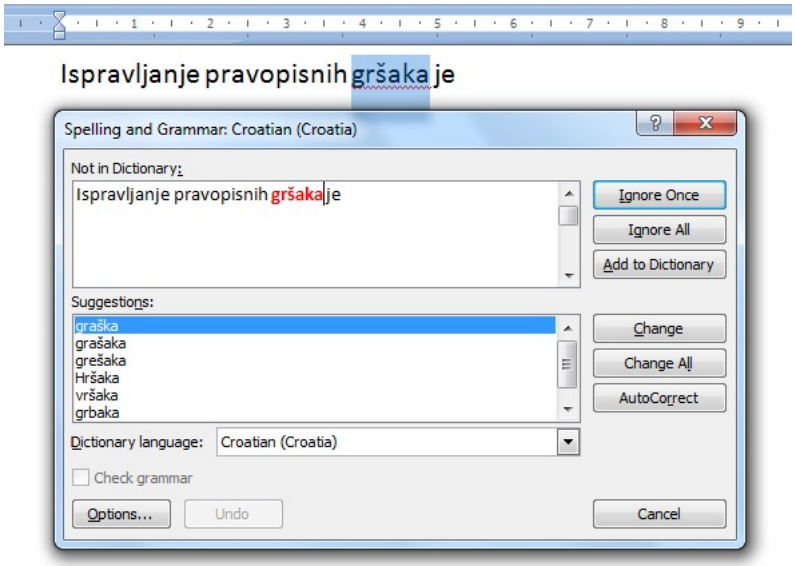
Uvod u obradu prirodnog jezika

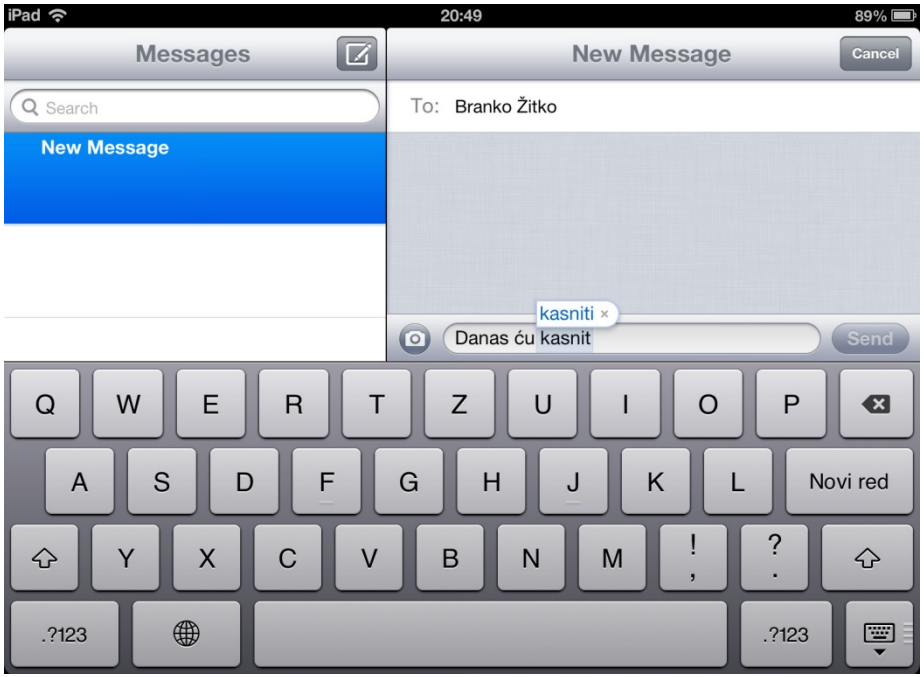
5.1. Ispravljanje pravopisnih grešaka (spelling correction)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Primjena ispravljanja pravopisnih grešaka





Pravopisni zadaci

- Detekcija pravopisnih grešaka
- Ispravljanje pravopisnih grešaka
  - automatsko ispravljanje
    - sma -> sam
  - predlaganje ispravke
  - lista prijedloga

