Uvod u obradu prirodnog jezika

9.1. Ekstrakcija informacija i prepoznavanje imenovanih entiteta (Information Extraction and Named Entity Recognition)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ekstrakcija informacija (IE)

- Sustavi za ekstrakciju informacija (IE)
  - pronalaženje i razumijevanje relevantnih dijelova teksta
  - skupljanje informacija iz mnogih izvora teksta
  - produkcija strukturne reprezentacije relevantnih informacija
    - relacije
    - baza znanja
  - Ciljevi:
    1. organizacija informacija tako da budu korisne ljudima
    2. postavljanje informacija u semantički preciznom obliku čime se omogućava daljnje zaključivanje uz pomoć računalnih algoritama

Ekstrakcija informacija (IE)

- IE sustavi ekstraktiraju čiste, činjenične informacije:
  - Ugrubo: Tko je učinio nešto nekome kada?
- Npr:
  - Prikupljanje zarade, profita, članova odbora, sjedišta, itd. iz izvještaja kompanije
    - Sjedišta ABC Trade d.o.o. i globalna sjedišta kombinirane ABC Trade Grupe, su locirane u Splitu, Hrvatska
    - sjedišta("ABC Trade d.o.o.", "Split Hrvatska")
  - Učenje lijek-gen interakcije iz znanstvene medicinske literature

IE na niskom nivou

- Dostupno – relativno popularno – u aplikacijama kao Apple ili Google mail i kod web indeksiranja

- Izgleda da su temeljena na regularnim izrazima i listama naziva

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:

  – odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:

  – odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:

  – odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Osoba
Datum
Lokacija
Organizacija

Prepoznavanje imenovanih entiteta (NER)

- Korištenje:
  - imenovani entiteti se mogu indeksirati, povezati, itd.
  - Sentiment se može pridružiti kompanijama ili produktima
  - Mnoge IE relacije su veze između imenovanih entiteta
  - Za odgovaranje na pitanja, odgovori su često imenovani entiteti
- Konkretno:
  - Mnoge Web stranice označavaju razne entitete, s vezama na biografiju, tematske stranice i slično
    - Reuter's OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction
  - Apple/Google/Microsoft/ … pametni prepoznavatelji za sadržaj dokumenta

Uvod u obradu prirodnog jezika

9.2. Evaluacija prepoznavanja imenovanih entiteta (Evaluation of Named Entity Recognition)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Zadatak NER prepoznavanja

- Zadatak: Predvidjeti entitete u tekstu

– govornik	O
– Ministarstva	ORG
– Vanjskih	ORG
– Poslova	ORG
– Ivan	PER
– Ivanić	PER
– rekao	O
– je	O
– Vjesniku	ORG
:	:

Standardna evaulacija je po entitetu,
ne po pojavnici (tokenu)

Zadatak NER prepoznavanja

- Zadatak: Predvidjeti entitete u tekstu

– govornik	O
– Ministarstva	ORG
– Vanjskih	ORG
– Poslova	ORG
– Ivan	PER
– Ivanić	PER
– rekao	O
– je	O
– Vjesniku	ORG
:	:

Standardna evaulacija je po entitetu,
ne po pojavnici (tokenu)

Zadatak NER prepoznavanja

- Zadatak: Predvidjeti entitete u tekstu

	točno	pogrešno
sustav detektirao	**2**	**0**
sustav nije detektirao	**1**	**0**

Preciznost: 100%

Odziv: 66%

Preciznost/Odziv/F1 za IE/NER

- Odziv i preciznost su odlične mjere za dohvat informacija (IR) i kategorizaciju teksta
- Mjera se ponaša čudno kod IE/NER kada ima **graničnih grešaka** (koje su **česte**):
  - Prva Banka za Splićane je proglasila…
- Ovim se obuhvaća i lažno pozitivne i lažno negativne vrijednosti
- Izbor ničega bi bilo bolje
- Neke druge metrike (npr. MUC bodovanje) daju djelomičan utjecaj (prema složenim pravilima)

IE na niskom nivou

- Dostupno – relativno popularno – u aplikacijama kao Apple ili Google mail i kod web indeksiranja

- Izgleda da su temeljena na regularnim izrazima i listama naziva

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:

  – odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:

  – odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Prepoznavanje imenovanih entiteta (NER)

- Važan podzadatak: pronalaženje i klasifikacija naziva u tekstu, npr:

  – odluka nezavisnog kandidata Ivana Mimača da obustavi njegovu podršku za manjinsku stranku Rada je zvučala dramatično, ali u buduće neće prijetiti stabilnosti. Kada su, nakon izbora 2010 godine, Ivan, Ante Jukić, Marija Anitovska i Milanić odlučili podržati Rad, dali su samo dvije garancije: povjerenje i opskrbu.

Osoba
Datum
Lokacija
Organizacija

Prepoznavanje imenovanih entiteta (NER)

- Korištenje:
  - imenovani entiteti se mogu indeksirati, povezati, itd.
  - Sentiment se može pridružiti kompanijama ili produktima
  - Mnoge IE relacije su veze između imenovanih entiteta
  - Za odgovaranje na pitanja, odgovori su često imenovani entiteti
- Konkretno:
  - Mnoge Web stranice označavaju razne entitete, s vezama na biografiju, tematske stranice i slično
    - Reuter's OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction
  - Apple/Google/Microsoft/ … pametni prepoznavatelji za sadržaj dokumenta

Uvod u obradu prirodnog jezika

9.3. Modeli sekvenci za prepoznavanje imenovanih entiteta (Sequence Models for Named Entity Recognition)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

NER i model sekvence iz strojnog učenja

Treniranje

1. Prikupi skup reprezentativnih dokumenata za treniranje
2. Označi svaku pojavnicu entitetskom klasom ili ostalo (O)
3. Oblikuj ekstraktore osobina prikladne za tekst i klase
4. Treniraj sekvencijski klasifikator za predviđanje oznaka iz podataka

Testiranje

1. Primi skup dokumenata za testiranje
2. Pokreni zaključivanje pomoću modela sekvence radi označavanja svake pojavnice
3. Prikladno vrati prepoznate entitete

Kodne klase za označavanje sekvence

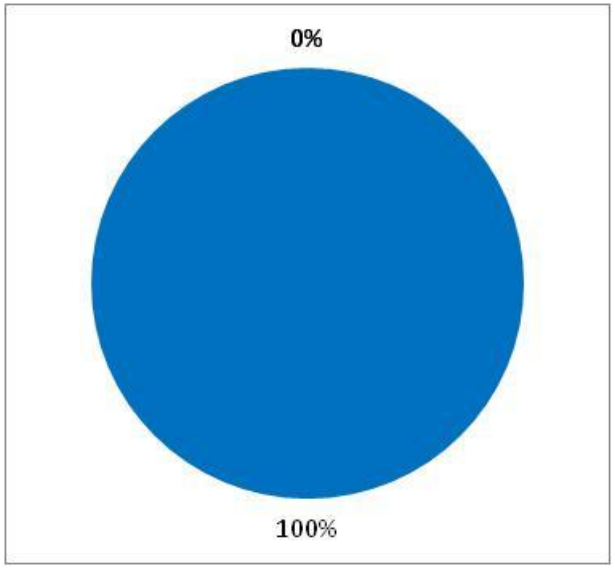
	IO kodiranje	**IOB kodiranje**
Luka	PER	B-PER
pokazuje	O	O
Sanji	PER	B-PER
Ivo	PER	B-PER
Ivičevu	PER	I-PER
novu	O	O
sliku	O	O

