FNLP 2016 RESIT Past Paper

PART A

1. Sparse data problem + why it is difficult to overcome + name 2 solutions.

Sparse data problem is when we don't have enough observations to estimate probabilities well.

When calculating probabilities we can assume that word probability depends only on short history (as in N-gram model for trigrams and bigrams, longer histories face the sparse data problem again) so alleviate sparse data.

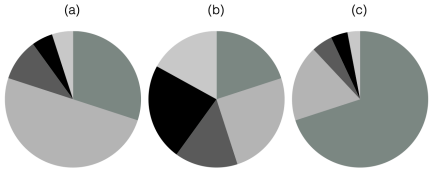
Other solutions: smoothing (allocate probabilities to unseen data from seen data), e.g. add-1, add-τ and Good-Turing
backoff
interpolation

2. What modelling framework is shared by statistical approaches to speech recognition, machine translation, and OCR? Using one of these tasks as an example, state the two key components of the framework and what they represent in that particular task.

Noisy channel model.

Spelling correction
2 key components: 1. Language Model (to give us a good approximation)
2. Error Model (likelihood of the output given the intention)

3. Pie charts to depict probability distribution over 5 events:



a. Rank the distributions in order of entropy, from LOWEST to HIGHEST.

c, a, b.

b. Explain how entropy is used in language modelling.

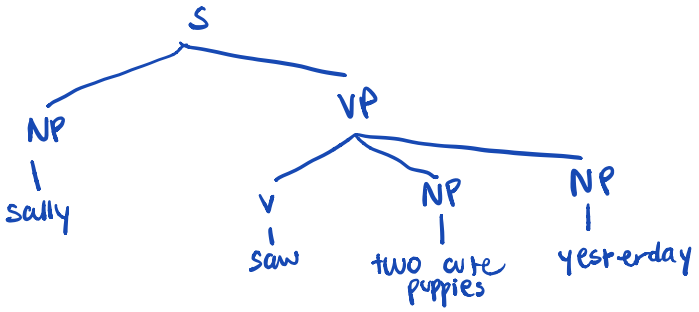
Perplexity ($2^{cross\_entropy}$) is used as intrinsic evaluation of language models - we want to have low uncertainty of what word comes next in a word sequence, hence we want to have a lower entropy as well.

4. What is the difference between a constituency parse and a dependency parse? Provide structural analysis of the following sentence using each method:

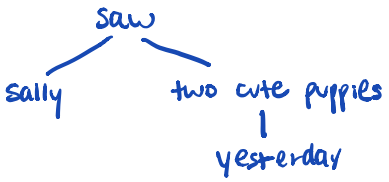
Sally saw two cute puppies yesterday.

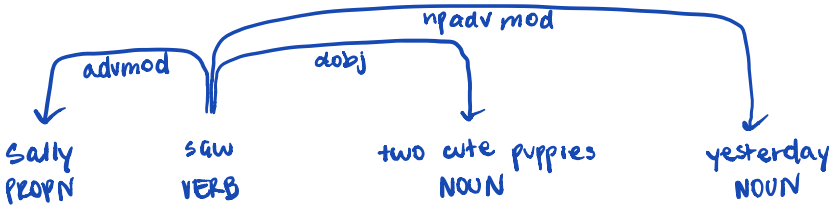
You do not need to include labels, only the structures themselves.

Constituency parse usually includes breaking text into sub-phrases of parse tree. The parse tree will use phrasal categories (non-terminals) and terminals.



Dependency parse, on the other hand, represents the tree as relationships / dependencies between words.





5. Two examples of structural ambiguity: one POS ambiguity and one not.

"I saw her duck"
→ POS tag ambiguity; knowing the correct POS tags can be disambiguated (since different words can get different meanings)

"I saw the man with the telescope"
→ Attachment ambiguity; cannot be disambiguated even with POS tags (parsing depends on where different phrases attach in the tree)

6. You are considering crowdsourcing as a way to evaluate the output of a machine translation system. Pros and cons + how to ensure useful results?

PROS - offer higher probabilities of success
- saves time and money

CONS - quality can be difficult to achieve when there are lots of contributors
- project management: it can get complicated to plan and manage projects with a large pool of people without system and process.

