Title (EN): Domain-Specific NER Adaptation

In the past years, the Named Entity Recognition (NER) technology has been under an active development and enjoy a significant increase in popularity and usage in the academic and industrial sphere. Nevertheless, vast majority of the developed NER systems have been developed as general-purpose systems. While they can perform well on multiple domains (macro level), on specific domains (micro level) their performance quality might be low. The ultimate goal of the thesis is to develop domain-specific NER models. Guidelines:

- Get familiar with the NER technology and available NER frameworks.
- Investigate possible datasets for domain-specific training of NER.
- Develop NER training datasets for several selected domains (e.g. sports, politics, music, etc.).
- Train a domain-specific NER model using existing frameworks, such as DBpedia Spotlight or StanfordNER.
- Validate and evaluate the developed domain-specific NER models.

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Master's thesis

Domain-specific Name Entity Recognition

Bc. Bogoljub Jakovcheski

Department of software engineering
Supervisor: Ing. Milan Dojchinovski, Ph.D.

May 17, 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 17, 2018 . . . .

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

List of Figures

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 that extracting information from a specific domain, for example politics domain or any other domain.

Thesis outline

Chapter 1

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

1.5 Evaluation/training datasets

1.6 Domain specific NER

1.7 Stanford NER

Chapter 2

Domain specific NER

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

2.1 Data pre-processing

âĂć Coarse grained âĂć 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

âĂć How we choose those domains (Politics, Sport, Transportation)?

2.4 Data transformation/output for Stanford

âĂć How we process data to be ready to use on Stanford NER system?

Chapter 3

Experiments

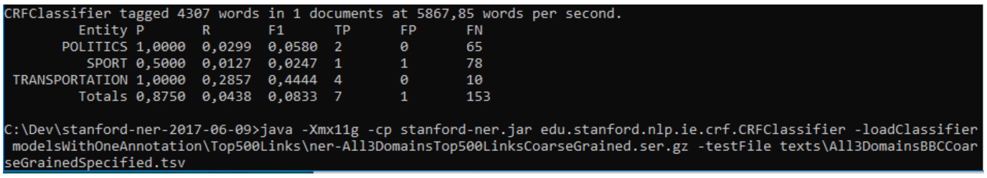
We have provide various types of experiments. In next sections we will discuss more about every provided experiment.

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

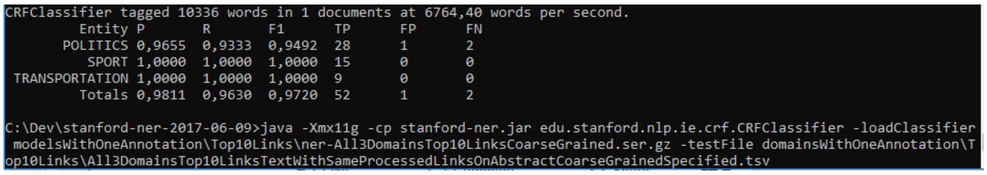


```
CRFClassifier tagged 4307 words in 1 documents at 5867,85 words per second.
         Entity P       R       F1      TP      FP      FN
       POLITICS 1,0000  0,0299  0,0580  2       0       65
          SPORT 0,5000  0,0127  0,0247  1       1       78
 TRANSPORTATION 1,0000  0,2857  0,4444  4  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

3.2.1 Top 10 Links

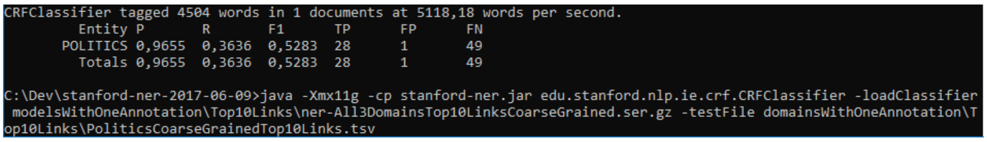
In this experiment we get the model that is trained with all 3 domains together with coarse grained and we make an experiment with the text that is trained, that means text from all 3 domains annotated with coarse grained. Here we have for every domain 10 abstracts, so in total 30 annotated abstracts. 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.



```
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

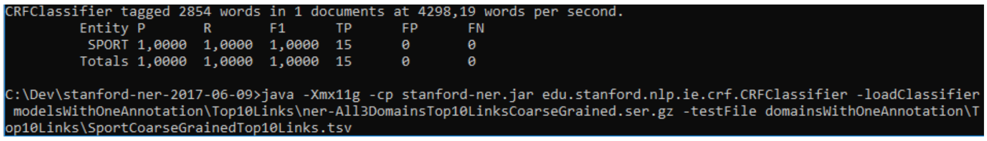
In this experiment we get the previous trained model and test with politics coarse grained text. Result is not satisfying, because we have a very low recall, so that results with bad F1 value and very big false negative words. We can clearly say that this is not the model that we want to use in annotating. 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.



```
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

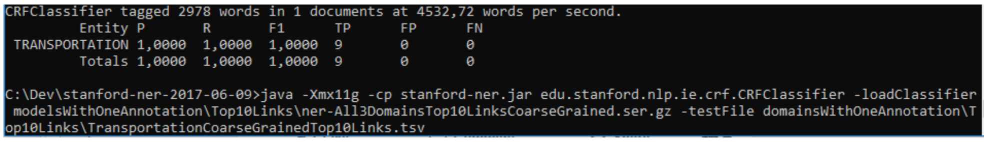
Then we get the same model, and like previous, we tested with sport coarse grained text. So how we can see from the picture, we have perfect result, there is no false positive or false negative words, there are only true positive which results with F1 value of 1.000 (no lost words).



