Johdatus luonnollisen kielen käsittelyyn

15.1.2018. Vastaa kaikkiin kolmeen tehtävään.

1. Selitä lyhyesti seuraavat termit: (10 p)
   a. Regularisointi
   b. Käänteinen dokumenttifrekvenssi
   c. IOB-malli
   d. Ekstraktiivinen tiivistäminen
   e. Sanojen semanttiset vektoriesitykset
   f. Morfologinen analyysi ja generointi
   g. Säännölliset lausekkeet
   h. Nimetty entiteetti
   i. Mielipidesanasto
   j. Web crawler

2. Kielimallit (10p)
   a. Mikä kielimallien tarkoitus on ja millaisissa sovelluksissa niitä voidaan hyödyntää?
   b. Mitä tarkoitetaan n-grammeilla? Miten ne liittyvät kielimalleihin?
   c. Miten kielimallien käyttäytyminen tekstin generoinnissa muuttuu, kun n:n arvoa kasvatetaan? Mistä tämä johtuu?
   d. Add-one-siloittelussa (smoothing) kaikkiin n-grammilukumääriin lisätään yksi. Millainen ongelma sillä pyritään ratkaisemaan?
   e. Miten kielimallien hyvyyttä voidaan arvioida?
   f. Sanojen semantiikkaa voidaan tarkastella word2vec-mallilla, joka koulutusvaiheessa pyrkii ennustamaan tarkasteltavan sanan kontekstissa esiintyviä sanoja. Ts. malli saa syötteenä tarkasteltavan sanan ja pyrkii ennustamaan yhden kontekstissa esiintyvän sanan kerrallaan. Miten muokkaisit kyseistä menetelmää, jotta se toimisi kielimallina?

3. Konekääntäminen (10 p)

a) Miten määrittelisit konekääntämisen? Miten se eroaa tietokoneavusteisesta kääntämisestä?

b) Minkälaista koulutusdataa tavallisesti käytetään konekääntämisjärjestelmän opettamiseen? Mitä koulutusdataa on olemassa esimerkiksi Suomi-Englanti -kieliparille?

c) Konekäännösjärjestelmää voidaan arvioida automaattisesti käyttämällä esimerkiksi BLEU-mittaa. Mihin tämä mitta perustuu ja mitä (dataa) täytyy olla, jotta voimme käyttää tätä arviota? Mitä mahdollisia ongelmia BLEU-mitassa on?

d) Neuroverkkomalleihin perustuva konekääntämisongelma voidaan esittää yleisessä muodossa 'muokataan input-sekvenssi joksikin muuksi output-sekvenssiksi'. Mitä input- ja output-sekvensseillä tarkoitetaan konekääntämisen yhteydessä? Mistä eri tekstin yksiköistä (pienistä tekstin osista) sekvenssit voivat koostua konekääntämisessä?

e) Mitä muita luonnollisen kielen käsittelyn ongelmia voi muokata yleisen konekääntämisongelman muotoon? Anna ainakin yksi esimerkki. Miten tässä tapauksessa määrittelisit input- ja output-sekvenssit ja mistä yksiköistä sekvenssit tällöin koostuvat?