Need to take measures to ensure annotators are qualified and taking the task seriously.
- Redundancy to combat noise: Elicit 5+ annotations per data point
- Embed data points with known answers, reject annotators who get them wrong.

7. Train a machine learning system to do word sense disambiguation. Name model for the task + 3 types of features for disambiguation.

Naive Bayes. (others: MaxEnt, Decision lists, Decision trees)

FEATURES - directly neighboring words (and/or their lemmas)
- any context words in a 50 word window
- syntactically related words
- syntactic role in sense
- topic of the text
- PoS tag, surrounding PoS tags

8. a. Distributional hypothesis in lexical semantics.

We infer the meaning of a word from the context that word is in. Similar context also implies similar word meaning.

b. Brown clustering algorithm ran on 847 tokens of English tweets. 4 clusters:

i. 01111110110 soon shortly sooon soooon soonish sooooon
ii. 0111111110 now noww nowww #now nowwww now-n0w
iii. 01111111110 there der dere ther thar therre thurr
iv. 11111111110 no n0 -no

Which pair are distributionally more similar?

i and ii because ii is an extension of i.

PART B

1. Hidden Markov Models

a. What independence assumptions does an HMM tagger make? Give an example of a linguistic phenomenon where these assumptions are too strong to capture the true dependencies in the language.

Markov Independence Assumption
- each tag/state only depends on fixed number of previous tags/states
  → here, just one!

Example: Long-range dependencies
→ Sam, the man with red hair who is my cousin, sleeps soundly.

These dependencies are often important for translation.

b. Consider two different algorithms for computing the best tag sequence for a given HMM and input sequence of words:

A1: Enumerate all possible tag sequences, compute the probability of each one, and return the highest probability sequence.

A2: The Viterbi Algorithm.

Give time complexity in big-O: N - no. of tokens
T - no. of distinct POS tags
V - vocabulary size

A1 → $O(T^N)$ A2 → $O(T^2N)$

c. We want to use an HMM POS tagger to tag the following sentence:

```
<s> one dog bit </s>
```

Our HMM has only five tags (plus beginning/end of sentence markers, <s> and </s>). Below are the transition probabilities (left) and output probabilities (right). We assume there are other possible output words not shown in the table, and that the <s> and </s> states output <s> and </s> words, respectively, with probability 1.

$t_{i-1}\backslash t_i$	CD	PRP	NN	VB	VBD	</s>
<s>	.5	.2	0	.3	0	0
CD	.2	0	.3	.2	.2	.1
PRP	.1	.1	0	.3	.4	.1
NN	.05	.15	.2	.25	.3	.05
VB	0	.2	.6	0	0	.2
VBD	0	.1	.6	0	0	.3

$t\backslash w$	one	cat	dog	bit	...
CD	.1	0	0	0	
PRP	.02	0	0	0	
NN	.05	.03	.04	.007	
VB	0	0	.03	0	
VBD	0	0	0	.06	

i. Compute $P(\vec{w}, \vec{t})$, where $\vec{w}$ is the given sentence and $\vec{t}$ is the tag sequence <s> CD NN NN </s>.

$\vec{w}$ = [<s>, one, dog, bit, </s>]

$\vec{t}$ = [<s>, CD, NN, NN, </s>]

$P(\vec{w}, \vec{t}) = \leq$

$$P(\vec{w}, \vec{t}) = \prod_{i=1}^{n} P(t_i | t_{i-1}) P(w_i | t_i)$$

$$= P(CD|<s>) \cdot P(one|CD) \cdot P(NN|CD) \cdot P(dog|NN) \cdot P(NN|NN) \cdot P(bit|NN) \cdot P(</s>|NN)$$

$$= [0.5 \times 0.1] \times [0.3 \times 0.04] \times [0.2 \times 0.007] \times 0.05$$

ii. Partially completed Viterbi chart for the above sentence.

	<s>	one	dog	bit	</s>
<s>	1	0			
CD	0	.05			
PRP	0	.004			
NN	0	0	*		
VB	0	0			
VBD	0	0			
</s>	0	0			

Write down the computation that needs to be done in order to fill in the cell * [NN, dog]. What does the value in this cell represent

The most probable tag for 'one' is CD.