Tipovi pravopisnih grešaka

- Greške ne-riječi (non-word errors)
  - žrafa -> žirafa
- Greške stvarnih riječi (real-word errors)
  - Tipografske greške
    - staklo -> stablo
  - Kognitivne greške (homonimi, homografi, homofoni)
    - pas -> pas,
    - luk -> luk
    - gore -> gore
    - knight -> night

Stope pravopisnih grešaka

26% Web upiti

Wang et al. 2003

13% Prekucavanje bez brisanja

Whitelaw et al. English&German

7% Riječi ispravljene prekucavanjem na malim uređajima (mobitel)

2% Neispravljene riječi na malim uređajima

Soukoreff &MacKenzie 2003

1-2% Prekucavanje

Kane & Wobbrock 2007, Gruden et al. 1983

Pravopisne greške ne-riječi

- Detekcija pravopisnih grešaka ne-riječi
  - svaka riječ koja nije u **rječniku** je greška
  - bolje je imati veliki rječnik
- Ispravljanje pravopisnih grešaka ne-riječi
  - generiranje **kandidata**: stvarne riječi koje su slične pogrešnoj riječi
  - izbor najboljeg kandidata:
    - najkraća težinska udaljenost riječi
    - najveća vjerojatnost kanala sa šumom

Pravopisne greške stvarnih riječi

- Za svaku riječ w generira se skup kandidata
  - pronađi kandidate sa sličnim **izgovorom**
  - pronađi kandidate sa sličnim **pravopisom**
  - uključi w u skup kandidata
- Izbor najboljeg kandidata
  - kanal sa šumom (noisy channel)
  - klasifikator

