Mérési feladatok

Nyelvi modellezés a beszédfelismerésben című méréshez

Tarján Balázs
BME-TMIT
2016

1. feladat – Ismerkedés a tanítószövegekkel

A feladat során megismerkedünk a későbbiekben nyelvi modellek készítéséhez használt két tanítószöveggel. Az első egy magyar nyelvű szövegkorpusz, mely Krúdy Gyula prózai műveit tartalmazza (*txt/krudy_train.txt*). A második korpusz szintén irodalmi gyűjtés, azonban angol nyelvű és William Shakespeare művei alapján készült (*txt/shakespeare_train.txt*).

a) Nyissa meg az egyes szövegeket (pl. gedit) és olvasson beléjük! Milyen nyelvezetűnek találja a szövegeket?

b) A „train_txt" könyvtárban a nyelvi modell tanításához előkészített „shakespeare_train.txt" nevű szöveg mellett megtalálja az eredeti „shakespeare_orig.txt" nevű korpuszt is. Hasonlítsa össze a két szöveget és röviden foglalja össze, hogy milyen módosításokat végeztünk rajta!

c) Számítsa ki mindkét korpuszban a token-ek (összes szó) és type-ok (összes fajta szóalak) számát!

d) Készítsen közös ábrát, melyben a magyar és angol nyelvű korpusz szótárbővülési sebessége hasonlítható össze (type-ok száma a token-ek függvényében). Mivel magyarázza a különbséget a magyar és angol korpusz görbéje között?

2. feladat – Nyelvimodell-fokszám

a) Készítsen egy 0 és egy 1-gram és egy 2-gram modellt a magyar korpuszon és vizsgálja meg a nyelvi modell fájlt (használja az alapértelmezett simítási módszert). Milyen különbséget lát a 0-gram és az 1-gram modell között? Milyen további információt tartalmaz a 2-gram modell az előbbiekhez képest?

b) Készítsen 0-gram nyelvi modellt az angol nyelvű korpuszon is. Mérje meg a 0-gram nyelvi modellek perplexitását és OOV arányát a hozzájuk tartozó tesztszövegeken (*test_txt*)! Miért pont ekkorára adódik a perplexitás értéke? Milyen különbség figyelhető meg a magyar és angol korpusz OOV aránya között? Mi ennek az oka?

c) Készítsen különböző fokszámú (N=1..5) nyelvi modelleket a magyar és angol nyelvű tanítószövegek felhasználásával, az alapértelmezett simítási technika használatával. Minden elkészült modellt teszteljen a hozzá tartozó tesztszövegen. Hogyan változik a perplexitás és az OOV arány a nyelvi modell fokszám növelésével? Készítsen ábrákat hasonlítsa össze őket és magyarázza meg a jelenséget!

3. feladat – Simítás a nyelvi modellezésben

a) Az alapértelemezett Good-Turing simítási eljárás helyett használjon Add-1 simítást és ismételje meg az előző feladat c) pontjában ismertetett feladatot. Ábrázolja közös grafikonon az eredeti Good-Turing és az Add-1 simítás perplexitását a fokszám függvényében. Ezt a feladatot elegendő a magyar nyelvű korpuszon elvégezni!

b) Add-1 simítás helyett most alkalmazzon Kneser-Ney simítást (interpolációval) és ezt vesse össze az alapértelmezett Good-Turing simítás hatékonyságával. Szintén elegendő a magyar korpuszon elvégezni!

4. feladat – OOV modellezés

a) Futtassa le újra egy tetszőleg nyelvi modell kiértékelését, azonban futtatás előtt kapcsolja be a részletes kiértékelés funkciót (-debug 2). Vizsgálja meg, hogy hagyományos nyelvi modell milyen valószínűséget társít a szótáron kívüli szavakhoz (<unk>)!

b) Készítsen OOV modellezést tartalmazó nyelvi modellt a magyar tanítószöveg alapján! Az ismeretlen szavakat modellező <unk> szimbólumot az egyszer előforduló szavak helyére helyettesítsük be. A szótár hány százaléka fordul elő egyszer a szövegben? Hogyan változik a tesztszövegben található OOV szavakhoz rendelt valószínűség?

5. feladat – Szöveggenerálás

A következő két feladatban a Krúdy Gyula és William Shakespeare műveiből összeállított nyelvi modelleken felül használhat egy időjárás-jelentések alapján tanított nyelvi modellt is, melyet a „blm" könyvtárban „meteo_10gram.blm" néven talál.

a) Generáljon mondatokat tetszőlegesen a magyar vagy angol nyelvű nyelvi modellek segítségével. Hogyan változnak a generált mondatok a nyelvi modell fokszám növekedésével?

b) Generáljon mondat befejezéseket tetszőlegesen megadott prefix és modell alapján. A „prefixes" könyvtárban talál két szöveges fájlt, mely ötletet ad a kísérletezéshez.

6. feladat – Kiegészítő feladatok

a) Rajzolja fel az angol és magyar nyelvű korpusz Zipf-görbéjét! Milyen hasonlóságok és különbségek figyelhetők meg?

b) A „train_txt" könyvtárban talál egy „shakespeare_decap_list.txt" nevű szöveges fájlt is, melyet a Shakespeare tanítószöveg mondatkezdő pozícióba került közneveinek kisbetűsítéséhez használtunk. Mit gondol, hogyan állítható elő egy ilyen lista? Javasoljon módszereket!