Xue Yang 19230949

Section 1:

(Question 1A)

① Morphology relation of "talk" and "talking" is inflection.
talking → talk + Verb + Progressive

② Morphlogy relation of "talk" and "talkative" is derivation.
talkative Adj derived from talk Verb

③ Morphlogy relation of "talk" and ~~the steel~~ "sweet-talk" is: compound. sweet-talk > sweet + talk

(Section 1B)

~~"the men meet"~~

- ① "the men meet the men"
~~"men meet"~~
- ② "the men meet men"
③ "men meet the men"

(Section 1C) A semantic lexicon defines word senses explicitly. Here, explicitly defined word senses mean that there are no ambiguity in the word sense like indirect sense, irony, etc.

(Section 1D)

In Lesk algorithm, a 'signature' is a set of words in the gloss and examples of sense. Normally it is from corpus dataset. We can compare signature and the context of the target word, and then to find common words they have. Thus, the 'overlap' can be generated as Overlap (signature, context). If the overlap is great than the max-overlap we have already known, the best-sense is the corresponding sense.

P1

Xue Yang 1923094 9

Section 2

(Question 2A)

$P(\text{sailor} \mid a) = \frac{C(a\ \text{sailor})}{C(a\ \text{sailor})} = 1$

$P(\text{went} \mid \text{sailor}) = \frac{C(\text{sailor went})}{C(\text{sailor went})} = 1$

$P(\text{to} \mid \text{went}) = \frac{C(\text{went to})}{C(\text{went to})} = 1$

$P(\text{sea} \mid \text{to}) = \frac{C(\text{to sea})}{C(\text{to sea}) + C(\text{to see})} = 0.5$

$P(\text{sea} \mid \text{sea}) = \frac{C(\text{sea sea})}{C(\text{sea sea})} = 1$

$P(\text{see} \mid \text{to}) = \frac{C(\text{to see})}{P(\text{to see}) + P(\text{to sea})} = \frac{1}{2} = 0.5$

$P(\text{what} \mid \text{see}) = \frac{C(\text{see what})}{C(\text{see what}) + C(\text{see see})} = \frac{1}{1+4} = 0.2$

$P(\text{the} \mid \text{what}) = \frac{C(\text{what he})}{C(\text{what he})} = 1$

$P(\text{could} \mid \text{he}) = \frac{C(\text{the could})}{C(\text{he could})} = 1$

$P(\text{see} \mid \text{could}) = \frac{C(\text{could see})}{C(\text{could see})} = 1$

$P(\text{see} \mid \text{see}) = \frac{C(\text{see see})}{C(\text{see see}) + C(\text{see what})} = \frac{4}{4+1} = 0.8$

$P(\text{all} \mid \text{but}) = \frac{C(\text{but all})}{C(\text{but all})} = 1$

$P(\text{that} \mid \text{all}) = 1$

$P(\text{he} \mid \text{that}) = 1$

$P(\text{could} \mid \text{he}) = 1$

$P(\text{the} \mid \text{was}) = \frac{C(\text{was the})}{C(\text{was the})} = 1$

$P(\text{bottom} \mid \text{the}) = \frac{C(\text{the bottom})}{C(\text{the bottom}) + C(\text{the deep})} = \frac{1}{2} = 0.5$

$P(\text{of} \mid \text{bottom}) = 1$

$P(\text{the} \mid \text{of}) = 1$

$P(\text{deep} \mid \text{the}) = \frac{C(\text{the deep})}{C(\text{the deep}) + C(\text{the bottom})} = 0.5$

Xue Yang 19230949

p(blue|deep) = 1

p(sea|blue) = 1

Thus, we can generate a probability table:

p(Sailor|a) 0.1	p(all|but) 1
p(went|sailor) 1	p(that|all) 1
p(to|went) 1	p(he|that) 1
p(sea|to) 0.5	p(could|he) 1
p(sea|sea) 1	p(the|was) 1
p(see|to) 0.5	p(bottom|the) 0.5
p(what|see) 0.2	p(of|bottom) 1
p(he|what) 1	p(the|of) 1
p(could|he) 1	p(deep|the) 0.5
p(see|see) 0.8	
p(see|could) 1	

Given a bigram language model, the formula is:

$$p(w_1 w_2 \cdots w_n) = \prod_{i=1,2,\cdots,n} p(w_k | w_{k-1})$$

And p("he could see what")

$= p(he) \times p(could|he) \times p(see|could) \times p(what|see)$

$= \frac{2}{23} \times 1 \times 1 \times 0.2 \approx 0.01212$

Xue Yang 19230949

(Question 2B)

p(he could see the sea) = p(he) × p(could|he) × p(see|could) × p(the|see) × p(sea|the)

$p(he) = \frac{2+1}{18+33}$ (the unique words in corpus is 18)

