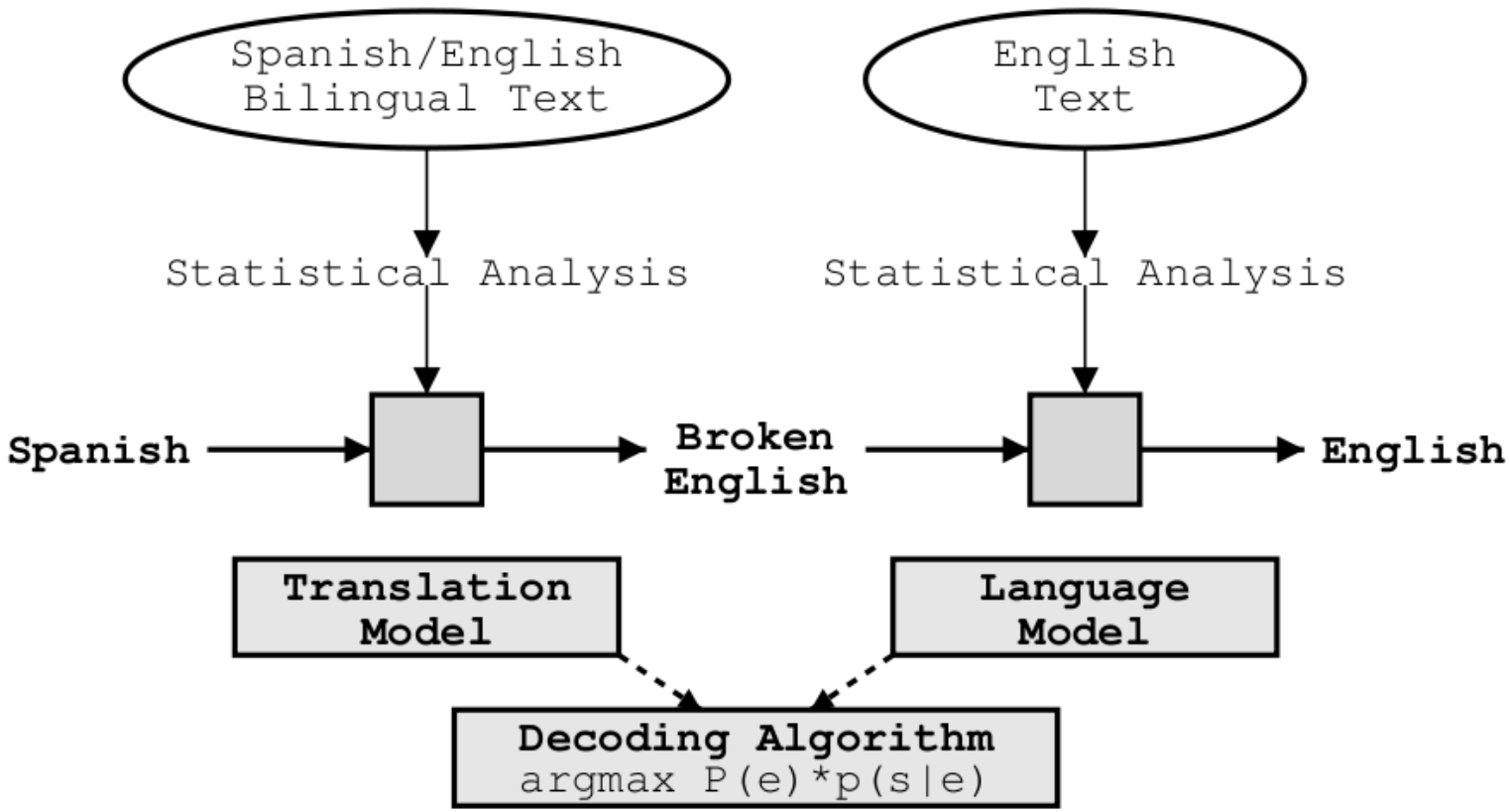
1 Úvod do statistického strojového překladu

Úvod do SMT

- pravidlové systémy motivovány lingvistikou
- SMT inspirován teorií informace a statistikou
- v současnosti mnoho společností se zaměřením na SMT: Google, IBM, Microsoft, Language Weaver (2002)
- 50 miliónů stránek denně přeložených pomocí SMT
- **gisting**: stačí, má-li překlad nějaký užitek, nepotřebujeme přesný význam; nejčastější užití MT na internetu

Úvod

Schéma SMT



Nástroje SMT

- GIZA++: trénování IBM modelů, zarovnávání na úrovni slov (word alignment pomocí HMM)
- SRILM: trénování jazykových modelů
- IRST: trénování velkých jazykových modelů
- Moses: frázový dekodér, trénování modelů
- Pharaoh: předchůdce Mosese
- Thot: trénování frázových modelů
- SAMT: tree-based modely

Data pro SMT – (paralelní) korpusy

- Linguistics Data Consorcium (LDC): paralelní korpusy pro páry arabština-angličtina, čínština-angličtina atd.
  Gigaword korpus (angličtina, 7 mld slov)
- Europarl: kolekce textů Evropského parlamentu (11 jazyků, 40 M slov)
- OPUS: paralelní texty různého původu (lokalizace software)
- Acquis Communautaire: právní dokumenty Evropské únie (20 jazyků)

