Korpusa analizo de dinamika evoluo de la slovaka lingvo

Radovan Garabík

SNK JÚĽŠ SAV

2016-11-18
Modra, Harmónia

Signifo de la „frekvenco“

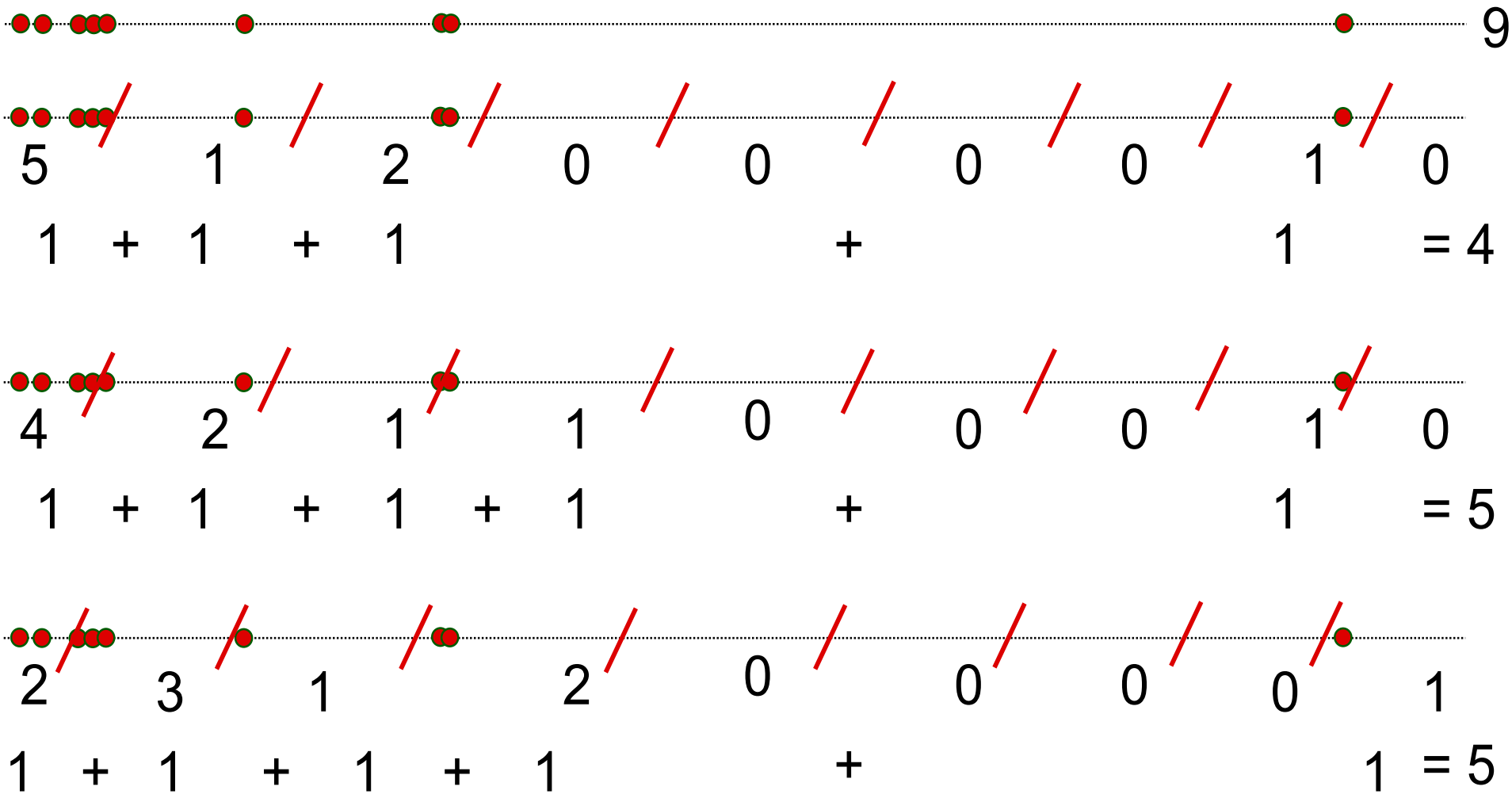
- korpuso, tekstaro
- reala nombro de la vorto (aŭ gramatikaĵo) en la korpuso

Signifo de la „frekvenco“

- korpuso, tekstaro
- reala nombro de la vorto (aŭ gramatikaĵo) en la korpuso – frekvenco
- sed kio okazus se iu uzas la vorton pli ofte?