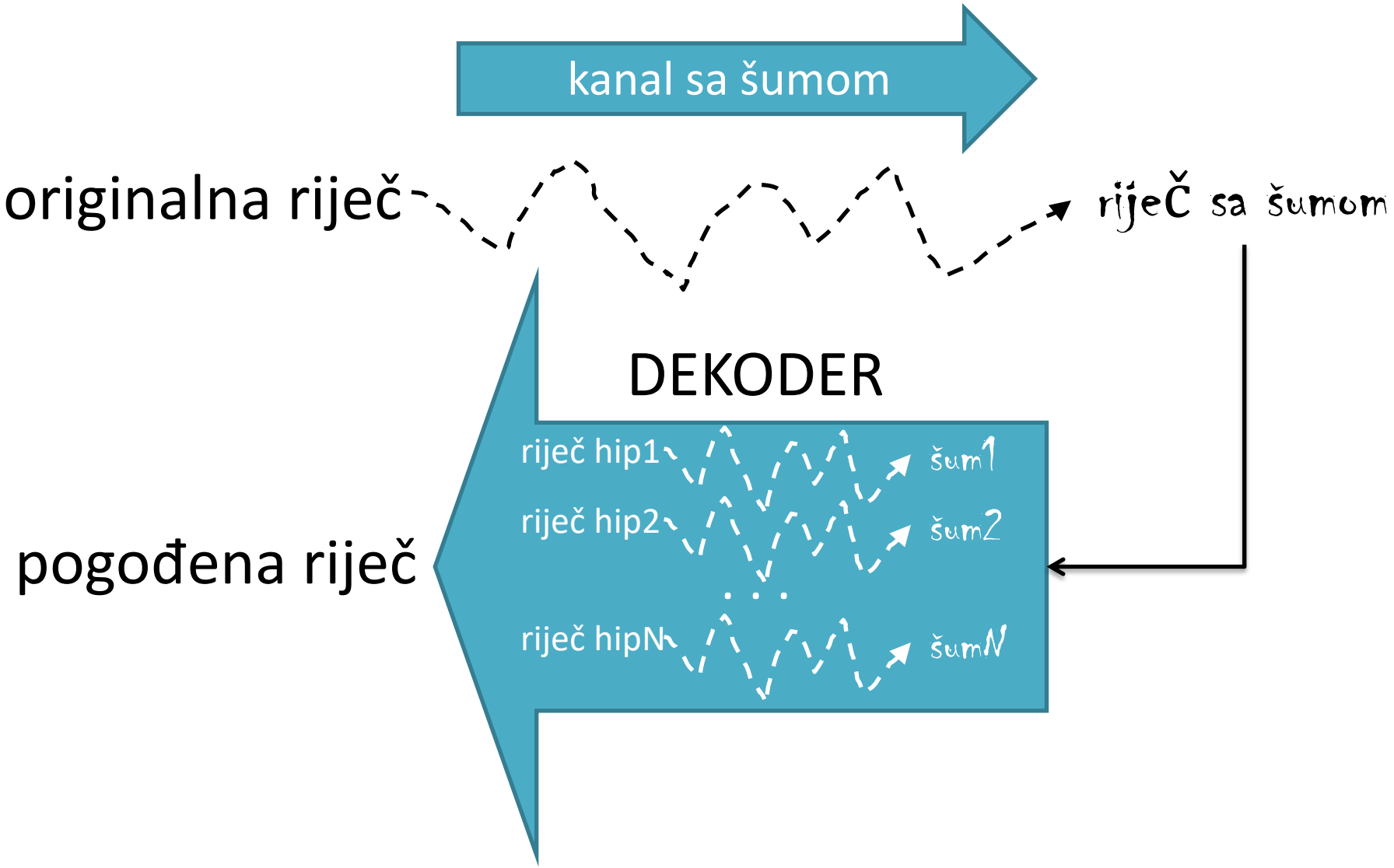
Uvod u obradu prirodnog jezika

5.2. Model ispravljanja pravopisnih grešaka korištenjem kanala sa šumom (the noisy channel model of spelling)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ideja kanala sa šumom



Kanal sa šumom

- za danu pogrešno napisanu riječ x
- pronađi točnu riječ w

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w|x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x|w)P(w)}{P(x)}$$

$$= \underset{w \in V}{\operatorname{argmax}} \; P(x|w)P(w)$$

Izglednost (model kanala)

prior

Povijest kanala sa šumom

- **IBM**

  Mays, Eric, Fred J. Damerau and Robert L. Mercer. 1991. Context based spelling correction. Information Processing and Management, 23(5), 517–522

- **AT&T Bell Labs**

  Kernighan, Mark D., Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. Proceedings of COLING 1990, 205-210

Primjer ispravljanja greške ne-riječi

acress

Generiranje kandidata

- Riječi sa sličnim pravopisom
  - mala udaljenost između riječi i pogreške
- Riječi sa sličnim izgovorom
  - mala udaljenost između izgovora riječi i pogreške

Damerau-Levenshtein udaljenost

- Minimalna udaljenost dva niza znakova koja uključuje sljedeće operacije:
  - Ubacivanje
  - Brisanje
  - Zamjena
  - Transpozicija dva susjedna znaka

- Damerau-Levenshtein udaljenost

Riječi udaljene za 1 od acress

Greška	Kandidat	Točni znak	Pogrešni znak	Operacija
acress	actress	t	-	brisanje
acress	cress	-	a	ubacivanje
acress	caress	ca	ac	transpozicija
acress	access	c	r	zamjena
acress	across	o	e	zamjena
acress	acres	-	s	ubacivanje
acress	acres	-	s	ubacivanje

Generiranje kandidata

- 80% grešaka su udaljene za 1
- Skoro sve greške su udaljene za 2

- Dozvoliti ubacivanje **razmaka** i **crtice**
  - ovaideja -> ova ideja
  - splitskodalmatinska -> splitsko-dalmatinska

Model jezika

- Koristiti bilo koji algoritam za modeliranje jezika
  - unigram, bigram, trigram
- Ispravljanje pravopisnih grešaka za velike sadržaje (Web)
  - glupo odustajanje (Stupid backoff)

