Comp550 a1

Yiran Mao

260850827

Question 1: Find Cases of Ambiguity

1. SMU's Collins Named 2018-19 Conference Swimming Chair

URL:

https://swimswam.com/smus-collins-named-2018-19-conference-swimming-chair/

Ambiguity: lexicon--- different meaning of one word

There are two words in this sentence that cause ambiguity, "named" and "chair".

"chair" can be the highest officer of an organized group, but it can also be a type of seat. "named" can be the past tense of the verb "name". Also, it can be the past particle. And the meanings are totally different. As a title of a news, this sentence is not complete.

So the meaning of the sentence can be: SMU's Collins gave a name to… chair (a type of seat). And it can also be: SMU's Collins has been named the 2018-19 conference swimming chair (the highest officer of an organized group).

To disambiguate the sentence, we(machine) need to know the relationship between "Collins" and "Chair". If they are the same thing, the meaning of the sentence should be the second one. And obviously we can find the relationship in context.

2. We are very relieved that there has been a positive outcome to the kidnapping and are very grateful for the excellent support we have received.

URL: https://www.telegraph.co.uk/news/2018/05/13/british-tourists-kidnapped-democratic-republic-congo-released/

Ambiguity: Part of Speech

“Relieved” can be an adjective or the past dense of the verb “relieve”.

To disambiguate the sentence, we need to know that the word “very” is always followed by adjectives.

3.Flying planes can be dangerous.

URL: https://www.quora.com/What-are-the-two-interpretations-of-the-sentence-Flying-planes-are-dangerous

Ambiguity: syntactic

The two meanings are:

"It can be dangerous for a person to fly planes" -- perhaps because they might have an accident, crash the plane, etc.

"Planes that are flying in the air can be dangerous" -- perhaps because they might come crashing down on you, run into other planes, etc.

To disambiguate this sentence, we need to find out the reason that causes the danger, as mentioned above.

4. Japan’s Abe finds himself on sidelines amid outreach with North Korea.

URL: https://www.washingtonpost.com/world/japans-abe-finds-himself-on-sidelines-amid-outreach-with-north-korea/2018/09/23/5dce8842-bdac-11e8-97f6-0cbdd4d9270e_story.html?noredirect=on&utm_term=.eae7929f7db1

Ambiguity: Phonological

The word “Korea” has the same pronunciation with the word “career”.

To disambiguate this sentence, we need to know the knowledge that North Korea is a proper noun which means a country.

