Ekstrakcija likova iz kratkih priča

Gorana Levačić Tomislav Horina

***Sažetak*—U ovom radu smo radili klasifikator za likove u kratkim pričama. Koristili smo kratke priče s Gutenberg projecta.Priče smostalno označavali jer postojeća riješenja ne prepoznaju na primjer kralja kao lika.Promatrali smo ponašanje klasifikanirao na 4 načina zadavanja tesnog skupa.Vrste su :**

1) **bez ikakve promjene**
2) **priče bez interpukcijski znakova**
3) **sve riječi u lowercaseu**
4) **sve riječi u lowercaseu i bez interpukcijski znakova**

Isprobali smo algoritme klasifikacije koji su preporučeni za izradu tog modela – skriveni Markovljev model(eng. Hidden Markov Model, HMM) i uvjetno slučajno polje(eng. Conditional Random Fields, CRF) te smo Stanford Ner učili na naše ozanke. (sada tu treba ici nesto rezzultatima to treba nadopisati.)

I. Uvod

U ovom projektu bavit ćemo se ekstrahiranjem likova iz kratkih priča, konkretno priča za djecu. Taj problem pripada problemu ekstrakcije, odnosno identifikacije entiteta u tekstu, poznatiji pod engleskim nazivom named-entity recognition (NER). NER je podvrsta zadaće crpljenja obavijesti (information extraction), u kojoj se svakom elementu teksta pridjeljuje neki atribut. U općem slučaju imamo više atributa, na primjer osoba, lokacija, vrijeme, iznos novca i drugi, te više riječi može činiti entitet kojem se pridjeljuje jedan atribut. Jasnije je iz sljedećeg primjera:
Jim bought 300 shares of Acme Corp. in 2006.
$|Jim|_{person}$ bought 300 shares of $|AcmeCorp.|_{organization}$ in $|2006|_{time}$.
U našem slučaju imamo samo jedan atributa, lik, koji određuje tko su sve likovi u priči.

II. Podaci

Tekstove priča smo skidali s stranice Project Gutenberg. Potom smo samostalno označavali likove. To smo radili pošto postojeća riješenja ne prepoznaju, na primjer kralj, kraljica, princ i tako dalje, kao likove. Svakoj riječi smo dodljeljivali klasu O ,što označava "$Other$", i klasu C ,što označava "$Charachter$".

U našim modelima gledat ćemo dali na njega utječu veličine slova i postojanje interpunkcijski znakova, to jest imat ćemo četri vrste tesnog skupa:

1) bez ikakve promjene
2) priče bez interpukcijski znakova
3) sve riječi u lowercaseu
4) sve riječi u lowercaseu i bez interpukcijski znakova

III. Opis korištenih meotda

Koristili smo algoritme klasifikacije koji su preporučeni za izradu tog modela :

1) Skriveni Markovljev model(eng. Hidden Markov Model, HMM)
2) Uvjetno slučajno polje(eng. Conditional Random Fields, CRF)

Dodatno smo učili Stanford Ner na naše ozanke.

A. Skriveni Markovljev model

Skriveni Markovljev model (HMM) prvog reda jest skup slučajnih varijabli koji se sastoji od dva podskupa, Q i O :

- $Q = Q_1, \ldots, Q_N$ – skup slučajnih varijabli koje poprimaju diskretne vrijednosti
- $O = O_1, \ldots, O_N$ – skup slučajnih varijabli koje poprimaju diskretne ili kontinuirane vrijednosti.

Te varijable zadovoljajvaju sljedeće uvjete:

1) $P(Q_t|Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) = P(Q_t|Q_{t-1})$ (1)
2) $P(O_t|Q_T, O_T, \ldots, Q_{t+1}, O_{t+1}, Q_t, Q_{t-1}, O_{t-1}, \ldots Q_1, O_1) = P(O_t|Q_t)$ (2)

Relacija (1) kaže da je vjerojatnost da se, za neko $t \in \{1, 2, \ldots, N\}$, nalazimo u stanju Q_t uz uvjet da su se dogodila prethodna stanja $Q_1, \ldots, Q_{t-1}$ i da su emitirani simboli $O_1, \ldots, O_{t-1}$ jednaka tranzicijskoj vjerojatnosti iz stanja Q_{t-1} u stanje Q_t.

Relacija (2) povlači da realizacija nekog opažanja u sadašnjem stanju ovisi samo o tom stanju . Vjerojatnosti iz relacije (2) nazivamo emisijske vjerojatnosti i kažemo da stanje Q_t emitira simbol O_t.

