The University of Texas at Dallas
CS 6320
Natural Language Processing
Spring 2020
Instructor: Dr. Sanda Harabagiu
Grader/ Teaching Assistant: Ramon Maldonado

Homework 3: 100 points (50 points extra-credit)
Issued April 14, 2020
Due April 30, 2020 before midnight

PROBLEM 1: Word Sense Disambiguation (WSD) (**40 points**)
In this problem you asked to use the WordNet thesaurus/semantic dictionary to disambiguate the sense of all the words from the following short text:

The singer will not be performing at this year's festival after all. An accomplished singer, songwriter, dancer, and style guru, she graced the festival's stage last year when she joined her big sister during her historic, headlining tour.

Because WordNet encodes the word senses of nouns, verbs, adjectives and adverbs, you are asked to identify the semantic senses only of the words or word expressions that have the roles of nouns, verbs, adjectives or adverbs. Note: a large percentage of the synset entries in WordNet are multi-word expressions, e.g. {"line up", "get hold", "come up", "find"} or {"girl scout"}. That entails that you need to search the senses of multi-word expressions as well, as encountered in the short text above, not only of single words.

Specify if you have used the online WordNet or if you have downloaded it, and in that case, which version of WordNet you have downloaded.

Present your annotations in the following format:

If the text is "Her brother killed the idea of a trip."
The word-sense annotations would be:
Her **brother:n#1 killed:v#2** the **idea:n#2** of a **trip:n#1**.

In this way you indicate the **word** from the text, i.e. **brother** immediately followed by **:POS,** i.e. **:n** (in this case the possible POS are: n – for noun, v- for verb; a – for adjective; r- - for adverb), Do not use the POS tag from the Penn Treebank – that would be incorrect! After that you immediately write #Number – to indicate the sense number of the **word**, as you find it in WordNet.

a/ Produce the manual disambiguation of the text (**10 points**)

b/ Implement the simple Lesk WSD algorithm that considers the glosses of each sense of the words in your text and combines them with the context in which they appear in the text to predict the semantic sense of the words or multi-word expressions. Compare the results against your manually coded senses – and compute the accuracy of your automatically generated word senses. (**30 points**)

Software Engineering (includes documentation for your programming assignments)

Your README file must include the following:

- Your name and email address.
- *Homework number* for this class (NLP CS6320), and the *number of the problem* it solves.
- A description of every file for your solution, the programming language used, supporting files, any NLP tools used, etc.
- How your code operates, in detail.
- A description of special features (or limitations) of your code.

Within Code Documentation:

- Methods/functions/procedures should be documented in a meaningful way. This can mean expressive function/variable names as well as explicit documentation.
- Informative method/procedure/function/variable names.
- Efficient implementation
- Don't hardcode variable values, etc

PROBLEM 2: Information Extraction (**60 points**)
a/ Identify manually the Named Entities in the following short text: (**10 points**)

Solange Knowles will not be performing at this year's Coachella Music and Arts Festival after all. The music festival's official Twitter account announced the news Sunday night. Knowles released her new album, "When I Get Home," last month. An accomplished singer, songwriter, dancer, DJ and style guru, Knowles graced the Coachella stage last year when she joined her big sister, Beyoncé, during her historic, headlining tour. Ariana Grande, Childish Gambino, and Tame Impala are headlining the festival this year. Coachella takes place annually over two weekends in April in Indio, California.

Use only the following Name Entity Categories and their respective acronyms:
PERSON (PER), LOCATION (LOC), ENTERTAINMENT-EVENT (ENT), ORGANIZATION (ORG), ALBUM-TITLE (ALB).

The document was posted on April 14th, 2019.

Use the following format to mark the named entities that you identify:

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

State how many mentions of Named entities you have found and note how many entities were Persons, how many were Locations, how many were Entertainment Events, how many were Organizations and how many were Album Titles.

b/ (**5 points**) Use the IOB notation to indicate the labels that you would assign to the text from Problem 2.a. To indicate the labels, use the following table formatting for each sentence:

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
.	O

Generate a different table for each sentence in the text, e.g. Table 1 for sentence 1, Table 2 for sentence 2 etc.

c/ Annotate manually the temporal expressions in the text from Problem 2.a, using the following format (**5 points**):

A fare increase initiated <TIMEX3>last week</TIMEX3> by UAL Corp's United Airlines was matched by competitors over <TIMEX3>the weekend</TIMEX3>, marking the second successful fare increase in <TIMEX3>two weeks</TIMEX3>.