Osobine za označavanje kod sekvenci

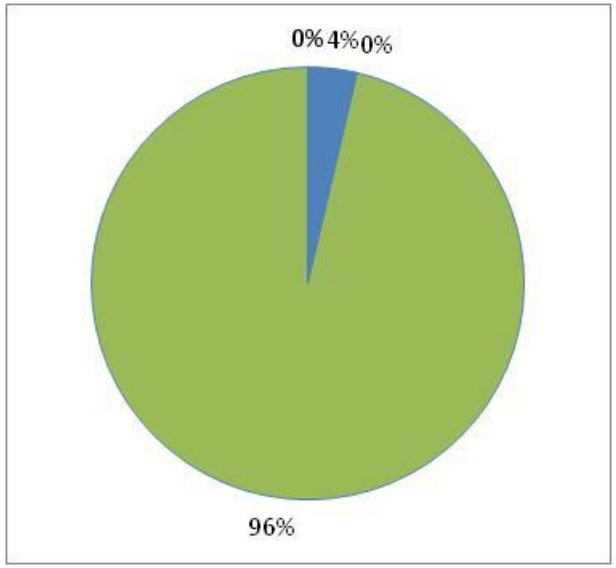
- Riječi
  - Trenutna riječ (kao naučeni rječnik)
  - Prethodna/sljedeća riječ (sadržaj)
- Druge vrste naslijeđenih lingvističkih klasifikacija
  - POS
- Sadržaj oznake
  - prethodna (i možda sljedeća) oznaka

Osobine: Podnizovi riječi

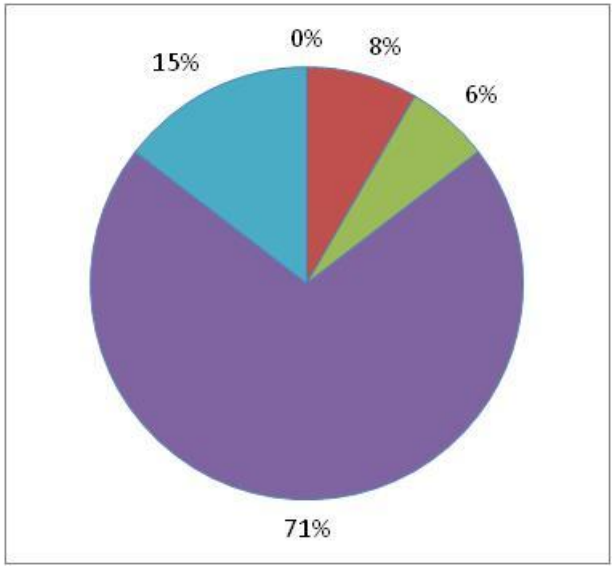
oxa



:



field



lijek
tvrtka
film
mjesto
osoba

Cotrimoxazole

Wethersfield

Rambo: First Blood

Osobine: Oblik riječi

- Oblik riječi
  - pridruživanje pojednostavljenog prikaza riječi koji kodira atribute kao što su duljina, velika/mala slova, brojevi, grčka slova, unutrašnje interpunkcije, itd.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

Uvod u obradu prirodnog jezika

9.4. Maksimalna entropija Markovljevog modela (Maximum Entropy Markov Models)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Problemi sekvenci

- Mnogi problemi OPJ imaju podatke kao sekvence znakova, riječi, fraza, linija, rečenica …
- Naš zadatak je označavanje svakog elementa sekvence

POS označavanje

N	V	C	V	A	N
Stručnjaci	navode	kako	će	metalurški	sektor

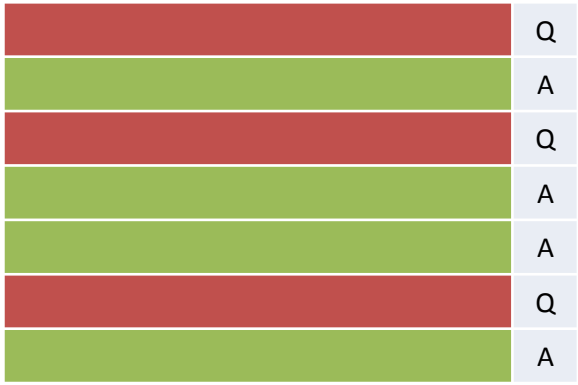
NER

PERS	O	O	O	ORG	ORG
Matić	diskutira	o	budućnosti	Fakulteta	strojarstva

Segmentacija riječi

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Segmentacija teksta



MEMM zaključivanje

- **Uvjetni Markovljev model** (Conditional Markov Model) tj. **Markovljev model maksimalne entropije** (MEMM) je klasifikator koji donosi odluku ovisno o opservacijama i prethodnim odlukama.
- Naš zadatak je označiti svaki element sekvence.

