Student Name: Student NetID:

University of Texas at Dallas
Department of Computer Science
CS6320 – Natural Language Processing
Spring 2021

Instructor: Dr. Sanda Harabagiu

Take-Home Final Exam

Issued: April 30th 2021

Due: May 7th 2021 – before Midnight

Instructions: Do not communicate with anyone in any shape or form. This is an independent exam. Do not delete this page or any problem formulation, just attach your answer in the space provided. If this page or any problem is deleted and you send only the answer, you shall receive ZERO points. If you do not write your name and netid, it will be considered that you did not submit your Final Exam and will obtain ZERO points. If you delete this page will obtain ZERO points.

Copy and paste the Final Exam into a Word document, enter your answers (either by typing in Word, or by inserting a VERY CLEAR picture of your hand-written solution) and transform the file of the exam into a PDF format. If we cannot clearly read the picture, you will get ZERO for that answer! If you create an enormous file for your final exam (i.e. larger than 5 Mbytes) you will receive ZERO for your entire exam. Please follow the instructions from the attached **instructions_submission_EXAM.pdf** file to make sure your final pdf file is of reasonable size.

If you will use a pencil instead of a black pen, you will receive a ZERO for the entire final exam.

Make sure that you insert EACH answer immediately after EACH question. Failure to do so will result in ZERO points for the entire exam! Submit the PDF file with the name **Final_Exam_netID.pdf**, where netID is your unique netid provided by UTD. If you submit your exam in any other format your will receive ZERO points.

The Final exam shall be submitted in eLearning before the deadline. No late submissions shall be graded! Any cheating attempt will determine the ENTIRE grade of the final exam to become ZERO.

Write your answers immediately after the problem statements. Write your answers immediately after the problem statements. If you enter multiple possible answers to the same problem – the most incorrect answer shall be selected and graded (as you are not sure about which answer is the correct one!)

Problem 1 (Information Extraction: Named Entity Recognition, Relation Extraction; Event and Temporal Inference) [TOTAL: **65 points**]

a) [**30 points**] Suppose you are building a Named Entity Recognition (NER) system that should identify three categories of names: (1) LOCATIONS; (2) ORGANIZATIONS and (3) PERSONS. However, your NER system will be providing fine-grained Location entity information by automatically recognizing different types of names of Locations, namely: (a) Countries, (b) States/Provinces, (c) Cities, (d) Oceans, (e) Rivers, (f) Seas and (g) Mountains. This entails that it will produce nested NER annotations, in which inside the [ ]$_{\text{LOCATION}}$ annotations you will nest fine grained location types as [ ]$_{\text{LOCATION:COUNTRY}}$, [ ]$_{\text{LOCATION:STATE}}$, [ ]$_{\text{LOCATION:CITY}}$, [ ]$_{\text{LOCATION:OCEAN}}$, [ ]$_{\text{LOCATION:RIVER}}$, [ ]$_{\text{LOCATION:SEA}}$, or [ ]$_{\text{LOCATION:MOUNTAIN}}$. Fine-grained location annotations will also be nested in the [ ]$_{\text{ORGANIZATION}}$ annotations, if needed. For example:

Instead of annotating: [San Francisco CA]$_{\text{LOCATION}}$ it will produce the nested annotations: [[San Francisco]$_{\text{LOCATION:CITY}}$ [CA]$_{\text{LOCATION:STATE}}$]$_{\text{LOCATION}}$.

Example 1: In this way, the NER annotation of the text

Naples, with its fantastic views of the Vesuvius and the Mediterranean, is a great tourist attraction of Italy.

-will become:

[[Naples]$_{\text{LOCATION:CITY}}$]$_{\text{LOCATION}}$, with its fantastic views of the [[Vesuvius]$_{\text{LOCATION:MOUNTAIN}}$]$_{\text{LOCATION}}$ and the [[Mediterranean]$_{\text{LOCATION:SEA}}$]$_{\text{LOCATION}}$, is a great tourist attraction of [[Italy]$_{\text{LOCATION:COUNTRY}}$]$_{\text{LOCATION}}$.

Example 2:

Johnny Arrow graduated from University of Texas at Dallas before moving to San Francisco CA. Johnny now enjoys the Pacific Ocean's views.

