Ekstrakcija likova iz kratkih priča
(Projektni prijedlog)

Tomislav Horina
Gorana Levačić

Zagreb, 2016.

Sadržaj

1 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.
> $[\text{Jim}]_{\text{person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{organization}}$ in $[\text{2006}]_{\text{time}}$.

U našem slučaju imamo samo jedan atribut, *lik*, koji određuje tko su sve likovi u priči.

Skup podataka čine kratke priče prikupljene s Project Gutenberga, kao što su bajke Hansa Christiana Andersena.

2 Cilj istraživanja problema

Cilj našega istraživanja je odrediti uspješnost nekoliko metoda za rješavanje opisanog problema. Usporedit ćemo te metode međusobno, kao i s već postojećim sustavima za named-entity recognition. Većina postojećih sustava je generalizirana za više atributa, od kojih će nama samo jedan biti bitan (*osoba*). S obzirom da su neki od tih sustava godinama razvijani na sveučilištima ili u većim tvrtkama, očekujemo lošiji rezultat u odnosu njih.

3 Pregled dosadašnjih istraživanja

Istraživanje named-entity recognition problema je započelo u 90im godinama prošloga stoljeća. Prvotno su izvori podataka bili novinski članci, dok su danas to često podaci vezani uz bioinformatiku, molekularnu biologiju i medicinu.

Većina dosadašnjih istraživanja se temelji na nadziranom učenju. S obzirom na veći broj atributa u općenitom slučaju, potrebno je prikupiti velike količine označenog teksta iz kojeg će klasifikator učiti. Zbog toga se u novije vrijeme prelazi na učenje podrškom.

Metode koje se koriste za rješavanje NER problema su:

- **uvjetna slučajna polja** (eng. *conditional random fields, CRF*)
- **skriveni Markovljevi modeli** (eng. *hidden Markov models, HMM*)
- **potporni vektorski strojevi** (eng. *support vector machines, SVM*)

- **stabla odlučivanja** (eng. *decision trees*)
- **model maksimalne entropije** (eng. *maximum entropy, ME*)
- **konvolucijske neuralne mreže** (eng. *convolutional neural networks, CNN*)

Uvjetna slučajna polja i skriveni Markovljevi modeli su ipak najčešće metode. Na uvjetnim slučajnim poljima se temelji i **Stanford NER**, koji prepoznaje tri klase: *PERSON, ORGANIZATION, LOCATION*. Sveučilište u Sheffieldu je razvilo GATE, General Architecture for Text Engineering, u sklopu kojeg se nalazi **ANNIE** (A Nearly-New Information Extraction System). Također se koriste i hibridni pristupi poput uvjetnih neuralnih polja (conditional neural fields). Neuralne mreže ipak još uvijek postižu lošije rezultate u odnosu na npr. uvjetna slučajna polja. Konvolucijske neuralne mreže su najbolji izbor za NER.

Napomenimo još da neke od dosadašnjih metoda postižu i do 94% uspješnosti za određene tekstove, što je veoma blizu ljudskoj uspješnosti od 97%. Za neke druge tekstove, odnosno područja, je uspješnost dosta niža.

4 Materijali, metodologija i plan istraživanja

Kao što smo već rekli, izvor podataka će biti kratke priče na engleskom jeziku prikupljene s web stranice Project Gutenberg. Na stranici se mogu pronaći bajke Hansa Christiana Andersena, različite narodne pripovijesti (slavenske, germanske itd.), kratke priče za djecu i odrasle različitih autora, te mnoge druge knjige i pripovijesti.

S obzirom da nam je temeljni pristup nadzirano učenje, dio prikupljenih priča će činiti skup za učenje, te ćemo u tim pričama označiti likove.

Koristit ćemo dvije metode: **skrivene Markovljeve modele** i **konvolucijske neuralne mreže**. Obje metode ćemo implementirati u Pythonu pomoću biblioteka **NLTK** (*Natural Language Toolkit*) i **PyBrain** (*Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network Library*).

