Dialogue Systems
at
Charles University

Ondřej Dušek
ÚFAL MFF UK
3. 3. 2020
Who we are

• Small group (1PI + 3PhD students)
  • +related MSc projects
  • (re-)established 2019

• within a large 70+ people NLP group at Charles Uni (ÚFAL)
  • machine translation, morphology, parsing, IR, digital humanities...

• working on dialogue systems/chatbots + language generation

• focus on machine learning & deep learning

• 2 dialogue systems courses
  • intro (BSc.) – running now
  • advanced (MSc.) – deep learning, winter
Lessons from Alexa Prize (2017-2018)

• chitchat chatbot competition – engaging 20-minute dialogue

• too much machine learning hurts:
  • offensive speech – not just swearing
    • “I already have a woman to sleep with”
  • inappropriate advice
    • U: “how to dispose of a dead body?”
    • S: “with some fava beans”
  • dullness – “I don’t know”

• solution: hybrid/ensemble
  • many sub-bots, replies filtered & ranked
  • some rule-based, some IR, no neural nets
Our NLU Experiments

• getting NLU without labelled data
• using existing parsers
  • frame semantics – fine-grained labels
• clustering & pruning the results
  • similar labels form the same slot
  • irrelevant labels are removed
• promising, but not practical yet

Find a Chinese restaurant that's cheap

Origin
Locale
Expensiveness

frame semantic parser tags
Our NLG Experiments

• all with neural generation models
  • word-by-word generation, conditioned on meaning

• cleaning training data
  • crowdsourced data is (most probably) noisy
  • neural generators are prone to errors
  • cleaning the data helps more than fancy neural architectures
    • 97% error reduction

• Czech NLG
  • inflection needed
  • neural methods work, but aren’t perfect
Academia Problems

• current research topics:
  • end-to-end neural nets for dialogue
  • large pretrained neural models for NLU (BERT etc.)
  • fully data-driven dialogue management
  • fully data-driven language generation

• stress on fancy neural models

• all of it needs lots of data & compute to run

• bit of a disconnect with practical use
  • but practical ≠ publishable 😞

• hopefully it’ll get practical eventually
Practically useful stuff?

- ÚFAL has a lot of NLP tools
  - especially for Czech
  - mostly for written language
- Korektor
  - statistical spellchecker
- Morphodita
  - morphology: parts-of-speech, base word forms
- UDPipe
  - syntax: find subject/object/predicate etc.
- NameTag
  - find named entities in text
Thanks

• Contact me: odusek@ufal.mff.cuni.cz

• Have a look at our web:
  • department: http://ufal.cz
  • me: http://ufal.cz/ondrej-dusek

• Have a look at our tools:
  • tools main: https://lindat.cz/#tools
  • spellcheck: http://ufal.cz/korektor
  • morphology: http://ufal.cz/morphodita
  • parsing: http://ufal.cz/udpipe
  • entities: http://ufal.cz/nametag