1 Coreference Resolution

1.1 Understanding the task and representation

1.1.1

1.1.2

1.1.3

The coreference chains in the text:

Chinese tourist: [Chinese tourist, A Chinese tourist, she, the 61-year-old woman, She, her, she, the woman, The ill Chinese woman, Her, The Chinese woman, Her],

Public Health Minister Anutin Charnvirakul: [Public Health Minister Anutin Charnvirakul, Mr Anutin, Mr Anutin, he, Mr Anutin, Mr Anutin, Mr Anutin, The Public Health Ministry,he],

59 people in China: [59 people in China, them, All, They],

Health officials:[Health officials, They]

The new strain of coronavirus: [the new strain of coronavirus, the new strain of the coronavirus, which, the virus, the virus, the new coronavirus,it, the new virus strain]

1.1.4

1.1.5

Table 1:

Real-world entity	Mentions	Singleton?	Anaphoras	cataphoras	bridging anaphora
A Chinese tourist	1.A Chinese tourist was found 2.the 61-year-old woman was 3.close to the woman 4.The ill Chinese woman.	No	1. when she arrived in Thailand		1. She now had no fever. 2.gave her a clearance she would be
the new strain of coronavirus	1.with the new strain of coronavirus. 2. infected with the new strain of the coronavirus. 3. transmission of the virus. 4.with the virus outside	No			
Thailand	1. arrived in Thailand.	Yes			
China	1. people in China 2. in central China 3.outside China	No			
Public Health Minister Anutin Charnvirakul	1.Public Health Minister Anutin Charnvirakul 2. Mr Anutin 3. said Mr Anutin. 4.Mr Anutin said.	No			1.he said
Sixteen other people who	1.Sixteen other people who	N0	1.the results were		
59 people	1. 59 people	No	1.which has been linked		1.One of them died 2.All had attended big markets 3.They were either workers
Wuhan city	1.Wuhan city	Yes			
Bamrasnaradura Infectious Diseases Institute in Nonthaburi province.	1.Bamrasnaradura Infectious Diseases Institute in Nonthaburi province.	Yes			

1.2 Towards designing a system

1.2.1 Analyzing the phenomenon

1)

Table 2:

Item	Antecedent	Reason
Mr Anutin	Public Health Minister Anutin Charnivirakul.	1. Match name Anutin 2. Mr is man and there is only male in sentence
She	61-year-old woman	1. She means female and there is only female in sentence. 2. Woman is close to she.
them	59 people	1. Them means many people and 59 people close to this word them.

2)

1. Kim and his brother bet that Kim could take a picture of himself while doing a handstand.

Antecedent: Kim. Because Kim is the closest words to himself. And self means it represents subject.

2. Kim and his brother bet that Kim could take a picture of him while doing a handstand.

Antecedent: his brother. Because Kim takes the action-"takes the picture" and doing a handstand. And if him means Kim, it should be himself.

3.I went with John to the market yesterday and then took a walk in the forest with Mark. He was in a cheerful mood.

Antecedent: Mark. He is close to Mark and then means that this is second action of I.

4. John met Mark for dinner. He felt like having company.

Antecedent: John. Because John takes the action - met. Thus, the subject of felt is John.

5. Kim was stopped by a policeman. He was not pleased.

Antecedent: Kim. Because stopped is a passive word and it emphasizes the subject Kim. Thus, he is Kim.

6. John invited Mark for dinner. He knew he was broke.

Antecedent: he is Mark. John takes the action invited. Thus subject of knew is John and according to the meaning subject of broke is Mark.

7. John invited Mark for dinner. He knew he could please him with his latest recipe for lava cake.

Antecedent: he is John. John takes the action invited and Thus John makes the dinner. And John makes dinner to please Mark.

8. John has a bike, but no car. It is red.

Antecedent: bike. Because there is a negation-no and it means John has no car. Thus it means John.

9. He brought flowers and chocolates. They tasted great.

Antecedent: flowers and chocolates. They means many and in this sentence it must be flowers and chocolates. And tasted is passive.