```
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

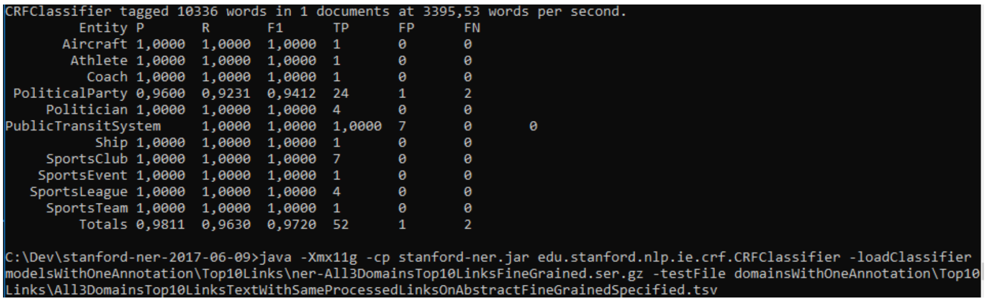
Finally we tested this model with our last domain, transportation domain with coarse grained text. Because of very small annotated words we have again perfect results without any false positive or false negative words, same as previous test, there are only true positive matches, which is what we want.



```
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

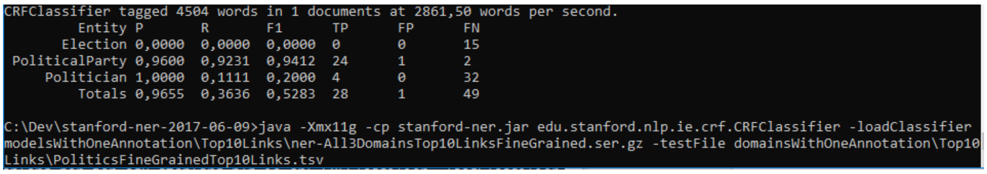
After we finish testing of all 3 domains trained with coarse grained texts, we move on with new type of experiment, model that contains all our domains and is trained with fine grained texts. We run the same experiment like in coarse grained, but now text is annotated with fine grained. So, in such small trained model is not surprising that we get the same results as in experiment with coarse grained annotated text.



```
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

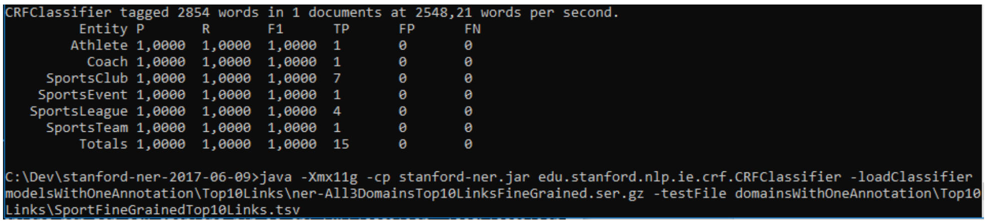
We have repeated the same experiments from previous, but now with fine grained annotation. So we tested our model with politics domain annotated in fine grained and we get exactly the same results like in coarse grained. In this case there is no difference if we use coarse or fine grained trained model, because the results are same.



```
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

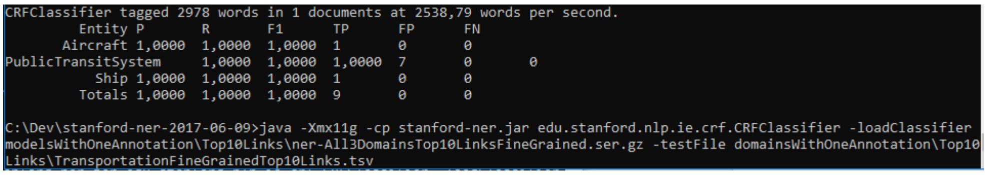
The sport domain text annotated in fine grained provide also same result as coarse grained annotated domain.



```
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

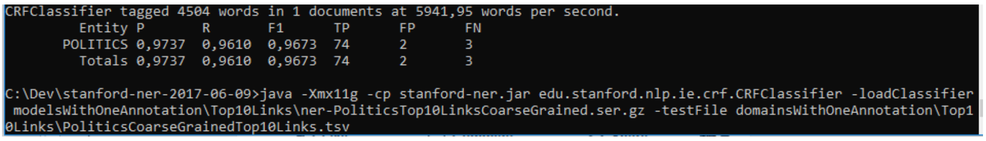
Same as sport domain, the transportation domain Figure 3.9 text annotated with fine grained, provide also same results.



```
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

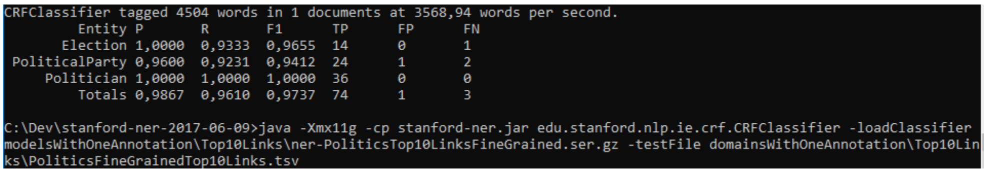
After experimenting with models that are trained with all domains together, we move on with experiments where we create specific models, both coarse and fine grained. First we tested the model that is created with politics text annotated in coarse grained. How we can see from Figure 3.10 we have a better results than that experiment that we provide in all 3 domains. The F1 value is very big which is the results of low false positive and false negative annotated words.



