Haladó Gépi Tanulás 3

Kornai András

HGT 2022/9/22

Óraterv

- Az öt (négy?) projekt-terv megbeszélése
- A tartalom-összefoglalás (summarization) alapjai
- Ha marad idő: a beszédfelismerés alapjai

„K” csoport:Text summarization using NLP methods (többféle tematikájú szövegen)

Miért esett erre a választás: Alapból valamilyen NLP problémakörrel akartunk foglalkozni, mert ezzel kevés tapasztalatunk van még, illetve a felhasználása igen sokrétű. A szövegösszefoglalást azért választottuk, mert egyrészt igen érdekes és kihívást jelentő, hogy pontosan milyen szemantikai tulajdonságokat felhasználva és milyen algoritmusokkal lehet a szövegből a lényegi tartalmakat kinyerni gépi tanulással, másrészt pedig a mai embernek sokszor nincs elég ideje több oldalas dokumentumokat átnéznie egy adott információ keresése végett, viszont Text Summarization-nel ezt a problémát valamelyest orvosolni lehet, így megéri kutatni.

K csoport menetrend (hetek)

3-5 irodalom és korábbi megoldások áttekintése, adatok begyűjtése

6-8 adatpreparáció, felderítő adatelemzés

9-10 Modellezés

11-12 hiperparaméter optimalizáció, modellek kiértékelése

13-14 Prezentáció bemutatása

Summarization on one slide

- Two major types: *extractive* – select words, phrases, clauses, sentences from the original text v. *abstractive* understand the text, generate summary based on this understanding
- What makes this problem hard is the evaluation: no clear automated figure of merit, so human labor in the eval loop. This is slow and expensive, unfit for gradient descent
- Many competing automated figures of merit, starting with ROGUE (2004) For a recent summary and a new proposal see https://www.aaai.org/AAAI22Papers/AAAI-4389.ColomboP.pdf for a comparison across several see `SummEval`
- The unbeatable baseline: take the first three sentences

K csoport kritika

- A téma (text summarization) fontos, hasznos
- Előre el kell dönteni mi az adat, mi a jóságmérték, utólag ez ciki (eredményhalászat)
- Van egy csomó standard adathalmaz, ld pl. `https://paperswithcode.com/task/text-summarization`
- Jövő hétre ezeket a döntéseket illene meghozni
- „felderítő adatelemzés” *ne már . . .*
- Helyette lehet párhuzamosan több benchmarkon dolgozni, az nem ciki (amíg az eredmények közt nem válogatunk)

„Tomi” csoport

Keywords: OLS Bootstrap, Resampling, Pairs, Residual, Wild Bootstrap, MacKinnon-White errors, White-test, Bootstrapped Sampling Distribution of Regression Coefficients, Bootstrapped Confidence Interval, Hungarian Balance Sheet Data, Cross-Sectional Data, Observational Unit Bootstrapping

Plan: My project is built on the ideas of MacKinnon and Efron to bootstrap linear regressions, whose standard errors are not constant, that is, heteroskedastic. The aim of the bootstrap in our case is to get an accurate standard errors and be able to build a CI on regression coefficients. I am gonna demonstrate it on a Hungarian Balance Sheet data of firms choosing a year. I have been building a Python package for several months and am gonna be deploy on this dataset, where the target variable is the log(sales). The regressors will be later determined. The main topic would be how accurately we can estimate the coefficients and the how accurate our CIs on those coefficients. These stuff will be shown with different sample sizes.

„Tomi” csoport kritika

- Nyelvi: csak formális nyelvezetet lehet használni, semmi I'm gonna. Nekem mindegy, de ez megakadályozza a szöveg újrafelhasználását (cikkben, disszertációban), mert ilyen helyeken garantált, hogy valamelyik bíráló beleköt.
- Nem értem rendesen. Se az adatok nincsenek leírva kellő alapossággal, se a javasolt módszer. Mi a CI?
- Ha már régóta építesz egy programcsomagot, tervezz be egy előadást ennek ismertetéséről
- Általában is kérek menetrendet
- Nincs specifikálva, hogy az eredmény hogy lesz kiértékelve. Nálunk (NLP) nem lehet elfogad(tat)ni olyan cikket, amiben nincs evaluation

Beszédfeldolgozás

- A két alapfeladat: beszédfelismerés (automatic speech recognition, ASR) és beszédszintésis (text-to-speech, TTS)
- Sokszor kell mindakettő (pl. speech-to-speech translation)
- A történet: először emberi szakértelemmel, aztán egyre automatikusabban
- The bitter lesson

Information content

- CD-ROM Audio – 700 kbps (44.1khz, 16 bits per sample)
- MPEG Audio – 112 kbps (44.1khz, 3 bits per sample)
- regular „toll" quality speech – 96 kbps
- ADPCM – 32 kbps (toll quality)
- LPC – 9.6 kbps (near toll quality)
- VQ homomorphic – 0.6 kbps
- symbolic – 0.2 kbps (0.05 kbps)

Excitation-Filter Model

- *Source (voiced):* glottal pulse train (half-rectified sine wave)
- *Source (unvoiced):* bandlimited white noise further up the vocel tract
- *Filter (oral):* hardwalled, lossless tube (only poles)
- *Filter (nasal):* two tubes (zeros as well)
- *Radiation losses:* (at lips, nostrils) mostly ignored

$e(t) \leftrightarrow E(\Omega)$
$v(t) \leftrightarrow V(\Omega)$
$s(t) = e(t) * v(t) \leftrightarrow S(\Omega) = E(\Omega)V(\Omega)$

Main encoding methods

- Filter Bank Analysis (Gabor-type filters)
- Subband Vocoders (wavelet-type analysis)
- Homomorphic (cepstral) Analysis
- (Delta) Pulse Code Modulation
- Linear Prediction Coding

A fonetikusok összeomlása

Taháfut-al-fonetika

- Kb. 40 éven át hazudták a fonetikusok azt, hogy ők értik a folyamatokat (akusztikus és auditorikus oldalról egyaránt): Potter (1945) kezdte és Makhoul and Schwartz (1986) lezárta (bár van aki máig hazudik)
- A szakértők nem voltak szakértők. Amit mondtak az nagyjából igaz volt, de a teljes igazságnak kevesebb mint a harmada
- M&S 1986: tudomásul kell vennünk, hogy pont a részletek tekintetében (pedig az ördög ott lakik) tudatlanok vagyunk
- Ha ez így van, akkor csak a tudatlanságon alapuló modellezés lehet hatékony (addigra a statisztikai modellek már péppé verték a tudás-alapúakat)
- A fonetikusoknak volt 35 évük arra hogy visszavágjanak
- Fool me once, shame on you, fool me twice, shame on me. Ma nincs az a majom, aki elhinné, hogy a fonológusok, szintakták, szemanták bármivel is küönbek lennének a fonetikusoknál