Vjerojatnost za prior kod unigrama

- od 404253213 riječi u Corpus of Contemporary English (COCA)

riječ	frekvencija riječi	P(riječ)
actress	9321	0.0000230573
cress	220	0.0000005442
caress	686	0.0000016969
access	37038	0.0000916207
across	120844	0.0002989314
acres	12874	0.0000318463

Vjerojatnost kanala sa šumom

- Vjerojatnost modela pogreške
  - Kernighan, Church, Gale 1990

- Pogrešno napisana riječ $x = x_1\, x_2\, x_3 \ldots\, x_m$
- Točna riječ $w = w_1\, w_2\, w_3 \ldots\, w_n$

- $P(x|w)$ = vjerojatnost operacije
  - (brisanje/ubacivanje/zamjena/transpozicija)

Izračun vjerojatnosti pogreške: matrica konfuzije

brisanje[x, y]:	broj(xy unesenih kao x)
ubacivanje[x, y]:	broj(x unesenog kao xy)
zamjena[x, y]:	broj(x unesenog kao y)
transpozicija[x, y]:	broj(xy unesenih kao yx)

Ubacivanje i brisanje ovisi o prethodnom znaku

Matrica konfuzije za pravopisne pogreške

zamjena[X,Y] = supstitucija od X(pogrešno) s Y(točno)

		Y(točno)																									
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
X	a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
	b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
	c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
	d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
	e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	14	0	1	0	18	0
	f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
	g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
	h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
	i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
	j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
	k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
	l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
	m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
	n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
	o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
	p	0	11	1	2	0	6	5	0	2	0	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
	q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
	s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
	t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
	u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
	v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
	w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
	x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
	y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
	z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Generiranje matrice konfuzije

Roger Mitton-ova lista pogrešaka

https://www.dcs.bbk.ac.uk/~ROGER/corpora.html

Peter Norvig-ova lista pogrešaka

http://norvig.com/ngrams/spell-errors.txt

Model kanala sa šumom

$$P(x|w) = \begin{cases} \dfrac{brisanje[x_{i-1}, w_i]}{broj[x_{i-1}w_i]} & \text{ako je brisanje} \\ \dfrac{ubacivanje[x_{i-1}, w_i]}{broj[x_{i-1}w_{i-1}]} & \text{ako je ubacivanje.} \\ \dfrac{zamjena[x_i, w_i]}{broj[w_i]} & \text{ako je zamjena.} \\ \dfrac{transpozicija[w_i, w_{i+1}]}{broj[w_i w_{i+1}]} & \text{ako je transpozicija} \end{cases}$$

Kernighan, Church, Gale 1990

Model kanala za acress

Kandidat	Točni znak	Pogrešni znak	x|w	P(x|w)
actress	t	-	c|ct	0.000117
cress	-	a	a|#	0.00000144
caress	ca	ac	ac|ca	0.00000164
access	c	r	r|c	0.000000209
across	o	e	e|o	0.0000093
acres	-	s	es|e	0.0000321
acres	-	s	ss|s	0.0000342

Vjerojatnost kanala sa šumom za acress

Kandidat	Točni znak	Pogrešni znak	x|w	P(x|w)	P(w)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	0.000117	0.0000231	2.7
cress	-	a	a|#	0.00000144	0.000000544	0.00078
caress	ca	ac	ac|ca	0.00000164	0.00000170	0.0028
access	c	r	r|c	0.000000209	0.0000916	0.019
across	o	e	e|o	0.0000093	0.000299	2.8
acres	-	s	es|e	0.0000321	0.0000318	1.0
acres	-	s	ss|s	0.0000342	0.0000318	1.0

Vjerojatnost kanala sa šumom za acress

Kandidat	Točni znak	Pogrešni znak	x|w	P(x|w)	P(w)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	0.000117	0.0000231	2.7
cress	-	a	a|#	0.00000144	0.000000544	0.00078
caress	ca	ac	ac|ca	0.00000164	0.00000170	0.0028
access	c	r	r|c	0.000000209	0.0000916	0.019
across	o	e	e|o	0.0000093	0.000299	2.8
acres	-	s	es|e	0.0000321	0.0000318	1.0
acres	-	s	ss|s	0.0000342	0.0000318	1.0