```
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\Top1
0Links\PoliticsCoarseGrainedTop10Links.tsv
```

Figure 3.10: Politics Coarse Top 10 Links Politics

We also provide the same experiment like previous, but now the model is annotated in fine grained. From Figure 3.11 is clearly that we have improve our model, here the F1 value is bigger that in coarse grained (see Figure 3.10) and also false positive annotated words are decreased by 1. That means even we have a complete review of our entities and can see the precision of each entity, this domain, in total, is more precise that coarse grained model and all 3 domains in coarse and fine grained.

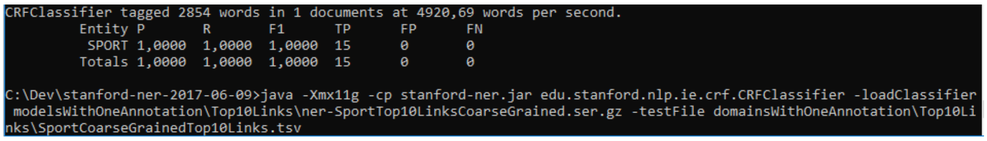


```
CRFClassifier tagged 4504 words in 1 documents at 3568,94 words per second.
         Entity P       R       F1      TP      FP      FN
       Election 1,0000  0,9333  0,9655  14      0       1
 PoliticalParty 0,9600  0,9231  0,9412  24      1       2
     Politician 1,0000  1,0000  1,0000  36      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\Top10Lin
ks\PoliticsFineGrainedTop10Links.tsv
```

Figure 3.11: Politics Fine Top 10 Links Politics

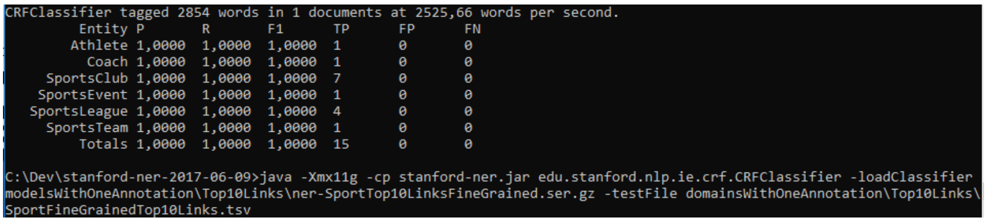
From here, with this small amount of data we can say that it's better to use a politics specific domain, in fine grained, than a global domain, because in global domain we have 28 words annotated as true positive, but in a politics specific coarse grained domain we have 74 words annotated as true positive, which is very big difference, if we want to be precise.

We repeated the experiment for sport specific domain, but in this case the results are exactly the same in coarse and fine grained (see Figure 3.12 and Figure 3.13 as well as in all 3 domains also in coarse and fine grained



```
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: Sport Coarse Top 10 Links Sport

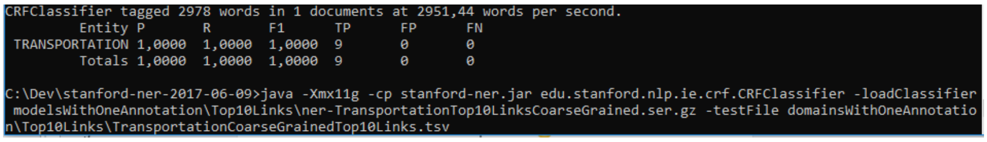


```
CRFClassifier tagged 2854 words in 1 documents at 2525,66 words per second.
         Entity P       R       F1      TP      FP      FN
        Athlete 1,0000  1,0000  1,0000  1       0       0
          Coach 1,0000  1,0000  1,0000  1       0       0
     SportsClub 1,0000  1,0000  1,0000  7       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\Top10Links\
SportFineGrainedTop10Links.tsv
```

Figure 3.13: Sport Fine Top 10 Links Sport

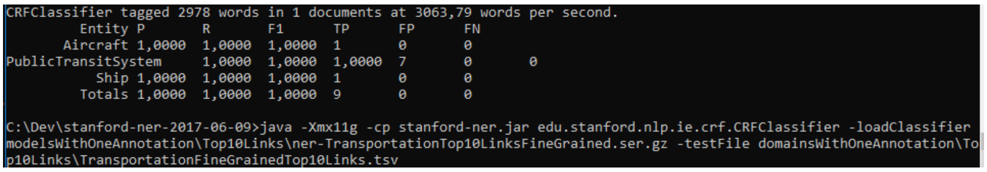
This model in really small, only 15 annotated words, that's because we cannot make a conclusion which model is better to use. But if we speak about performance, of coarse the better solution is specified domain.

The transportation model, like a sport domain, is also a very small, only 9 annotated words. We provide the experiments for coarse grained (see Figure 3.14) and we get same results as in all 3 domains (see Figure 3.9). Same is happening for the fine grained domain (see Figure 3.15). In such a small trained domain, we cannot clearly say which model is better to use.



```
CRFClassifier tagged 2978 words in 1 documents at 2951,44 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-TransportationTop10LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotatio
n\Top10Links\TransportationCoarseGrainedTop10Links.tsv
```

Figure 3.14: Transportation Coarse Top 10 Links Transportation



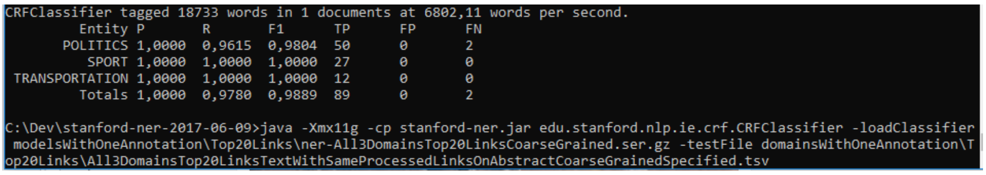
```
CRFClassifier tagged 2978 words in 1 documents at 3063,79 words per second.
         Entity P       R       F1      TP      FP      FN
       Aircraft 1,0000  1,0000  1,0000  1       0       0
PublicTransitSystem     1,0000  1,0000  1,0000  7       0       0
           Ship 1,0000  1,0000  1,0000  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-TransportationTop10LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\To
p10Links\TransportationFineGrainedTop10Links.tsv
```

Figure 3.15: Transportation Fine Top 10 Links Transportation

In conclusion of this subsection, we can say that our strategy for creating a specific domain pays off, because in those domains we have a better results that in global domains. But let see in subsection 3.2.2 how domains will behave with a bigger data(text).