Signifo de la „frekvenco“

- korpuso, tekstaro
- reala nombro de la vorto (aŭ gramatikaĵo) en la korpuso – frekvenco
- sed kio okazus se iu uzas la vorton pli ofte?
- aŭ se la vorto aperas ĉiam dufoje?
- Ĉe konsideri pli ofta uzado de la vorto en la sama dokumento (de la sama aŭtoro...)?


9
5 1 2 0 0 0 0 1 0
1 + 1 + 1 + 1 = 4
4 2 1 1 0 0 0 1 0
1 + 1 + 1 + 1 + 1 = 5
2 3 1 2 0 0 0 0 1
1 + 1 + 1 + 1 + 1 = 5

⋮

ARF

$$ARF = \frac{1}{v} \sum_{i=1}^{f} min\{d_i, v\}$$

$$v = \frac{N}{f}$$

Atributoj

- homogene uzada vorto – $ARF \to f$

Atributoj

- homogene uzada vorto – $ARF \to f$
- vorto uzada multfoje nur en unu loko en la korpuso

$$ARF = 1 + \frac{f^2 - f}{N} \to 1$$

Atributoj

- homogene uzada vorto – $ARF \to f$
- vorto uzada multfoje nur en unu loko en la korpuso

$$ARF = 1 + \frac{f^2 - f}{N} \to 1$$

- $$1 \leq ARF \leq N$$

Atributoj

- homogene uzada vorto – $ARF \to f$
- vorto uzada multfoje nur en unu loko en la korpuso

$$ARF = 1 + \frac{f^2 - f}{N} \to 1$$

- $$1 \leq ARF \leq N$$

- tre taŭgaj intuiciaj atributoj – funkcias kiel frekvenco, sed se la vorto estas plidensa, ARF malplias

Problemoj

- neadiciebla
- kelkoblaj aperoj de la vorto malpliigas la ARF same kiel onaj aperoj

Evoluo de la lingvo

- la korpuso enhavas sufiĉajn datumojn el lastaj jardekoj

Evoluo de la lingvo

- la korpuso enhavas sufiĉajn datumojn el lastaj jardekoj, ĉu?

Evoluo de la lingvo

- la korpuso enhavas sufiĉajn datumojn el lastaj jardekoj, ĉu?
- Slovaka nacia korpuso – 1250 milionoj da vortoj (kun interpunkcio)

Evoluo de la lingvo

- la korpuso enhavas sufiĉajn datumojn el lastaj jardekoj, ĉu?
- Slovaka nacia korpuso – 1250 milionoj da vortoj (kun interpunkcio)
- sed ni ne havas sufiĉe por bona statistiko
- averaĝo tra kelkaj jaroj

Evoluo de la lingvo

- la korpuso enhavas sufiĉajn datumojn el lastaj jardekoj, ĉu?
- Slovaka nacia korpuso – 1250 milionoj da vortoj (kun interpunkcio)
- sed ni ne havas sufiĉe por bona statistiko
- averaĝo tra kelkaj jaroj
- glitaveraĝa metodo (SMA)

$$\bar{f} = \frac{1}{n} \sum_{i}^{n} f_i$$

- kiu vidas la problemon?

