Upotreba jezičnih tehnologija u digitalizaciji teksta i njegovoj daljnjoj obradi

Nikola Ljubešić

http://nlp.ffzg.hr
Odsjek za informacijske i komunikacijske znanosti
Filozofski fakultet, Sveučilište u Zagrebu

Četvrti festival hrvatskih digitalizacijskih projekata
10. travnja 2014.

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Motivacija

- pozvan prikazati rezultate rada u prepoznavanju naziva
- razvijamo niz drugih jezičnih tehnologija koje također mogu biti vrlo korisne u digitalizaciji i obogaćivanju teksta
- većina jezičnih tehnologija danas počiva na strojnom učenju
- strojno učenje dobiva izrazito na važnosti unutar svih područja
- konačna tema:
  - prikaz jezičnih tehnologija korisnih u digitalizaciji i kasnijoj obradi tekstnih podataka
  - pod kišobranom principa strojnog učenja

Što su to jezične tehnologije?

- tehnologije koje omogućuju obradu prirodnog jezika

1. govoreni jezik
   - sinteza govora – *text to speech*
   - analiza govora – *speech to text*
2. pisani jezik
   - pravopisni i gramatički provjernici
   - ...
   - sustavi za pretraživanje informacija / prikupljanje, obradu i prikaz znanja

1. jednojezične
   - ...
2. višejezične
   - strojno prevođenje
   - višejezično pretraživanje informacija

Što su to jezične tehnologije?

- tehnologije koje omogućuju obradu prirodnog jezika

1. govoreni jezik
   - sinteza govora – *text to speech*
   - analiza govora – *speech to text*
2. pisani jezik
   - pravopisni i gramatički provjernici
   - ...
   - sustavi za pretraživanje informacija / prikupljanje, obradu i prikaz znanja

1. jednojezične
   - ...
2. višejezične
   - strojno prevođenje
   - višejezično pretraživanje informacija

Što su to jezične tehnologije?

- tehnologije koje omogućuju obradu prirodnog jezika

1. govoreni jezik
   - sinteza govora – *text to speech*
   - analiza govora – *speech to text*
2. pisani jezik
   - pravopisni i gramatički provjernici
   - ...
   - sustavi za pretraživanje informacija / prikupljanje, obradu i prikaz znanja

1. jednojezične
   - ...
2. višejezične
   - strojno prevođenje
   - višejezično pretraživanje informacija

Što su to jezične tehnologije?

- tehnologije koje omogućuju obradu prirodnog jezika

1. govoreni jezik
   - sinteza govora – *text to speech*
   - analiza govora – *speech to text*
2. pisani jezik
   - pravopisni i gramatički provjernici
   - ...
   - sustavi za pretraživanje informacija / prikupljanje, obradu i prikaz znanja

1. jednojezične
   - ...
2. višejezične
   - strojno prevođenje
   - višejezično pretraživanje informacija

Što je to strojno učenje?

- revolucija u dolasku!
- disciplina iz područja umjetne inteligencije koje razvija algoritme koji omogućuju računalima da uče na temelju dostupnih podataka
- vrlo blisko području statističkog modeliranja – na temelju dostupnih podataka gradi se statistički model koji omogućuje predviđanje
- klasični primjeri – filter za neželjenu e-poštu i optičko prepoznavanje znakova
- u slučaju da računalo uči na jezičnim podacima bliska su područja
  - računalna lingvistika
  - obrada prirodnog jezika

Što je to strojno učenje?

- revolucija u dolasku!
- disciplina iz područja umjetne inteligencije koje razvija algoritme koji omogućuju računalima da uče na temelju dostupnih podataka
- vrlo blisko području statističkog modeliranja – na temelju dostupnih podataka gradi se statistički model koji omogućuje predviđanje
- klasični primjeri – filter za neželjenu e-poštu i optičko prepoznavanje znakova
- u slučaju da računalo uči na jezičnim podacima bliska su područja
  - računalna lingvistika
  - obrada prirodnog jezika

Što je to strojno učenje?

- revolucija u dolasku!
- disciplina iz područja umjetne inteligencije koje razvija algoritme koji omogućuju računalima da uče na temelju dostupnih podataka
- vrlo blisko području statističkog modeliranja – na temelju dostupnih podataka gradi se statistički model koji omogućuje predviđanje
- klasični primjeri – filter za neželjenu e-poštu i optičko prepoznavanje znakova
- u slučaju da računalo uči na jezičnim podacima bliska su područja
  - računalna lingvistika
  - obrada prirodnog jezika

Što je to strojno učenje?

- revolucija u dolasku!
- disciplina iz područja umjetne inteligencije koje razvija algoritme koji omogućuju računalima da uče na temelju dostupnih podataka
- vrlo blisko području statističkog modeliranja – na temelju dostupnih podataka gradi se statistički model koji omogućuje predviđanje
- klasični primjeri – filter za neželjenu e-poštu i optičko prepoznavanje znakova
- u slučaju da računalo uči na jezičnim podacima bliska su područja
  - računalna lingvistika
  - obrada prirodnog jezika

Što je to strojno učenje?

- revolucija u dolasku!
- disciplina iz područja umjetne inteligencije koje razvija algoritme koji omogućuju računalima da uče na temelju dostupnih podataka
- vrlo blisko području statističkog modeliranja – na temelju dostupnih podataka gradi se statistički model koji omogućuje predviđanje
- klasični primjeri – filter za neželjenu e-poštu i optičko prepoznavanje znakova
- u slučaju da računalo uči na jezičnim podacima bliska su područja
  - računalna lingvistika
  - obrada prirodnog jezika

Što je to strojno učenje?