5. Can I marry a girl of my same age?

URL: https://www.quora.com/Can-I-marry-a-girl-of-my-same-age

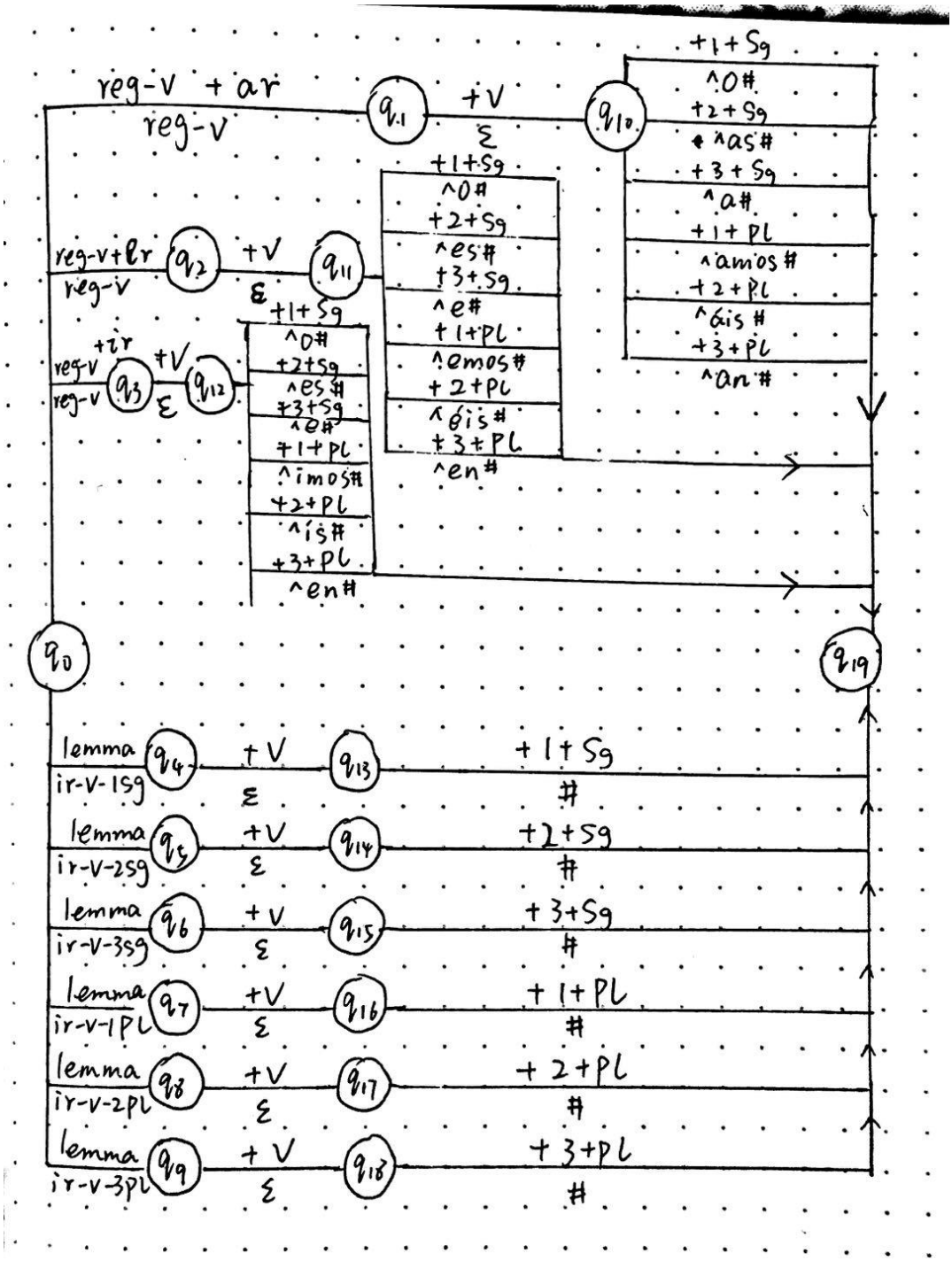
Ambiguity: Semantics

“a girl of my same age” can be construed as either specific or nonspecific.

To disambiguate this sentence, we need to know if the girl here is specific. We need to find it in context.

