Introduction to Natural Language Processing I [Statistické metody zpracování přirozených jazyků I] (NPFL067)

http://ufal.mff.cuni.cz/courses/npfl067

prof. RNDr. Jan Hajič, Dr. / doc. RNDr. Pavel Pecina, Ph.D.
ÚFAL MFF UK
{hajic,pecina}@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/jan-hajic
http://ufal.mff.cuni.cz/~pecina/index.html

Intro to NLP

- Instructors: Jan Hajič / Pavel Pecina
  - ÚFAL MFF UK, office: 420 / 422 MS
  - Hours: J. Hajic: Mon 10:00-11:00
  - preferred contact: **{hajic,pecina}@ufal.mff.cuni.cz**
- Room & time:
  - lecture: room S1, Tue 12:20-13:50
  - seminar [cvičení] room S1, Tue 14:00-15:30
  - Oct 2, 2018 – Jan 8, 2019
  - Final written exam (probable) date: Jan 15, 2019

Textbooks you need

- Manning, C. D., Schütze, H.:
  - *Foundations of Statistical Natural Language Processing*. The MIT Press. 1999. ISBN 0-262-13360-1. **[required]**
- Jurafsky, D., Martin, J.H.:
  - *Speech and Language Processing.* Prentice-Hall. 2000. ISBN 0-13-095069-6 and later editions. **[recommended].**

Other reading

- Charniak, E:
  - *Statistical Language Learning*. The MIT Press. 1996. ISBN 0-262-53141-0.
- Cover, T. M., Thomas, J. A.:
  - *Elements of Information Theory*. Wiley. 1991. ISBN 0-471-06259-6.
- Jelinek, F.:
  - *Statistical Methods for Speech Recognition*. The MIT Press. 1998. ISBN 0-262-10066-5
- Proceedings of major conferences:
  - ACL (Assoc. of Computational Linguistics)
  - EACL/NAACL/IJCNLP (European/American/Asian Chapter of ACL)
  - EMNLP (Empirical Methods in NLP)
  - COLING (Intl. Committee of Computational Linguistics)

Course requirements

- Grade components: requirements & weights:
  - Homeworks (1): 50%
  - Final Exam: 50%
- Exam:
  - approx. 4 questions:
    - mostly explanatory answers (1/4 page or so),
    - algorithms
    - only a few multiple choice questions

Homeworks

- Homework:
  - Entropy, Language Modeling
- Organization
    - (little) paper-and-pencil exercises, lot of programming
    - turning-in mechanism: see the web
    - no plagiarism!
- Deadline
  - Jan. 31, 2018
  - Late penalty: 5% of grade (0-100) per day (max. 50%)

Course segments

- Intro & Probability & Information Theory
  - The very basics: definitions, formulas, examples.
- Language Modeling
  - n-gram models, parameter estimation
  - smoothing (EM algorithm)
- Words and the Lexicon
  - word classes, mutual information, bit of lexicography
- Hidden Markov Models
  - background, algorithms, parameter estimation

NLP: The Main Issues

- Why is NLP difficult?
  - many "words", many "phenomena" --> many "rules"
    - **OED: 400k words; Finnish lexicon (of forms): ~2 . 10^7**
    - **sentences, clauses, phrases, constituents, coordination, negation, imperatives/questions, inflections, parts of speech, pronunciation, topic/focus, and much more!**
  - irregularity (exceptions, exceptions to the exceptions, ...)
    - **potato -> potato es (tomato, hero,...); photo -> photo s,** and even: both **mango -> mango s** or **-> mango es**
    - Adjective / Noun order: **new book, electrical engineering, general regulations, flower garden, garden flower, ...:** but **Governor General**

Difficulties in NLP (cont.)

- ambiguity
  - **books: NOUN** or **VERB?**
    - you **need** many books *vs.* she **books** her flights online
  - **No left turn weekdays 4-6 pm / except transit vehicles (***Charles Street at Cold Spring***)**
    - when may transit vehicles turn: Always? Never?
  - **Thank you for not smoking, drinking, eating or playing radios without earphones.** *(MTA bus)*
    - Thank you for not eating without earphones??
    - or even: Thank you for ~~not~~ drinking without earphones!?
  - **My neighbor's hat was taken by wind. He tried to catch it.**
    - ...catch the wind *or* ...catch the hat ?

(Categorical) Rules or Statistics?

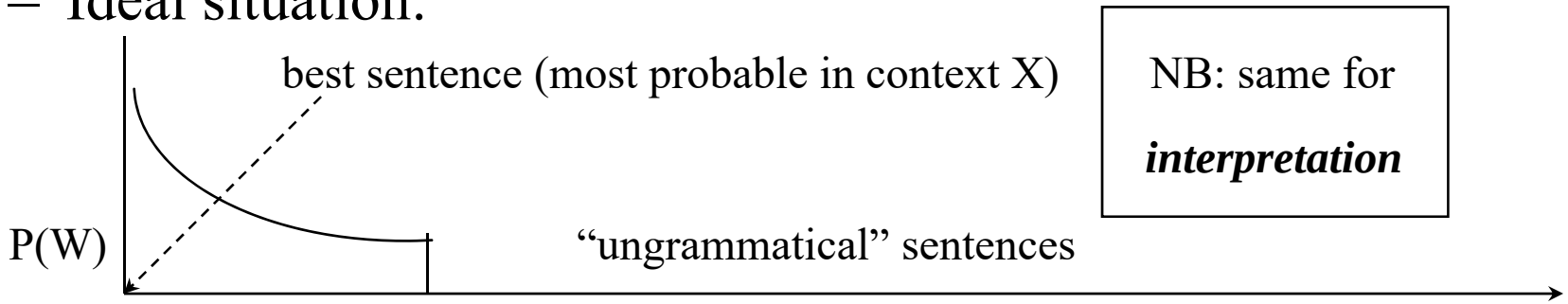
- Preferences:
  - clear cases: context clues: she books --> books is a verb
    - rule: if an ambiguous word (verb/nonverb) is preceded by a matching personal pronoun -> word is a verb
  - less clear cases: pronoun reference
    - she/he/it refers to the most recent noun or pronoun (?) (but maybe we can specify exceptions)
  - selectional:
    - catching hat >> catching wind (but why not?)
  - semantic:
    - never thank for drinking in a bus! (but what about the earphones?)

Solutions

- Don’t guess if you know:
  - **morphology (inflections)**
  - **lexicons (lists of words)**
  - **unambiguous names**
  - **perhaps some (really) fixed phrases**
  - **syntactic rules?**
- Use statistics (based on real-world data) for preferences (only?)
  - **No doubt: but this is the big question!**

Statistical NLP

- Imagine:
  - Each sentence W = { w_1, w_2, ..., w_n } gets a probability P(W|X) in a context X (think of it in the intuitive sense for now)
  - For every possible context X, sort all the imaginable sentences W according to P(W|X):
  - Ideal situation:



Real World Situation

- Unable to specify set of grammatical sentences today using fixed "categorical" rules (maybe never, cf. arguments in MS)
- Use statistical "model" based on **REAL WORLD DATA** and care about the best sentence only (disregarding the "grammaticality" issue)