- revolucija u dolasku!
- disciplina iz područja umjetne inteligencije koje razvija algoritme koji omogućuju računalima da uče na temelju dostupnih podataka
- vrlo blisko području statističkog modeliranja – na temelju dostupnih podataka gradi se statistički model koji omogućuje predviđanje
- klasični primjeri – filter za neželjenu e-poštu i optičko prepoznavanje znakova
- u slučaju da računalo uči na jezičnim podacima bliska su područja
  - računalna lingvistika
  - obrada prirodnog jezika

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- izgradnja statističkog modela vjerojatnosti slijeda riječi (ili znakova, gramatičkih kategorija) na temelju dostupnih podataka
- potrebna nam je samo (velika) količina tekstova reprezentativnih za uzorak koji obrađujemo
- služi procjeni vjerojatnosti nekog neviđenog slijeda
- najčešće se koristi kada imamo više hipoteza te želimo odabrati najvjerojatniju
  - analiza govora
    P(I saw a van) >> P(eyes awe of an)
  - pravopisni ispravnik
    P(Sutra predajem rad.)>P(Sutra predajem tad.)
  - strojno prevođenje
    P(Puhat će jak vjetar) >> P(Puhat će čvrst vjetar)
  - prepoznavanje znakova
    P(On nije mjerodavan)>P(On nije rnjerodavan)

Jezično modeliranje

- velike količine tekstnih podataka? – mreža
- mrežni korpusi se grade puzanjem vršnih internetskih domena te obradom prikupljenih podataka
  - prepoznavanje kodiranja
  - crpljenje tekstnog dijela dokumenta
  - uklanjanje (bliskih) duplikata
  - prepoznavanje jezika
- hrvatski mrežni korpus hrWaC u verziji 2.0 sadrži 2 milijarde pojavnica
- `http://nlp.ffzg.hr/resources/corpora/hrwac/`
- CC-BY-SA
- mrežni korpusi susjednih jezika
  - bsWaC 429 milijuna pojavnica
  - slWaC 525 milijuna pojavnica
  - srWaC 894 milijuna pojavnica

Jezično modeliranje

- velike količine tekstnih podataka? – mreža
- mrežni korpusi se grade puzanjem vršnih internetskih domena te obradom prikupljenih podataka
  - prepoznavanje kodiranja
  - crpljenje tekstnog dijela dokumenta
  - uklanjanje (bliskih) duplikata
  - prepoznavanje jezika
- hrvatski mrežni korpus hrWaC u verziji 2.0 sadrži 2 milijarde pojavnica
- `http://nlp.ffzg.hr/resources/corpora/hrwac/`
- CC-BY-SA
- mrežni korpusi susjednih jezika
  - bsWaC 429 milijuna pojavnica
  - slWaC 525 milijuna pojavnica
  - srWaC 894 milijuna pojavnica

Jezično modeliranje

- velike količine tekstnih podataka? – mreža
- mrežni korpusi se grade puzanjem vršnih internetskih domena te obradom prikupljenih podataka
  - prepoznavanje kodiranja
  - crpljenje tekstnog dijela dokumenta
  - uklanjanje (bliskih) duplikata
  - prepoznavanje jezika
- hrvatski mrežni korpus hrWaC u verziji 2.0 sadrži 2 milijarde pojavnica
- `http://nlp.ffzg.hr/resources/corpora/hrwac/`
- CC-BY-SA
- mrežni korpusi susjednih jezika
  - bsWaC 429 milijuna pojavnica
  - slWaC 525 milijuna pojavnica
  - srWaC 894 milijuna pojavnica

Jezično modeliranje

- velike količine tekstnih podataka? – mreža
- mrežni korpusi se grade puzanjem vršnih internetskih domena te obradom prikupljenih podataka
  - prepoznavanje kodiranja
  - crpljenje tekstnog dijela dokumenta
  - uklanjanje (bliskih) duplikata
  - prepoznavanje jezika
- hrvatski mrežni korpus hrWaC u verziji 2.0 sadrži 2 milijarde pojavnica
- `http://nlp.ffzg.hr/resources/corpora/hrwac/`
- CC-BY-SA
- mrežni korpusi susjednih jezika
  - bsWaC 429 milijuna pojavnica
  - slWaC 525 milijuna pojavnica
  - srWaC 894 milijuna pojavnica

Jezično modeliranje

- velike količine tekstnih podataka? – mreža
- mrežni korpusi se grade puzanjem vršnih internetskih domena te obradom prikupljenih podataka
  - prepoznavanje kodiranja
  - crpljenje tekstnog dijela dokumenta
  - uklanjanje (bliskih) duplikata
  - prepoznavanje jezika
- hrvatski mrežni korpus hrWaC u verziji 2.0 sadrži 2 milijarde pojavnica
- `http://nlp.ffzg.hr/resources/corpora/hrwac/`
- CC-BY-SA
- mrežni korpusi susjednih jezika
  - bsWaC 429 milijuna pojavnica
  - slWaC 525 milijuna pojavnica
  - srWaC 894 milijuna pojavnica

Nadzirano strojno učenje

- kod jezičnog modeliranja koristimo samo (velike) količine podataka te računalo uči vjerojatnost slijeda događaja
- problem prepoznavanja neželjene pošte – računalo će vrlo teško (nenadzirano) naučiti što je neželjena pošta, već želimo nadzirati postupak učenja pružanjem pozitivnih i negativnih primjera
- nadzirano učenje – skup podataka na kojemu se uči je označen vrijednošću koju želimo moći predvidjeti (poruka je željena / neželjena)
- terminologija
  - instance – primjeri na kojima učimo
  - značajke – varijable pomoću kojih pokušavamo predvidjeti traženu varijablu
  - zavisna varijabla – varijabla koju pokušavamo predvidjeti

Nadzirano strojno učenje

- kod jezičnog modeliranja koristimo samo (velike) količine podataka te računalo uči vjerojatnost slijeda događaja
- problem prepoznavanja neželjene pošte – računalo će vrlo teško (nenadzirano) naučiti što je neželjena pošta, već želimo nadzirati postupak učenja pružanjem pozitivnih i negativnih primjera
- nadzirano učenje – skup podataka na kojemu se uči je označen vrijednošću koju želimo moći predvidjeti (poruka je željena / neželjena)
- terminologija
  - instance – primjeri na kojima učimo
  - značajke – varijable pomoću kojih pokušavamo predvidjeti traženu varijablu
  - zavisna varijabla – varijabla koju pokušavamo predvidjeti

