Learning research embeddings

Benoit Favre <benoit.favre@lif.univ-mrs.fr>

Aix-Marseille Université

February 17, 2015

What's keeping me busy

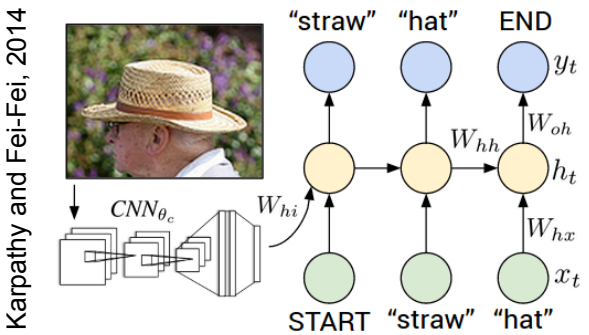
- BOLT: Speech-to-speech translation with clarification sub-dialogs
  - Merging rephrased segment into original utterance
  - Lattice alignment in the WFST framework (Mickael Rouvier)
- Rocio: Flamenco performance supported by speech recognition
  - Adaptation to non-native semi-scripted speech
- Orfeo: Unified corpora annotation and tools for linguists
  - Segmentation of spontaneous speech in sentence-like units
- Asfalda: Semantic frame parsing of French
  - Next-generation semantic parser (Olivier Michalon)
  - Low-effort semantic annotation of speech (Jeremy Trione)
- SENSEI: Conversation summarization
  - Cross-domain adaptation of NLP through deep learning (Jeremie Tafforeau)
  - Synopsis generation from call-center conversations (Jeremy Trione)
  - Coreference resolution (Elisabeth)
- ADNVideo: Video understanding
  - Deep learning for multimodal video characterization (Meriem Bendris)

Deeper dive into three topics

- Abstractive summarization
  - Overcoming the comfort of extracting sentences
- Social media summarization
  - What's in that million tweets?
- Programming with speech
  - Can we teach SIRI new tricks?

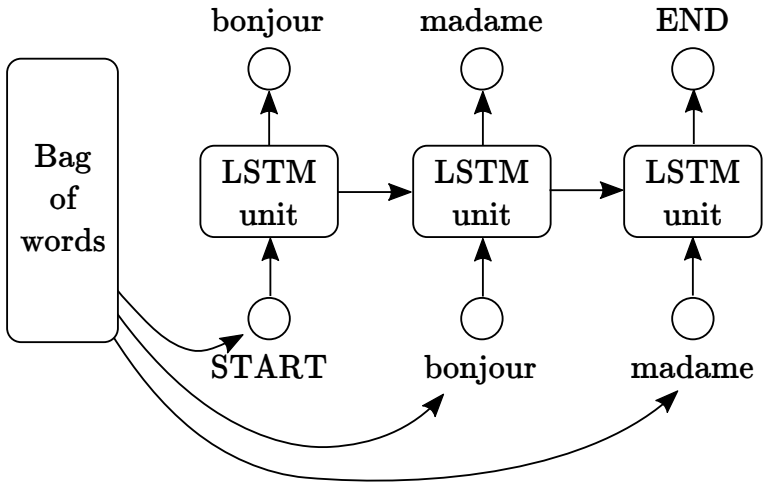
Abstractive summarization

- Idea: conditioned language model (Mehdy Bouaziz)
  - $P(\text{word}|\text{history}, \text{document representation})$
- Estimation with RNN+LSTM
  - Image captioning community
- Open questions
  - Data scarcity
    - ⋆ Will start with title generation
  - What is a good representation for document content?
    - ⋆ "Doc2Vec" family of embeddings are not there yet
    - ⋆ Can we take advantage of NLP analysis?
  - How to scale to large vocabularies?
    - ⋆ Word embeddings in output + language model
  - How to deal with specific words?
    - ⋆ Template generation and then filling with entities
  - Long term predictions
    - ⋆ Where do RNN go after 100 words?
    - ⋆ How to learn a long-term structure?
  - Learning through the embedding?



Initial experiment

- Task: generate a sentence from a bag of the words it should contain
- Corpus
  - Decoda (train with 19k words, test with 2.3k words)
  - Vocabulary size: 2116
  - Representation: one-of-n + bag-of-word



- RNN training: Currennt (`http://sourceforge.net/projects/currennt/`)
  - 200 Epochs (8h)
  - Learning rate 10^{-4}
- Results: 63% of accuracy

Summarizing social media

- Problem: finding trends in masses of online comments
  - What are humans interested in?
  - How can we extract meaning from 1 million tweets?
- Study with EJCM journalism students
  - Collect user needs
    - Track the source
  - Students are asked to summarize 50-300 comments from Le Monde articles
  - Methodology
    - Annotate comments with free-form topical description
    - Regroup descriptions to create comment clusters
    - Write a summary of trends
- Social-media sentiment analysis
  - Can we find the polarity, valence and target of opinions in tweets?
  - Shared task in April (Deft challenge)
- At a larger temporal scale (Balamurali A R)
  - Can we show the existence of linguistic pattern differences?
  - Separate topical shift from linguistic shift?

Slashdot corpus

- Case study: Slashdot corpus "stuff that matters"
  - 30 million comments over 15 years, 2.5B words
  - Relatively homogeneous community (engineers and scientists)
  - Elaborate community-driven moderation system
- Methodology
  - Create year-level word embeddings
  - Align embeddings with linear transform
    - $Op = (V_2 \times V_1^{-1})$
  - Look at words which move in the representation
  - Look at frequency differences across years
  - 2D representation with t-SNE

Programming with speech

- Can we teach SIRI to do new tricks? by speech only?
- Objectives:
  - Hands-free, eyes-free programming
  - The computer does the programming, the human just states his problem
- NLP can help programming
  - Program dictation / synthesis
  - Code generation from comments
  - Refactoring of variable names, make functions from code
- Open problems
  - How do humans deal with task definition?
  - How to navigate a program?
  - How to generate a program by analogy?
  - Can we refactor methods automatically?
  - How to introduce new words, new concepts?
  - Can be invent a language that will maximize ASR performance?
  - Can we make all that fit a coffee maker?

Some examples

- Tavis Rudd: using python to code by voice
  - Repetitive strain injury
  - `https://www.youtube.com/watch?v=8SkdfdXWYaI` (9:15)
- Programming by Voice: enhancing adaptivity and robustness of spoken dialogue systems [Georgila 2006]
  - Dialog system with user macros
- Mining source code repositories at massive scale using language modeling [Allamanis 2013]
  - N-gram LM for code suggestion
- Structured Generative Models of Natural Source Code [Maddison 2014]
  - PCFG for source-code generation
- Predicting Program Properties from "Big Code" [Raychev 2015]
  - Conditional Random Fields for predicting variables type and name
  - `http://www.jsnice.org/`
- Demo
  - Python language model + lexicon in Kaldi
  - Gtk3 UI