Korištenje digram modela jezika

- "a stellar and versatile **acress** whose combination of sass and glamour…"

- Frekvencije iz Corpus of Contemporary American English s dodaj 1 izglađivanjem

  P(actress|versatile)=0.000021 P(whose|actress) = 0.0010

  P(across|versatile) =0.000021 P(whose|across) = 0.000006

  P("versatile actress whose") = 0.000021*0.0010 = 210 $x10^{-10}$

  P("versatile across whose") = 0.000021*0.000006 = 1 $x10^{-10}$

Korištenje digram modela jezika

- "a stellar and versatile **acress** whose combination of sass and glamour…"

- Frekvencije iz Corpus of Contemporary American English s dodaj 1 izglađivanjem

  P(actress|versatile)=0.000021 P(whose|actress) = 0.0010

  P(across|versatile) =0.000021 P(whose|across) = 0.000006

  P("versatile actress whose") = 0.000021*0.0010 = 210 $x10^{-10}$

  P("versatile across whose") = 0.000021*0.000006 = 1 $x10^{-10}$

Evaluacija

- Neki skupovi za testiranje pravopisnih pogrešaka (Engleski)
  - Wikipedia's list of common English misspelling
  - Aspell filtered version of that list
  - Birkbeck spelling error corpus
  - Peter Norvig's list of errors (includes Wikipedia and Birkbeck, for training or testing)
- Neki Hrvatski rječnici za pravopisne pogreške
  - Croatian Dictionary (Hrvatski Rjecnik) for Mozilla Firefox, Thunderbird and SeaMonkey
  - Croatian dictionary and hyphenation patterns

Uvod u obradu prirodnog jezika

5.3. Ispravljanje pravopisnih pogrešaka stvarnih riječi (real-word spelling correction)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Pravopisne greške stvarnih riječi

- …odlazim za petnaest **minueta** kako bi stigao na vlak.
- Razbio sam vazu **trome** mužu.
- Možeš li me **nosivi**?
- Računalo ima procesor **a** memoriju.