Točka odluke

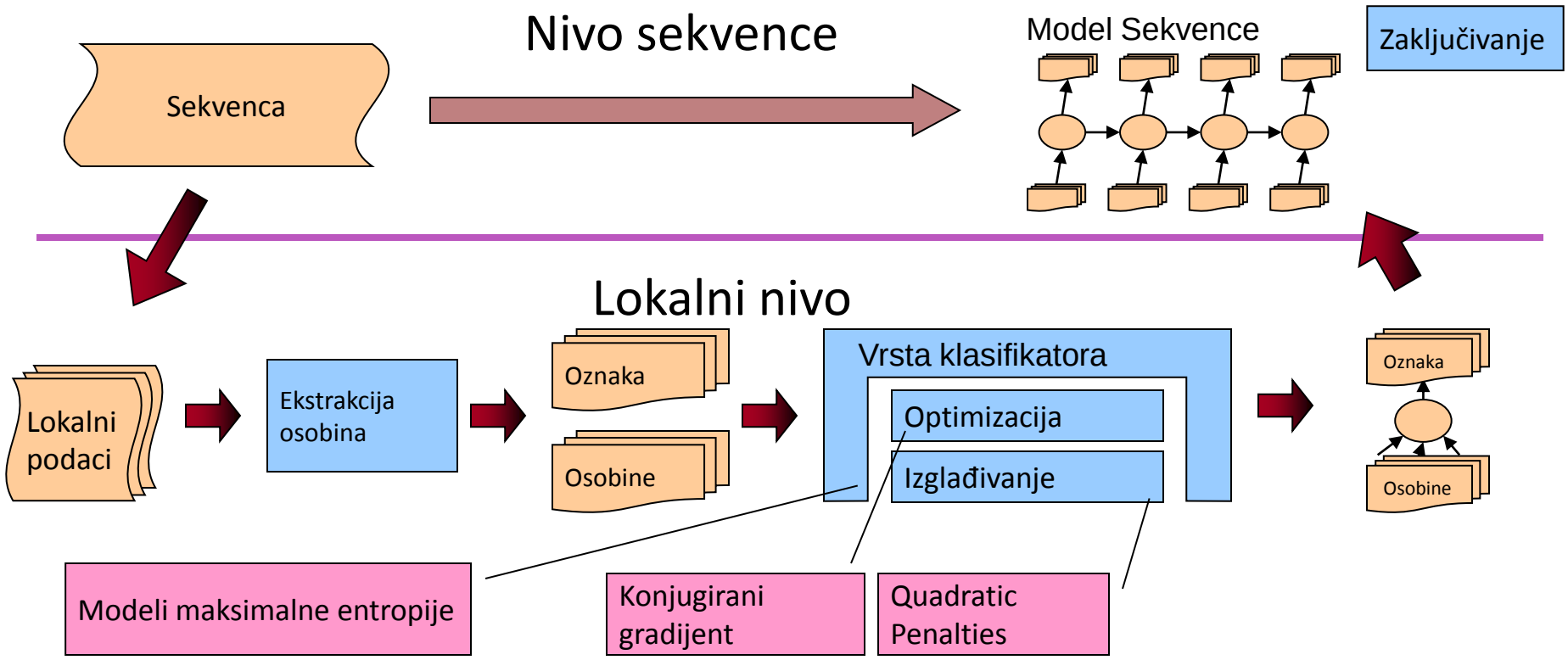
Lokalni sadržaj

-3	-2	-1	0	1
N	V	V	???	???
Dionice	su	pale	22.6	%

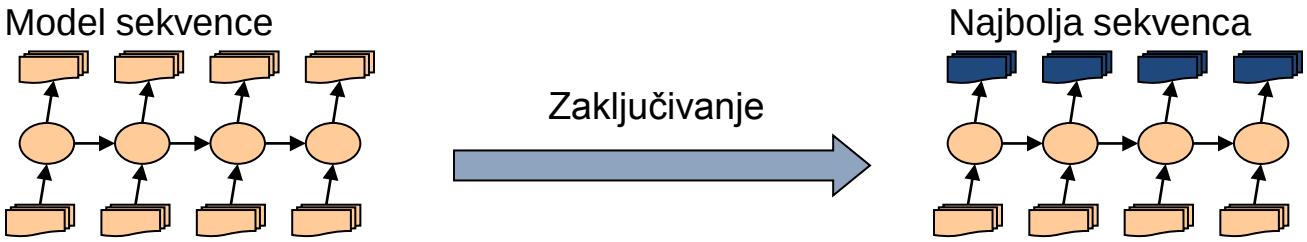
Osobine

W_0	22.6
W_{+1}	%
W_{-1}	pale
T_{-1}	V
T_{-1} T_{-2}	V V
imaBroj?	da
…	…

Sustav za zaključivanje

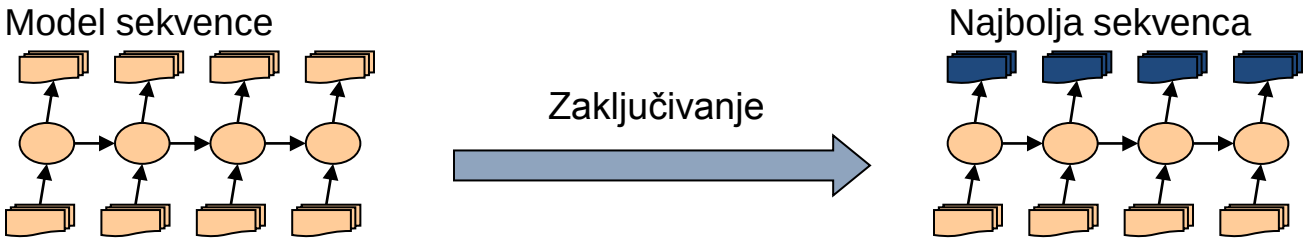


Pohlepno (greedy) zaključivanje



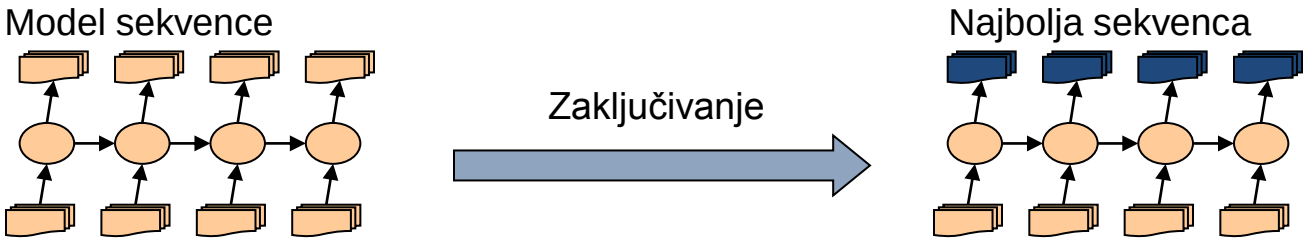
- Pohlepno zaključivanje
  - Počinjemo s lijeva i koristimo klasifikator na svakoj poziciji kako bi pridružili oznaku
  - Klasifikator može ovisiti o prethodnoj odluci kao i o promatranom podatku
- Prednosti
  - Brz, ne zahtjeva dodatnu memoriju
  - Jednostavan za implementaciju
  - Obogaćivanjem osobina tako da uključuju opservacije s desna mogu se postići dobri rezultati
- Mane
  - Pohlepan. Rade se greške od kojih se ne može oporaviti.

Zaključivanje zrakama (Beam)



- Zaključivanje zrakama
  - Na svakoj poziciji zadrži najboljih K kompletiranih sekvenci
  - Proširivanje sekvence se vrši lokalno
  - Proširivanjem s oznakom se dobiva novi skup K kompletiranih sekvenci
- Prednosti
  - Brz, zrake veličine 3-5 su u većini slučajeva dobre kao i egzaktno zaključivanje
  - Jednostavno za implementirati (ne zahtjeva dinamičko programiranje)
- Mane
  - Nije egzaktno: globalno najbolje sekvence mogu ispasti sa zrake

Viterbi zaključivanje



- Viterbi zaključivanje
  - Dinamičko programiranje ili memoizacija
  - Zahtjeva mali prozor utjecaja stanja (npr. prethodna dva stanja su relevantna)
- Prednosti
  - Egzaktan: Globalno najbolja sekvenca se dobiva
- Mane
  - Teže za implementirati duže interakcije stanja (ali zaključivanje zrakama ne dopušta duže interakcije)

Uvjetna slučajna polja

- Još jedan model sekvenci: Conditional Random Fields (CRF)
- Uvjetni model cijele sekvence u odnosu na ulančavanje lokalnih modela

$$P(\vec{c} \mid \vec{d}, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

- Prostor od C je sada prostor sekvenci
  - Ako osobine f_i ostaju lokalne, onda se uvjetna vjerodostojnost može izračunati dinamičkim programiranjem
- Treniranje je sporije, ali CRF izbjegava natjecanje pristranosti
- U praksi obično rade dobro kao i MEMM

Uvod u obradu prirodnog jezika

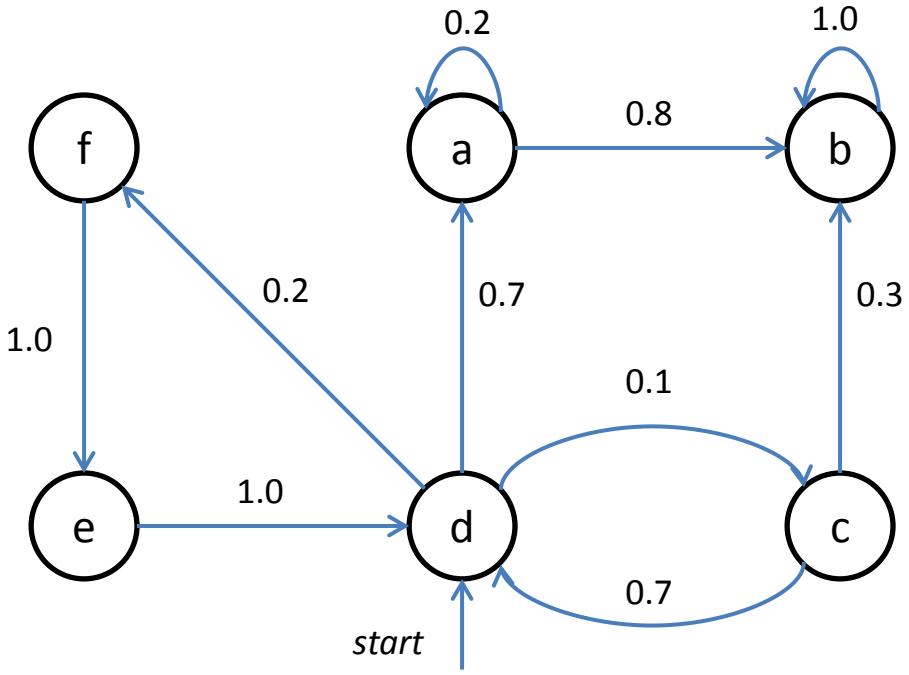
9.5. Markovljev model (Markov model)