Thus in [NN, dog] = $P(one|CD) \times P(NN|CD) \times P(dog|NN)$

$= 0.05 \times 0.3 \times 0.04$

This value represents the most likely path up to dog where dog is a noun NN.

→ choosing PRP/one to NN/dog will result to 0 because $P(NN|PRP) = 0$ so we go with CD/one.

2. Text Authorship

A manuscript has been discovered in the basement of a disused rectory in York- shire, bound into the back of a mid 19th-century diary. The diary itself describes it as “a faithful copy, in my own hand, of a composition by the daughter of my predecessor here as curate, of which the original is now lost.” The manuscript has no titlepage, or any other indication of authorship. The possibility that this is a hitherto unknown work by Charlotte or Emily Brontë sets the literary world buzzing.

But is it? And if so, which of the famous sisters wrote it?

Drawing on the language modelling technologies discussed in lectures and labs, design an experiment to answer these questions, assuming you can digitize the manuscript, and that you can also obtain digital versions of both Charlotte’s Jane Eyre and Emily’s Wuthering Heights, along with a wide range of other contemporary fiction.

a. Set out your background assumptions and hypotheses you would be trying to test in order to answer the questions.

b. Describe the experiments you'd perform, modelling techniques, how to train models to confirm/reject hypothesis.

Use N-gram model, trained on Charlotte Brontë's Jane Eyre and Emily Brontë's Wuthering Heights.

If we used the trained models to generate new sentences by sampling words from its probability distribution, we can measure how similar are those sentences from the sentences in the diary.

We can work out the cross-entropy of the model on each of the texts and from the scores, we can determine if the diary was written by one of the Brontë sisters.

Control: Measure Charlotte's other book with Jane Eyre (xent)
Measure Emily's other book with Wuthering Heights (xent)

c. What factors will determine the reliability of your results? In general, which is likely to be a more reliable conclusion in this sort of experiment:
   - These two are similar
   - These two are different

   Why?

3. Lexical Semantics

a. Explain the difference between homonymy and polysemy, giving examples of each to illustrate your answer.

HOMONYM - two words of the same spelling can have different meanings
- idiosyncratic, unrelated, not predictable (senses)
- sense ambiguity from accidents

e.g. "I put my money in the bank" and "I rest on the bank of the river."

POLYSEMY - words that exhibit co-existent meanings
- related and predictable (senses)
- sense ambiguity from language regularity

e.g. Newspaper → company that publishes recent news
→ single physical item published by the company
→ an edited work in a specific format

b. WordNet lists the following six senses for the noun *table*, with example usages in italics:

S1: table, tabular array (a set of data arranged in rows and columns)
see table 1

S2: table (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs)
it was a sturdy table

S3: table (a piece of furniture with tableware for a meal laid out on it)
I reserved a table at my favorite restaurant

S4: mesa, table (flat tableland with steep edges)
the tribe was relatively safe on the mesa but they had to descend into the valley for water

S5: table (a company of people assembled at a table for a meal or game)
he entertained the whole table with his witty remarks

S6: board, table (food or meals in general)
she sets a fine table; room and board

Cluster these senses using the definitions of homonymy and polysemy you gave in part (a). For any senses that are polysemous, give an argument as to how the senses are related, and whether the relationship between the senses is systematic (does it happen with other words?)

Polysemy
- S2 - S5
- S3 - S6

Homonym
- S1 - S4
→ same spelling, very different meanings

c. Suppose you want to disambiguate the different senses of the word *table*. You are considering three different approaches: a supervised Word Sense Disambiguation (WSD) system, an unsupervised WSD system, or a supervised supersense tagging system. Briefly discuss some of the pros and cons of each approach, including an explanation of how supersense tagging would be applied to this problem.

Supervised WSD
- ⊖ Loads of training data needed, and expensive
- ⊕ More accurate

Application: compare to the gold standard in the training set

Unsupervised WSD
- ⊖ Not to accurate if there is no gold standard
- ⊕ Infer from context so less data required

Application: infer from context around the word

Supervised Supersense Tagging
- ⊕ Similar to supervised WSD
- ⊖ Too general for polysemies
- * Finding the hypernym of a word while NER is categorising a word to a pre-defined set of categories.
  - i.e. PERSON, PHENOMENON