Natural Language Annotation for Machine Learning

January 12, 2018
Professor Meteer

+ Course overview

- Schedule and assignments
  - CS140.mmeteer.com
- Learn by doing
  - Course is centered around group annotation projects
  - We will walk through every step of the process
- Textbook:

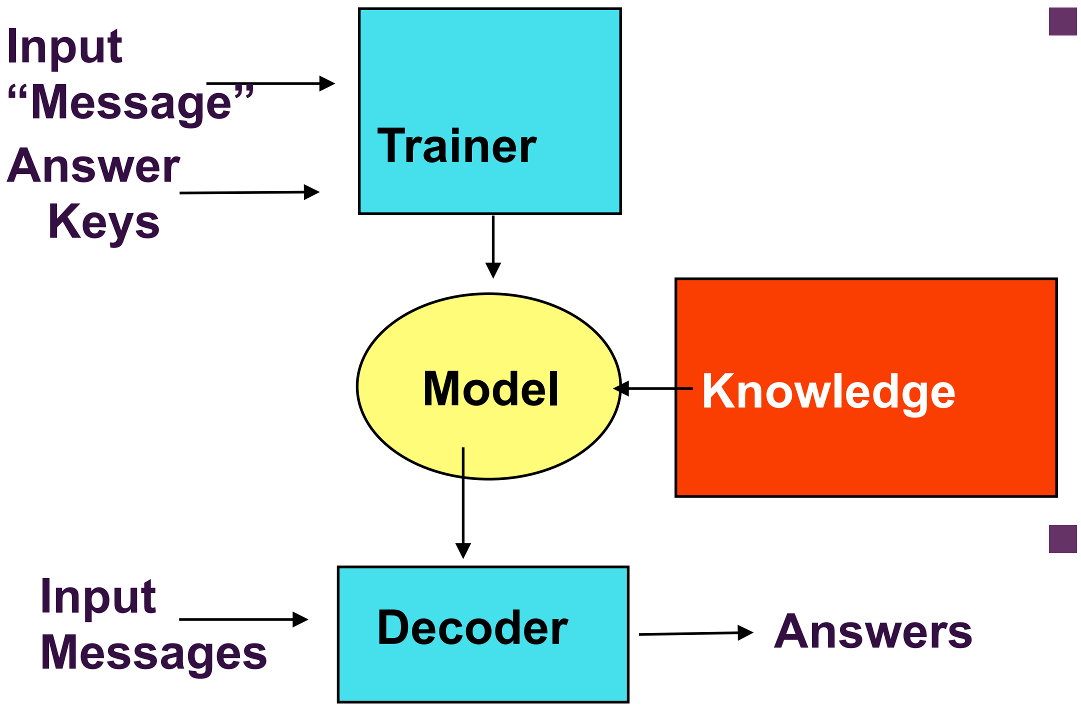
  Natural Language Annotation for Machine Learning

  Pustejovksy & Stubbs, O'Reilly Press

+ Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Variously referred to as the “corpus based,” “statistical,” or “empirical” approach.
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s.
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

+ Speech and NL Paradigm



- **Requirements:**
  - Annotation of messages with keys
  - Linguistic and Domain Knowledge
  - Statistical Model
  - Training Algorithm
  - Decoding Algorithm
- **Benefits:**
  - Statistical model can combine multiple kinds of information
  - Degrades "softly", finding the most likely answer
  - Learns what information is important to make a decision

+

Supervised Learning for Language Technologies

Technology	Input	Answers
Speech Recognition	Audio	Transcription
Optical Character Recognition	Image	Characters
Topic classification	Document	Topic labels
Information retrieval	Query	Document
Named entity extraction	Text or speech	Names and categories

+ Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

+ The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - “The a are of I” is a valid English noun phrase (Abney, 1996)
    - “a” is an adjective for the letter A
    - “are” is a noun for an area of land (as in hectare)
    - “I” is a noun for the letter I
  - “Time flies like an arrow” has 4 parses, including those meaning:
    - Insects of a variety called “time flies” are fond of a particular arrow.
    - A command to record insects’ speed in the manner that an arrow would.