- 25-40% pravopisnih pogrešaka su stvarne riječi Kukich 1992

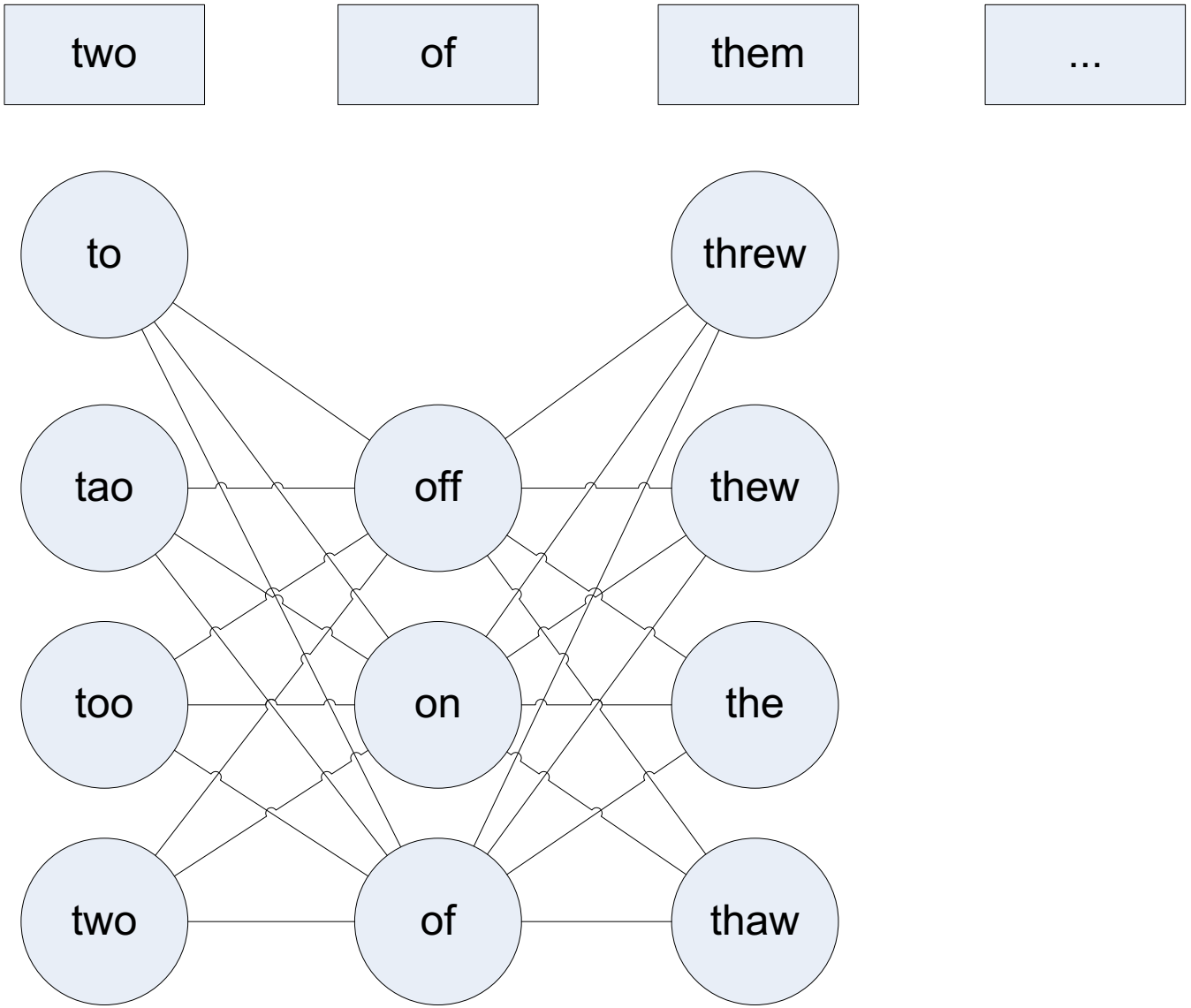
Ispravljanje pravopisnih grešaka stvarnih riječi

- Za svaku riječ u rečenici
  - generiraj skup kandidata
    - sama riječ
    - sve riječi iz rječnika koje su nastale provedbom jedne operacije (ubacivanje, brisanje, zamjena, transpozicija)
    - riječi koje su homonimi
- Izaberi najboljeg kandidata
  - model kanala sa šumom
  - specijalizirani klasifikator