Branko Žitko

Markovljev model

- Sekvenca slučajnih varijabli koja nije nezavisna
- Primjer
  - prognoza vremena
  - tekst
- Svojstva:
  - $P(X_{t+1} = s_k | X_1,\ldots,X_t) = P(X_{t+1} = s_k | X_t)$
  - Vremenska invarijanta
  - $P(X_2 = s_k | X_1)$
- Definicija:
  - u terminima tranzicijske matrice A i vjerojatnosti početnog stanja Π

(Vidljivi) Markovljev model (VMM)



$P(X1, \ldots, XT) = P(X1)\, P(X2|X1)\, P(X3|X1,X2) \ldots P(XT|X1,X2,\ldots,XT\text{-}1)$

$= P(X1)\, P(X2|X1)\, P(X3|X2) \ldots P(XT,XT\text{-}1)$

$=$

$P(d, a, b) = P(X_1=d)\, P(X_2=a|X_1=d)\, P(X_3=b|X_2=a)$

$= 1.0 * 0.7 * 0.8$

$= 0.56$

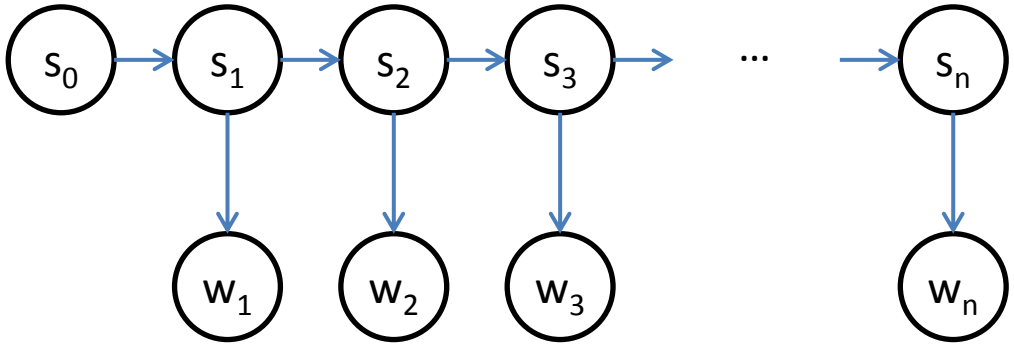
Skriveni Markovljev model

- Hidden Markov Model (HMM)
  - Promatra sekvencu simbola
  - Sekvenca stanja koja vodi do generiranja simbola je skrivena
- Definicija
  - Q = skup stanja
  - O = skup opservacija, napravljena iz rječnika
  - q_0, q_f = specijalna stanja (početno i završno stanje)
  - A = matrica vjerojatnosti tranzicija stanja
  - B = matrica vjerojatnosti emisije simbola
  - Π = vjerojatnosti početnog stanja
  - $\mu = (A, B, \Pi)$ = potpuni probabilistički model

Skriveni Markovljev model

- Koristi se za modeliranja sekvence stanja i sekvence opservacija
- Primjer:

$$P(S|W) = \Pi_i P(s_i|s_{i-1}) P(w_i|s_i)$$

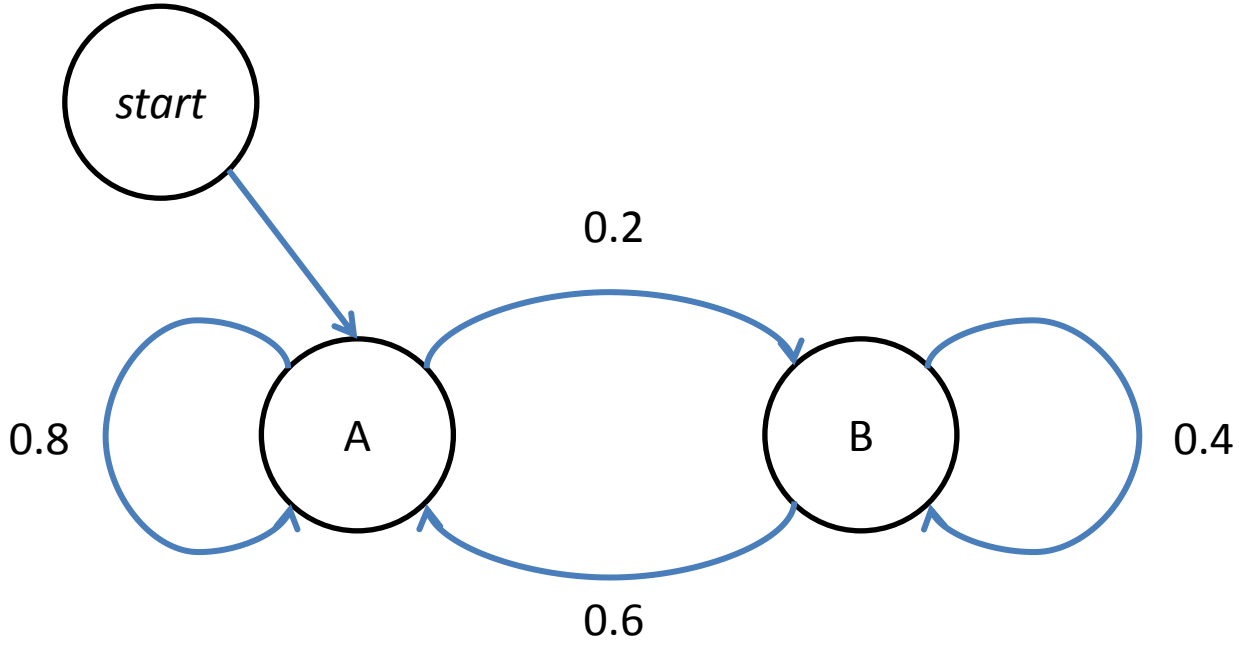


Generativni algoritam

1. Izaberi početak iz Π
2. za t = 1..T
    1. Prijeđi u sljedeće stanje temeljem A
    2. Emitiraj opservaciju temeljem B

Vjerojatnosti skrivenog Markovljevog modela

Vjerojatnosti prijelaza



Emisijske vjerojatnosti

	x	y	z
A	0.7	0.2	0.1
B	0.3	0.5	0.2

Svi parametri skrivenog Markovljevog modela

- **Početak**
  - $P(A|start) = 1.0 \quad P(B|start) = 0.0$
- **Tranzicije**
  - $P(A|A) = 0.8 \quad P(A|B) = 0.6$
  - $P(B|A) = 0.2 \quad P(B|B) = 0.4$
- **Emisije**
  - $P(x|A) = 0.7 \quad P(y|A) = 0.2 \quad P(z|A) = 0.1$
  - $P(x|B) = 0.3 \quad P(y|B) = 0.5 \quad P(z|B) = 0.2$

Opservacijska sekvenca "yz"

- Počevši u stanju A, koliki je P(yz)?
- Moguće sekvence stanja
  - AA
  - AB
  - BA
  - BB
- $P(yz) = P(yz|AA)+P(yz|AB)+P(yz|BA)+P(yz|BB)$