Nadzirano strojno učenje

- kod jezičnog modeliranja koristimo samo (velike) količine podataka te računalo uči vjerojatnost slijeda događaja
- problem prepoznavanja neželjene pošte – računalo će vrlo teško (nenadzirano) naučiti što je neželjena pošta, već želimo nadzirati postupak učenja pružanjem pozitivnih i negativnih primjera
- nadzirano učenje – skup podataka na kojemu se uči je označen vrijednošću koju želimo moći predvidjeti (poruka je željena / neželjena)
- terminologija
  - instance – primjeri na kojima učimo
  - značajke – varijable pomoću kojih pokušavamo predvidjeti traženu varijablu
  - zavisna varijabla – varijabla koju pokušavamo predvidjeti

Nadzirano strojno učenje

- kod jezičnog modeliranja koristimo samo (velike) količine podataka te računalo uči vjerojatnost slijeda događaja
- problem prepoznavanja neželjene pošte – računalo će vrlo teško (nenadzirano) naučiti što je neželjena pošta, već želimo nadzirati postupak učenja pružanjem pozitivnih i negativnih primjera
- nadzirano učenje – skup podataka na kojemu se uči je označen vrijednošću koju želimo moći predvidjeti (poruka je željena / neželjena)
- terminologija
  - instance – primjeri na kojima učimo
  - značajke – varijable pomoću kojih pokušavamo predvidjeti traženu varijablu
  - zavisna varijabla – varijabla koju pokušavamo predvidjeti

Nadzirano strojno učenje

- kod jezičnog modeliranja koristimo samo (velike) količine podataka te računalo uči vjerojatnost slijeda događaja
- problem prepoznavanja neželjene pošte – računalo će vrlo teško (nenadzirano) naučiti što je neželjena pošta, već želimo nadzirati postupak učenja pružanjem pozitivnih i negativnih primjera
- nadzirano učenje – skup podataka na kojemu se uči je označen vrijednošću koju želimo moći predvidjeti (poruka je željena / neželjena)
- terminologija
  - instance – primjeri na kojima učimo
  - značajke – varijable pomoću kojih pokušavamo predvidjeti traženu varijablu
  - zavisna varijabla – varijabla koju pokušavamo predvidjeti

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

- Naporno sam radio.
- instance – riječi (u kontekstu)
- zavisna varijabla – morfosintaktička kategorija svake riječi
- problem klasifikacije u 630 različitih klasa!
- značajke
  - riječ
  - nastavak riječi duljine $1..m$
  - riječ $n-1$
  - riječ $n-2$
- skup podataka za učenje – korpus (kolekcija tekstova) u kojemu je svaka riječ ručno označena odgovarajućom morfosintaktičkom kategorijom

Morfosintaktičko označavanje

Festival	Ncmsn
hrvatskih	Agpmpg
digitalizacijskih	Agpmpg
projekata	Ncmpg
okuplja	Vmr3s
predstavnike	Ncmpa
arhivske	Agpfsg
,	Z
knjižnične	Agpfsg
i	Cc
muzejske	Agpfsg
zajednice	Ncfsg

- trenutni je skup za učenje veličine 130k riječi, MSD točnost 85%, POS točnost 95%
- http://nlp.ffzg.hr/resources/models/tagging/
- CC-BY-SA

Morfosintaktičko označavanje

Festival	Ncmsn
hrvatskih	Agpmpg
digitalizacijskih	Agpmpg
projekata	Ncmpg
okuplja	Vmr3s
predstavnike	Ncmpa
arhivske	Agpfsg
,	Z
knjižnične	Agpfsg
i	Cc
muzejske	Agpfsg
zajednice	Ncfsg

- trenutni je skup za učenje veličine 130k riječi, MSD točnost 85%, POS točnost 95%
- http://nlp.ffzg.hr/resources/models/tagging/
- CC-BY-SA

Morfosintaktičko označavanje

Festival	Ncmsn
hrvatskih	Agpmpg
digitalizacijskih	Agpmpg
projekata	Ncmpg
okuplja	Vmr3s
predstavnike	Ncmpa
arhivske	Agpfsg
,	Z
knjižnične	Agpfsg
i	Cc
muzejske	Agpfsg
zajednice	Ncfsg

- trenutni je skup za učenje veličine 130k riječi, MSD točnost 85%, POS točnost 95%
- http://nlp.ffzg.hr/resources/models/tagging/
- CC-BY-SA

Morfosintaktičko označavanje

Festival	Ncmsn
hrvatskih	Agpmpg
digitalizacijskih	Agpmpg
projekata	Ncmpg
okuplja	Vmr3s
predstavnike	Ncmpa
arhivske	Agpfsg
,	Z
knjižnične	Agpfsg
i	Cc
muzejske	Agpfsg
zajednice	Ncfsg

- trenutni je skup za učenje veličine 130k riječi, MSD točnost 85%, POS točnost 95%
- http://nlp.ffzg.hr/resources/models/tagging/
- CC-BY-SA

Lematizacija

- instance – riječi
- zavisna varijabla – kanonski (osnovni) oblik svake riječi
- značajke – nastavak riječi duljine $1..m$, morfosintaktička kategorija

Festival	Ncmsn	festival
hrvatskih	Agpmpg	hrvatski
digitalizacijskih	Agpmpg	digitalizacijski
projekata	Ncmpg	projekt
okuplja	Vmr3s	okupljati
predstavnike	Ncmpa	predstavnik
arhivske	Agpfsg	arhivski
,	Z	,
knjižnične	Agpfsg	knjižničan
i	Cc	i
muzejske	Agpfsg	muzejski
zajednice	Ncfsg	zajednica

- točnost 97%, CC-BY-SA

Lematizacija

- instance – riječi
- zavisna varijabla – kanonski (osnovni) oblik svake riječi
- značajke – nastavak riječi duljine $1..m$, morfosintaktička kategorija

Festival	Ncmsn	festival
hrvatskih	Agpmpg	hrvatski
digitalizacijskih	Agpmpg	digitalizacijski
projekata	Ncmpg	projekt
okuplja	Vmr3s	okupljati
predstavnike	Ncmpa	predstavnik
arhivske	Agpfsg	arhivski
,	Z	,
knjižnične	Agpfsg	knjižničan
i	Cc	i
muzejske	Agpfsg	muzejski
zajednice	Ncfsg	zajednica

