N-Gram: Part 1
ICS 482 Natural Language Processing

Lecture 7: N-Gram: Part 1

Husni Al-Muhtaseb

بسم الله الرحمن الرحيم
ICS 482 Natural Language Processing

Lecture 7: N-Gram: Part 1

Husni Al-Muhtaseb

NLP Credits and Acknowledgment

These slides were adapted from presentations of the Authors of the book

SPEECH and LANGUAGE PROCESSING:

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

and some modifications from presentations found in the WEB by several scholars including the following

NLP Credits and Acknowledgment

If your name is missing please contact me
muhtaseb
At
Kfupm.
Edu.
sa

NLP Credits and Acknowledgment

Husni Al-Muhtaseb
James Martin
Jim Martin
Dan Jurafsky
Sandiway Fong
Song young in
Paula Matuszek
Mary-Angela Papalaskari
Dick Crouch
Tracy Kin
L. Venkata Subramaniam
Martin Volk
Bruce R. Maxim
Jan Hajič
Srinath Srinivasa
Simeon Ntafos
Paolo Pirjanian
Ricardo Vilalta
Tom Lenaerts

Heshaam Feili
Björn Gambäck
Christian Korthals
Thomas G. Dietterich
Devika Subramanian
Duminda Wijesekera
Lee McCluskey
David J. Kriegman
Kathleen McKeown
Michael J. Ciaraldi
David Finkel
Min-Yen Kan
Andreas Geyer-Schulz
Franz J. Kurfess
Tim Finin
Nadjet Bouayad
Kathy McCoy
Hans Uszkoreit
Azadeh Maghsoodi

Khurshid Ahmad
Staffan Larsson
Robert Wilensky
Feiyu Xu
Jakub Piskorski
Rohini Srihari
Mark Sanderson
Andrew Elks
Marc Davis
Ray Larson
Jimmy Lin
Marti Hearst
Andrew McCallum
Nick Kushmerick
Mark Craven
Chia-Hui Chang
Diana Maynard
James Allan

Martha Palmer
julia hirschberg
Elaine Rich
Christof Monz
Bonnie J. Dorr
Nizar Habash
Massimo Poesio
David Goss-Grubbs
Thomas K Harris
John Hutchins
Alexandros Potamianos
Mike Rosner
Latifa Al-Sulaiti
Giorgio Satta
Jerry R. Hobbs
Christopher Manning
Hinrich Schütze
Alexander Gelbukh
Gina-Anne Levow
Guitao Gao
Qing Ma
Zeynep Altan