  $= 0.8 \times 0.2 \times 0.8 \times 0.1$

  $+ 0.8 \times 0.2 \times 0.2 \times 0.2$

  $+ 0.2 \times 0.5 \times 0.4 \times 0.2$

  $+ 0.2 \times 0.5 \times 0.6 \times 0.1$

  $= 0.0128 + 0.0064 + 0.0080 + 0.0060 = 0.0332$

HMM zadaci

- Zadaci
  - Za dani model $\mu=(A,B,\Pi)$ pronađi vjerojatnost opservacija $P(O|\mu)$
  - Za dane opservacije O, koji je slijed stanja $(X_1, \ldots, X_{T+1})$
  - Za dane opservacije O i sve moguće modele μ, odredi model koji najbolje opisuje O
- Dekodiranje
  - označiti svaku pojavnicu oznakom
- Vjerodostojnost opservacije
  - klasificiraj sekvencu
- Učenje
  - treniraj model da odgovara empiričkim podacima

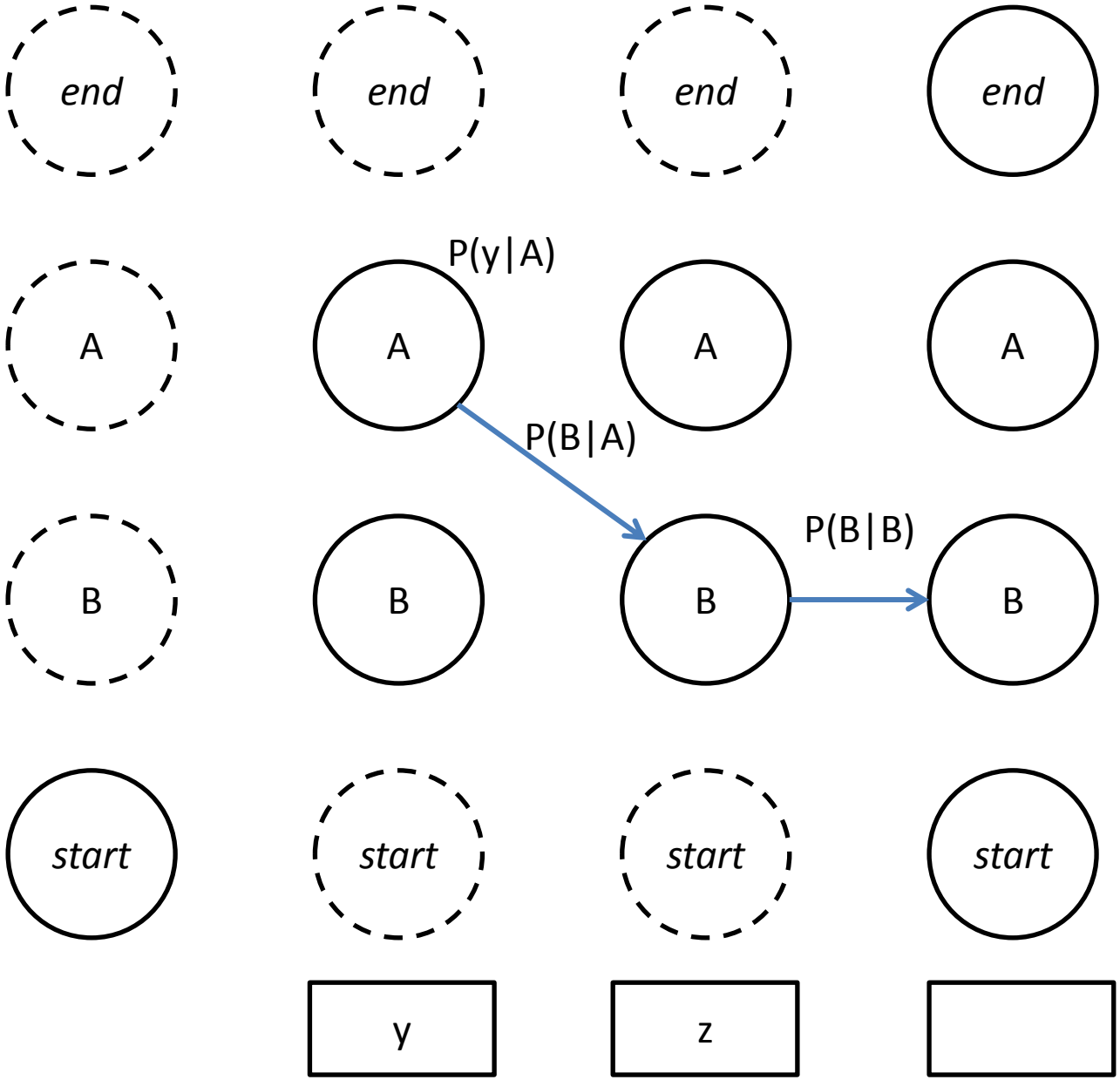
Zaključivanje

- Pronađi najvjerojatniji slijed oznaka, za dani slijed riječi
  - $t^* = \text{argmax}_t P(t|w)$
- Za dani model μ jeli moguće izračunati P(t|w) za sve vrijednosti od t
- U praksi, postoji previše kombinacija
- Moguća rješenja:
  - koristiti pretraživanje po zrakama (beam search) – djelomična hipoteza
  - U svakom stanju, čuvati k najboljih hipoteza do sada
  - Ne mora dobro raditi