10. She poured water from the pitcher into the cup until it was full.

Antecedent: Cup. Because the water is from pitcher into cup. And full means it is cup

11. She poured water from the pitcher into the cup until it was empty

Antecedent: pitcher. Because the water is from pitcher into cup. And empty means it is pitcher

3)

Sentence 1, 2,3,4,5. are Syntactically impossible. Because if Kim want to take a picture of Kim. The format should be himeself, otherwise it is him.

Sentence 6,7,8 are Semantically impossible. Because 8 there is negation.

Sentence 9 are world knowledge because it should like this and it is more logical.

Sentence 10,11 should be structure because from something to something.

1.2.2 Towards system design

System

(a)

For syntactic, semantic and structure, I think the most important thing is the structure of sentence. First, tokenization need to be done to split the words and to find the coreference words. Then dividing the sentence to word phrases using dependency plot to find the. For example using verb phrases to find noun phrase.

For word experience based. It is kind of difficult.

(b)

The first step is searching the noun phrase, because noun phrase can be the candidate of antecedent. Besides, the mentions are usually largest unit noun phrase. The next step is matching constraints for pronouns. Using method to backward search the candidates to match the agreement constraints, such as singular, plural agreement. Another contraints is from syntax. The main idea is dividing the sentence into x and y phrases. And using government and binding theory to determine the reference. The the next step can apply heuristics to select the candidates. The main theory behind it are *breadth-first, left-to-right* or *centering theory*.

(c)

1. Token-Noun phrases. Noun phrases are mainly features for the whole process.

2. Named entities. Name entities can identify special Proper Nouns. And in this way it can help to find reference.

3. Dependency. Using dependency tree can find the relationship for words phrases.

1.3 Rounding up system design

1.3.1

Sieves

Pass 1-Speaker Identification. This process is about identifying speaker. For example, In direct quotations: Sam asked Mike, " Where did you buy your jacket? I really like it." I is the speaker, so it matches with Sam and Mike will corefer with the addressee you. So first person pronouns will match the speaker. And second person pronouns will match addressee.

Pass 2 and 3-Exact Match and Relaxed String Match. Only when the two mentions are exactly the same, they will be matched. For example *Apple CEO Tim Cook* and *Apple CEO Tim Cook*.After deleting the content following the head words, if these two mentions are the same, they are coreference. For example, *Tim Cook* and *Tim Cook, who is Apple CEO*.

Pass 4 – Precise Constructs. There are some precise steps to match the two mentions. When two mentions are in same appositive construction (Apple CEO, Tim Cook said), when they have exact subject–object relation (Tim Cook is Apple CEO), when a noun is ahead of antecedent (CEO Tim), when is a relative pronoun, When they are acronym or demonym, they are coreference.

Pass 5, 6, 7, 8, 9 are about head mach. Strict head math is mainly about only when head words are exact same words, has compatible modifiers, has same stop-words (*The Unites States, The States*), are not in i-within-i, don't have different locations and numbers(*China, South China*, *people, about 200 people*), the head words will match. Or for relaxed head math, if there is one words match, the head words will match. For example, *Tim Cook, Tim, Apple CEO Tim Cook*, they all have *Tim*.

Pass 10-Pronominal Coreference Resolution. Number, gender, person, animacy, NER label, Pronoun distance(the distance of its antecedent is smaller than 3) are used in Pronominal Coreference Resolution.

1.3.2

Pass 10-number can apply to sentence i. Because word they refers to pular agreement. Other sentence I really don't know how to apply**********

1.3.3

POS tagging can be added into the features. For example, for some relative pronoun *which* can help to determine the correference. Based on Class 4.