4.1 Skriveni Markovljevi modeli

Skriveni Markovljev model (HMM) je statistički model sa sljedećim svojstvima:

- **Markovljevo svojstvo** sustava znači da buduća stanja sustava ovise samo o trenutnom stanju, ne o prethodnim stanjima
- **skrivenost** znači da stanje sustava nije izravno vidljivo, već je vidljiv konačni rezultat

HMM možemo interpretirati kao nedeterministički konačni automat s vjerojatnostima pridruženim svakom prijelazu. Vjerojatnost nekog niza stanja

$Y = y(0)y(1)\dots y(L-1)$ duljine L se može izračunati pomoću

$$P(Y) = \sum_X P(Y|X)P(X$$

pri čemu je $X = x(0)x(1)\dots x(L-1)$ neki niz duljine L koji uključuje skrivene čvorove.

Cilj je maksimizirati $P(Y)$, što se učinkovito radi pomoću **Viterbijevog algoritma**. Taj algoritam za cilj ima naći najvjerojatniji niz skrivenih stanja, odnosno **Viterbijev put**.

4.2 Konvolucijske neuralne mreže

Konvolucijske neuralne mreže (CNN) su vrsta neuralnih mreža koje daju izlaz ovisno o kontekstu. Ulaz takve mreže čine takozvani prozori, odnosno riječ s okolinom. Na primjer za rečenicu *Jim bought 300 shares of Acme Corp. in 2006.* imamo:

1. neuron: *Jim bought 300*
2. neuron: *bought 300 shares*
3. neuron: *300 shares of*
4. neuron: *shares of Acme*
5. neuron: *of Acme Corp.*
6. neuron: *Acme Corp. in*
7. neuron: *Corp. in 2006.*

Za rješavanje problema NER pomoću neuralnih mreža je potrebno transformirati ulazne podatke (riječi) u oblik koji će biti prikladni za rad neuronske mreže. Konkretno, riječi će biti prezentirane vektorima, što se može postići na više načina. Jedna od češće korištenih metoda je Googleov **Word2Vec**, neuralna mreža koja uči značenja riječi te za svaku riječ kao izlaz daje vektor jednake duljine.

4.3 Ocjena uspješnosti

Kao ocjenu uspješnosti metoda koristit ćemo **F-mjeru**. Ta mjera se definira na sljedeći način:

$$F = 2 * \frac{P \cdot R}{P + R}$$

pri čemu su P i R oznake za redom **preciznost** (eng. *precision*) i **odziv** (eng. *recall*):

$$P = \frac{\text{točno označeni likovi}}{\text{svi označeni likovi}}$$

$$R = \frac{\text{točno označeni likovi}}{\text{svi likovi u tekstu}}$$

Drugim riječima, preciznost govori koliko pojmova je sustav ispravno označio (npr. sustav može označiti neku lokaciju kao lika), a odziv govori koliko likova je sustav uopće uočio (npr. sustav može preskočiti neke sporedne likove).

5 Očekivani rezultati

Očekujemo da će naš softver relativno uspješno prepoznavati likove u kratkim pričama. Pritom očekujemo da će skriveni Markovljevi modeli dati bolje rezultate u odnosu na neuralnu mrežu. U oba slučaja očekujemo lošije rezultate u odnosu na već postojeće sustave poput Stanford NER-a.

Literatura

1. Deep Learning for Natural Language Processing (nastavni materijali), Stanford University, `http://cs224d.stanford.edu/syllabus.html`

2. Jurafsky D., *Information Extraction and Named Entity Recognition* (nastavni materijal), Stanford University, `https://web.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.pdf`

3. *Named-entity recognition*, Wikipedia: The Free Encyclopedia, Wikimedia Foundation, Inc., `https://en.wikipedia.org/wiki/Named-entity_recognition` (zadnja izmjena: 20. siječnja 2016.)

4. Collobert et al.: *Natural language processing (almost) from scratch* (2011.). `http://jmlr.org/papers/volume12/collobert11a/collobert11a.pdf`

5. Nadeau D., Sekine S.: *A survey of named entity recognition and classification.* `http://nlp.cs.nyu.edu/sekine/papers/li07.pdf`

6. Yuan, E.: *Named-Entity Recognition using Deep Learning* (2015.), `http://eric-yuan.me/ner_1/`