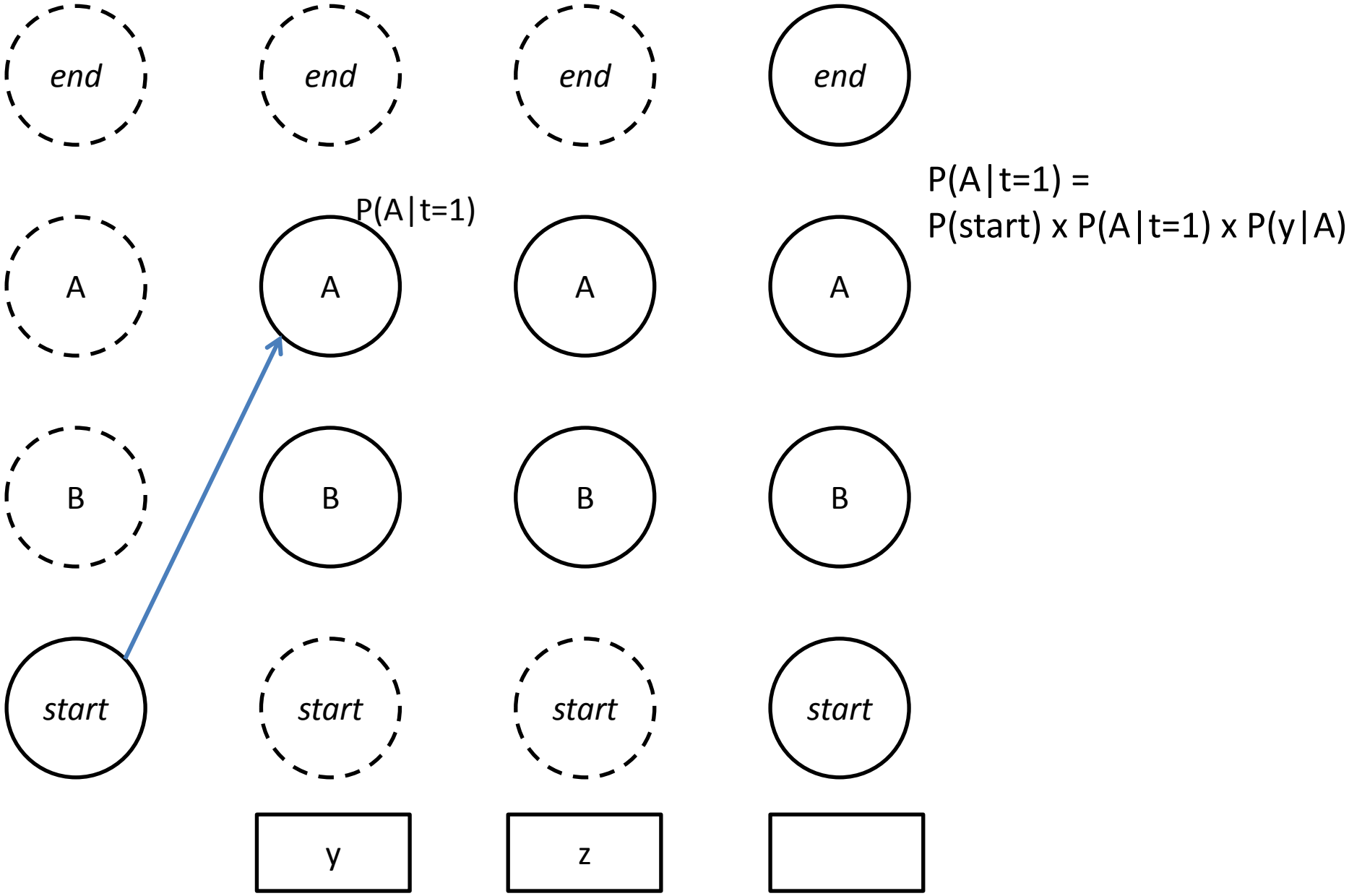
Viterbi algoritam

- Pronađi najbolju putanju do opservacije O i stanja S
- Karakteristike
  - koristi dinamičko programiranje
  - memoizacija
  - praćenje unatrag

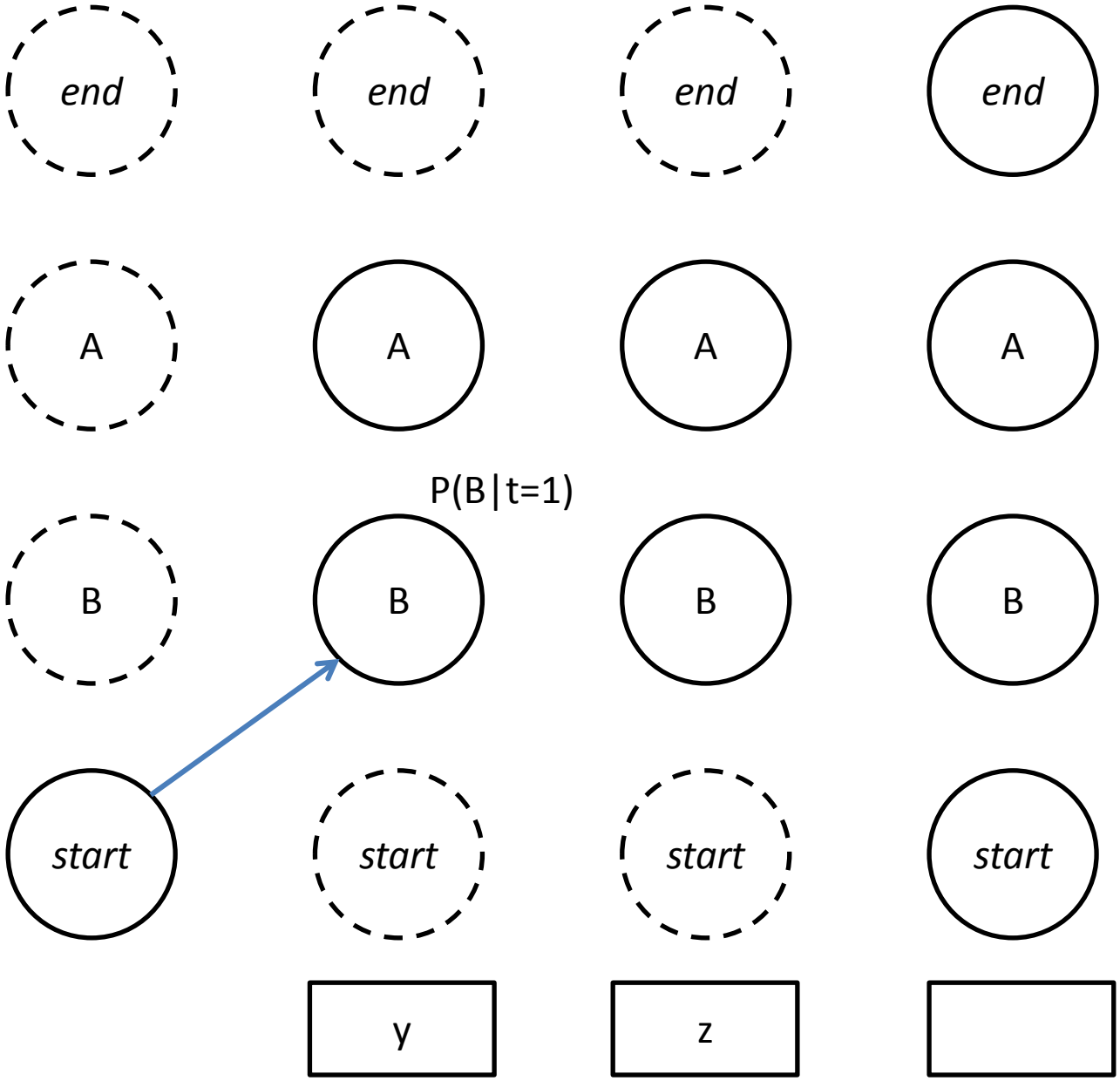
HMM rešetka (trellis)



HMM rešetka (trellis)

