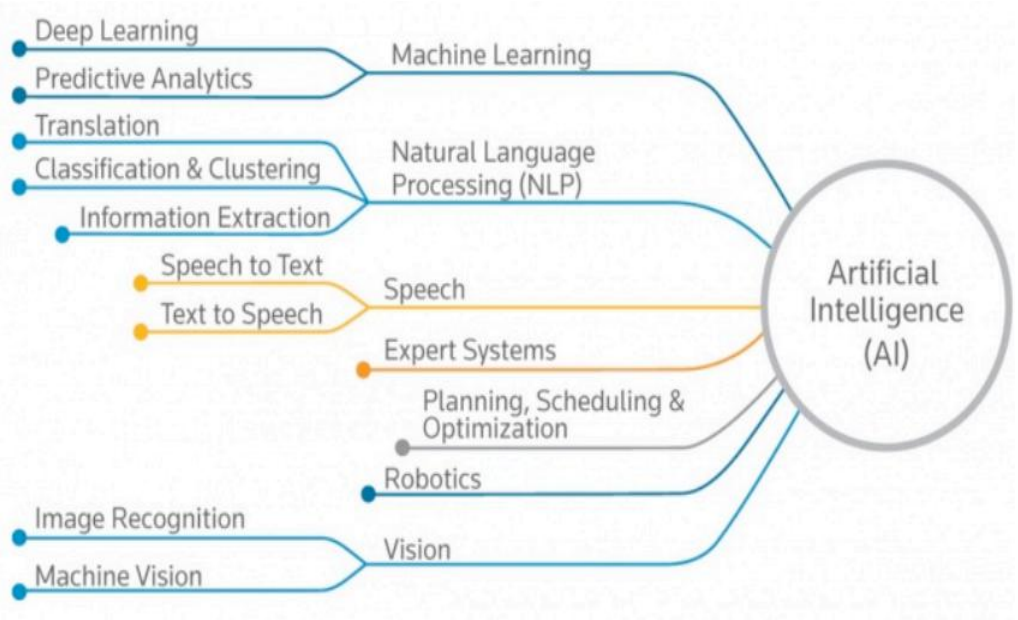
Ders01 - Intro. To NLP:

- ✓ Languages that are used by human beings are called 'natural languages'.
- ✓ Natural language processing (NLP) is a field of **computer science** and **artificial intelligence**, and it is concerned with the interactions between **computers** and **human (natural) languages.**
- ✓ **Scope of Application:**
  - ❖ **Question Answering,**
  - ❖ **Machine Translation,**
  - ❖ **Web Search,**
  - ❖ **Word Processing,**
  - ❖ **AI Bots(Siri, Google Asistan),**
  - ❖ **Text Processing(Categorization, Clustering),**
  - ❖ **Text Summarization,**
  - ❖ **Chat-Bot**
- ✓ **Example: Eliza** was written at MIT by Joseph Weizenbaum between 1964 and 1966. **Eliza** is rule-based. It has gained popularity in the area of psychotherapy.



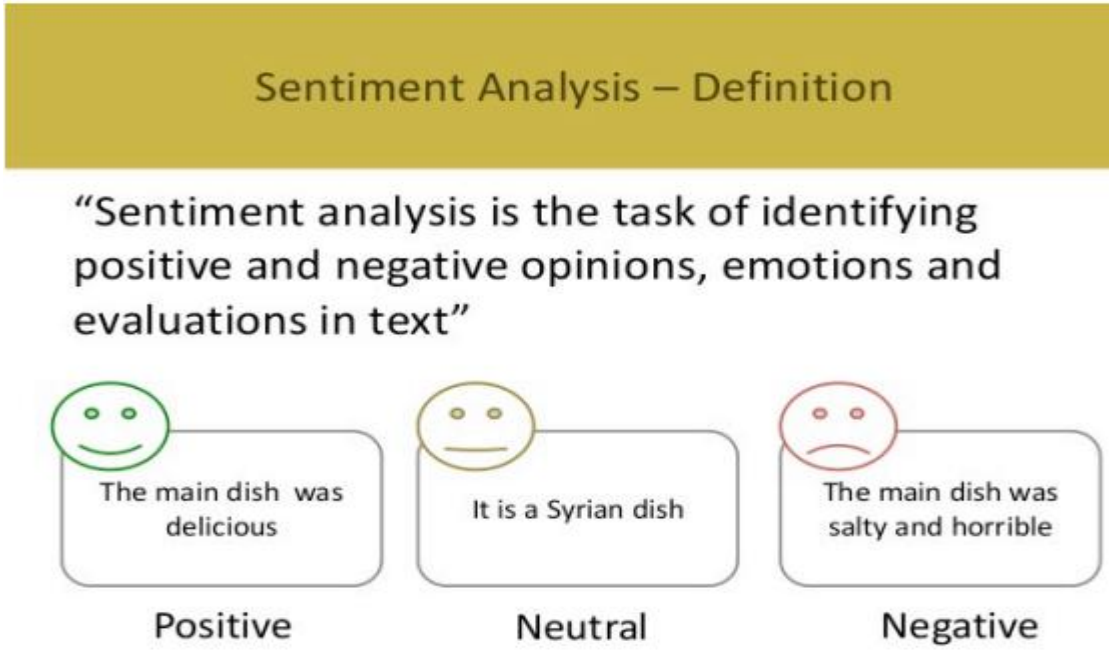
- ✓ **Text Classification:**
  - Male or female author:

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

  - Positive or negative movie review

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

✓ **Sentiment Analysis:** Extraction of emotions from a text.



✓ **Text Summarization:** Generation of a small text from a given long document.

- **Article:** With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee....
- **Summary:** Senator Barack Obama was declared the presumptive Democratic presidential nominee.

✓ **Information Extraction:** is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.

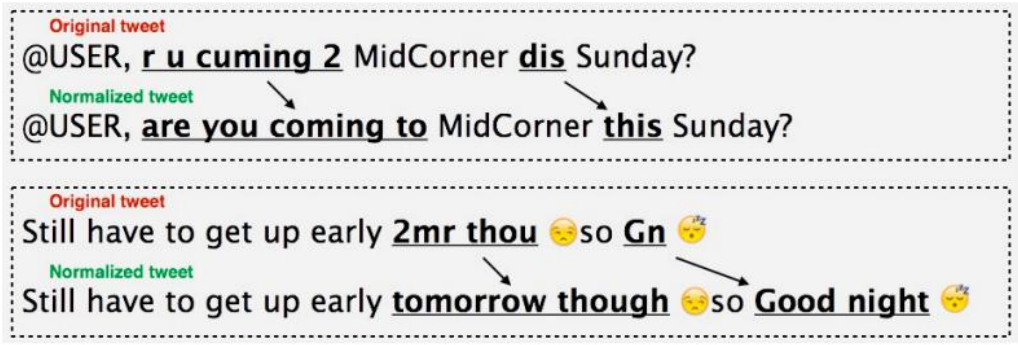
- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

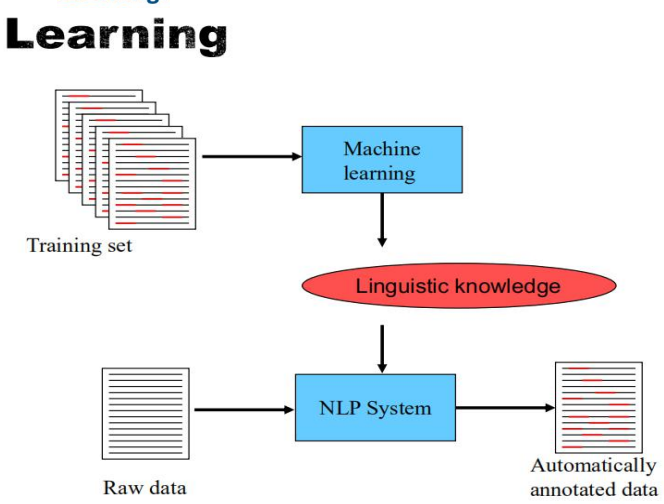
Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 80% accuracy for multi-sentence templates, 90% + for single easy fields
- But remember: information is redundant!

✓ **Text Normalization:** is the process of transforming a text into a canonical (standard) form. For example, the word "gooood" and "gud" can be transformed to "good", its canonical form. Another example is mapping of near identical words such as "stopwords", "stop-words" and "stop words" to just "stopwords".



✓ Learning:



✓ What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Unknown: human-level comprehension (is this just NLP?)

✓ Why Study NLP?

- Text is the largest repository of human knowledge and is growing quickly.
  - emails, news articles, web pages, IM, scientific articles, insurance claims, customer complaint letters, transcripts of phone calls, technical documents, government documents, patent portfolios, court decisions, contracts, ……
  - Are we reading any faster than before?