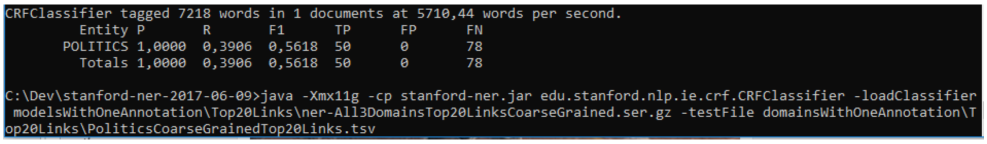
3.2.2 Top 20 Links

In subsection 3.2.1 we provide various experiments with different models. Now in this subsection we will repeat those experiments, but now we doubled the number of text to every domain, so we have 20 abstracts for every domain, in total 60 abstract. Those models, like previous, are trained in both coarse and fine grained. The first experiment in Figure 3.2 is now rerunned with a bigger model and text. From Figure 3.16 we can see that there are more annotated words and surprisingly better results than in previous domain. Then we tested the model with politics domain text in coarse grained annotation. Again the results are quite better than previous experiment (see Figure 3.3), also now there is no false positive words, but this is still not satisfying. After that we tested this model with sport domain text in coarse grained annotation (see Figure 3.18). And like in previous experiment (see Figure 3.4) we have a perfect results, so we can say that this kind of model with very small domain annotated words is good to use. And finally we tested the model with transportation domain text in coarse grained annotation (see Figure 3.19). Figure 3.19 show us that the model annotated some false positive word with SPORT, which even we have a perfect F1 value for transportation domain, in overall calculation, because of that word we have a lower F1 value than experiment in Figure 3.5.



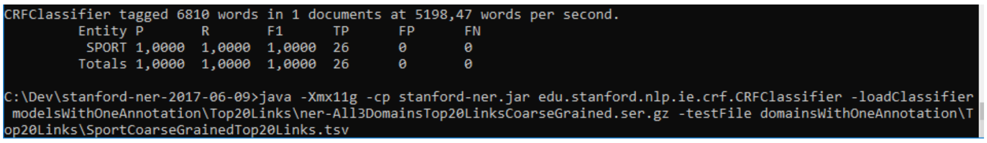
```
CRFClassifier tagged 18733 words in 1 documents at 6802,11 words per second.
         Entity P       R       F1      TP      FP      FN
       POLITICS 1,0000  0,9615  0,9804  50      0       2
          SPORT 1,0000  1,0000  1,0000  27      0       0
 TRANSPORTATION 1,0000  1,0000  1,0000  12      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 3.16: All 3 Domains Coarse Top 20 Links



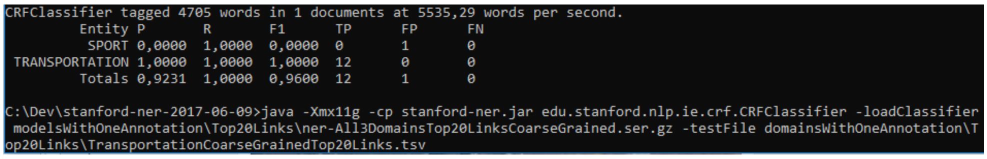
```
CRFClassifier tagged 7218 words in 1 documents at 5710,44 words per second.
         Entity P       R       F1      TP      FP      FN
       POLITICS 1,0000  0,3906  0,5618  50      0       78
         Totals 1,0000  0,3906  0,5618  50      0       78

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 3.17: 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
```

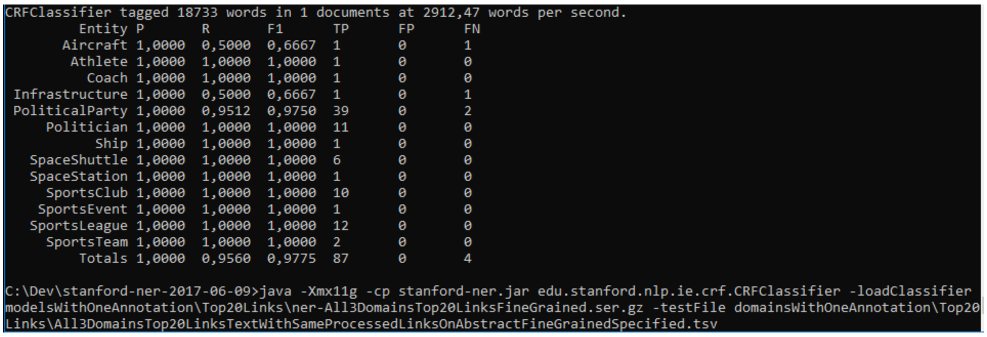
Figure 3.18: All 3 Domains Coarse Top 20 Links Sport