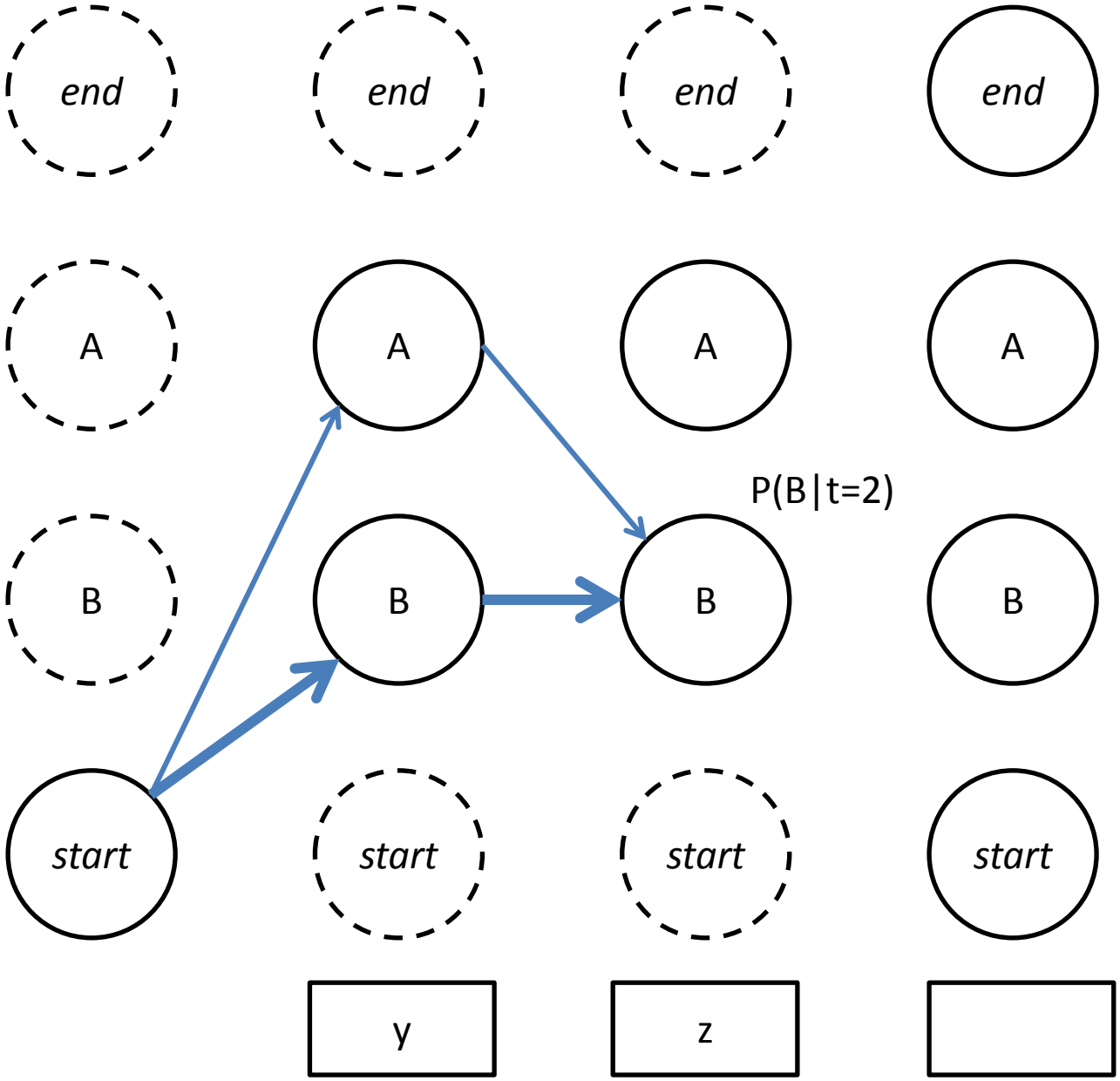


HMM rešetka (trellis)



P(B|t=1) =
P(start) x P(B|t=1) x P(y|B)

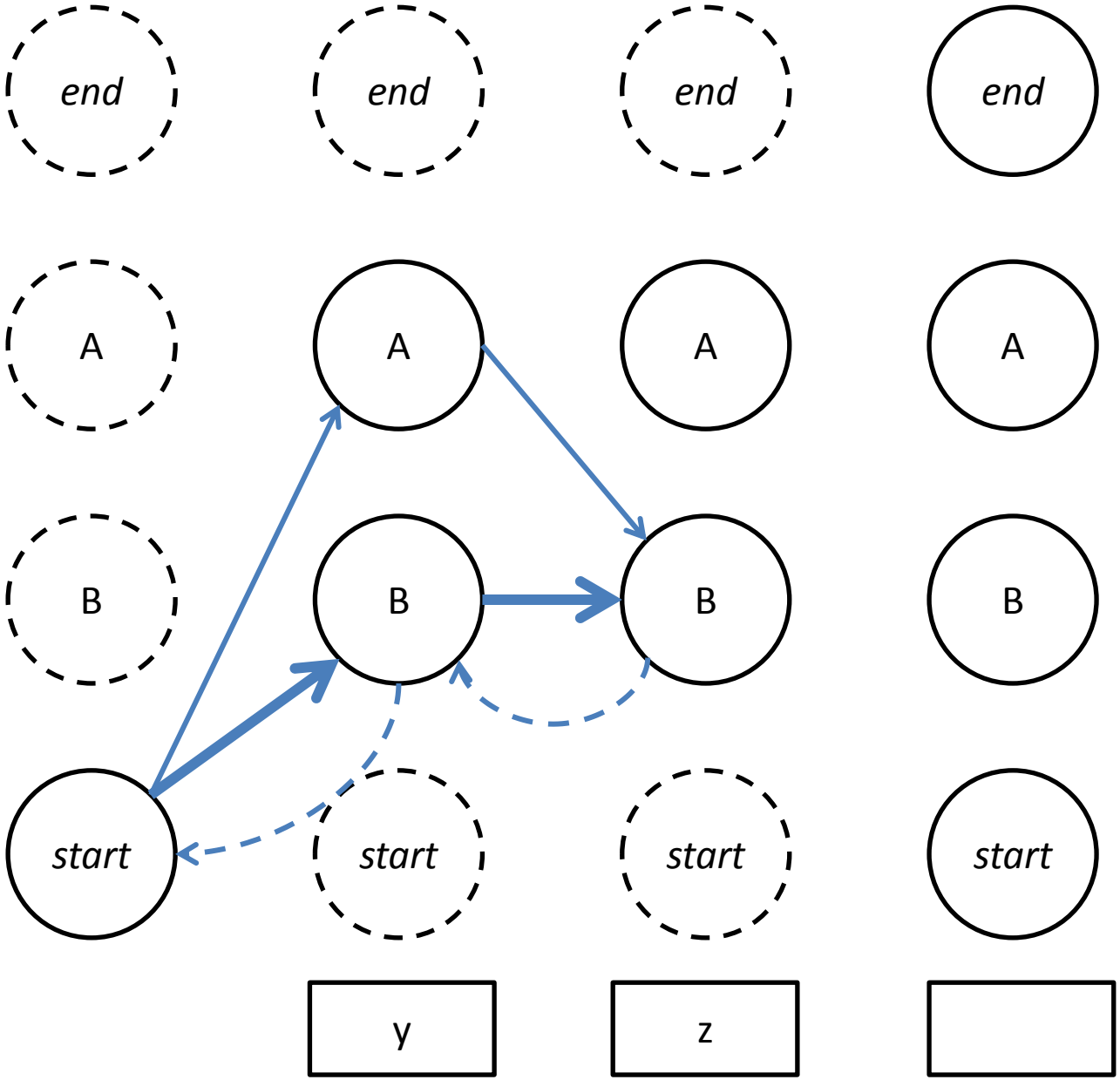
HMM rešetka (trellis)



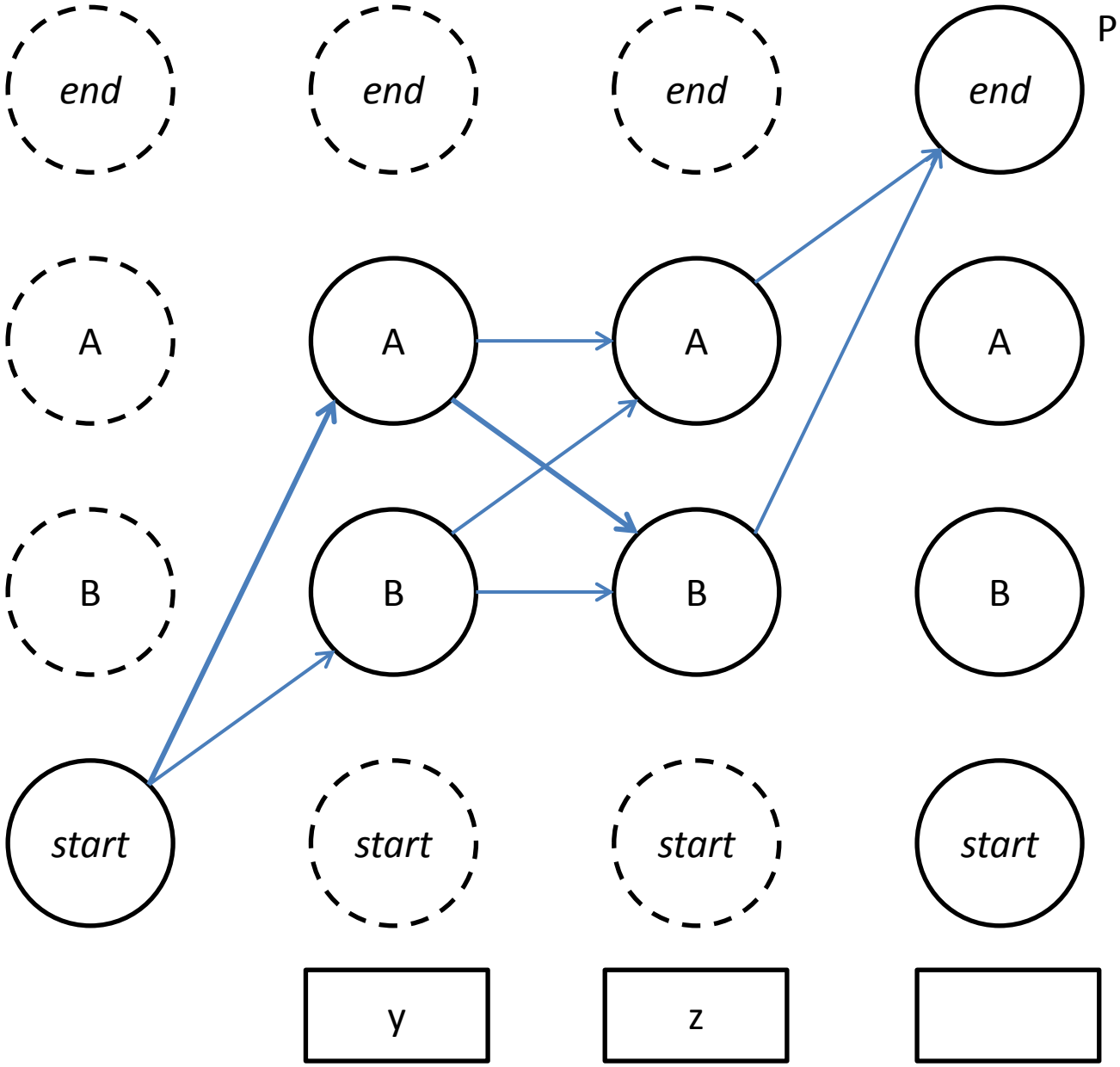
P(B|t=2)