✓ NLP Can Do: Many companies would make a lot of money if they could use computer programmes that understood text or speech. Just imagine if a computer could be used for:

- ❖ answering the phone, and replying to a question
- ❖ understanding the text on a Web page to decide who it might be of interest to
- ❖ translating a daily newspaper from Japanese to English
- ❖ understanding text in journals / books and building an expert system based on that understanding

✓ Why is NLP Hard?

- Language = Words + rules + exceptions..
- Ambiguity at all levels..
- We speak different languages..
- And language is a cultural entity..
- Highly systematic but also complex..
- Keeps changing.. New words, new rules and new exceptions..
- Source : Electronic texts / Printed texts / Acoustic Speech Signal.. they are noisy..
- Language looks obvious to us.. But it is a Big Deal☺!

✓ Levels of Language Analysis:

- ❖ **Phonology**
- ❖ **Morphology**
- ❖ **Syntax**
- ❖ **Semantics**
- ❖ **Pragmatics**
- ❖ **Discourse**

✓ Task: Segmentation

Text is split into words and sentences.

- Languages like Chinese do not have spaces between words.

original, un-segmented text

再往远些看，随着汉字识别和语音识别技术的发展，
中文计算机用户将跨越语言差异的鸿沟，
在录入上走向中西文求同的道路。

separated word entities after segmentation

再 往 远 些 看 ， 随着 汉字 识别 和 语音 识别 技术 的 发展 ，
中文 计算机 用户 将 跨越 语言 差异 的 鸿沟 ，
在 录入 上 走 向 中 西文 求 同 的 道路 。

✓ Task: Part-of-Speech Tagging

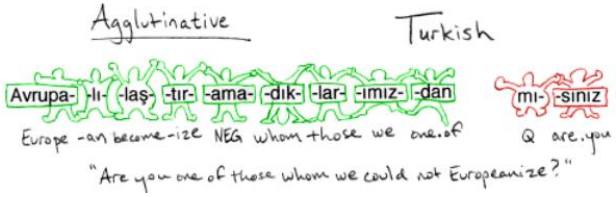
- Tagging the words in a sentence with their syntactic roles.

- John saw the saw and decided to take it to the table
- PN V Det N Con V Part V Pro Prep Det N

✓ Task: Morphological Segmentation

The longest word in Turkish:

- *muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine*
- *"As though you are from those whom we may not be able to easily make into a maker of unsuccessful ones"*



✓ Ambiguity

"I saw the man with the telescope": 2 meanings

"I saw the man on the hill with the telescope.": 5 meanings

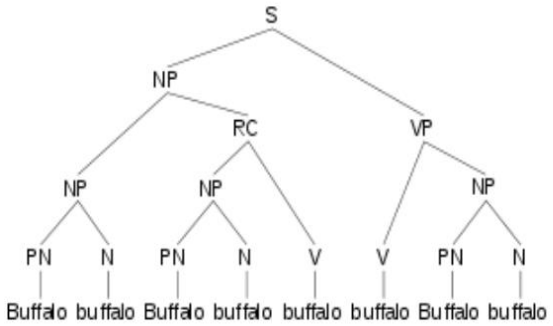
"I saw the man on the hill in Texas with the telescope": 14 meanings

"I saw the man on the hill in Texas with the telescope at noon.": 42 meanings

"I saw the man on the hill in Texas with the telescope at noon on Monday" 132 meanings

✓ Syntactic Parsing

Örnek: Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo



Diğer buffalo bizonlarının korkuttuğu bufalo bizonları, yine aynı buffalo bizonlarını korkutmaktadır

✓ Pragmatics

- Uses context of utterance
  - Where, by who, to whom, why, when it was said
  - Intentions: *inform, request, promise, criticize, …*
- Handling Pronouns
  - "Mary eats apples. She likes them."
    - She="Mary", them="apples".
- Handling ambiguity
  - Pragmatic ambiguity: "you're late": What's the speaker's intention: informing or criticizing?
- "I saw the man with binoculars"

✓ Discourse Analysis

Text units beyond sentences — examples

- A story (such as a fairy tale, a drama, ...).
- A news item.
- Dialogue.
- Technical text (manual, textbook, documentation).
- A document in a document base (abstract, patent description, ...).