Question 2: FST for Spanish Verbal Conjugation

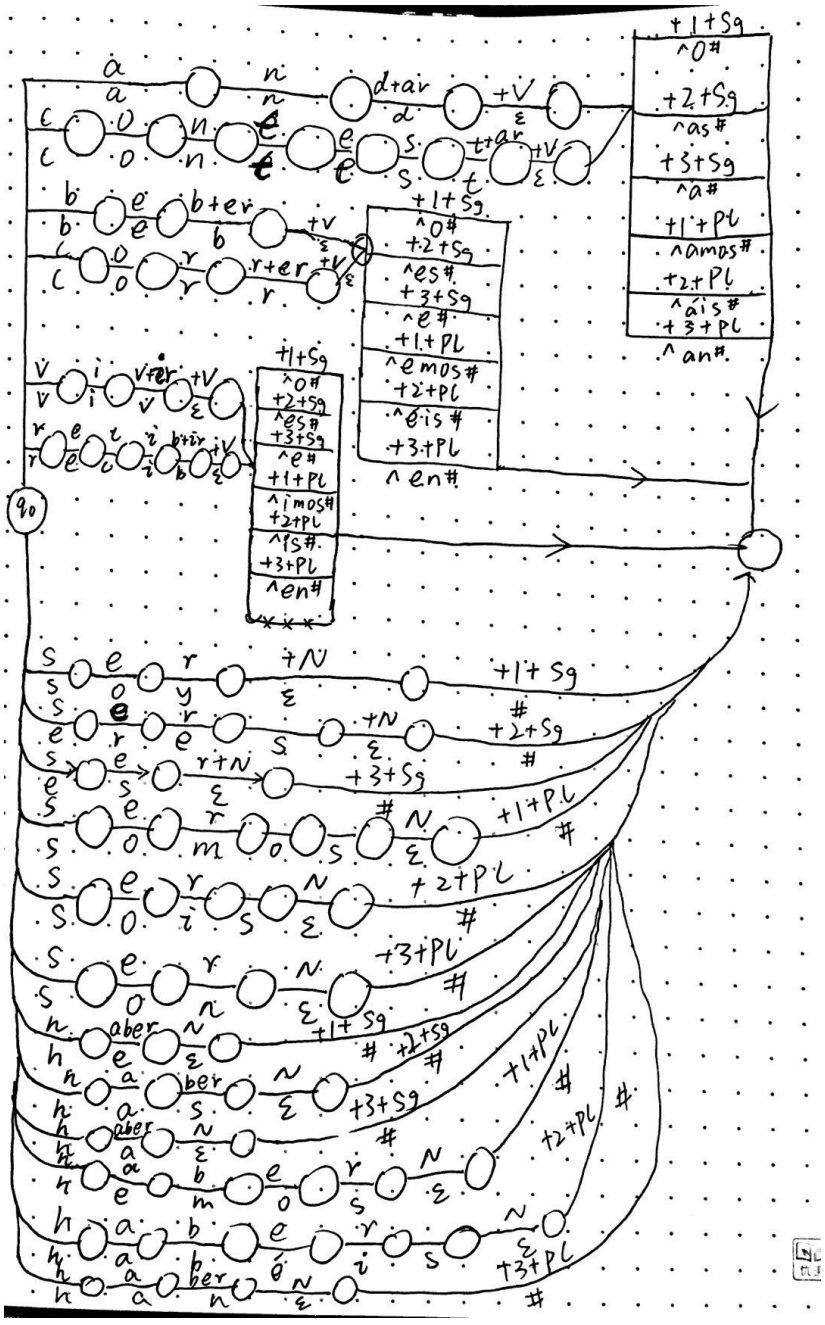
(1) A schematic transducer:



(2) A lexicon table

reg-V	irreg-1-Sg	irreg-2-Sg	irreg-3-Sg	irreg-1-Pl	irreg-2-Pl	irreg-3-Pl
andar	soy	eres	es	somos	sois	son
contestar	~~he~~	has	ha	hemos	habéis	han
beber						
correr						
vivir						
recibir						

(3) A "fleshed-out" FST



Question 3: Sentiment Analysis

Experiment Setup:

Python3, scikit-learn v0.19.2

Experiment Procedure:

To preprocess the input document, I use unigram method (CountVectorizer in sk-learn) to extract features from sentences. And the label for each sentence is ‘pos’ for positive data and ‘neg’ for negative data.

Then I complemented and compared three main classifiers: Logical Regression, Naïve Bayes and Support Vector Machine.

At the end I tuned the parameters in class CountVectorizer. (See Parameter setting for detail.)

Parameter Setting:

Set stop words as ‘English’. However, the performance of the model became worse.

Set max_df to 2500, the performance of LR became better but that of NB became worse.

```
Logistic Regression:
0.7711927981995499
```

Set min_df from 0 to 4, the performance of NB became better, the performance of LR became worse; from 4 to ~, both of them became worse.

```
Logistic Regressior
0.7426856714178545
GaussianNBClass:
0.7123030757689423
```

Conclusion:

At most time when setting different parameters, the performance of LR remains the best. But all of them are much better than the random baseline (accuracy:0.51).

```
C:\Users\MaoMao>python "D:\work\Mcgill\comp550\a1\rt-polaritydata\new 1.py"
Logistic Regression:
0.7595648912228057
GaussianNBClass:
0.6691672918229558
SVM:
0.7520630157539385
```

When set max_df to 2500, the confusion matrix and accuracy of LR are:

```
Logistic Regression:
[[1051  282]
 [ 328 1005]]
0.7711927981995499
```