,0000 1,0000 0,0000 0      2      0
Ship 0,0000 1,0000 0,0000 0      1      0
SportsLeague 0,0000 1,0000 0,0000 0      1      0
Totals 0,9717 0,4066 0,5733 618    18     902

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\PoliticsFineGrainedTop300Links.tsv
```

Figure 3.123: All 3 Domains Fine Grained Top 500 Links Runned With Politics Fine Grained Top 300 Links

```
CRFClassifier tagged 2854 words in 1 documents at 2679,81 words per second.
Entity P      R      F1      TP      FP      FN
Athlete 1,0000 1,0000 1,0000 1      0      0
Coach 1,0000 1,0000 1,0000 1      0      0
SportsClub 0,8750 1,0000 0,9333 7      1      0
SportsEvent 1,0000 1,0000 1,0000 1      0      0
SportsLeague 1,0000 1,0000 1,0000 4      0      0
SportsTeam 1,0000 1,0000 1,0000 1      0      0
Totals 0,9375 1,0000 0,9677 15     1      0

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top10Links\SportFineGrainedTop10Links.tsv
```

Figure 3.124: All 3 Domains Fine Grained Top 500 Links Runned With Sport Fine Grained Top 10 Links

3.2. List of experiments

```
CRFClassifier tagged 31872 words in 1 documents at 4022,72 words per second.
Entity P R F1 TP FP FN
Athlete 1,0000 1,0000 1,0000 36 0 0
Coach 1,0000 1,0000 1,0000 3 0 0
SportsClub 0,9524 0,9756 0,9639 80 4 2
SportsEvent 1,0000 0,8837 0,9383 38 0 5
SportsLeague 0,9275 0,9697 0,9481 64 5 2
SportsManager 1,0000 1,0000 1,0000 6 0 0
SportsTeam 1,0000 1,0000 1,0000 32 0 0
Totals 0,9664 0,9664 0,9664 259 9 9

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\SportFineGrainedTop100Links.tsv
```

Figure 3.125: All 3 Domains Fine Grained Top 500 Links Runned With Sport Fine Grained Top 100 Links

```
CRFClassifier tagged 89224 words in 1 documents at 3135,95 words per second.
Entity P R F1 TP FP FN
Athlete 1,0000 0,9802 0,9900 198 0 4
Coach 1,0000 1,0000 1,0000 4 0 0
Politician 0,0000 1,0000 0,0000 0 2 0
SportsClub 0,9707 0,9622 0,9664 331 10 13
SportsEvent 1,0000 0,8333 0,9091 55 0 11
SportsLeague 0,9582 0,9707 0,9644 298 13 9
SportsManager 1,0000 0,9615 0,9804 50 0 2
SportsTeam 1,0000 0,9804 0,9901 150 0 3
Train 0,0000 1,0000 0,0000 0 1 0
Totals 0,9766 0,9628 0,9696 1086 26 42

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\SportFineGrainedTop300Links.tsv
```

Figure 3.126: All 3 Domains Fine Grained Top 500 Links Runned With Sport Fine Grained Top 300 Links

```
CRFClassifier tagged 2978 words in 1 documents at 2749,77 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 1 0 0
PublicTransitSystem 0,8333 0,7143 0,7692 5 1 2
Ship 1,0000 1,0000 1,0000 1 0 0
Totals 0,8750 0,7778 0,8235 7 1 2

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\TransportationFineGrainedTop100Links.tsv
```

Figure 3.127: All 3 Domains Fine Grained Top 500 Links Runned With Transportation Fine Grained Top 10 Links

```
CRFClassifier tagged 24544 words in 1 documents at 3821,86 words per second.
Entity P R F1 TP FP FN
Aircraft 1,0000 1,0000 1,0000 23 0 0
Automobile 1,0000 1,0000 1,0000 3 0 0
Infrastructure 1,0000 0,9545 0,9767 21 0 1
PublicTransitSystem 0,9796 0,9231 0,9505 48 1 4
Ship 1,0000 1,0000 1,0000 5 0 0
SpaceShuttle 1,0000 1,0000 1,0000 6 0 0
SpaceStation 1,0000 1,0000 1,0000 1 0 0
SportsClub 0,0000 1,0000 0,0000 0 1 0
SportsTeam 0,0000 1,0000 0,0000 0 1 0
Train 1,0000 1,0000 1,0000 1 0 0
Totals 0,9730 0,9558 0,9643 108 3 5

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\TransportationFineGrainedTop100Links.tsv
```

Figure 3.128: All 3 Domains Fine Grained Top 500 Links Runned With Transportation Fine Grained Top 100 Links

3. EXPERIMENTS