- točnost 97%, CC-BY-SA

Lematizacija

- instance – riječi
- zavisna varijabla – kanonski (osnovni) oblik svake riječi
- značajke – nastavak riječi duljine $1..m$, morfosintaktička kategorija

Festival	Ncmsn	festival
hrvatskih	Agpmpg	hrvatski
digitalizacijskih	Agpmpg	digitalizacijski
projekata	Ncmpg	projekt
okuplja	Vmr3s	okupljati
predstavnike	Ncmpa	predstavnik
arhivske	Agpfsg	arhivski
,	Z	,
knjižnične	Agpfsg	knjižničan
i	Cc	i
muzejske	Agpfsg	muzejski
zajednice	Ncfsg	zajednica

- točnost 97%, CC-BY-SA

Lematizacija

- instance – riječi
- zavisna varijabla – kanonski (osnovni) oblik svake riječi
- značajke – nastavak riječi duljine $1..m$, morfosintaktička kategorija

Festival	Ncmsn	festival
hrvatskih	Agpmpg	hrvatski
digitalizacijskih	Agpmpg	digitalizacijski
projekata	Ncmpg	projekt
okuplja	Vmr3s	okupljati
predstavnike	Ncmpa	predstavnik
arhivske	Agpfsg	arhivski
,	Z	,
knjižnične	Agpfsg	knjižničan
i	Cc	i
muzejske	Agpfsg	muzejski
zajednice	Ncfsg	zajednica

- točnost 97%, CC-BY-SA

Lematizacija

- instance – riječi
- zavisna varijabla – kanonski (osnovni) oblik svake riječi
- značajke – nastavak riječi duljine $1..m$, morfosintaktička kategorija

Festival	Ncmsn	festival
hrvatskih	Agpmpg	hrvatski
digitalizacijskih	Agpmpg	digitalizacijski
projekata	Ncmpg	projekt
okuplja	Vmr3s	okupljati
predstavnike	Ncmpa	predstavnik
arhivske	Agpfsg	arhivski
,	Z	,
knjižnične	Agpfsg	knjižničan
i	Cc	i
muzejske	Agpfsg	muzejski
zajednice	Ncfsg	zajednica

- točnost 97%, CC-BY-SA

Lematizacija

- instance – riječi
- zavisna varijabla – kanonski (osnovni) oblik svake riječi
- značajke – nastavak riječi duljine $1..m$, morfosintaktička kategorija

Festival	Ncmsn	festival
hrvatskih	Agpmpg	hrvatski
digitalizacijskih	Agpmpg	digitalizacijski
projekata	Ncmpg	projekt
okuplja	Vmr3s	okupljati
predstavnike	Ncmpa	predstavnik
arhivske	Agpfsg	arhivski
,	Z	,
knjižnične	Agpfsg	knjižničan
i	Cc	i
muzejske	Agpfsg	muzejski
zajednice	Ncfsg	zajednica

- točnost 97%, CC-BY-SA

Generalizacija

- "Pladna duska brči po tumi" – rečenica koja samo "zvuči" hrvatski

```
Pladna    Agpfsn   pladan
duska     Ncfsn    duska
brči      Vmr3s    brčiti
po        Sl       po
tumi      Ncfsl    tuma
.         Z        .
```

- zahvaljujući generalizaciji, strojno učenje / statističko modeliranje može donositi odluke i na "neviđenim" podacima

Generalizacija

- "Pladna duska brči po tumi" – rečenica koja samo "zvuči" hrvatski

Pladna	Agpfsn	pladan
duska	Ncfsn	duska
brči	Vmr3s	brčiti
po	Sl	po
tumi	Ncfsl	tuma
.	Z	.

- zahvaljujući generalizaciji, strojno učenje / statističko modeliranje može donositi odluke i na "neviđenim" podacima

Generalizacija

- "Pladna duska brči po tumi" – rečenica koja samo "zvuči" hrvatski

Pladna	Agpfsn	pladan
duska	Ncfsn	duska
brči	Vmr3s	brčiti
po	Sl	po
tumi	Ncfsl	tuma
.	Z	.

- zahvaljujući generalizaciji, strojno učenje / statističko modeliranje može donositi odluke i na "neviđenim" podacima

Generalizacija

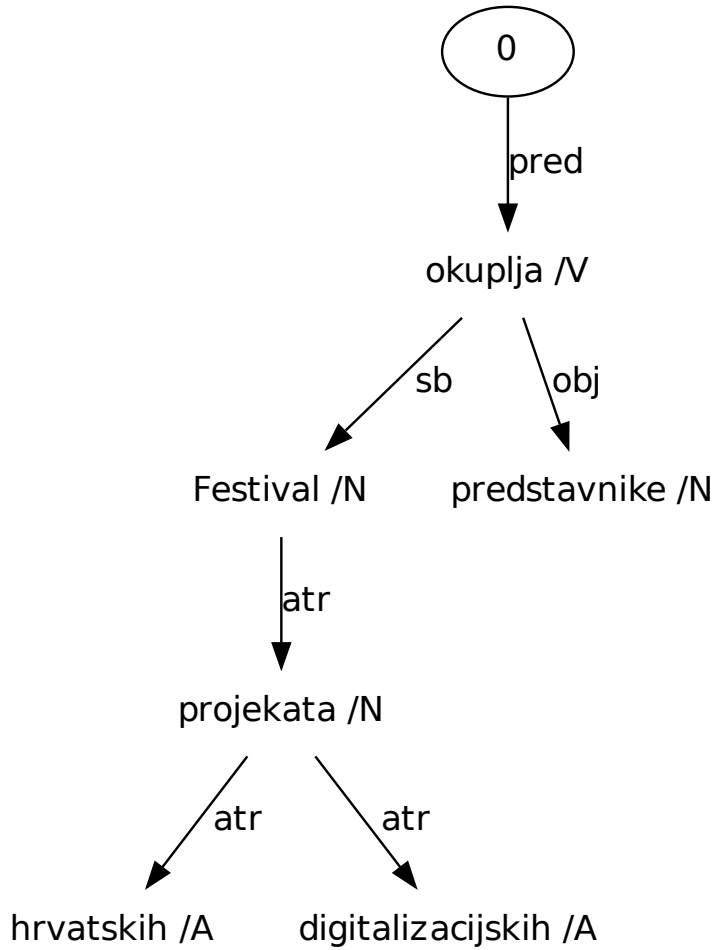
- "Pladna duska brči po tumi" – rečenica koja samo "zvuči" hrvatski

```
Pladna    Agpfsn   pladan
duska     Ncfsn    duska
brči      Vmr3s    brčiti
po        Sl       po
tumi      Ncfsl    tuma
.         Z        .
```