```
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
```

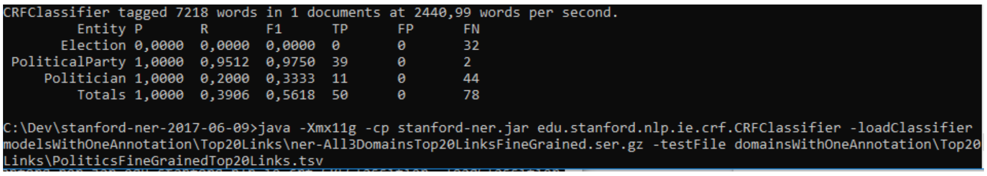
Figure 3.19: All 3 Domains Coarse Top 20 Links Transportation

Similarly like experiment with a coarse grained model, we repeated the same experiments, but now, of coarse, with fine grained model. We test the trained model with text from all 3 domains together in fine grained annotation, and how we can see from Figure 3.20 the results are quite worst that in experiment with coarse grained (see Figure 3.16), but better that the experiment where we have only 10 abstract (see Figure 3.6). Furthermore we test the model with politics text, and how we can see from Figure 3.17 again results are not even close to previous experiment with all 3 domains together where we have better results like overall also in politics domain. Then we move on with a sport text (see Figure 3.22) and again we have perfect results like in coarse grained (see Figure 3.18) and the experiment with 10 abstract (see Figure 3.8). And final experiment for this model is transportation text. Figure 3.23 shows us that results are worst that in coarse grained (see Figure 3.19). Now we do not have any false positive word annotated with SPORT domain, but we have a lower number of annotated words which results with a lower F1 value in total.



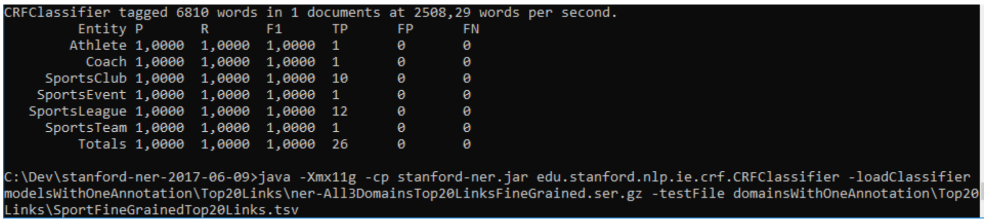
```
CRFClassifier tagged 18733 words in 1 documents at 2912,47 words per second.
         Entity P       R       F1      TP      FP      FN
       Aircraft 1,0000  0,5000  0,6667  1       0       1
        Athlete 1,0000  1,0000  1,0000  1       0       0
          Coach 1,0000  1,0000  1,0000  1       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 3.20: All 3 Domains Fine Top 20 Links



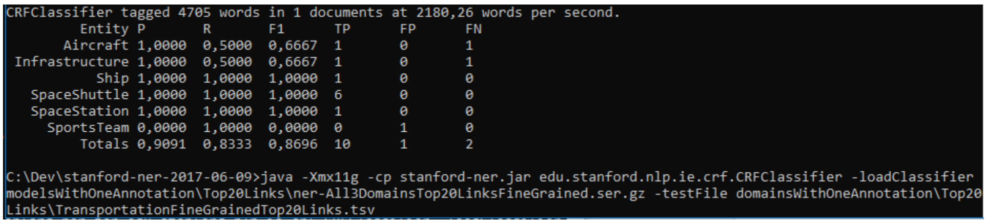
```
CRFClassifier tagged 7218 words in 1 documents at 2440,99 words per second.
         Entity P       R       F1      TP      FP      FN
       Election 0,0000  0,0000  0,0000  0       0       32
 PoliticalParty 1,0000  0,9512  0,9750  39      0       2
     Politician 1,0000  0,2000  0,3333  11      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 3.21: 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
```

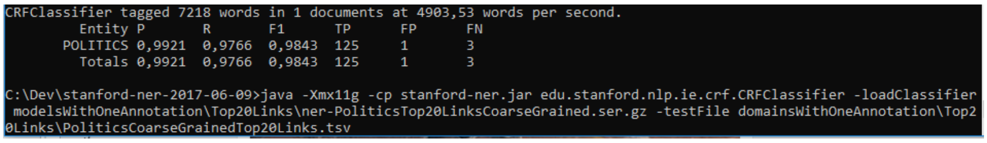
Figure 3.22: All 3 Domains Fine Top 20 Links Sport



```
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
```

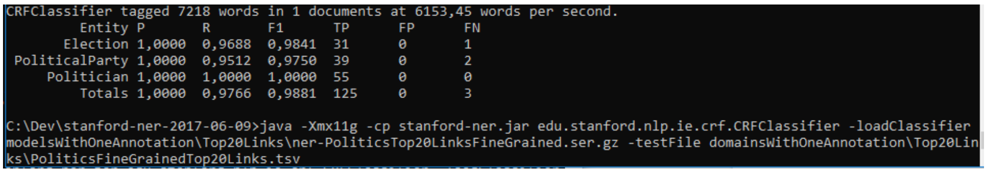
Figure 3.23: All 3 Domains Fine Top 20 Links Transportation

After we finish the experiments with models that are trained with texts from every domain, we do tests with models that are trained with texts from a particular domain (politics, sport and transportation). Figure 3.24 shows the output of politics coarse grain specified domain, where we can see that we have a way more better results than in experiment with a model trained with texts from all 3 domains (see Figure 3.16 and Figure 3.17). The politics specific domain finds 125 true positive words unlike the experiments with all 3 domains that gives us only 50 true positive annotated words. However we repeated this experiment, but now with model trained in fine grained. Figure 3.25 shows the result of experiment, where we can see that there is no false positive annotated word like in previous experiment (see Figure 3.24), which result with a better F1 value in total.



```
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\Top2
0Links\PoliticsCoarseGrainedTop20Links.tsv
```