```
CRFClassifier tagged 73646 words in 1 documents at 3209,68 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 1,0000 1,0000 1,0000 102    0      0
Automobile 1,0000 1,0000 1,0000 20     0      0
Infrastructure 1,0000 0,9640 0,9817 107    0      4
Politician 0,0000 1,0000 0,0000 0       1      0
PublicTransitSystem 0,9917 0,9715 0,9815 239    2      7
Ship 1,0000 1,0000 1,0000 14     0      0
SpaceShuttle 1,0000 1,0000 1,0000 6       0      0
SpaceStation 1,0000 1,0000 1,0000 1       0      0
SportsClub 0,0000 1,0000 0,0000 0       3      0
SportsTeam 0,0000 1,0000 0,0000 0       1      0
Train 1,0000 1,0000 1,0000 5       0      0
Totals 0,9860 0,9782 0,9821 494    7      11

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-All3DomainsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\TransportationFineGrainedTop300Links.tsv
```

Figure 3.129: All 3 Domains Fine Grained Top 500 Links Runned With Transportation Fine Grained Top 300 Links

```
CRFClassifier tagged 97089 words in 1 documents at 7883,80 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 0,0000 0,0000 0       0      23
Athlete 0,0000 0,0000 0,0000 0       0      36
Automobile 0,0000 0,0000 0,0000 0       0      3
Coach 0,0000 0,0000 0,0000 0       0      3
Election 0,0000 1,0000 0,0000 0      136    0
Infrastructure 0,0000 0,0000 0,0000 0       0      22
PoliticalParty 0,7831 0,7291 0,7551 148    41     55
Politician 0,1541 0,9273 0,2642 51     280    4
PublicTransitSystem 0,0000 0,0000 0,0000 0       0      52
Ship 0,0000 0,0000 0,0000 0       0      5
SpaceShuttle 0,0000 0,0000 0,0000 0       0      6
SpaceStation 0,0000 0,0000 0,0000 0       0      1
SportsClub 0,0000 0,0000 0,0000 0       0      83
SportsEvent 0,0000 0,0000 0,0000 0       0      43
SportsLeague 0,0000 0,0000 0,0000 0       0      66
SportsManager 0,0000 0,0000 0,0000 0       0      6
SportsTeam 0,0000 0,0000 0,0000 0       0      33
Train 0,0000 0,0000 0,0000 0       0      1
Totals 0,3034 0,3105 0,3069 199    457    442

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-PoliticsTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.130: Politics Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 100 Links

```
CRFClassifier tagged 97089 words in 1 documents at 5175,60 words per second.
Entity P      R      F1      TP      FP      FN
Aircraft 0,0000 0,0000 0,0000 0       0      23
Athlete 0,0000 0,0000 0,0000 0       0      36
Automobile 0,0000 0,0000 0,0000 0       0      3
Coach 0,0000 0,0000 0,0000 0       0      3
Election 0,0000 1,0000 0,0000 0     163    0
Infrastructure 0,0000 0,0000 0,0000 0       0      22
PoliticalParty 0,9652 0,9557 0,9604 194    7       0
Politician 0,1471 1,0000 0,2564 55     319    0
PublicTransitSystem 0,0000 0,0000 0,0000 0       0      52
Ship 0,0000 0,0000 0,0000 0       0      5
SpaceShuttle 0,0000 0,0000 0,0000 0       0      6
SpaceStation 0,0000 0,0000 0,0000 0       0      1
SportsClub 0,0000 0,0000 0,0000 0       0      83
SportsEvent 0,0000 0,0000 0,0000 0       0      43
SportsLeague 0,0000 0,0000 0,0000 0       0      66
SportsManager 0,0000 0,0000 0,0000 0       0      6
SportsTeam 0,0000 0,0000 0,0000 0       0      33
Train 0,0000 0,0000 0,0000 0       0      1
Totals 0,3374 0,3885 0,3611 249    489    392

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-PoliticsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.131: Politics Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 100 Links

3.2. List of experiments

```
CRFClassifier tagged 270361 words in 1 documents at 4572,62 words per second.
Entity P R F1 TP FP FN
Aircraft 0,0000 0,0000 0,0000 0 0 102
Athlete 0,0000 0,0000 0,0000 0 0 202
Automobile 0,0000 0,0000 0,0000 0 0 20
Coach 0,0000 0,0000 0,0000 0 0 4
Election 0,0000 1,0000 0,0000 0 337 0
Infrastructure 0,0000 0,0000 0,0000 0 0 111
PoliticalParty 0,9569 0,9550 0,9559 488 22 23
Politician 0,1518 0,9784 0,2628 136 760 3
PublicTransitSystem 0,0000 0,0000 0,0000 0 0 248
Ship 0,0000 0,0000 0,0000 0 0 15
SpaceShuttle 0,0000 0,0000 0,0000 0 0 6
SpaceStation 0,0000 0,0000 0,0000 0 0 1
SportsClub 0,0000 0,0000 0,0000 0 0 347
SportsEvent 0,0000 0,0000 0,0000 0 0 66
SportsLeague 0,0000 0,0000 0,0000 0 0 307
SportsManager 0,0000 0,0000 0,0000 0 0 52
SportsTeam 0,0000 0,0000 0,0000 0 0 154
Train 0,0000 0,0000 0,0000 0 0 6
Totals 0,3580 0,2724 0,3094 624 1119 1667