Will be annotated as:

[Johnny Arrow]$_{\text{PERSON}}$ graduated from [University of [Texas]$_{\text{LOCATION:STATE}}$ at [Dallas]$_{\text{LOCATION:CITY}}$]$_{\text{ORGANIZATION}}$ before moving to [[San Francisco]$_{\text{LOCATION:CITY}}$ [CA]$_{\text{LOCATION:STATE}}$]$_{\text{LOCATION}}$. [Johnny]$_{\text{PERSON}}$ now enjoys the [[Pacific Ocean]$_{\text{LOCATION:OCEAN}}$]$_{\text{LOCATION}}$'s views.

a.i) You are asked to annotate the following text with Named Entities, in which the fine-grained nested Location annotations like those provided in the two examples above may be produced. [**5 points**]

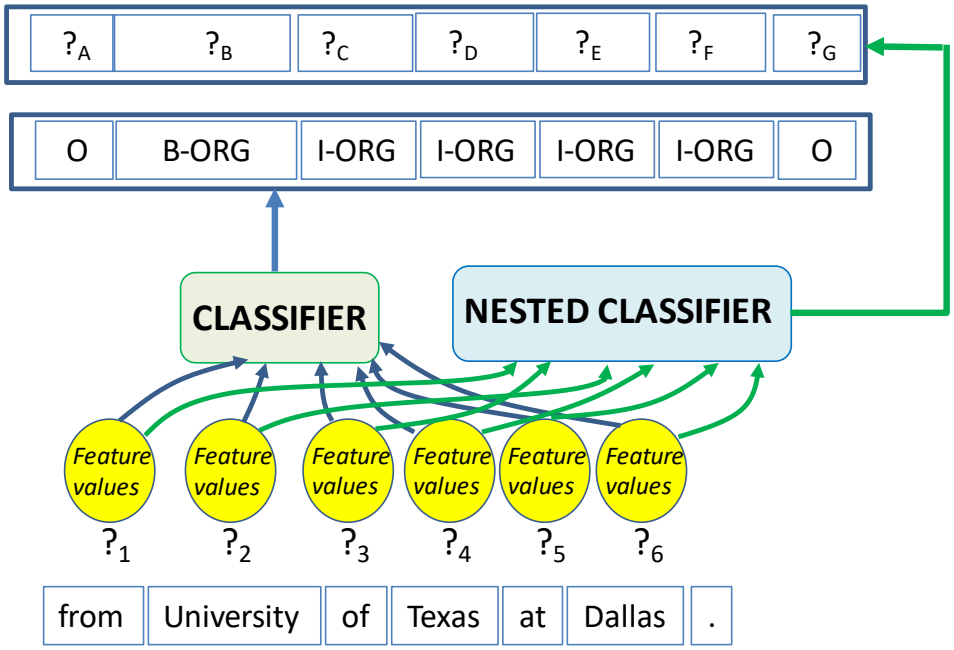
Paris offers incredible views from the Seine, with France's most famous monuments facing it. The French Office of Tourism is promoting cruises on the river and offers free guides. Anna Hidalgo, the Mayor of Paris, recently solved the problem of crowded museums by opening new attractions and providing cheap trips to the Alps or to Dauville, by the Atlantic.

a.ii) How would you design the sequence labeling for the nested location named entities? Be specific. Show how your tags will operate on Example 2. [**10 points**] Use the following format:

Words	IOB Label
American	B-ORG
Airlines	I-ORG
,	O
a	O
unit	O
of	O
AMR	B-ORG
Corp.	I-ORG
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B-PER
Wagner	I-PER
said	O

a.iii) Gazetteers provide names of many locations, such as (a) Countries, (b) States/Provinces, (c) Cities, (d) Oceans, (e) Rivers, (f) Seas and (g) Mountains. Show which features you would use to train an NER capable of identifying nested Location names in order to take advantage of Gazetteers. [**2 points**]

a.iv) (**13 points**) Consider that you train jointly a classifier and a nested-classifier to assign the NER tags, as in the following Figure.



[**1 point**] Fill 2 feature values examples for $?_1$:

[**1 point**] Fill 2 feature values examples for $?_2$:

[**1 point**] Fill 2 feature values examples for $?_3$:

[**1 point**] Fill 2 feature values examples for $?_4$:

[**1 point**] Fill 2 feature values examples for $?_5$:

[**1 point**] Fill 2 feature values examples for $?_6$:

[**1 point**] What is the tag for the annotation $?_A$:

[**1 point**] What is the tag for the annotation $?_B$:

[**1 point**] What is the tag for the annotation $?_C$:

[**1 point**] What is the tag for the annotation $?_D$:

[**1 point**] What is the tag for the annotation $?_E$:

[**1 point**] What is the tag for the annotation $?_F$:

[**1 point**] What is the tag for the annotation $?_G$:

b) [**10 points**] Suppose you are building an information extraction system to identify the city and state in which a person was born. You want to use bootstrapping to do this. You know where Barack Obama was born (Honolulu, Hawaii) but you don't know where any other famous person was born. Describe how you could use this information, along with a combination of Google and Wikipedia, to find patterns that could be used in general to determine place of birth. Be specific. Give at least 4 examples.

c) [**25 points**] You need to produce fine-grained annotations for events, temporal expressions and three forms of temporal links, namely TLINK, ALINK and SLINk.