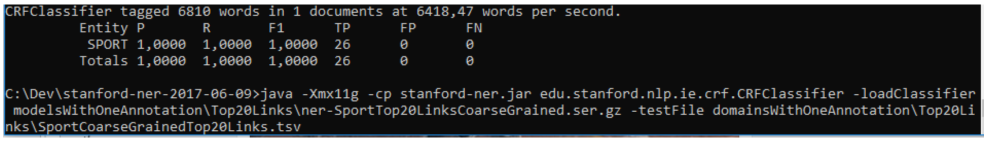
Figure 3.24: Politics Coarse Top 20 Links Politics



```
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
```

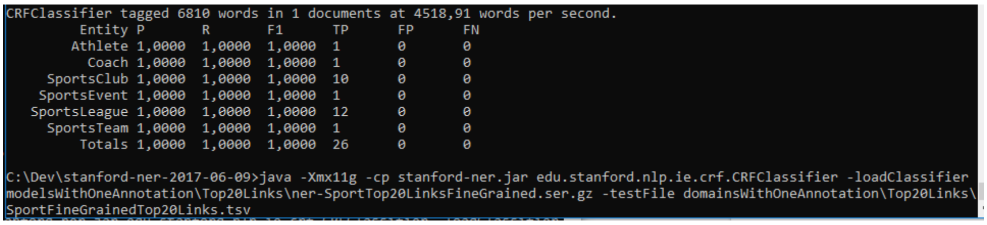
Figure 3.25: Politics Fine Top 20 Links Politics

Sport specific domain do not disappointed us. Figure 3.26 show the output of experiment in coarse grained and Figure 3.27 show the output of experiment in fine grained. Both experiment same as experiments with all 3 domains together (see Figure 3.16, Figure 3.18, Figure 3.20, Figure 3.22) give us perfect results. But if you want to be more precise, we recommend to use the specific model, because is fastest of reading words.



```
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
```

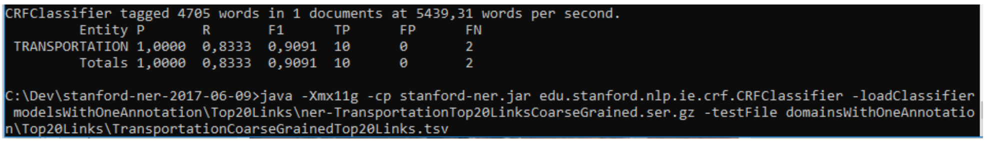
Figure 3.26: Politics Coarse Top 20 Links Politics



```
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\Top20Links\
SportFineGrainedTop20Links.tsv
```

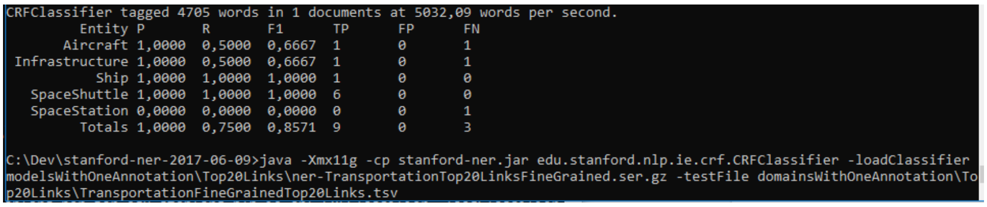
Figure 3.27: Sport Fine Top 20 Links Sport

At the end we make experiments with a transportation specific domain. In Figure 3.28 we can see that our model gives us a worst results that experiment in all 3 domains. For instance experiment in Figure 3.16 gives us a perfect annotation with 1.0000 value at F1, also experiment in Figure 3.5 gives the same result like previous experiment, but here because of false positive annotated word in SPORT domain in total we have a lower value in F1. In overall experiments in all 3 domains gives a better result that the experiment in specific domain. Of coarse we make an experiment with model that is trained with fine grained text. How Figure 3.29 shows that this model even gives worst result that the experiment with coarse grained model, and also worst results than the experiment in Figure 3.19.



```
CRFClassifier tagged 4705 words in 1 documents at 5439,31 words per second.
        Entity P       R       F1      TP      FP      FN
 TRANSPORTATION 1,0000  0,8333  0,9091  10      0       2
        Totals 1,0000  0,8333  0,9091  10      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-TransportationTop20LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotatio
n\Top20Links\TransportationCoarseGrainedTop20Links.tsv
```

Figure 3.28: Politics Coarse Top 20 Links Politics

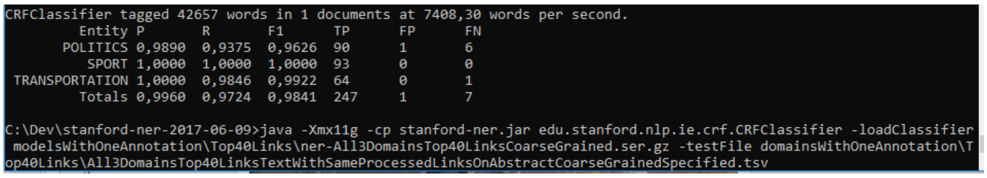


```
CRFClassifier tagged 4705 words in 1 documents at 5032,09 words per second.
        Entity P       R       F1      TP      FP      FN
      Aircraft 1,0000  0,5000  0,6667  1       0       1
 Infrastructure 1,0000  0,5000  0,6667  1       0       1
          Ship 1,0000  1,0000  1,0000  1       0       0
  SpaceShuttle 1,0000  1,0000  1,0000  6       0       0
  SpaceStation 0,0000  0,0000  0,0000  0       0       1
        Totals 1,0000  0,7500  0,8571  9       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-TransportationTop20LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\To
p20Links\TransportationFineGrainedTop20Links.tsv
```