Links between sentences/phrases in a larger text

- Textual ordering.
- Temporal link (for example, an event precedes another event).
  - *Jim saw the bus. He ran to catch it.*
    - "saw" precedes "ran"

Ders02 - Language Models:

Machine Translation:

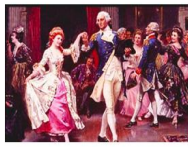
- P(high winds tonite) > P(large winds tonite)

Spell Correction:

- The office is about fifteen minuets from my house
  - P(about fifteen minutes from) > P(about fifteen minuets from)

Speech Recognition:

- P(I saw a van) >> P(eyes awe of an)



Summarization, question-answering, etc.,

✓ Completion Prediction

A language model also supports predicting the completion of a sentence.

- Please turn off your cell _____
- Your program does not _____

Predictive text input systems can guess what you are typing and give choices on how to complete it.

✓ Language & Probability

1. Colorless green ideas sleep furiously.
2. Furiously sleep ideas green colorless.

✓ Autocomplete

✓ Moving Towards Language

What's the probability of drawing a 2 from a deck of 52 cards?

$$P(drawing\ \ a\ \ two) = \frac{4}{52} = \frac{1}{13} = .077$$

What's the probability of a random word (from a random dictionary page) being a verb?

$$P(drawing\ a\ verb) = \frac{\#\ of\ \ ways\ to\ get\ a\ verb}{all\ words}$$

✓ Probabilities on Text (Conditioning on the Previous Word)

- English
  - ALICE was beginning to get very tired of sitting by her sister on the bank and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice, "without pictures or conversations?'
- Word salad
  - beginning by, very ALICE but was and? reading no tired of to into sitting sister the, bank, and thought of without her nothing: having conversations Alice once do or on she it get the book her had peeped was conversation it pictures or sister in, 'what is the use had twice of a book"pictures or' to

P(English)≫P(word salad)

$P(w_i{+}1 = \text{of} \mid w_i = \text{tired}) = 1$
$P(w_i{+}1 = \text{of} \mid w_i = \text{use}) = 1$
$P(w_i{+}1 = \text{sister} \mid w_i = \text{her}) = 1$
$P(w_i{+}1 = \text{beginning} \mid w_i = \text{was}) = ½$
$P(w_i{+}1 = \text{reading} \mid w_i = \text{was}) = ½$

$P(w_i{+}1 = \text{bank} \mid w_i = \text{the}) = 1/3$
$P(w_i{+}1 = \text{book} \mid w_i = \text{the}) = 1/3$
$P(w_i{+}1 = \text{use} \mid w_i = \text{the}) = 1/3$

✓ Language Model

In statistical language applications, the knowledge of the source is referred as Language Model.

We use language models in the various NLP applications:

- speech recognition
- spelling correction
- machine translation
- .....

N-GRAM models are the language models which are widely used in NLP domain.

✓ N-Grams

To collect statistics to compute the functions in the following forms is difficult (sometimes impossible):

$P(w_n|w_1^{n-1})$

Here we are trying to estimate the probability of seeing w_n after seeing w_1^{n-1}.

We may approximate this computation just looking N previous words:

$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N}^{n-1})$

So, an N-GRAM model

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-N}^{k-1})$$

✓ Chain Rule

The probability of a word sequence $w_1w_2\ldots w_n$ is:

$P(w_1w_2\ldots w_n)$

We can use the chain rule of the probability to decompose this probability:

$P(w_1^n) = P(w_1)P(w_2|w_1)\ P(w_3|w_1^2)\ \ldots\ P(w_n|w_1^{n-1})$

$$p(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

Example:

P(the man from jupiter) =
P(the)P(man|the)P(from|the man)P(jupiter|the man from)

✓ N-Grams (Cont.)

Unigrams -- $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k)$

Bigrams -- $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1})$

Trigrams -- $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1}w_{k-2})$

Quadrigrams -- $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1}w_{k-2}w_{k-3})$

✓ Examples

Unigram

P(the man from jupiter) ≈ P(the)P(man)P(from)P(jupiter)

Bigram

P(the man from jupiter) ≈ P(the|*s*)P(man|the)P(from|man)P(jupiter|from)

Trigram

P(the man from jupiter) ≈
P(the|*s* *s*)P(man|*s* the)P(from|the man)P(jupiter|man from)

✓ Markov Model

The assumption that the probability of a word depends only on the last n words is called Markov assumption.

Markov models are the class of probabilistic models that assume that we can predict the probability of some future unit without looking too far into the past.