P(B|t=2) =
max(P(A|t=1) x P(B|A)xP(z|B),
P(B|t=1) x P(B|B)xP(z|B))

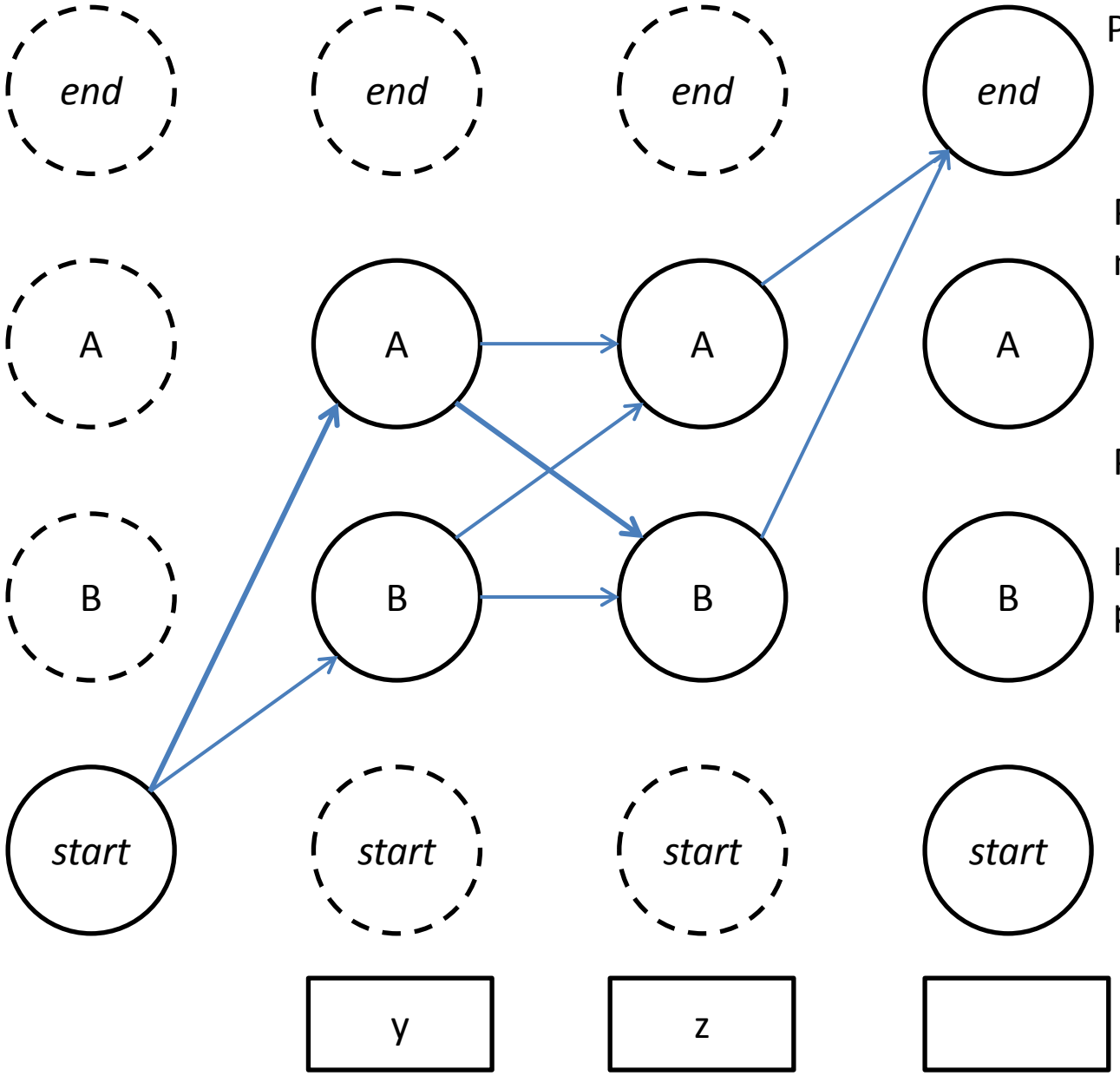
HMM rešetka (trellis)



HMM rešetka (trellis)



HMM rešetka (trellis)



P(end|t=3)

P(end|t=3) =
max(P(A|t=2) x P(end|A),
P(B|t=2) x P(end|B))

P(end|t=3) = najbolja putanja

koristiti povratnu vezu za pronalaženje sekvence stanja

HMM učenje

- Nadzirano
  - Sekvence za treniranje su označene
- Nenadzirano
  - Sekvence za treniranje nisu označene
  - Poznat broj stanja
- Polunadzirano
  - Neke sekvence za treniranje su označene

Nadzirano HMM učenje

- Procjeni vjerojatnosti tranzicija koristeći maksimalnu vjerodostojnost

$$a_{i,j} = \frac{broj(q_t = s_i, q_{t+1} = s_j)}{broj(q_t)}$$

- Procjeni vjerojatnosti opservacije koristeći maksimalnu vjerodostojnost

$$b_j(k) = \frac{broj(q_i = s_j, o_i = v_k)}{broj(q_i = s_j)}$$

- Koristi izglađivanje

Nenadzirano HMM učenje

- Dano
  - slijed opservacija
- Cilj
  - izgraditi HMM
- Koristiti metodu maksimizacije očekivanja (EM - Expectation Maximization)
  - naprijed-nazad (forward-backward) (Baum-Welch) algoritam
  - Baum-Welch pronalazi približno rješenje za $P(O|\mu)$

Baum-Welch

- Algoritam
  - Postavi slučajnim izborom parametre za HMM
  - Dok parametri konvergiraju ponavljaj
    - E korak (očekivanje) – odredi vjerojatnosti za razne sekvence stanja koje generiraju opservaciju (forward-backward)
    - M korak (maksimizacija) – ponovno procjeni parametre za HMM temeljem dobivenih vjerojatnosti
- Rezultati
  - algoritam garantira da će se kod svake iteracije vjerodostojnost od $P(O|\mu)$ povećavati
  - može se zaustaviti bilo kada i dobiti djelomično rješenje
  - konvergira prema lokalnom maksimumu