- zahvaljujući generalizaciji, strojno učenje / statističko modeliranje može donositi odluke i na "neviđenim" podacima

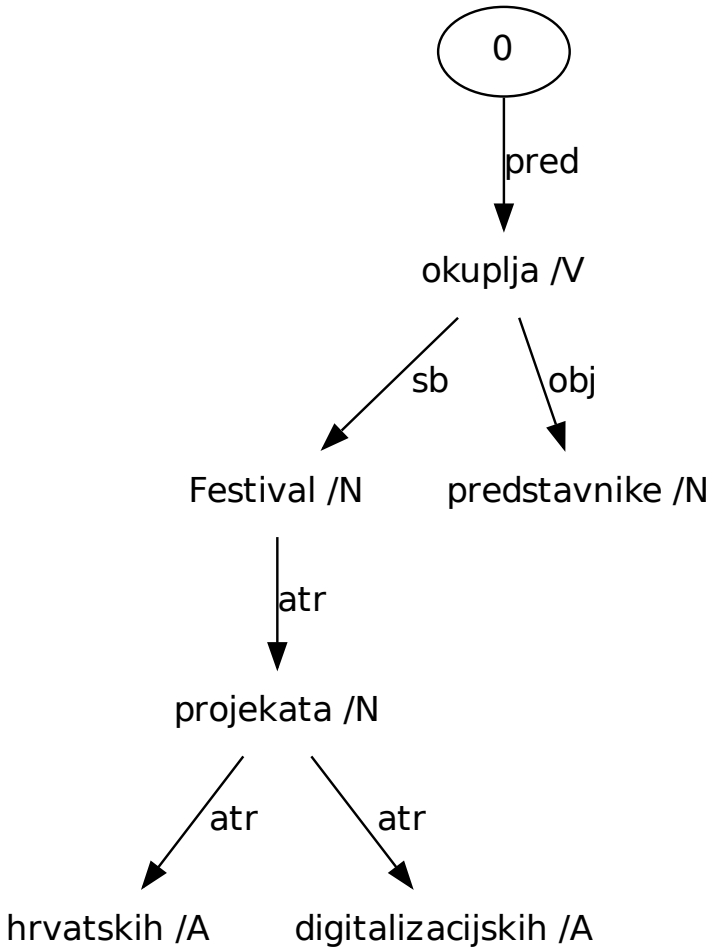
Ovisnosno sintaktičko označavanje

- instance – riječi
- zavisna varijabla – glava i vrsta veze
- značajke – površinski oblik, lema, morfosintaktičke kategorije, riječi u okolini...
- točnost 80%
- `http://nlp.ffzg.hr/resources/models/dependency-parsing/`
- CC-BY-SA



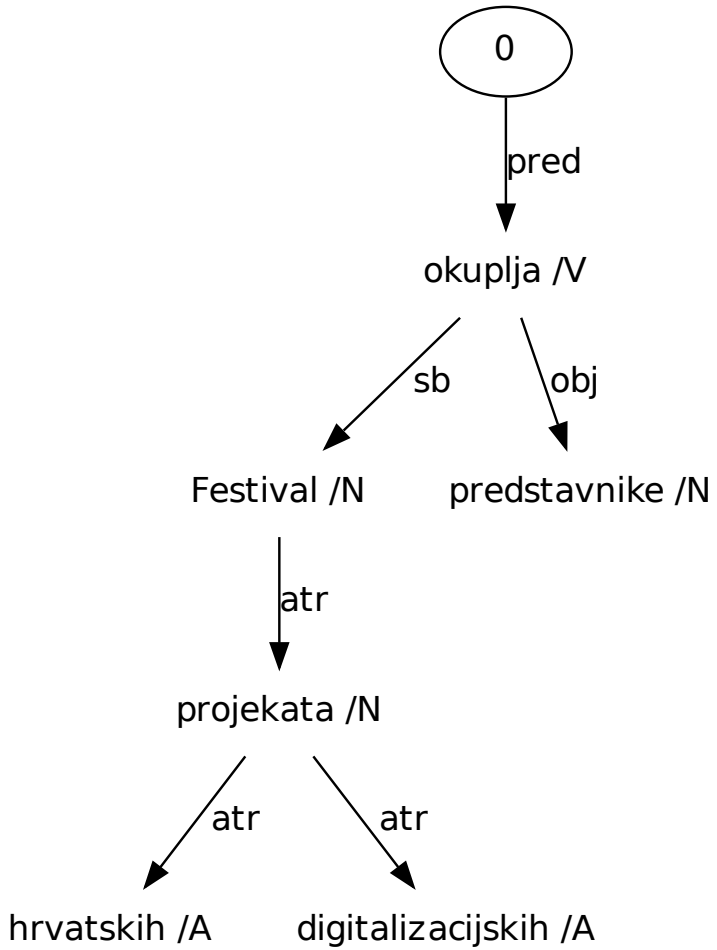
Ovisnosno sintaktičko označavanje

- instance – riječi
- zavisna varijabla – glava i vrsta veze
- značajke – površinski oblik, lema, morfosintaktičke kategorije, riječi u okolini...
- točnost 80%
- `http://nlp.ffzg.hr/resources/models/dependency-parsing/`
- CC-BY-SA