$= \frac{1}{17}$

$p(could|he) = \frac{c(he\ could)+1}{c(he\ could)+18} = \frac{2+1}{2+18} = \frac{3}{20}$

$p(see|could) = \frac{c(could\ see)+1}{c(could\ see)+18} = \frac{2+1}{2+18} = \frac{3}{20}$

$p(the|see) = \frac{c(see\ the)+1}{c(see\ what)+18} = \frac{0+1}{1+4+18} = \frac{1}{23}$

$p(sea|the) = \frac{c(the\ sea)+1}{c(the\ deep)+18} = \frac{0+1}{1+18} = \frac{1}{19}$

Thus, $p = \frac{1}{17} \times \frac{3}{20} \times \frac{3}{20} \times \frac{1}{23} \times \frac{1}{19} \approx 0.303 \times 10^{-5}$

Xue Yang 19230949

(Question 2C)

In machine translation, we wish to get the best translation given a foreign text, say, we wish to find $\arg\max(t|f)$ where t is translation and f is foreign text.

Given Bayes's Law:

$$p(t_1|f) = p(f|t_1) \times p(t_1) / p(f)$$
$$p(t_2|f) = p(f|t_2) \times p(t_2) / p(f)$$

It shows that $p(t_1|f) > p(t_2|f)$ if $p(f|t_1)p(t_1) > p(f|t_2)p(t_2)$

Thus, to find the maximum $p(t_i|f)$ is equivalent to find:

$\arg\max p(f|t) * p(t)$, where $p(f|t)$ is translation model, $p(t)$ is language model.

This is how we use language model to resolve ambiguities in machine translation.

Section 3

(Question 3A)

The Hearst patterns are:

① Y including X

② Y such as X

∴ The hypony my extraction ~~rules~~ with Hearst Patterns rules are:

P5 ① If NP$_0$ such as {NP$_1$, NP$_2$, ..., (and/or)} NP$_n$ THEN for all NP$_i$ $|\leq i \leq n$, hyponym (NP$_i$, NP$_0$)

~~Thus, hypernym (NP$_0$, NP$_i$)~~

① Thus, we can get hyponym { association football, sports}
{ rugby, sports)
(horse racing, sports)
(golf sports)

② "IF 'Y' including [X1, ..., Xn]" THEN {X1, Y} ... {Xn, Y}

Thus, we can get (English language, culture)
(sports, culture)

Xue Yang 19230949

(Question 3B)

Geopolitically (O), Ireland (B-LOC) is (O) divided (O) between (O) the (O) Republic (B-LOC) of (I-LOC) Ireland (I-LOC) (officially (O) named (O) Ireland (B-LOC)), which (O) covers (O) five-sixths (O) of (O) the (O) island (O), and (O) Northern (B-LOC) Ireland (I-LOC), which (O) is (O) part (O) of (O) the (O) United (B-LOC) Kingdom (I-LOC).

The Location entities are: "Ireland", "Republic of Ireland", "Northern Ireland", and "United Kingdom".
According to IOB tagging scheme, the beginning of the entity is annotated with "B", the words inside of entity is annotated with 'I', and the words outside of an entity is annotated with 'O'.

(Question 3C)

·A knowledge graph can indicate the properties of objects, the relationship between objects, or events. ~~I~~ The knowledge graphs put data in context via ling linking and semantic metadata.

Thus, in addition to textual context information, we can also utilize the links in knowledge graphs to build entity linking. For example, we can use some algorithms like PageRank algorithm to rank pages, or we can generate authority score or hub score (in/out degree) for the link to see how central this entity is.

Section 4

(Question 4A)

"nice" has a positive sentiment score: $1 \times 2 = 2$

: Lovely, has a positive sentiment score: $1 \times 2 = 2$

: recommend" has a positive sentiment score: 1

~~Thus, the~~ : There are no words ~~that~~ have negative sentiment.

Thus, the Positive (POS) score is 5,

the Negative (NEG) sentiment score is 0.

(Question 4B)

Aspect-based sentiment analysis:

aspect	sentiment
hotel	POS
bar	POS
restaurant	POS
food	POS
staff	POS

~~hotel~~

~~(Question 4C)~~

Question (4C)

~~We we~~ d When we do NLP tasks, we need to ~~a~~ extract information from text and analyse it. For example, we ~~can~~ analyze the data from social media ~~to observe~~ for NLP projects.

But it is important to pay attention to data privacy and data protection issues associated with the data collection for a specific (NLP) task. For example, ~~we~~ sometimes we need to inform the affected individuals what we are doing with their data; we may need to establish the maximum time period required for keeping the data; or we need to develop appropriate data anonymization strategie, etc.

This is because language data is also personal data as individuals can be identified by their language data use. NLP can lead to 'fingerprinting' of individuals. So, data privacy of individuals need to be protected in NLP tasks