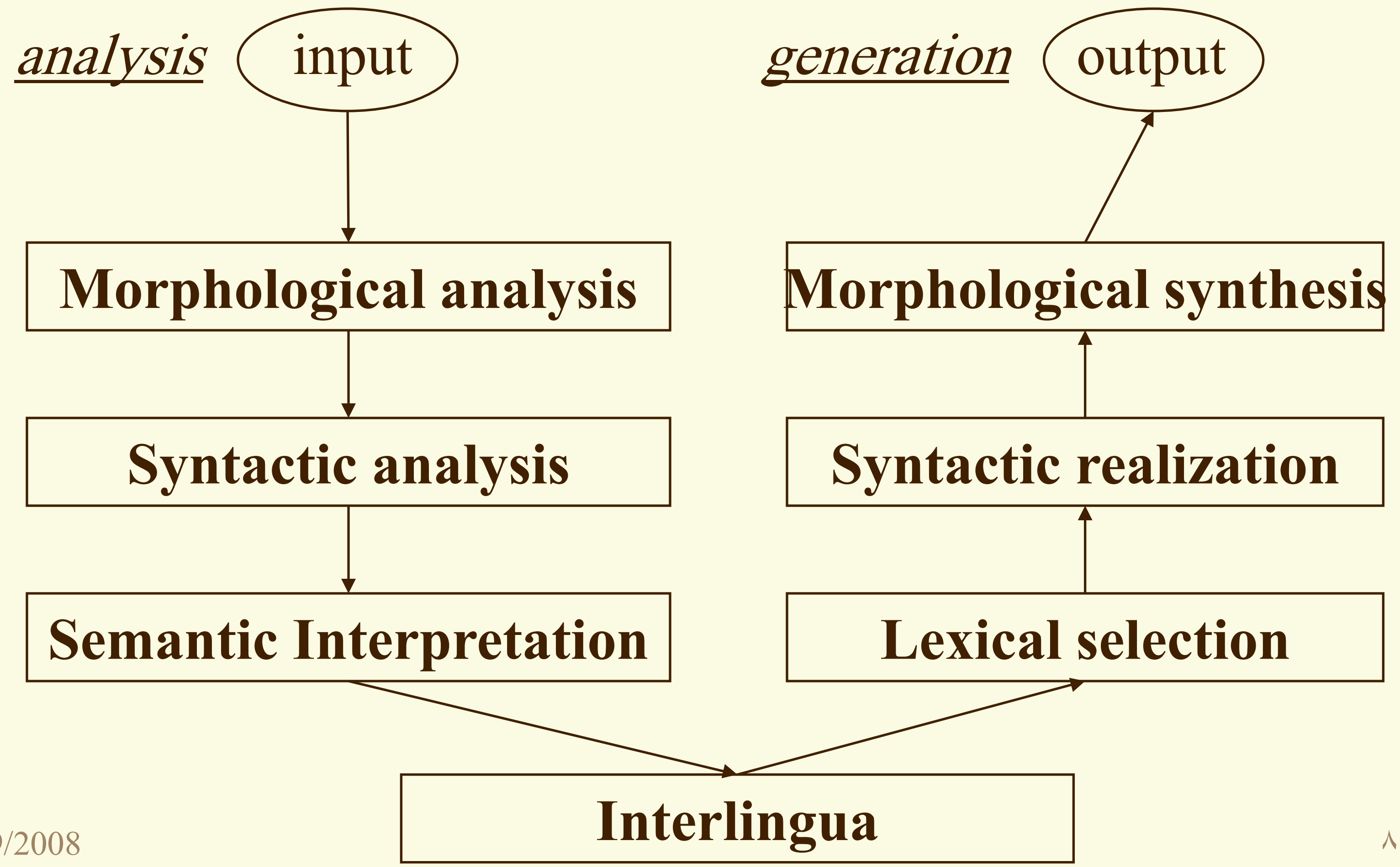
Previous Lectures

- Pre-start questionnaire
- Introduction and Phases of an NLP system
- NLP Applications - Chatting with Alice
- Regular Expressions, Finite State Automata, and Regular languages
- Deterministic & Non-deterministic FSAs
- Morphology: Inflectional & Derivational
- Parsing and Finite State Transducers
- Stemming & Porter Stemmer

Today's Lecture

- 20 Minute Quiz
- Words in Context
- Statistical NLP – Language Modeling
- N Grams

NLP – Machine Translation



Where we are?

- Discussed individual words in isolation
- Start looking at words in context
- An artificial task: predicting next words in a sequence

Try to complete the following

- The quiz was ------
- In this course, I want to get a good -----
- Can I make a telephone -----
- My friend has a fast -----
- This is too -------
- الوقت كالسيف إن لم تقطعه -------
- لا إله إلا أنت سبحانك إني كنت من -------

Human Word Prediction

- Some of us have the ability to predict future words in an utterance
- How?
  - Domain knowledge
  - Syntactic knowledge
  - Lexical knowledge

Claim

- A useful part of the knowledge is needed to allow Word Prediction (guessing the next word)
- Word Prediction can be captured using simple statistical techniques
- In particular, we'll rely on the notion of the probability of a sequence (e.g., sentence) and the likelihood of words co-occurring

Why to predict?

- Why would you want to assign a probability to a sentence or…
- Why would you want to predict the next word…
- Lots of applications

Lots of applications

- Example applications that employ language models:
  - Speech recognition
  - Handwriting recognition
  - Spelling correction
  - Machine translation systems
  - Optical character recognizers

Real Word Spelling Errors

- Mental confusions (cognitive)
  - Their/they're/there
  - To/too/two
  - Weather/whether
- Typos that result in real words
  - Lave for Have

Real Word Spelling Errors

- They are leaving in about fifteen minuets to go to her horse. *horse: house, minuets: minutes*
- The study was conducted mainly be John Black. *be: by*
- The design an construction of the system will take more than a year. *an: and*
- Hopefully, all with continue smoothly in my absence. *With: will*
- I need to notified the bank of…. *notified: notify*
- He is trying to fine out. *fine: find*

Real Word Spelling Errors

- Collect a set of common pairs of confusions
- Whenever a member of this set is encountered compute the probability of the sentence in which it appears
- Substitute the other possibilities and compute the probability of the resulting sentence
- Choose the higher one

Mathematical Foundations

Reminder

Motivations

- Statistical NLP aims to do statistical inference for the field of NL
- *Statistical inference* consists of taking some data (generated in accordance with some unknown *probability distribution*) and then making some inference about this distribution.