d/ (**10 points**) Perform temporal normalization of the temporal expressions you have identified in the text from Problem 2.a and produce an annotation of the normalized temporal expressions in the text.

e/ (**25 points**) Identify in the text from Problem 2.a all the relations of the following types:

TYPE 1: PERSON → *Entertainer* → ENTERTAINMENT-EVENT (PER → *Entertainer* → ENT)

TYPE 2: PERSON → *Family* → PERSON (PER → *Family* → PER)

TYPE 3: PERSON → *Involved-in* → ALBUM-TITLE (PER → *Involved-in* → ALB)

TYPE 4: ENTERTAINMENT-EVENT → *Happens-in* → LOCATION (ENT → *Happens-in* → LOC)

TYPE 5: ENTERTAINMENT-EVENT → *Communicates-through* → ORGANIZATION (ENT → *Communicates-through* → ORG)

When presenting the relations that you manually discover, use the following format:
$[\text{Tim Warner}]_{\text{PER}} \rightarrow \text{Spokesman} \rightarrow [\text{American Airlines}]_{\text{ORG}}$

Indicate how many relations of each type that you have discovered.

f/ (**5 points**) Provide examples of five classes of feature you would use to discover the five types of relations you have identified manually. Select one of your relations and show the feature values that you extract from the text from Problem 2.a for 2 of the five feature you selected. Discuss the classifier you would use for recognizing automatically these relations.

EXTRA-CREDIT PROBLEM 2 (50 points):

The main goal of the extra-credit problem 2 is test your ability to manually identify the event timeline that can be inferred from the text from Problem 2.a.

a/ (**20 points**) Identify the events in the text and annotate them in the TimeBank format, e.g.:

```
<TIMEX3 tid="t57" type="DATE" value="1989-10-26"  functionInDocument="CREATION_TIME">
10/26/89  </TIMEX3>

Delta Air Lines earnings <EVENT eid="e1" class="OCCURRENCE"> soared </EVENT> 33% to a
record in  <TIMEX3 tid="t58" type="DATE" value="1989-Q1" anchorTimeID="t57"> the
fiscal first quarter </TIMEX3>, <EVENT eid="e3"  class="OCCURRENCE">bucking</EVENT>
the industry trend toward <EVENT eid="e4" class="OCCURRENCE">declining</EVENT>
profits.
```

Each event must receive its event ID and EVENT CLASSES in the annotation.

TimeML considers “events” (and the corresponding tag) as a cover term for situations that happen or occur. Events can be punctual or last for a period of time. TimeML also considers as events those predicates describing states or circumstances in which something obtains or holds true. Not all stative predicates are marked up, however, as only those states which participate in an opposition structure in a given text are marked up.

Events are generally expressed by means of tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases. The specification of EVENT in TimeML is shown below:

attributes ::= eid class tense aspect

eid ::= ID {eid ::= EventID EventID ::= e}

class ::= ’OCCURRENCE’ | ’PERCEPTION’ | ’REPORTING’ | ’ASPECTUAL’ | ’STATE’ | ’I_STATE’ | ’I_ACTION’ | ’MODAL’ tense ::= ’PAST’ | ’PRESENT’ | ’FUTURE’ | ’NONE’ aspect ::= ’PROGRESSIVE’ | ’PERFECTIVE’ | ’PERFECTIVE_PROGRESSIVE’ | ’NONE’

Examples of each of these event types are given below:

1. **Occurrence**: die, crash, build, merge, sell
2. **State**: on board, kidnapped, love, ..
3. **Reporting**: say, report, announce, 4. I-Action: attempt, try, promise, offer
5. **I-State**: believe, intend, want
6. **Aspectual**: begin, finish, stop, continue.
7. **Perception:** See, hear, watch, feel

b/ (**20 points**) Using Allen's definitions of temporal relations, establish the temporal relations in the text from Problem 2.a. shared between pairs of events or events and the normalized temporal expressions you have discovered in Problem 2.d.
Provide your temporal relations in the following format:

Soaring$_{e1}$ is **included** in the fiscal first quarter$_{t58}$
Soaring$_{e1}$ is **before** 1989-10-26$_{t57}$
Soaring$_{e1}$ is **simultaneous** with the bucking$_{e3}$
Declining$_{e4}$ **includes** soaring$_{e1}$

c/ (**10 points**) Draw the graphical representation of the timeline of the events and temporal expressions in the text from Problem 2.a., in the format: