Introduction to Computational Linguistics
Session 11: Recap

Denise Löfflad, Stephen Bodnar

Universität Tübingen

January 31, 2024

Overview

Session 1 and Definition

- Defining of Computational Linguistics
- Aims of the field
- Different characteristics and applications
- Different perspectives (theory, methods, tasks)
- Different methodologies (rule-based, statistical approaches)
- Relation to General Linguistics
- Main subtasks in NLP (NLU, NLG)

Session 2:Encoding Language

- Writing Systems
  - Alphabetic
  - Syllabic
  - Logographic
- How is language encoded on computers? (bits & bytes)
- Binary and Hexadecimal System
  - Binary system is base 2, using the symbols 0 and 1
  - $01101 = 0 * 2^4 + 1 * 2^3 + 1 * 2^2 + 0 * 2^1 + 1 * 2^0 = 13$
  - Hexadecimal system is base 16, the numbers 0 to 9 and the letters A (=10) to F(=15) are used
  - $0xA8E = 10 * 16^2 + 8 * 16^1 + 14 * 16^0 = 2702$
  - Convertion: 4 positions in binary can be represented with 1 position in hexadecimal

Session 2: Encoding Language

- ASCII
  - ASCII uses 8 bits to store 256 characters, was developed for the Latin Alphabet (esp. English)
  - Served as basis for future encoding systems
  - You should be able to read and use the ASCII chart
  - You should be able to discuss the necessity of a standardized coding table & the limitations of ASCII
- Unicode
  - Uses (up to) 32 bits
  - has a single representation for every character
  - UTF-8 allows for backwards compatibility with ASCII
  - UTF-8 allows for variable length (first bits show how many bytes are used)
    - 0 ... $\rightarrow$ one byte
    - 110... $\rightarrow$ two bytes (following: 10..)
    - ...
  - You should be able to discuss the advantages and limitations of unicode

Session 3: Writers' Aids

- Spelling variation (regional differences, different genres)
- Spelling error types (non-word, real-word errors)
- Error sources (opaque writing systems, L1 transfer, etc.)
- Basic edit operations (I, D, S, T)
- High-level stages in method for automatic spell checking (detection, generation, ranking)
- Information sources for different stages:
  - Detection: word lists
  - Generation: a) rules b) list of valid words & similarity measures
  - Ranking: explain why ranking is needed
- Minimal String Edit Distance (concept and algorithm)
  - Operatilising string similarity with naive examples
  - Modelling string similarity after human typist operations
  - Why DAGs are useful tool for representing search space

Session 4: Text as Data

- What is a Corpus?
  - Structured collection of texts collected with a specific question in mind. Usually contains linguistic annotation and metadata
- What is Metadata?
  - Data describing the primary, ra data (creation date, information and the speaker/writer/author, tags, ...)
- What are Linguistic Annotation
  - POS tags, sentence/word boundaries, parse trees, ...
- TTR $= \frac{\#types}{\#tokens}$
- Limitations & applications of corpora
- You should be able to discuss whether a certain corpus is appropriate for a specific research question, what the steps would be to collect data, what kind of (meta) data would be necessary to answer a certain RQ...

Session 5: Text Classification

- How computers learn
  - Feature vectors
  - Training, Testing, Validation Set
- Supervised vs. Unsupervised ML
  - Supervised: Labeled data through expert annotators. 1) Split the data, 2) train the model, 3) test and cross-validate, 4) evaluate
  - Unsupervised: 1) feature extraction, 2) apply algorithm on dataset, 3) inspect resulting structure
- How to evaluate a model
  - Recall, Precision, Accuracy, TNR
  - How are these calculated?
  - When should you use which measure? $\rightarrow$ Is it more important to catch everyone with the disease and maybe treat healthy patients, or is it better to miss some patients but be sure not to treat healthy people? Is it better to censor non-harmful tweets and make sure to censor every hate speech tweet ... ?

Session 5: Text Classification

- NLP segmentation
  - Preprocessing
  - Tokenization: What is a token, what are the challenges (contracted forms, hyphenated forms, periods (St.), special characters, NER, ...)
  - Sentence segmentation: How are sentences defined, what are the challenges in sentence segmentation (esp. with regards to other languages than English¸)

Session 9: POS tagging & Parsing

- POS tags
  - Information on word type (noun, verb, etc.)
  - Different granularity depending on the corpus
  - Used to encode morphological and syntactical information of tokens, can be used to measure linguistic properties such as complexity or pronunciation
  - Often prerequisite for other NLP tasks
- Methods for POS tagging
  - Rule-based
    - Use constraint grammars and lexicons to retrieve all candidate POS tags and the associated rules
    - Try each rule on input sentence and exclude POS tag if rule does not fit
  - Statistical
    - Input a token sequence, assign a POS tag to each toekn, handle ambbiguity by exploring different possible sequences
    - Rank each sequence by its statistical probability using supervised ML

Session 9: POS tagging & Parsing

- Challenges in POS tagging
  - Ambiguity
  - Sometimes, various POS tags are possible (I can can a can of tuna) → Context necessary
  - You should be able to annotate short sequences and discuss possible ambiguities including ways to resolve the ambiguity

Session 9: POS tagging & Parsing

- Parsing
  - Represent grammatical structures or relationships
- Constituency Parsing
  - Constituents: Group of words forming syntactic units (e.g. NPs)
  - Parsing assigns syntactic structures
  - Identification of multi-word units and nested structures
  - Visualizations: Trees and parentheses notation. You should be able to read & create both forms
  - Challenge: Ambiguities, e.g. I saw the man with the telescope
- Dependency Parsing
  - Aim: Capture the syntactic relations between words in a sentences using directed grammatical relations between pairs of words
  - Relations have governors (head) and dependent
  - Challenges: Attachment and coordination ambiguity

Session 6: ASR

- Brief introduction to speech (phones and phonemes, vocal tract)
- How speech is represented with computers
- Different speech processing applications
  - Which ones can you think of ?
  - What task does each one perform ?
  - What language technology components are required in the pipeline to make make these applications work ?