Figure 3.29: Sport Fine Top 20 Links Sport

3.2.3 Top 40 Links

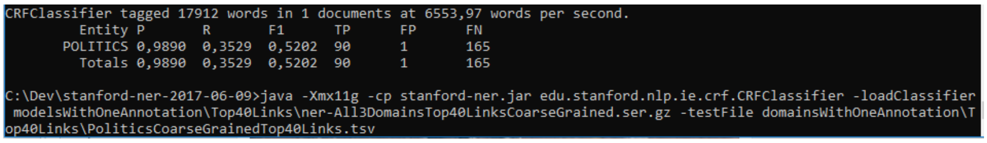
3.2.3.1 All 3 domains model annotated with coarse grained texts



```
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 3.30: All 3 Domains Coarse Top 40 Links

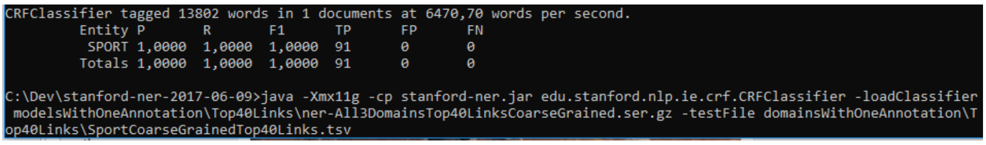
TEXT



```
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 3.31: All 3 Domains Coarse Top 40 Links Politics

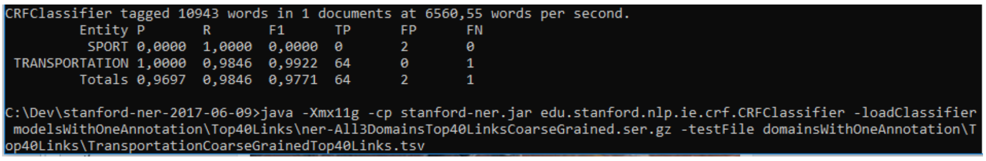
TEXT



```
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 3.32: All 3 Domains Coarse Top 40 Links Sport

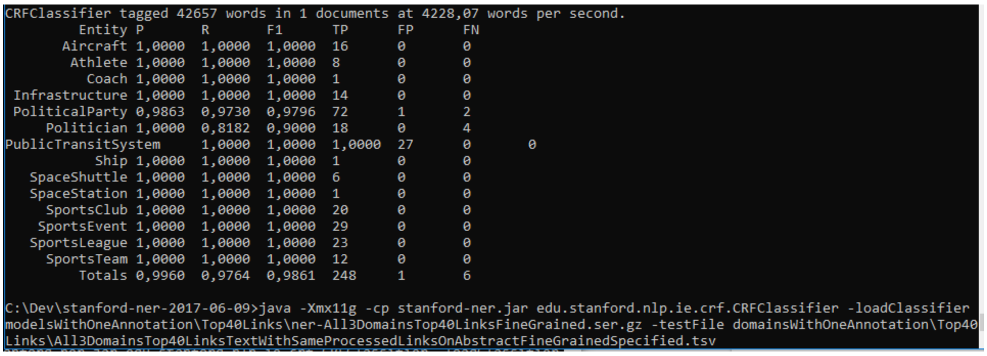
TEXT



```
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 3.33: All 3 Domains Coarse Top 40 Links Transportation

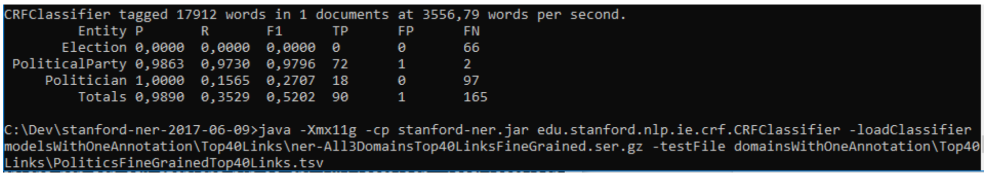
3.2.3.2 All 3 domains model annotated with fine grained texts



```
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 3.34: All 3 Domains Fine Top 40 Links

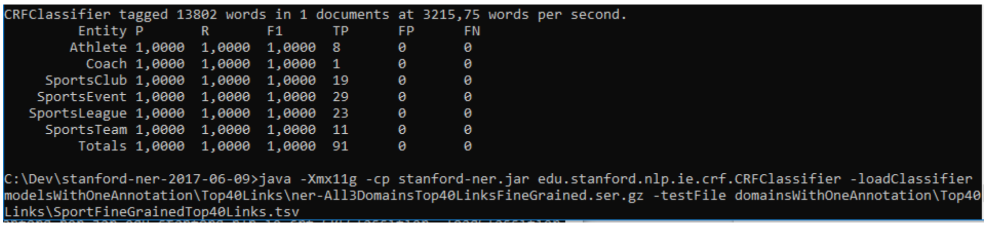
TEXT



```
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 3.35: All 3 Domains Fine Top 40 Links Politics

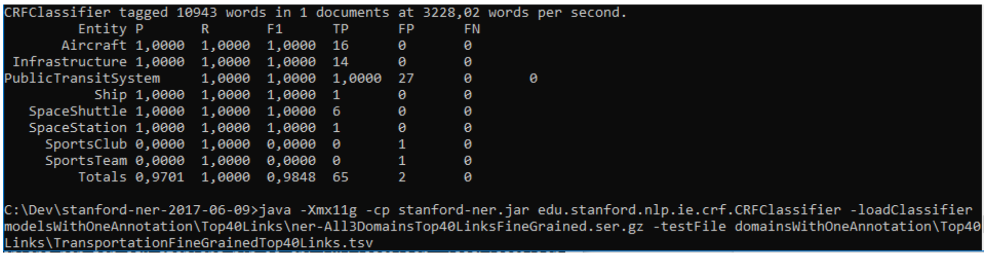
TEXT



```
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 3.36: All 3 Domains Fine Top 40 Links Sport

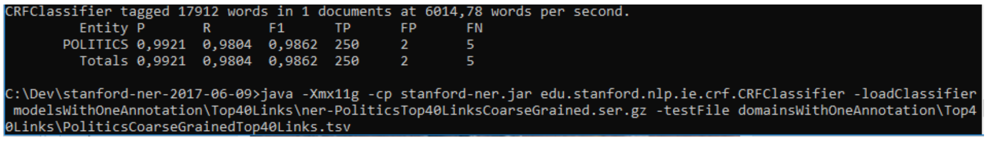
TEXT



