

Bayes' Theorem & Machine Translation

(10/11/09)

Wouter Beek

me@wouterbeek.com

<http://www.wouterbeek.com>

Overview

1. Natural language processing (NLP) overview.
2. Bayes' theorem explained.
3. Bayes' theorem applied to NLP.
4. Questions.

Part I: Natural Language Processing (NLP) overview

AI Goals of NLP

- **Systems that extract information from textual or spoken media.** (Information retrieval, information extraction, data mining.)
- **Systems that transform text to speech (TTS) and speech to text (STT).** (Translation systems, summarization, dictation, reading.)
- **Systems that communicate with people through language.** (Dialogue systems, voice-based UI.)

Subdivisions of NLP

- **Speech:** Acoustics and recognition.
- **Words:** Structure of words (morphology), categories, meanings.
- **Sentences:** Structure of sentences (grammar) and their meanings (semantics).
- **Meaning/conceptualization:** sense disambiguation, semantic representations, Translation equivalence,
- **Text/dialogue:** How texts and dialogs are structured (e.g. turn-taking).
- **Conventions:** Cultural preferences, world knowledge, translation habits.

Research questions of NLP

- **Scientific:** Build models of the human use of language and speech.
 - **Technological:** Build models that serve in technological applications e.g. machine translation, speech systems, information extraction.
1. What are the kinds of things that people say and write?
 2. What do these things mean?
 3. What is the algorithmic structure that describes linguistic behavior?

What makes NLP so difficult

- **Ambiguity:** Both verbal and structural.
- **Robustness/gradedness:** Humans can process typing errors, spelling errors, grammatical errors, narrative reorderings.
- **World-knowledge:** Human processing involves many extra-linguistic factors.
- **Cultural deviation:** Different speech communities.
- **Language creativity**

Part II: Bayes' theorem

Thomas Bayes (1702-1761)

- 1731 - *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures.*
- 1736 - *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst.*
- 1742 – Elected as a Fellow to the Royal Society.
- Bayes' theorem, published posthumously.



Conditional probability

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

- $P(A|B)$, conditional probability of A given B.
- $P(A \wedge B)$, probability that both A and B occur.
- $P(B)$, probability of B irrespective of A.
- $P(B) > 0$
- A and B can be any proposition.

Example of conditional probability

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

- Domain = people
- A = isBald
- B = isOld
- A^B = isBald and isOld
- P(A) = 0.25
- P(B) = 0.25
- P(A^B) = 0.20
- P(A|B) = 0.80

Bayes' theorem

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

- $P(A)$, prior probability of A. (Does not consider B.)
- $P(A|B)$, conditional probability of A given B.
- $P(B|A)$, conditional probability of B given A.
- $P(B)$, prior of B. (Does not consider A.)
- $P(B) > 0$
- A and B can be any proposition.

Cond. prob. \rightarrow Bayes' theorem

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

- $P(A \wedge B) = P(A|B)P(B)$
- $P(B \wedge A) = P(B|A)P(A)$
- $P(A \wedge B) = P(B \wedge A)$, commutativity

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Example of Bayes' theorem

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

- Knowledge: disease A has as a complication B, about half of the time.
- Knowledge: the probability that someone has disease A is 1/50.000
- Knowledge: the probability that someone has B is 1/20
- Infer: is B a reliable indicator for disease A? $P(A|B) = 0.0002$, so not really...

Part III: Bayes' theorem applied to NLP

Machine Translation (MT), levels of complexity

- **Rough translation**
- **Restricted-source translation:** specific domain, vocabulary, and styling.
- **Preempted translation:** Preliminary human intervention.
- **Literary translation**

“The vodka is good but the meat is rotten”

- “The spirit is willing, but the flesh is weak.”
(allusion to Mark 14:38)
- 1960s ambition of building MT.
- From the perspective of Turing-style code-breaking.
- During the decennium it became clear that MT requires an understanding of the meaning of the message.

Noisy channel metaphor

- MT in terms of code-breaking.
- Translating English to French is like deciphering an English message into French.
- Which sentence F gives the highest value for $P(F|E)$?

Bayes' theorem & MT

- Translate English sentence E into French sentence F .
- $P(F)$, **language model**. How probable is sentence F in French?
- $P(E|F)$, **translation model**. How probable is E a translation of F ?
- $P(E) = 1.0$, since E is given.

$$\begin{aligned} & \operatorname{argmax}(F, P(F|E)) \\ &= \operatorname{argmax}(F, P(E|F)P(F)/P(E)) \\ &= \operatorname{argmax}(F, P(E|F)P(F)) \end{aligned}$$

Language model, $P(F)$

- Count occurrences of $F = f_1, \dots, f_n$ in a corpus.
- But, most sentences occur 0 times.
- Therefore, use bigrams:

$$P(f_1, \dots, f_n) = \prod_{i=1..n} P(f_i | f_{i-1})$$

Translation model, $P(E|F)$

- Count translation of F as E in a bilingual corpus.
- But, most sentences occur 0 times, and bilingual corpora are hard to come by.
- Therefore, we simplify things a bit:

$$P(e_1, \dots, e_n | f_1, \dots, f_n) = \prod_{i=1..n} P(e_i | f_i)$$

Part IV: Questions