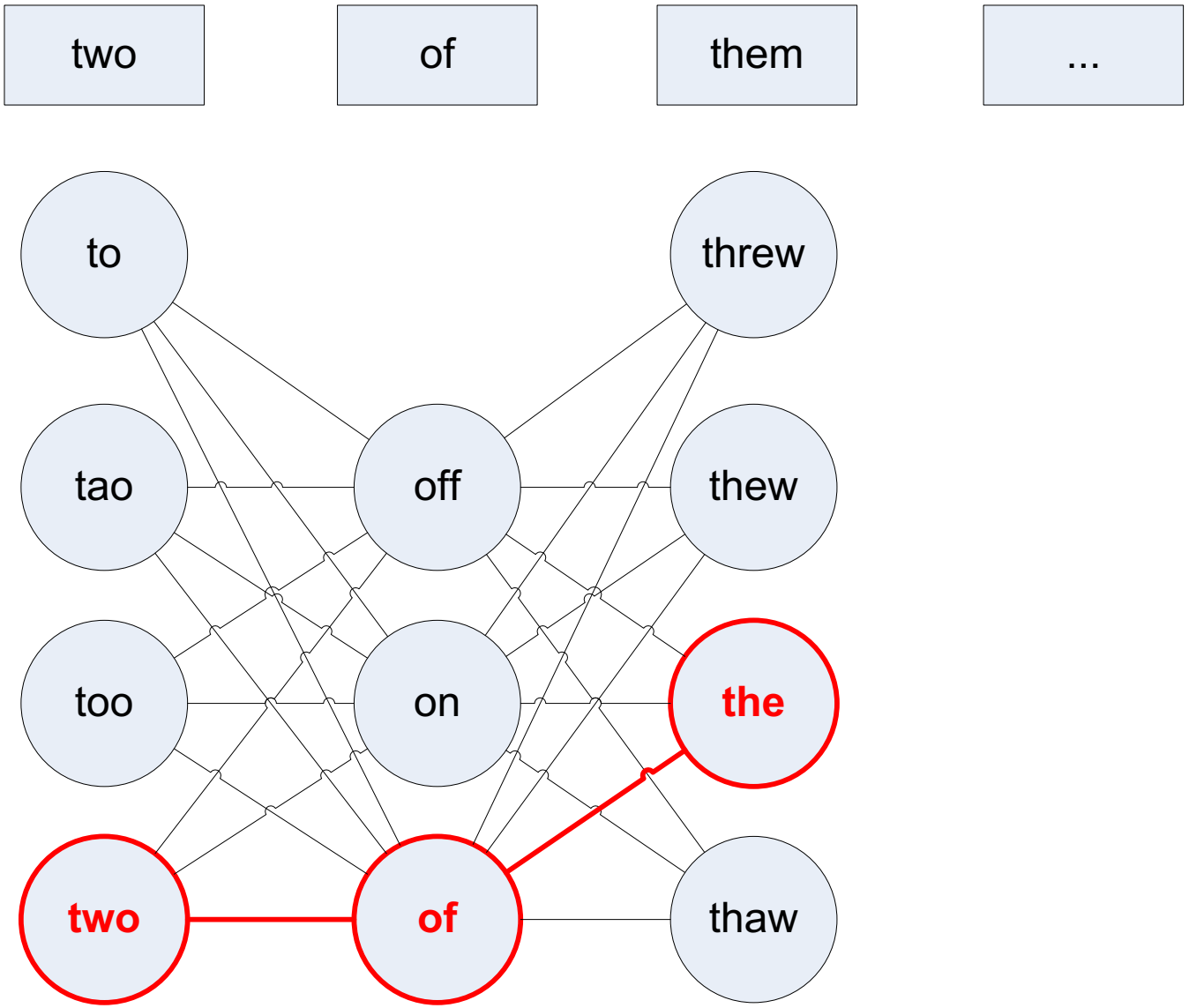
Kanal sa šumom za stvarne riječi

- Za danu rečenicu $w_1, w_2, w_3 \ldots w_n$
- Generiraj skup kandidata za svaku riječ w_i
  - Kandidat(w_1) = {w_1, w'_1 , w''_1 , w'''_1 ,…}
  - Kandidat(w_2) = {w_2, w'_2 , w''_2 , w'''_2 ,…}
  - Kandidat(w_n) = {w_n, w'_n , w''_n , w'''_n ,…}
- Izaberi slijed riječi W koji maksimizira $P(W)$

Kanal sa šumom za stvarne riječi



Kanal sa šumom za stvarne riječi



Pojednostavljenje: Jedna greška po rečenici

- Od svih mogućih rečenica s jednom pogrešnom riječi
  - w_1, $\mathbf{w''_2}$, w_3, w_4 two off thew
  - w_1, w_2, $\mathbf{w'_3}$, w_4 two of the
  - $\mathbf{w'''_1}$, w_2, w_3, w_4 too of thew
  - …
- Izaberi slijed W koji maksimizira $P(W)$