```
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 3.37: All 3 Domains Fine Top 40 Links Transportation

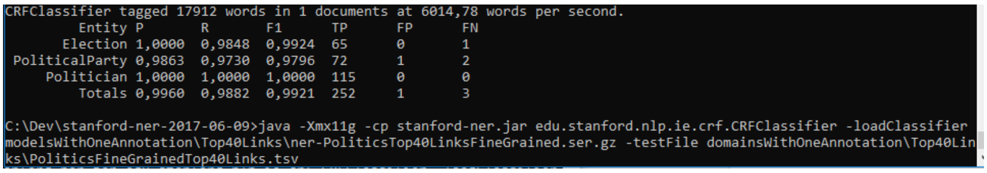
3.2.3.3 Politics specific domain



```
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 3.38: Politics Coarse Top 40 Links Politics

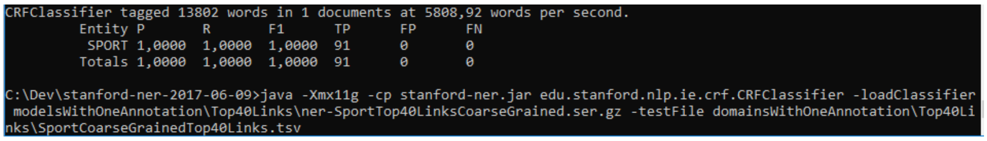
TEXT



```
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\Top40Lin
ks\PoliticsFineGrainedTop40Links.tsv
```

Figure 3.39: Politics Fine Top 40 Links Politics

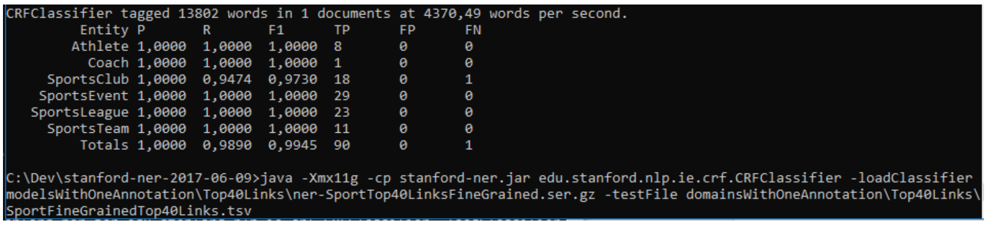
3.2.3.4 Sport specific domain



```
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 3.40: Sport Coarse Top 40 Links Sport

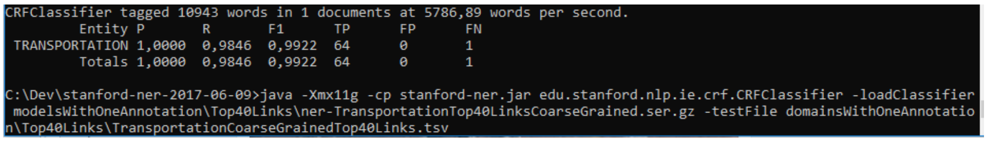
TEXT



```
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 3.41: Sport Fine Top 40 Links Sport

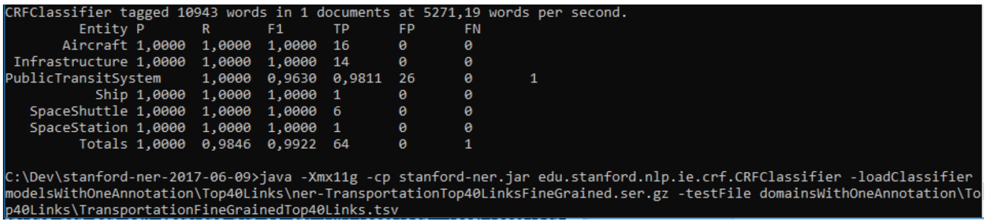
3.2.3.5 Transportation specific domain



```
CRFClassifier tagged 10943 words in 1 documents at 5786,89 words per second.
         Entity P       R       F1      TP      FP      FN
 TRANSPORTATION 1,0000  0,9846  0,9922  64      0       1
         Totals 1,0000  0,9846  0,9922  64      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-TransportationTop40LinksCoarseGrained.ser.gz -testFile domainsWithOneAnnotatio
n\Top40Links\TransportationCoarseGrainedTop40Links.tsv
```

Figure 3.42: Transportation Coarse Top 40 Links Transportation

TEXT



```
CRFClassifier tagged 10943 words in 1 documents at 5271,19 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  0,9630  0,9811  26      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
         Totals 1,0000  0,9846  0,9922  64      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-TransportationTop40LinksFineGrained.ser.gz -testFile domainsWithOneAnnotation\To
p40Links\TransportationFineGrainedTop40Links.tsv
```

Figure 3.43: Transportation Fine Top 40 Links Transportation

3.2.4 Top 100 Links

3.2.5 Top 300 Links

3.2.6 Top 400 Links

3.2.7 Top 500 Links

3.2.8 MIXED

3.2.9 With lower links on model

3.2.10 With higher links on model

3.3 Summary of results

3.3.1 Graphs

Conclusion

Appendix A

Contents of CD