Skriveni Markovljev model zadan je sljedećim parametrima:

- N – broj stanja u kojima se proces može nalaziti, $S = \{1, \ldots, N\}$, S – kup svih stanja procesa
- M – broj mogućih opažanja ,$B = \{b_1, \ldots, b_M\}$, B – skup svih opaženih vrijednosti
- L – duljina opaženog niza , $X = (x_1, \ldots, x_L)$, X – opaženi niz
- A – matrica tranziciskih vjerojatnosti , $A = \{a_{ij}\}, a_{ij} = P(Q_{t+1} = j|Q_t = i), 1 \leq i, j \leq N$
- E – matrica emisijskih vjerojatnosti $E\{e_j(k), e_j(k) = P(O_t = b_k|Q_t = j), 1 \leq j \leq N, 1 \leq k \leq M$

B. Uvjetno slučajno polje

Uvjetna slučajna polja (CFR) je diskriminativni vjerojatnosni model strojnog učenja za struktuiranu predikciju koja je temeljena na modelima neusmjerenih grafova. Neka je dan vektor $x = \{x_1, x_2, \ldots, x_T\}$ gdje je T dužina sekvence, a svaki x_i predstavlja vektor karakteristika podataka na poziciji i. Za svaki od tih vektora porebno je odrediti oznaku (tag) y_i. U slučaju određevianja vrste riječi, T bi bio broj riječima u tekstu, y_i vrsta riječi na poziciji i , a za svaki x_i bi bio

vektor informacija o i-toj riječi kojima se raspolaže sama riječ, informacije o prefiksima, sufiksima i veličini slova ....

Neka su x i Y slučajne varijable, $w = \{w_k\}$ realni vektor parametara i $F = \{f_k(y_t y_{t-q}, x_t)\}_{k=1}$ skup realnih karakterističnih funkcija. Linearni uvjeta slučajna polja definiraju se uvjetnom raspodjelom

$$P(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp \left[\sum_{k=1}^{K} w_k f_k(y_t, y_{t-q}, x_t) \right]$$

pri čemu je

$$Z(x) = \sum_{y} \prod_{t=1}^{T} exp \left[\sum_{k=1}^{K} w_k f_k(y_t, y_{t-q}, x_t) \right]$$

Karakteristična funkcija može se promatrati proizvoljne karakteristike sekvence , tako da se vektor x_t, zapravo može zamijneiti cjelokupnom sekvencom opservacija x. Linearna uvjetna slučajna polja omogućavaju promatranje samo dva uzastopna taga, zbog čega se model može posmatrati kao lanac. Linarna uvjetna slučajna polja mogu se definirati i preko grafa nad skupom čvorova $U = X \cup Y$ kao

$$P(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \Psi_t(y_t, y_{t-1}, x_t)$$

$$\Psi_t(y_t, y_{t-1}, x_t) = exp \left[\sum_{k=1}^{K} w_k f_k(y_t, y_{t-q}, x_t) \right]$$

Neka je G faktor graf nad slučajnim varijablama X i y. (X, Y) je uvjentno slučajno polje ako se za neke vrijednosti x varijabli X, uvjetna vjerojatnost $P(y|x)$ faktorira prema faktoru G. Svako uvjetno slučajno polje se faktoria prema nekom faktoru grafa G. Dati faktor graf G i njegov skup faktora $\{\Psi_a\}$ definiraju raspodjelu

$$P(y|x) = \frac{1}{Z(x)} \prod_{i=1}^{T} \Psi_a(x_a, y_a)$$

gdje je

$$\Psi_a(x_a, y_a) = exp \left[\sum_{k=1}^{K} w_{ak} f_{ak}(x_a, y_a) \right]$$

gdje u indeksiranju karakterisinih funkcija i odgovarajućih parametara sudjeluje i oznaka faktora a jer svaki fakotr može da ima svoj skup parametara. Faktor Ψ_a zavisi od unije nekih podskupova $X_a \subseteq X$ i $Y_a \subseteq Y$. Uslučaju linearnih uvjetnih slučajnih polja, skup Y_a je mogao da sadrži samo dva susjedna taga. Skup svih parametara može se podjeliti na klase $C = \{C_1, C_2, \ldots, C_p\}$ gdje svaka klasa C_p koristi isti skup karakterističnih funkcija $\{f_{pk} 8 x_c, y_c)\}_{k=1,\ldots,K(p)}$ i vektor parametara w_{kp} veličine $K(p)$. Raspodjela vjerojatnosti se može napisati kao

$$P(y|x) = \frac{1}{Z(x)} \prod_{i=1}^{T} \Psi_a(x_a, y_a)$$

$$\Psi_c(x_c, y_c; w_p) = exp \left[\sum_{k=1}^{K} w_{pk} f_{pk}(x_c, y_c) \right]$$

C. Stanford Ner

IV. Rezultati

hmm nevalja

Literatura

[1] M. Rudman, *Kompleksnost skrivenih Markovljevih modela*, (diplomski rad). Prirodoslovno-matematiki fakultet, 2014.

[2] I. Medic, *Uslovna slučajna polja*, (master rad). Univerzitet u Beogradu, Matematički fakultet, 2013