2 Distributional Semantics

2.1 Measuring Similarity: Intrinsic Evaluation

2.1.1

2.1.2

We used predictive model-word2vec in the experiment because it is efficient and has high performance. The input of this model is text corpus and the output are word vectors. This vectors mean semantic similarity of words pairs. And these word vectors can be used as features in the machine learning model. For example, When we use the model to generate the most similar words to *cry*, the results of model show *('crying', 0.6610245704650879), ('cries', 0.6551704406738281)* . . . . The first is most similar words and the number is score of similarity. And it also can compare the word pairs' similarity. For example, the similarity of *'man', 'woman'* is *0.76640123*. While the similarity of *'man', 'dog'* is *0.3088647*.

In this part, we use *SimLex999*, which is human judgments of similarity scores of word pairs. And then we used our model to calculate the same word pairs' similarity scores. And Spearman Rho is used to compare the results of human and model. Theory of Spearman Rho is comparing of rank of same word pair in human and model scores. For example, *vanish-disappear* word pair ranks first in human score and ranks third place in model score. Thus the distance $d_i^2 = (3-1)^2 = 4$. Thus, the Spearman Rho is give by 1, where n is sample size.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

And the Spearman Rho is 0.4419 and p-value is $5.6 * 10^{-49}$, which means two scores have high correlation.

2.1.3

Many word pairs rank in human scores higher than computer scores. For example, *'shore-coast',-179*, *'teacher-instructor', -175*, *'motor-engine',-212* and so on. On the other hand, *'winter-summer',710*, *'bathroom-kitchen',590*, *'mother-wife', 653* and so on rank higher than human scores.

The reason why *'motor-engine'* etc words ranks higher than model is that these words pairs appear in the text for a certain aim, for example, sea, school and motor. But it is difficult for model to detect the relationship and dependency because the number of these words are usual small and even appear only once. But human can detect the relationship because we know the actual meaning behind them. In this case, model will give them low rank. For computer rank is higher than human rank. It is because model simply put them into same category, for example, action, animal and so on. But human's rank not only depend on category.

2.2 Machine Learning: traditional features and embeddings

2.2.1

In this part, we start to explore how the individual features influence the machine learning system. The main task is to predict the named entity recognition. *'Token','Prevtoken','Cap','Pos','Chunklabel'* are used as features in th model. Token is spliting the whole sentence or text into single words. Prevtoken is further token after first Token. Cap means Capitalization, which contain lower letter, upper letter and so on. Pos is part-of-speech of each word. Chunklabel is phrases (or constituents).

At first, we just use one feature *Token* to train the classifier, and the using confusion matrix to estimate the precision, recall, and F1 score. And the result shows *P: 0.774 R: 0.445 F1: 0.529*. It is clear to see that precision is high but the recall and f1 score are quite low.

Furthermore, we did ablation analysis of features selection. Table 3 shows selection of label. It is clear to see that when the features are Token, Prevtoke, there is high precision 0.870. But the recall and f1 score are really low. And we can see the best combination of features are *Token, Prevtoken, Cap, Pos*. Chunklabel decreases the precision. Thus we can remove it.

Table 3: Best feature selection

Number of feature	Best features	Best precision	Best recall	Best f1
1	Token	0.774	0.445	0.529
2	Token, Prevtoken	0.870	0.596	0.663
3	Token, Prevtoken, Cap	0.865	0.707	0.735
4	Token, Prevtoken, Cap, Pos	0.774	0.723	0.740
5	Token, Prevtoken, Cap, Pos, Chunklabel	0.771	0.723	0.740

2.2.2

Additionally, we use word embeddings to represent the token and to do the task of predicting the named entity recognition. One-hot representation has disadvantage that it can not descirbe the dependency and relation between words. While word embeddings has this features. And first we use vector representation of tokens. The result shows when there is only Token in word embeddings the three scores are still not that great. While when adding Prevtoke in in word embeddings the result show greatly improvement and it is better than traditional features. And the best performance shows in mixed system.

Table 4: Outcome of word embeddings

Best features	Best precision	Best recall	Best f1
Token	0.680	0.621	0.647
Token, Prevtoken	0.789	0.754	0.770
A mixed system	0.784	0.784	0.785

Conclusion: traditional One-hot representation can not show the dependency and relation of words. But word embeddings have this features and when it contains mixed features the result is better.

References