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top500Links\ner-PoliticsTop500LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top300Links\All3DomainsTop300LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.132: Politics Fine Grained Top 500 Links Runned With All 3 Domains Fine Grained Top 300 Links

```
CRFClassifier tagged 97089 words in 1 documents at 5454,44 words per second.
Entity P R F1 TP FP FN
Aircraft 0,0000 0,0000 0,0000 0 0 23
Athlete 0,2817 0,5556 0,3738 20 51 16
Automobile 0,0000 0,0000 0,0000 0 0 3
Coach 1,0000 0,6667 0,8000 2 0 1
Infrastructure 0,0000 0,0000 0,0000 0 0 22
PoliticalParty 0,0000 0,0000 0,0000 0 0 203
Politician 0,0000 0,0000 0,0000 0 0 55
PublicTransitSystem 0,0000 0,0000 0,0000 0 0 52
Ship 0,0000 0,0000 0,0000 0 0 5
SpaceShuttle 0,0000 0,0000 0,0000 0 0 6
SpaceStation 0,0000 0,0000 0,0000 0 0 1
SportsClub 0,8795 0,8795 0,8795 73 10 10
SportsEvent 0,9706 0,7674 0,8571 33 1 10
SportsLeague 0,8852 0,8182 0,8504 54 7 12
SportsManager 0,5000 1,0000 0,6667 6 6 0
SportsTeam 0,9583 0,6970 0,8070 23 1 10
Train 0,0000 0,0000 0,0000 0 0 1
Totals 0,7352 0,3292 0,4547 211 76 430

C:\Dev\stanford-ner-2017-06-09>java -Xmx11g -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier modelsWithOneAnnotation\Top300Links\ner-SportTop300LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\Top100Links\All3DomainsTop100LinksTextWithSameProcessedLinksOnAbstractFineGrainedSpecified.tsv
```

Figure 3.133: Sport Fine Grained Top 300 Links Runned With All 3 Domains Fine Grained Top 100 Links