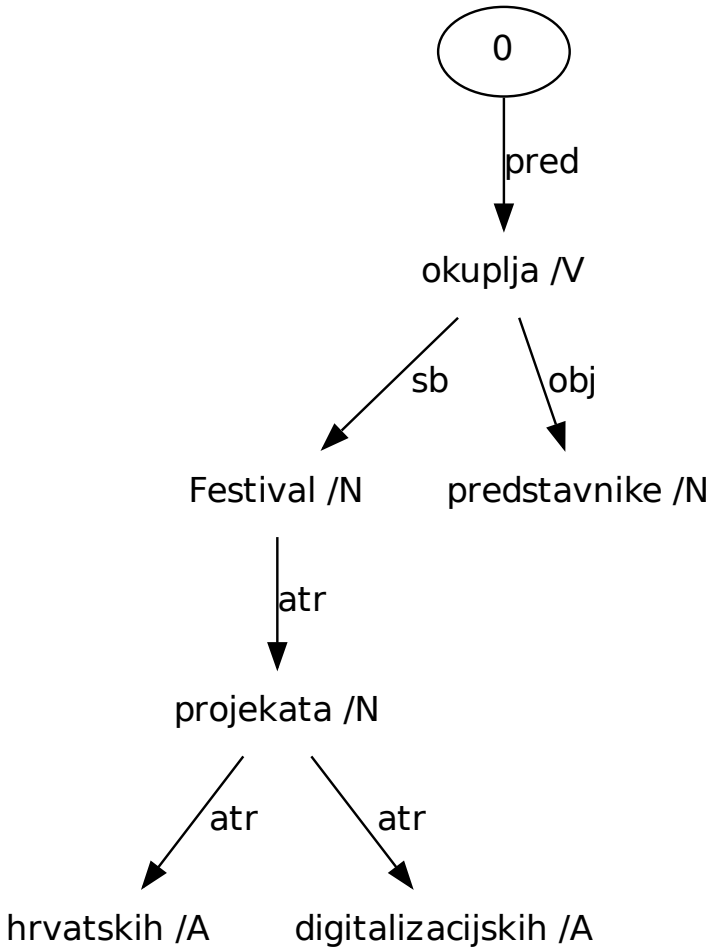
Ovisnosno sintaktičko označavanje

- instance – riječi
- zavisna varijabla – glava i vrsta veze
- značajke – površinski oblik, lema, morfosintaktičke kategorije, riječi u okolini...
- točnost 80%
- `http://nlp.ffzg.hr/resources/models/dependency-parsing/`
- CC-BY-SA



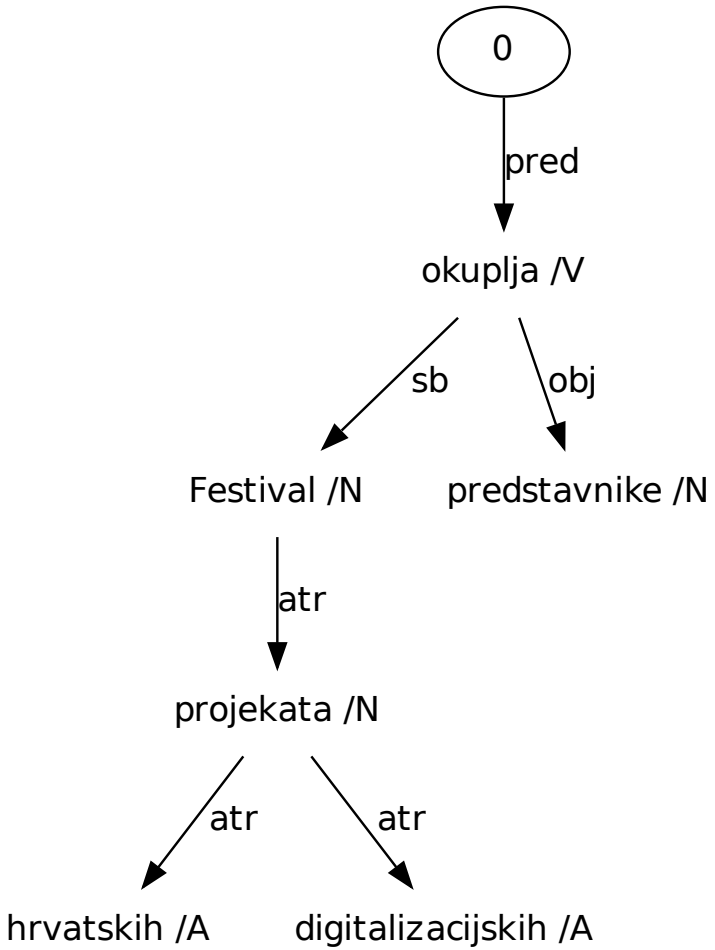
Ovisnosno sintaktičko označavanje

- instance – riječi
- zavisna varijabla – glava i vrsta veze
- značajke – površinski oblik, lema, morfosintaktičke kategorije, riječi u okolini...
- točnost 80%
- http://nlp.ffzg.hr/resources/models/dependency-parsing/
- CC-BY-SA