A bigram is called a first-order Markov model (because it looks one token into the past); A trigram is called a second-order Markov model;

In general an N-Gram is called a N-1 order Markov model.

✓ Markov Assumptions

Simplifying assumption:

P(the | its water is so transparent that) ≈ p(the|that)

Or maybe:

P(the | its water is so transparent that) ≈ p(the|transparent that)

✓ N-Gram Models

We can extend to trigrams, 4-grams, 5-grams

In general this is an insufficient model of language because language has long-distance dependencies:

- "The computer which I had just put into the machine room on the fifth floor crashed."

But we can often get away with N-gram models.

✓ Unigram Model

$$p(w_1 w_2 ... w_n) \approx \prod_i p(w_i)$$

Some automatically generated sentences from a unigram model:

fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass

thrift, did, eighty, said, hard, 'm, july, bullish

that, or, limited, the

✓ Estimating N-Gram Probabilities

Estimating bigram probabilities:

Where C is the count of that pattern in the corpus

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

$$= \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Estimating n-gram probabilities

$$P(w_n \mid w_{n-N}^{n-1}) = \frac{C(w_{n-N}^{n-1} w_n)}{C(w_{n-N}^{n-1})}$$

✓ Biagram Model

Condition on the previous word:

$$p(w_i \mid w_1 w_2 ... w_{i-1}) \approx \prod_i p(w_i \mid w_{i-1})$$

Some automatically generated sentences from a bigram model:

Texaco rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr. gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november

✓ Example

- $p(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$

<s>I am Sam</s>

<s>Sam I am</s>

<s>I do not like green eggs and ham</s>

$P(\text{I} \mid \text{<s>}) = \frac{2}{3} = .67$ $P(\text{Sam} \mid \text{<s>}) = \frac{1}{3} = .33$ $P(\text{am} \mid \text{I}) = \frac{2}{3} = .67$

$P(\text{</s>} \mid \text{Sam}) = \frac{1}{2} = 0.5$ $P(\text{Sam} \mid \text{am}) = \frac{1}{2} = .5$ $P(\text{do} \mid \text{I}) = \frac{1}{3} = .33$

✓ Which N-Grams?

- Unigram model $P(w_1)\ P(w_2)...P(w_i)$
- Bigram model $P(w_1)\ P(w_2|w_1)...P(w_i|w_{i-1})$
- Trigram model $P(w_1)\ P(w_2|w_1)P(w_3|w_2w_1)...P(w_i|w_{i-2}w_{i-1})$
- N-gram model $P(w_1)\ P(w_2|w_1)P(w_3|w_2w_1)...P(w_i|w_{i-n-1}...w_{i-1})$

- N-gram models assume each word (event) depends only on the previous n-1 words (events). Such independence assumptions are called Markov assumptions (of order n-1).

- Bigger N, the model will be more accurate.
  - But we may not get good estimates for n-gram probabilities.
  - The n-gram tables will be more sparse.
- Smaller N, the model will be less accurate.
  - But we may get better estimates for n-gram probabilities.
  - The n-gram tables will be less sparse.
- In reality, we do not use higher than trigram (not more than bigram).
- How big are n-gram tables with 10,000 words?
  - Unigram -- 10,000
  - Bigram -- 10000*10000 = 100,000,000
  - Trigram -- 10000*10000*10000 = 1,000,000,000,000

✓ Example

Raw bigram counts

- Out of 9222 sentences

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Raw bigram probabilities

- Normalize by unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Result:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Bigram estimates of sentence probabilities

- P(<s> I want english food </s>) =
- p(I | <s>)
- × p(want | I)
- × p(english | want)
- × p(food|english)
- × p(</s> | food)
- =0.000031

Practical issues

- We do everything in log space:
  - Avoid underflow
  - Also adding is faster than multiplying.

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

✓ Zeroes

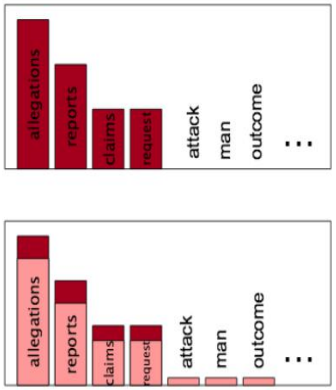
- Training set
  - …denied the allegations
  - …denied the reports
  - …denied the claims
  - …denied the request
- Test set
  - …denied the offer
  - …denied the loan

P("offer" | denied the) = 0

- Bigrams with zero probability
  - Mean that we will assign 0 probability to the test set!