3. EXPERIMENTS

3.3 Summary of results

3.3.1 Graphs

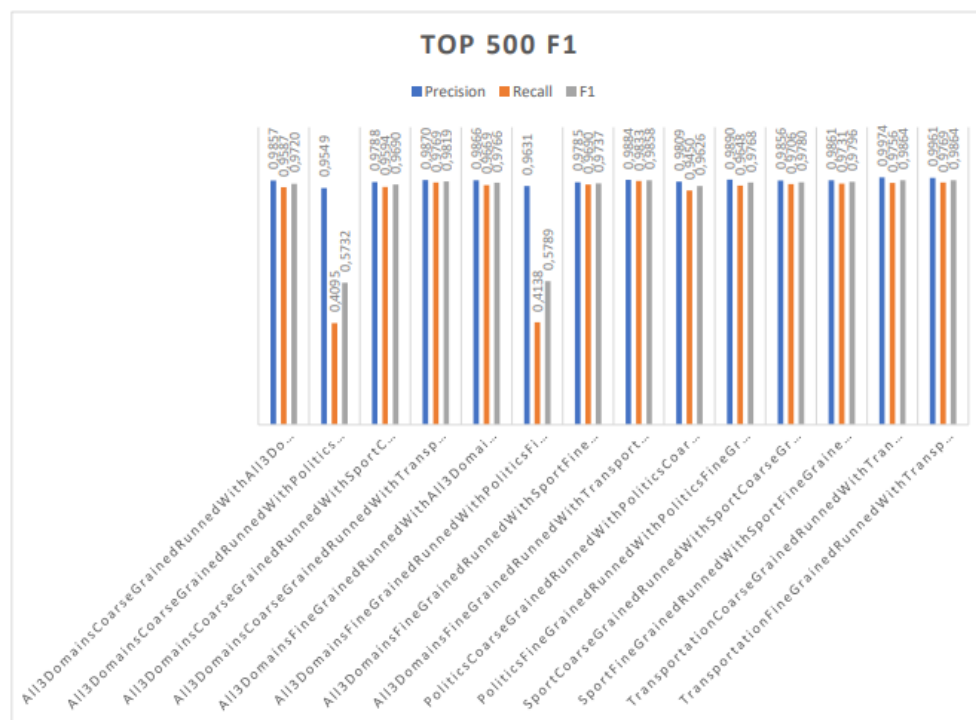


Figure 3.134: Top 500 Links F1

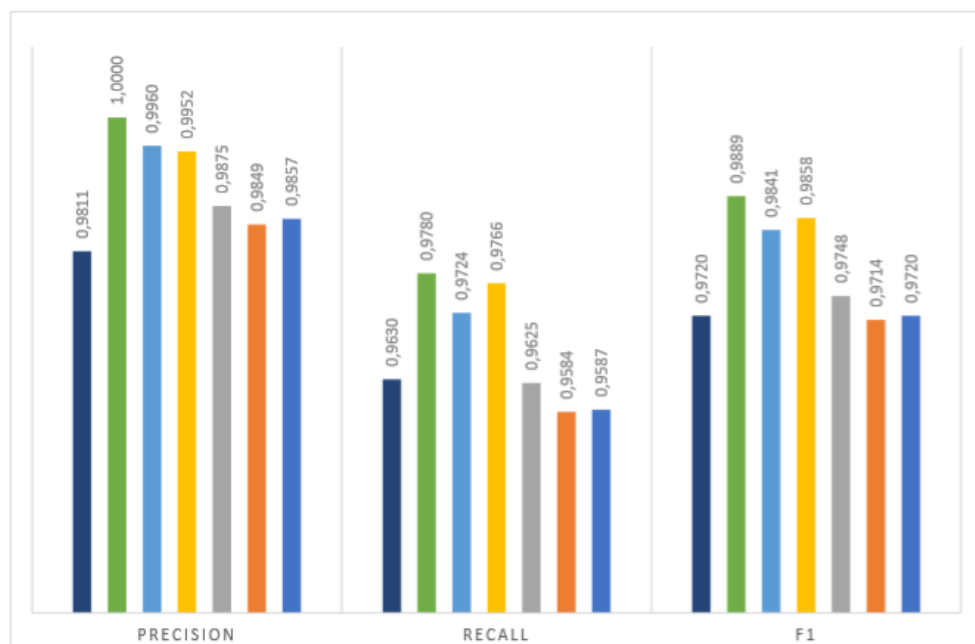


Figure 3.135: All 3 Domains Coarse Grained Runned With All 3 Domains Coarse Grained-tsv F1

Conclusion

The word-based dictionary data compression algorithms (a part of lossless data compression) are the subject of this thesis. The lossless data compression is a very important field of research because the data compression allows to reduce the amount of space needed to store data.

The background of data compression field was presented in Chapter 1. There are basic notions and definitions followed by description of character-based dictionary algorithms. The word-based dictionary compression methods were investigated and discussed at the end of this chapter too.

The testing of memory used during compression and/or decompression process is one of the possibilities of further research. The experiments with files of greater size or multilingual files could be also good opportunity to gain new improvements of algorithms. The static part of dictionaries could improve the compression efficiency too.

The implemented methods achieve fairly good compression ratio (25–30% at large files) with acceptable compression and decompression time. There are possibilities of further improvements especially at semi-adaptive methods. However, the gain of these improvements is not good enough to top the compression efficiency of other lossless data compression methods (context methods from PPM family). The results of implemented algorithms were not as good as it was expected but the work on this thesis showed new ways of possible further research—word-based version of grammar-based compression algorithms and another possibilities in the field of word-based context methods of data compression.

The Gnuplot 4.2 utility was very useful for generation of graphs in this thesis. There was the drawing editor Ipe 6.0 used for figures creation.

Contents of CD

Visualise the contents of enclosed media. Use of `dirtree` is recommended. Note that directories `src` and `text` with appropriate contents are mandatory.

```
├── readme.txt ..... the file with CD contents description
├── data ..... the data files directory
│   ├── graphs ..... the directory of graphs of experiments
│   │   ├── *.eps ..... the B/W graphs
│   │   ├── *.png ..... the color graphs
│   │   └── *.dat ..... the graphs data files
├── exe ..... the directory with executable WBDCM program
│   ├── wbdcm ..... the WBDCM program executable (UNIX)
│   └── wbdcm.exe ..... the WBDCM program executable (Windows)
├── src ..... the directory of source codes
│   ├── wbdcm ..... the directory of WBDCM program
│   │   └── Makefile ..... the makefile of WBDCM program (UNIX)
│   ├── thesis ..... the directory of LATEX source codes of the thesis
│   │   ├── figures ..... the thesis figures directory
│   │   └── *.tex ..... the LATEX source code files of the thesis
├── text ..... the thesis text directory
│   ├── thesis.pdf ..... the Diploma thesis in PDF format
│   └── thesis.ps ..... the Diploma thesis in PS format
```