Kako izračunati vjerojatnosti

- Model jezika
  - unigram
  - bigram
  - …
- Model kanala
  - Isto kao i za greške ne-riječi
  - dodatno, potrebno je izračunati vjerojatnost $P(w|w)$

$$\hat{W} = \underset{W \in C(X)}{\operatorname{argmax}} P(X|W)P(W)$$

```
...
only two of thew apples
oily two of thew apples
only too of thew apples
only to of thew apples
only tao of the apples
only two on thew apples
only two off thew apples
only two of the apples
only two of threw apples
only two of thew applies
only two of thew dapples
...
```

Vjerojatnost da nije greška

- Koja je vjerojatnost kanala točno upisane riječi?
- P("the"|"the")

- Pretpostavimo $P(w|w) = \alpha$
- Onda jednostavan model kanala je

$$p(x|w) = \begin{cases} \alpha & \text{ako } x = w \\ \dfrac{1-\alpha}{|C(x)|} & \text{ako } x \in C(x) \\ 0 & \text{inače} \end{cases}$$

- Biranje α , ovisno o aplikaciji
  - 0.90 (1 greška od 10 riječi)
  - 0.95 (1 greška od 20 riječi)
  - 0.99 (1 greška od 100 riječi)
  - 0.995 (1 greška od 200 riječi)

Peter Norvig primjer s "thew"

Složeniji model kanal koristit će matrice konfuzije

x	w	x|w	P(x|w)	P(w)	10^9 P(x|w)P(w)
thew	the	ew|e	0.000007	0.02	144
thew	thew		0.95	0.00000009	90
thew	thaw	e|a	0.001	0.0000007	0.7
thew	threw	h|hr	0.000008	0.000004	0.03
thew	thwe	ew|we	0.000003	0.00000004	0.0001

Uvod u obradu prirodnog jezika

5.4. Najsuvremeniji sustavi (state-of-the-art systems)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Treba li riječ ispravljati?

- Model kanala je sklon ispravljanju točnih riječi (rijetka imena) s frekventnijom riječi

- Proširenje modela kanala:
  - riječ potrebno ispravljati ili ne

- Primjeri:
  - korištenjem crne liste
  - zabrana promjene određenih pojavnica (brojevi, interpunkcija, riječi s jednim znakom)
  - razlika između vjerojatnosti promjene ili ne

$$\log P(w|x) - \log P(x|x) > \Theta$$

HCI pitanja kod pravopisnih grešaka

- Ako smo veoma uvjereni kod ispravljana
  - automatsko ispravljanje
- Manje uvjereni
  - daj najbolju ispravku
- Još manje uvjereni
  - daj listu ispravaka
- Nismo uvjereni
  - samo označi kao grešku

Odluka najčešće ovisi o klasifikatoru

Suvremeni kanal sa šumom

- Nikad se ne množi model jezika s modelom kanala
- Pretpostavka nezavisnosti → vjerojatnost nije primjerena
- Umjesto toga ih izmjeri i skaliraj

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)^{\lambda}$$

- Nauči λ iz razvojnog skupa

Fonetski model pogreške

- Metafon, kod GNU aspell
  - Konvertiraj pogrešku u metafonski izgovor
    - Izbaci duple susjedne znakove, osim C
    - Ako riječ počinje s KN, GN, PN, AE, WR; izbaci prvi znak
    - Izbaci B ako je iza M i ako je na kraju riječi
    - …
  - Pronađi riječi čiji je izgovor udaljen za 1-2 od pogrešne riječi
  - boduj rezultate
    - težinska udaljenost između kandidata i pogreške
    - udaljenost između izgovora kandidata i izgovora pogreške

Poboljšanja na kanalu za šum

- Dozvoli bogatije operacije (Brill & Moore 2000)
  - ent -> ant
  - ph -> f
  - le -> al
- Integriranje izgovora u kanal (Toutanova & Moore 2002)

Model kanala

- Faktori koji mogu utjecati na P(pogreška|word)
  - početni znak
  - krajnji znak
  - Okolni znakovi
  - pozicija znaka u riječi
  - susjedne tipke na tipkovnici
  - izgovor
  - transformacije sličnih morfema
  - …

Susjedne tipke