✓ The Intuition of Smoothing

- When we have sparse statistics
  - p(w | denied the)
  - 3 allegations
  - 2 reports
  - 1 claims
  - 1 request
  - 7 total
- Steal probability mass to generalize better
  - p(w | denied the)
  - 2.5 allegations
  - 1.5 reports
  - 0.5 claims
  - 0.5 request
  - 2 other
  - 7 total



Since N-gram tables are too sparse, there will be a lot of entries with zero probability (or with very low probability).

- The reason for this, our corpus is finite and it is not big enough to get that much information.
- The task of re-evaluating some of zero-probability and low-probability n-grams is called smoothing.

Smoothing Techniques:

- add-one smoothing -- add one to all counts.
- Witten-Bell Discounting -- use the count of things you have seen once to help estimate the count of things you have never seen.
- Good-Turing Discounting -- a slightly more complex form of Witten-Bell Discounting
- Backoff -- using lower level n-gram probabilities when n-gram probability is zero.

✓ Add-one Smoothing (Laplace)

- Pretend we saw each word one more time than we did.
- Just add one to all counts!
- MLE estimate:

$$p(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

- Add-1 estimate:

$$p(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i) + 1}{C(w_{i-1}) + V}$$

✓ Advanced Smoothing

Many advanced techniques have been developed to improv smoothing for language models.

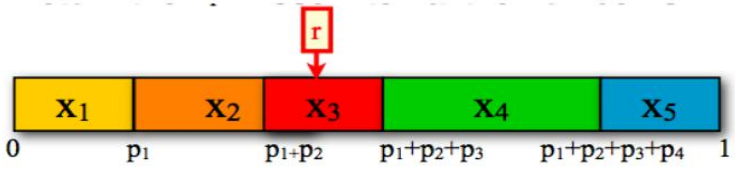
- Good-Turing
- Interpolation
- Backoff
- Kneser-Ney
- Class-based (cluster) N-grams

✓ Maximum Likelihood Estimate (MLE)

- Suppose the word "bagel" occurs 400 times in a corpus of a million words
- What is the probability that a random word from some other text will be "bagel"?
  - MLE estimate is 400/1,000,000 = .0004
- This may be a bad estimate for some other corpus
  - But it is the estimate that makes it most likely that "bagel" will occur 400 times in a million word corpus.

✓ Using N-Gram Models to Generate Languages

- How do you generate text from an n-gram model?
- That is, how do you sample from a distribution P(X |Y=y) ?
- Assume X has n possible outcomes (values): $\{x_1,\ldots,x_n\}$ and $P(x_i \mid Y=y) = p_i$
- Divide the interval [0,1] into n smaller intervals according to the probabilities of the outcomes
- Generate a random number r between 0 and 1.
- Return the x1 whose interval the number is in.



✓ Evaluation: How Good Is Our Model?

- Does our language model prefer good sentences to bad ones?
  - Assign higher probability to "real" or "frequently observed" sentence
  - Than "ungrammatical" or "rarely observed" sentences?
- We train parameters of our model on a training set
- We test the model's performance on data we haven't seen.
  - A **test set** is an unseen dataset that is different from our training set, totally unused.
  - An evaluation metric tells us how well our model does on the test set.

✓ Perplexity

The Shannon Game:

- How well can we predict the next word?

  I always order pizza with cheese and ____

  The 33rd President of the US was ____

  I saw a ____

  mushrooms 0.1
  pepperoni 0.1
  anchovies 0.01
  ....
  fried rice 0.0001
  ....
  and 1e-100

- Unigrams are terrible at this game. (Why?)

A better model of a text

- is one which assigns a higher probability to the word that actually occurs

The best language model is one that best predicts an unseen test set

- Gives the highest P(sentence)

Perplexity is the inverse probability of the test set, normalized by the number of words:

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

Chain rule:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

For bigrams:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

Minimizing perplexity is the same as maximizing probability

✓ Summary

- Language models assign a probability that a sentence is a legal string in a language.
- They are useful as a component of many NLP systems, such as ASR, OCR, and MT.
- Simple N-gram models are easy to train on unsupervised corpora and can provide useful estimates of sentence likelihood.
- MLE gives inaccurate parameters for models trained on sparse data.
- Smoothing techniques adjust parameter estimates to account for unseen (but not impossible) events.