Evoluo de la lingvo

- la korpuso enhavas sufiĉajn datumojn el lastaj jardekoj, ĉu?
- Slovaka nacia korpuso – 1250 milionoj da vortoj (kun interpunkcio)
- sed ni ne havas sufiĉe por bona statistiko
- averaĝo tra kelkaj jaroj
- glitaveraĝa metodo (SMA)

$$\bar{f} = \frac{1}{n} \sum_{i}^{n} f_i$$

- kiu vidas la problemon? ARF ne estas adiciebla

Evoluo de la lingvo

- frekvenco, ARF: IMG – beletro, INF – ĵurnalismo, PRF – fakaj tekstoj
- ne nur vortoj, sed ankaŭ gramatikaĵoj
- ± 5 jaroj

Evoluo de la lingvo

- kiel kalkuli nivelon de vorta „plidensiĝo“?

Evoluo de la lingvo

- kiel kalkuli nivelon de vorta „plidensiĝo“?
- homogeneco:

$$h(w) = \frac{arf(w) - 1}{f(w) - 1}$$

- kial -1 en la numeratoro? se $arf \to 1$ – la vorto aperas nur en unu loko – $h \to 0$
- kial -1 en la denominatoro? se $arf = f \Rightarrow h = 1$

Evoluo de la lingvo

- kiel kalkuli nivelon de vorta „plidensiĝo“?
- homogeneco:

$$h(w) = \frac{arf(w) - 1}{f(w) - 1}$$

- kial -1 en la numeratoro? se $arf \to 1$ – la vorto aperas nur en unu loko – $h \to 0$
- kial -1 en la denominatoro? se $arf = f \Rightarrow h = 1$
- ... sed se $f = 1 \Rightarrow arf = 1$ kaj $h = \frac{0}{0}$

Evoluo de la lingvo

- kiel kalkuli nivelon de vorta „plidensiĝo“?
- homogeneco:

$$h(w) = \frac{arf(w) - 1}{f(w) - 1}$$

- kial -1 en la numeratoro? se $arf \to 1$ – la vorto aperas nur en unu loko – $h \to 0$
- kial -1 en la denominatoro? se $arf = f \Rightarrow h = 1$
- ... sed se $f = 1 \Rightarrow arf = 1$ kaj $h = \frac{0}{0}$ – ĝuste tion ni volas
- intuicie:
- $h = 0$, tute nehomogena, vorto aperas en unu loko
- $h = 1$, vorto aperas homogene en la korpuso

Evoluo de la lingvo

- kiel kalkuli nivelon de „fidindeco“ de la statistiko?

Evoluo de la lingvo

- kiel kalkuli nivelon de „fidindeco“ de la statistiko?
- intervalo de konfido

Evoluo de la lingvo

- kiel kalkuli nivelon de „fidindeco“ de la statistiko?
- intervalo de konfido
- dunomiala distribuo: populacio de la subkorpuso estas n, nombro de vortoj estas nombro de pozitivaj aperoj k

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathrm{E}(X) = np$$

$$\mathrm{Var}(X) = np(1-p)$$

Evoluo de la lingvo

- $< c_1, c_2 >$ – 95 %-a intervalo de konfido de dunomiala distribuo

Evoluo de la lingvo

- $< c_1, c_2 >$ – 95 %-a intervalo de konfido de dunomiala distribuo

$$c_1 = B\left(\frac{\alpha}{2}; k, n-k+1\right)$$

$$c_2 = B\left(1-\frac{\alpha}{2}; k+1, n-k\right)$$

Evoluo de la lingvo

- $< c_1, c_2 >$ – 95 %-a intervalo de konfido de dunomiala distribuo

$$c_1 = B\left(\frac{\alpha}{2}; k, n-k+1\right)$$

$$c_2 = B\left(1-\frac{\alpha}{2}; k+1, n-k\right)$$

- fidindeco:

$$r(w) = 1 - \frac{c_2 - c_1}{c_2 + c_1}$$

- se $c_2 - c_1 \ll c_2 + c_1 \Rightarrow r \to 1^-$
- se $c_2 = 0 \Rightarrow r = 0^+$

ekzemploj

Dankon por via atento