- Some combinations of words are more likely than others:
  - “vice president Gore” vs. “dice precedent core”

- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

+ Course Project

- Form Groups
- Define annotation goal
- Group Contract
- Task Description and Corpus Selection
- Initial Annotation Spec
- Full Annotation Spec
- Adjudication and precision and recall
- Train Machine Learning Algorithm
- Write-up
- Presentation
- Peer Evaluation

+ Projects from 2015

- Clickbait
- Habeas Corpus
- NLP4SLE
- OKML
- RxML
- TrollML
- Project folders
  https://drive.google.com/drive/u/1/folders/0B6z1otdg2OZuSXpPek1ZZi1oOG8

+ Projects from 2016

- Hatespeech
- L1ML
- MojiSem
- SoccEval
- WriteRec
- Yelp
- Project deliverables: https://drive.google.com/drive/u/1/folders/0B4mlJQRTyYwgODhkTndSS2tUVUU

Projects from 2017

- SwitchML: Annotated Yelp reviews of cell phone providers to predict whether customers would switch carriers

- Topic Changes: Annotated changes in topic in dialogs to predict when the topic changed. goal: improve topic change in Alexa Prize Chatbot

- Yelp Travel Review: Annotate yelp travel reviews to build profiles of travelers by what they talk about.

- SpeechAct ML: Classify speech acts in conversational dialog. Longer goal: helping Alexa Prize Chatbot

- https://drive.google.com/drive/u/1/folders/0B6z1otdg2OZuRkVLS2Q5RDhTaFk

+ The Switchboard Corpus

- Corpus description
  - Conversations between people about randomly assigned topics
  - “Conversation without intention”
    - Speakers weren’t invested in the topics
    - No history and not future
  - Aside from that “natural conversation”¯
- Original goal
  - “Conversational” speech recognition (in contrast to broadcast news
  - Wide variety of speakers, accents, vocabulary
  - Fully available for research (no strings)¯

+ The many levels of Switchboard

from Univ. of Edinburgh “Switchboard in NXT”

- Text: Words, punctuation, silence, other noisesPart of speech and syntactic structure (ala Penn Treebank)
- Movement: Traces and antecedents
- Turns and liner order, including overlap
- Disfluencies: hesitations, false starts, repair, repetition
- Active voice
- Information status, animacy
- Coreference
- Dialog act
- Kontrast & trigger

+ Many Levels of Switchboard (cont.)

- Word timing (automatically derived with alignment)
- Syllables
- Phones
- Accent
- Phrase
- ToBi intonation (phrase and boundary tone)
- Prosody notes

+ SWBD Dialog Act Corpus

- Includes the following representations
  - Call/caller attributes
  - Text (with dysfluencies marked)
  - Dialog act per utterance
  - POS tags
  - Syntactic tree
- This version comes with a set of python readers and classes
- http://compprag.christopherpotts.net/swda.html

+ Levels of annotation

Phonemes	Syllables	Wprd
0.054626 121 h# 0.126670 121 l 0.245484 121 ay_cr 0.312623 121 k 0.463527 121 f 0.611510 121 ay 0.673819 121 n 0.739049 121 ih_n 0.787725 121 ng 0.917473 121 ax_cr 1.207598 121 h# 1.344802 121 p 1.364688 121 r 1.467519 121 aa 1.540537 121 p 1.583020 121 er 1.693388 121 n 1.828292 121 er 1.939090 121 s 2.000064 121 ih_n 2.041928 121 ng 2.150000 121 hh 2.281267 121 ow_cr 2.380000 121 m 2.464436 121 h#	0.054626 121 h# 0.312623 121 l ay_cr k 0.611510 121 f ay 0.787725 121 n ih_n ng 0.917473 121 ax_cr 1.207600 121 h# 1.467520 121 p r aa 1.583020 121 p er 1.828290 121 n er 2.041930 121 s ih_n ng 2.380000 121 hh ow_cr m 2.464440 121 h#	0.054191 121 H# 0.312720 121 LIKE 0.787725 121 FINDING 0.918446 121 A 1.207500 121 H# 1.583020 121 PROPER 2.041930 121 NURSING 2.380150 121 HOME 2.464440 121 H#

Discourse

b	A.15 utt1: Uh-huh. /
+	B.16 utt1: -- the different, - /
qy	B.16 utt2: do you have kids? /
na	A.17 utt1: I have three. /
bh	B.18 utt1: {F Oh, } really? /
ny	A.19 utt1: Uh-huh. /
x	B.20 utt1: <Laughter>.
b	A.21 utt1: Yeah, /
sd	A.21 utt2: I do <laughter>. /
%	A.21 utt3: Yes, {F uh, } /
sd	A.21 utt4: I don't work, though, /
sd	A.21 utt5: {C but } I used to work [and, +] when I had two children. /