Motivations (Cont)

- An example of statistical inference is the task of *language modeling* (ex how to predict the next word given the previous words)
- In order to do this, we need a *model* of the language.
- Probability theory helps us finding such model

Probability Theory

- How likely it is that an A Event (something) will happen
- Sample space Ω is listing of all possible outcome of an experiment
- Event A is a subset of Ω
- Probability function (or distribution)

$$P : \Omega \rightarrow [0,1]$$

Prior Probability

- *Prior (unconditional) probability*: the probability before we consider any additional knowledge

$$P(A)$$

Conditional probability

- Sometimes we have partial knowledge about the outcome of an experiment
- Conditional Probability
- Suppose we know that event B is true
- The probability that event A is true given the knowledge about B is expressed by

$$P(A|B)$$

Conditionals Defined

- Conditionals

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

- Rearranging

$$P(A \wedge B) = P(A \mid B) P(B)$$

- And also

$$P(A \wedge B) = P(B \mid A) P(A)$$

$$P(A \wedge B) = P(B \wedge A) = P(B \mid A) P(A)$$

Conditional probability (cont)

$$P(A,B) = P(A \mid B)P(B)$$
$$= P(B \mid A)P(A)$$

- Joint probability of A and B.

Bayes' Theorem

- Bayes' Theorem lets us swap the order of dependence between events
- We saw that $P(A|B) = \frac{P(A,B)}{P(B)}$
- Bayes' Theorem:

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

Bayes

- We know…

$$P(A \wedge B) = P(A \mid B)P(B)$$

and

$$P(A \wedge B) = P(B \mid A)P(A)$$

- So rearranging things

$$P(A \mid B)P(B) = P(B \mid A)P(A)$$

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

Bayes

- "Memorize" this

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

Example

- S:stiff neck, M: meningitis
- P(S|M) =0.5, P(M) = 1/50,000 P(S)=1/20
- Someone has stiff neck, should he worry?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$= \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

More Probability

- The probability of a sequence can be viewed as the probability of a conjunctive event
- For example, the probability of "the clever student" is:

$$P(the \wedge clever \wedge student)$$

Chain Rule

conditional probability: $$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B)P(B)$$
and
$$P(A \wedge B) = P(B \mid A)P(A)$$

$$P(A \wedge B) = P(B \mid A)P(A)$$

"the student":

$$P(The \wedge student) = P(student \mid the)P(the)$$

"the student studies": $$P(The \wedge student \wedge studies) =$$
$$P(The)P(student \mid The)P(studies \mid The \wedge student)$$

Chain Rule

the probability of a word sequence is the probability of a conjunctive event.

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

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

Unfortunately, that’s really not helpful in general. Why?

Markov Assumption

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

- $P(w_n)$ can be approximated using only N-1 previous words of context
- This lets us collect statistics in practice
- 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
- Order of a Markov model: length of prior context

Corpora

- Corpora are (generally online) collections of text and speech
- e.g.
  - Brown Corpus (1M words)
  - Wall Street Journal and AP News corpora
  - ATIS, Broadcast News (speech)
  - TDT (text and speech)
  - Switchboard, Call Home (speech)
  - TRAINS, FM Radio (speech)

Counting Words in Corpora

- Probabilities are based on counting things, so ….
- What should we count?
- Words, word classes, word senses, speech acts …?
- What is a word?
  - e.g., are cat and cats the same word?
  - September and Sept?
  - zero and 0?
  - Is seventy-two one word or two? AT&T?
- Where do we find the things to count?

Terminology

- Sentence: unit of written language
- Utterance: unit of spoken language
- Wordform: the inflected form that appears in the corpus
- Lemma: lexical forms having the same stem, part of speech, and word sense
- Types: number of distinct words in a corpus (vocabulary size)
- Tokens: total number of words

Training and Testing

- Probabilities come from a training corpus, which is used to design the model.
  - narrow corpus: probabilities don't generalize
  - general corpus: probabilities don't reflect task or domain
- A separate test corpus is used to evaluate the model, typically using standard metrics
  - held out test set
  - cross validation
  - evaluation differences should be statistically significant

Simple N-Grams

- An N-gram model uses the previous N-1 words to predict the next one:
  - $P(w_n \mid w_{n-1})$
  - Dealing with *P*(<word> | <some prefix>)
- unigrams: *P*(*student*)
- bigrams: *P*(*student* | *clever*)
- trigrams: *P(student* | *the clever*)
- quadrigrams: *P*(*student* | *the clever honest*)

السلام عليكم ورحمة الله