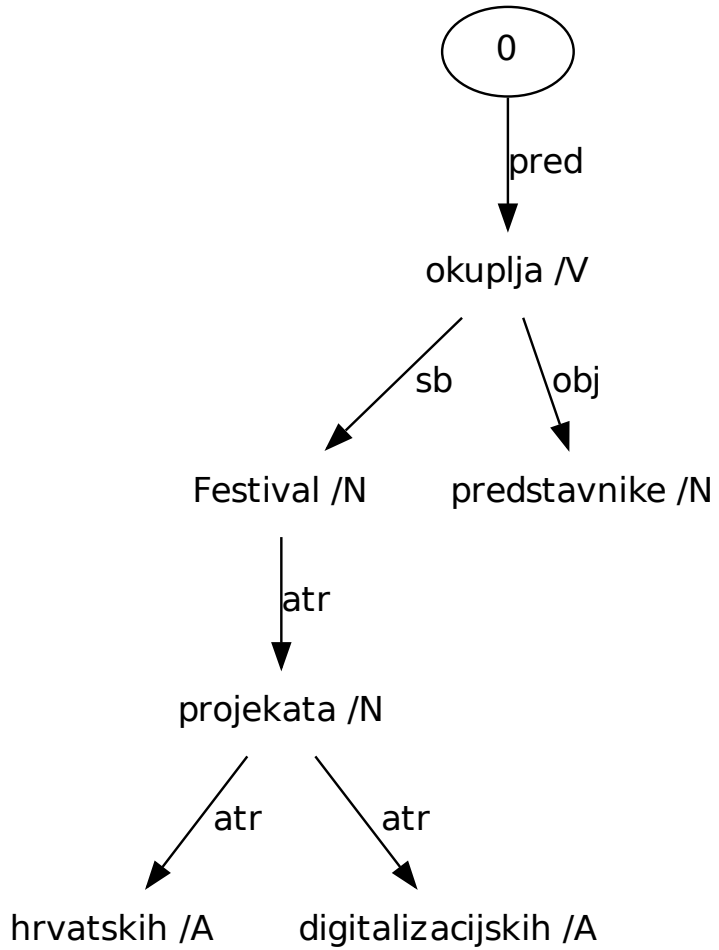
Ovisnosno sintaktičko označavanje

- instance – riječi
- zavisna varijabla – glava i vrsta veze
- značajke – površinski oblik, lema, morfosintaktičke kategorije, riječi u okolini...
- točnost 80%
- http://nlp.ffzg.hr/resources/models/dependency-parsing/
- CC-BY-SA



Ovisnosno sintaktičko označavanje

- instance – riječi
- zavisna varijabla – glava i vrsta veze
- značajke – površinski oblik, lema, morfosintaktičke kategorije, riječi u okolini...
- točnost 80%
- `http://nlp.ffzg.hr/resources/models/dependency-parsing/`
- CC-BY-SA



Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- instance – riječi
- zavisna varijabla – vrsta naziva
- domena zavisne varijable
  - O
  - PERS
  - LOC
  - ORG
  - MISC
  - NUMEX i TIMEX
- značajke
  - riječ
  - počinje li riječ velikim slovom
  - trigrami znakova riječi
  - riječ $n-1$
  - riječ $n+1$
  - vrsta riječi
  - kontekstna sličnost s poznatim nazivom
  - ...

Prepoznavanje naziva

- model izgrađen na tri osnovne kategorije PERS, ORG i LOC

```
<PERS>Nada Mihaljević</PERS>, spisateljica i
    djelatnica <ORG>Nacionalne i sveučilišne knjiž
    nice u Zagrebu</ORG>, dobitnica je Nagrade
    Grigor Vitez za 2013. godinu.
Nagrada Grigor Vitez najstarija je književna
    nagrada u <LOC>Hrvatskoj</LOC>, a njezinu
    dodjelu ove godine poduprli su <ORG>
    Ministarstvo znanosti, obrazovanja i sporta</
    ORG> i <ORG>Grad Zagreb</ORG>. Od 2005. godine
     Nagrada se ostvaruje u suradnji sa Zagrebač
    kim kazalištem lutaka.
```

- F1 91%
- http://nlp.ffzg.hr/resources/models/ner/
- CC-BY-SA

Prepoznavanje naziva

- model izgrađen na tri osnovne kategorije PERS, ORG i LOC

```
<PERS>Nada Mihaljević</PERS>, spisateljica i
    djelatnica <ORG>Nacionalne i sveučilišne knjiž
    nice u Zagrebu</ORG>, dobitnica je Nagrade
    Grigor Vitez za 2013. godinu.
Nagrada Grigor Vitez najstarija je književna
    nagrada u <LOC>Hrvatskoj</LOC>, a njezinu
    dodjelu ove godine poduprli su <ORG>
    Ministarstvo znanosti, obrazovanja i sporta</
    ORG> i <ORG>Grad Zagreb</ORG>. Od 2005. godine
     Nagrada se ostvaruje u suradnji sa Zagrebač
    kim kazalištem lutaka.
```

- F1 91%
- http://nlp.ffzg.hr/resources/models/ner/
- CC-BY-SA

Prepoznavanje naziva

- model izgrađen na tri osnovne kategorije PERS, ORG i LOC

```
<PERS>Nada Mihaljević</PERS>, spisateljica i
    djelatnica <ORG>Nacionalne i sveučilišne knjiž
    nice u Zagrebu</ORG>, dobitnica je Nagrade
    Grigor Vitez za 2013. godinu.
Nagrada Grigor Vitez najstarija je književna
    nagrada u <LOC>Hrvatskoj</LOC>, a njezinu
    dodjelu ove godine poduprli su <ORG>
    Ministarstvo znanosti, obrazovanja i sporta</
    ORG> i <ORG>Grad Zagreb</ORG>. Od 2005. godine
     Nagrada se ostvaruje u suradnji sa Zagrebač
    kim kazalištem lutaka.
```

- F1 91%
- http://nlp.ffzg.hr/resources/models/ner/
- CC-BY-SA

Prepoznavanje naziva

- model izgrađen na tri osnovne kategorije PERS, ORG i LOC

```
<PERS>Nada Mihaljević</PERS>, spisateljica i
    djelatnica <ORG>Nacionalne i sveučilišne knjiž
    nice u Zagrebu</ORG>, dobitnica je Nagrade
    Grigor Vitez za 2013. godinu.
Nagrada Grigor Vitez najstarija je književna
    nagrada u <LOC>Hrvatskoj</LOC>, a njezinu
    dodjelu ove godine poduprli su <ORG>
    Ministarstvo znanosti, obrazovanja i sporta</
    ORG> i <ORG>Grad Zagreb</ORG>. Od 2005. godine
     Nagrada se ostvaruje u suradnji sa Zagrebač
    kim kazalištem lutaka.
```