Session 6: ASR cont.

- Automatic speech recognition
  - Inputs / outputs
  - Why ASR is a difficult problem
  - Noisy channel model and application to ASR
  - System architecture of 'classical' ASR system; role of components (acoustic model, lexicon, language model, search lattice, decoder)
  - Purpose of language model
    - Using a *language model* to rank different candidates ("wooden boets")
    - A very high-level look at probabilities (coin flipping) and n-gram language models
  - Steps in training a speech recognizer (dataset requirements, feature extraction, training, testing, evaluation, error analysis)

Session 7: Text Search

- Why search can be challenging, with example(s)
- Different search tasks (hint: Google, Quora, Netflix)
- Key terms, such as *information need*, *intent*, *query*
- Formulating queries: general user vs. specialist user

Session 7: Text Search cont. - Regular Expressions

- Definition of regular expressions
- Basic building blocks:
  - literals
  - wild cards
  - escaping
  - modes (/g, or /i)
  - character sets and ranges
  - quantified repetition
  - groups using parentheses
  - anchors

Session 7: Text Search cont. - Multiple Documents

- Examples of applications searching multiple documents
- What document indexing is, why it is important, and different approaches (term-by-document matrix, inverted indices list)
- Evaluating search quality
  - User experience surveys
  - Objective user interaction logs
  - Precision, Recall (and challenges with computing recall), F-measure
    - An understanding of the limitations of these measures in the context of Text Search
  - An understanding of why Ranking is important in text search

Session 8: CALL

- Characteristics of learner language (in contrast to native-like language) and implications for NLP technology
- Know what an Intelligent Tutoring System (ITS) is
- Be able to talk about NLP applications in the context of ITS for language learning
  - NLP applied to well-formed language (e.g. FLAIR)
  - NLP applied to malformed learner language
  - For each, be able to describe:
    - The needs of the users (e.g., input, corrective feedback)
    - How NLP can be used in an ITS to meet those needs
  - Using knowledge from other lectures (e.g., text search), make suggestions for how to evaluate the quality of an NLP-enhanced ITS system

Session 9: Dialogue Systems

- Terminology
  - speech acts, common ground, turn taking, adjacency pairs
  - how are these important for dialogue systems?
  - you should be able to define, identify, and discuss these
- Grice's Maxims
  - what are they, how are they defined?
  - why are they important in dialogue systems? what happens if they are not met?

Session 9: Dialogue Systems

- Task Specific Systems
  - are designed to help accomplish one or more specific tasks
  - intents containing slots, intent recognition
- Task Independent Systems
  - chatbots
  - different implementations (brute force, rule-based, corpus trained, generative)
  - you should be able to explain the previously mentioned phenomena on chatbot examples (e.g. explain the rule based approach on the example of ELIZA)
- How are chatbots evaluated? (Turing test, Winograd Schemas)

Session 10: LLMs

- Technical Aspects
  - You do not need to be able to explain what neural networks are or what a transformer is, but you need to know the vocabulary!
  - You should be able to discuss what was new about LLMs (one model for all, parallelization, attention)
  - **Large** Language models
- Applications of LLMs
  - You should be able to name a few applications of LLMs

Session 10: LLMs

Ethical Aspects

- Bias
  - difficulty of unbalanced, restricted, and biased corpora
  - how are social biases inscribed in the data?
  - why is this problematic?
- Fine tuning
  - why is it difficult to ensure that only 'ethical' utterances are generated?
  - how did chatbots try to solve this problem? how is it done nowadays?
- Environmental impact
  - You should be able to shortly discuss possible environmental impacts (also considering the large datasets, long learning, etc.)

Exam

- Feb. 7^{th}, start at 4.15pm and end at 5.45pm
- Hörsaal 0.02VG
- Be here at least 15min early to make sure we can start on time!
- Bring your student ID
- You're allowed one (!) cheat-sheet
- If you are sick, get a doctor's note!! Otherwise, you'll automatically fail the exam
- Please let us know if you plan to write the exam in this survey so we can plan accordingly
- There will be a retake, likely at the end of the Vorlesungsfreie Zeit. You can only participate if you achieved 60% on the assignments
- The deadline to de-register for the exam is Feb. 2^{nd}!

Exam

- If you still have questions, do not hesitate to post on moodle! We, the tutor, and the other student could help you!
- If that is no option, write us an e-mail (posting on moodle is better)
- If you need a Schein, please register on Moodle


GOOD LUCK ON YOUR
YOUR JOURNEY!

References and Acknowledgments I