As in TimeML, you are considering the following seven classes of events: (1) Occurences; (2) States; (3) Reporting; (4) I-Actions; (5) I-States; (6) Aspectuals and (7) Perceptions.

Consider the text:

The examinations started at the beginning of May 2020. Faculty were all on board to request distance-learning exams. It was said that students living on campus since January 1, 2020 tried to cancel their studies before 5pm today. They watched their friends packing their stuff since March 12 2020.

1. **[12 points]** Given the above text, identify all events from the seven classes considered in TimeML. Mark up the identified events in the following way:

Once they [find]OCCURENCE their friends, they [hope]I-STATE that they will [leave]OCCURENCE together.

Solution:

2. **[3 points]** Identify all temporal expressions and temporal signals in the text:

   The examinations started at the beginning of May 2020. Faculty were all on board to request distance-learning exams. It was said that students living on campus since January 1, 2020 tried to cancel their studies before 5pm today. They watched their friends packing their stuff since March 12 2020.

   Mark up the identified temporal expressions using the TIMEX notations and the temporal signals in the following way:

<TSIGNAL>Since</TSIGNAL> <TIMEX3 ID=t1, TYPE=DATE, VALUE="2020-02-02">Monday, February 2nd 2020 </TIMEX3>, Jane has disappeared. This is not the first time. She was gone <TSIGNAL>before</TSIGNAL> for a whole <TIMEX3 ID=t2, TYPE=DURATION, VALUE=P4W>month</TIMEX3>.

Solution:

3. **[10 points]** Indicate the TLINKS, SLINKS and ALINKS that you identify in the above text. Use the following format for the temporal relations:

   [EVENT 1] →TLINK/SLINK/ALINK[Type] →[EVENT 2]

   For example, in the text:

Once they [find]OCCURENCE their friends, they [hope]I-STATE that they will [leave]OCCURENCE together.

You have the following temporal relations:

   [*find*] →TLINK[Before] →[*hope*]

   [*hope*] →SLINK[Modal] →[*leave*]

Solution:

Problem 2 (Semantic Role Labeling) (**25 points**)

a] Given the sentence:

S1: A new book explores the amazing national parks that are protected against development and ruin for decades now.

Using the PropBank definition of predicates and arguments available from: *http://verbs.colorado.edu/verb-index/index.php*, identify manually the predicates and their arguments in the above sentence. (**10 points**)

Use the following format as shown for the sentence:

S2: PS of New Hampshire has proposed an internal reorganization plan.

Predicate 1 [S2]: Propose
Arg0: *proposer:* [PS of New Hampshire]
Rel: [has proposed]
Arg1: *proposition:* [an internal reorganization plan]

Predicate 2 [S2]: ????

Solution:

b] Draw the dependency parse of the sentence S3 (**5 points**):

S3: A new book explores the amazing national parks and shows the beauty of the nature.

and explain how you would use stack-pointer dependency networks for generating the dependency parse (**10 points**).

Solution:

Problem 3 (Word Sense Disambiguation) (**10 points**)

Consider the noun "cloud" which is semantically ambiguous. Three different senses from WordNet are provided to you along with their glosses:

SENSE 1: cloud -- (a visible mass of water or ice particles suspended at a considerable altitude)

SENSE 2: cloud -- (out of touch with reality; "his head was in the clouds")

SENSE 3: cloud -- (a cause of worry or gloom or trouble; "the only cloud on the horizon was the possibility of dissent by the French")

Use the semantic definitions of the senses of the noun "cloud" in the following three texts:

TEXT 1: *Everything in Brazil can seem bigger than elsewhere - and the same goes for its crises. The country has barely emerged from its worst-ever recession and now it is under the cloud[1] of economic crisis and the possibility of a turn towards populism seems real enough.*

TEXT 2: *For the record, I am not anti-romance or love at first sight, for that matter. It is just that, I find that when people have their heads in the clouds[2] they often don't see the crash coming. I have seen the same thing happen in relationships and jobs. One minute they will tell you how wonderful everything is and the next why they had no choice but to leave.*

TEXT 3: *All clouds[3] are made up of basically the same thing: water droplets or different ice crystal that float in the sky. But all clouds[4] look a little bit different from one another, and sometimes these differences can help us predict a change in the weather.*

3.a.[**4 points**]: What are the semantic senses of :

Cloud[1]= Sense ???

Cloud[2]= Sense ???

Cloud[3]= Sense ???

Cloud[4]= Sense ???

3.b. [**3 points**] For disambiguating “cloud” in Text 1, show the feature vector that would be used if you were considering collocational features of a window of size + or -3.

3.c. [**3 points**] For disambiguating “cloud[4]” in Text 3, show the feature vector that would be used if you were considering the bag of words approach with a vocabulary of 10 words: {blue, river, ice, mountain, sky, sun, water, weather, wind, zen}.