- F1 91%
- http://nlp.ffzg.hr/resources/models/ner/
- CC-BY-SA

Strojno prevođenje za standardizaciju jezika

- proces prevođenja se uči iz dostupnih prijevoda – biteksta
- osnovni elementi procesa
  - prijevodni model iz ishodišnog u ciljni jezik
    P("I was"|"Ja sam")=0.83
  - jezični model ciljnog jezika
- statističko strojno prevođenje na razini znakova
  P("cz"|"c")=0.27
  - prevođenje između bliskih jezika i narječja
  - prevođenje s nestandardnog na standardni jezik
  - prevođenje povijesnog jezika na suvremeni

sm	sem	dervesiza	drevesca
tolk	toliko	kapliz	kapljic
skor	skoraj	leshala	ležala
nč	nič	drushbo	družbo
nism	nisem	stergalu	strgalu
rečt	reči	hzherko	hčerko
dobr	dobro	bojy	boji

Strojno prevođenje za standardizaciju jezika

- proces prevođenja se uči iz dostupnih prijevoda – biteksta
- osnovni elementi procesa
  - prijevodni model iz ishodišnog u ciljni jezik
    P("I was"|"Ja sam")=0.83
  - jezični model ciljnog jezika
- statističko strojno prevođenje na razini znakova
  P("cz"|"c")=0.27
  - prevođenje između bliskih jezika i narječja
  - prevođenje s nestandardnog na standardni jezik
  - prevođenje povijesnog jezika na suvremeni

sm	sem	dervesiza	drevesca
tolk	toliko	kapliz	kapljic
skor	skoraj	leshala	ležala
nč	nič	drushbo	družbo
nism	nisem	stergalu	strgalu
rečt	reči	hzherko	hčerko
dobr	dobro	bojy	boji

Strojno prevođenje za standardizaciju jezika

- proces prevođenja se uči iz dostupnih prijevoda – biteksta
- osnovni elementi procesa
  - prijevodni model iz ishodišnog u ciljni jezik
    P("I was"|"Ja sam")=0.83
  - jezični model ciljnog jezika
- statističko strojno prevođenje na razini znakova
  P("cz"|"c")=0.27
  - prevođenje između bliskih jezika i narječja
  - prevođenje s nestandardnog na standardni jezik
  - prevođenje povijesnog jezika na suvremeni

sm	sem	dervesiza	drevesca
tolk	toliko	kapliz	kapljic
skor	skoraj	leshala	ležala
nč	nič	drushbo	družbo
nism	nisem	stergalu	strgalu
rečt	reči	hzherko	hčerko
dobr	dobro	bojy	boji

Strojno prevođenje za standardizaciju jezika

- proces prevođenja se uči iz dostupnih prijevoda – biteksta
- osnovni elementi procesa
  - prijevodni model iz ishodišnog u ciljni jezik
    P("I was"|"Ja sam")=0.83
  - jezični model ciljnog jezika
- statističko strojno prevođenje na razini znakova
  P("cz"|"c")=0.27
  - prevođenje između bliskih jezika i narječja
  - prevođenje s nestandardnog na standardni jezik
  - prevođenje povijesnog jezika na suvremeni

sm	sem	dervesiza	drevesca
tolk	toliko	kapliz	kapljic
skor	skoraj	leshala	ležala
nč	nič	drushbo	družbo
nism	nisem	stergalu	strgalu
rečt	reči	hzherko	hčerko
dobr	dobro	bojy	boji

Strojno prevođenje za standardizaciju jezika

- proces prevođenja se uči iz dostupnih prijevoda – biteksta
- osnovni elementi procesa
  - prijevodni model iz ishodišnog u ciljni jezik
    P("I was"|"Ja sam")=0.83
  - jezični model ciljnog jezika
- statističko strojno prevođenje na razini znakova
  P("cz"|"c")=0.27
  - prevođenje između bliskih jezika i narječja
  - prevođenje s nestandardnog na standardni jezik
  - prevođenje povijesnog jezika na suvremeni

sm	sem	dervesiza	drevesca
tolk	toliko	kapliz	kapljic
skor	skoraj	leshala	ležala
nč	nič	drushbo	družbo
nism	nisem	stergalu	strgalu
rečt	reči	hzherko	hčerko
dobr	dobro	bojy	boji

Strojno prevođenje za standardizaciju jezika

- proces prevođenja se uči iz dostupnih prijevoda – biteksta
- osnovni elementi procesa
  - prijevodni model iz ishodišnog u ciljni jezik
    P("I was"|"Ja sam")=0.83
  - jezični model ciljnog jezika
- statističko strojno prevođenje na razini znakova
  P("cz"|"c")=0.27
  - prevođenje između bliskih jezika i narječja
  - prevođenje s nestandardnog na standardni jezik
  - prevođenje povijesnog jezika na suvremeni

sm	sem	dervesiza	drevesca
tolk	toliko	kapliz	kapljic
skor	skoraj	leshala	ležala
nč	nič	drushbo	družbo
nism	nisem	stergalu	strgalu
rečt	reči	hzherko	hčerko
dobr	dobro	bojy	boji

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- http://nl.ijs.si/imp/index-en.html

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- http://nl.ijs.si/imp/index-en.html

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- http://nl.ijs.si/imp/index-en.html

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- http://nl.ijs.si/imp/index-en.html

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- http://nl.ijs.si/imp/index-en.html

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- http://nl.ijs.si/imp/index-en.html

Strojno prevođenje za standardizaciju jezika

- standardizacija slovenskih Twitter poruka (Ljubešić et al. 2014)

```
dons boš pa lohk gledu, a?
danes boš pa lahko gledal, a?
```

- prepolovljuje pogrešku pri lematizaciji standardiziranog teksta (točnost 84%) s obzirom na nestandardizirani (75%) te ručno standardizirani tekst (92%)

- osuvremenjavanje povijesnih tekstova (Scherrer i Erjavec, 2013)
- točnost nad tekstovima 18. stoljeća 72%, 19. stoljeća 92%
- značajno bolji rezultati nego indukcijom ili pisanjem pravila
- `http://nl.ijs.si/imp/index-en.html`

Upotreba jezičnih tehnologija u digitalizaciji teksta i njegovoj daljnjoj obradi

Nikola Ljubešić

http://nlp.ffzg.hr
Odsjek za informacijske i komunikacijske znanosti
Filozofski fakultet, Sveučilište u Zagrebu

Četvrti festival hrvatskih digitalizacijskih projekata
10. travnja 2014.