Pravidelné události v oblasti SMT, soutěže

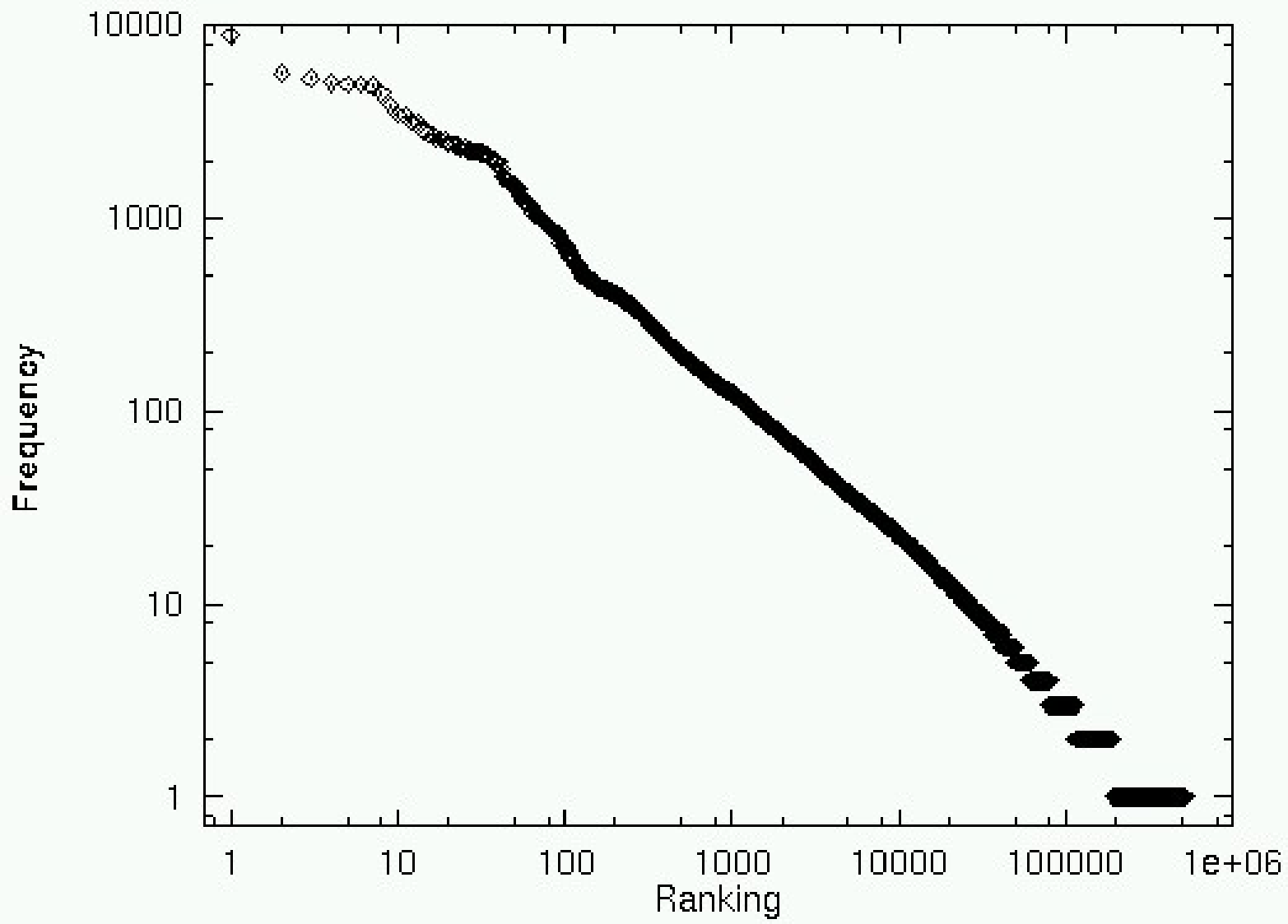
Většinou roční vyhodnocování kvality SMT. Tvorba testovacích sad, manuální vyhodnocování dat, referenční systémy.

- NIST: National Institute of Standards and Technology; nejstarší, prestižní; hodnocení překladu arabštiny, čínštiny
- IWSLT: mezinárodní workshop překladu mluveného jazyka; překlad řeči; asijské jazyky
- WMT: Workshop on SMT; překlady mezi evropskými jazyky

Slova

- pro SMT v drtivé většině případů základní jednotka = slovo
- v mluvené řeči slova neoddělujeme: jak je od sebe oddělíme?
- SMT systémy provádí de-tokenizaci
- překlad samotný je většinou s lowercase textem
- jaká slova má angličtina $\rightarrow$ jaká slova jsou v anglických korpusech
- *the* tvoří 7 % anglického textu
- 10 nejčastějších slov (tokenů) tvoří 30 % textu (!)
- *Zipfův zákon*: r rank (pořadí ve frekvenčním seznamu slov), f = frekvence výskytu slova, c = konstanta; platí $r \times f = c$
- překlepy, čísla, vlastní jména, názvy a cizí slova

Zipfův zákon



Věty

- syntaktická struktura se v jazycích liší
- vkládání funkčních slov, která jsou typická pro daný jazyk (*the*, interpunkce)
- přerovnávání: *er wird mit uns gehen* $\rightarrow$ *he will go with us*
- některé jevy nelze přeložit na úrovni věty: anafory
- úroveň celého dokumentu: téma (topic) může pomoci při volbě vhodného překladového ekvivalentu
- v textu o jeskynních živočiších zřejmě nebude překládat *bat* jako *pálka*

Paralelní korpusy

- základní datový zdroj pro SMT
- volně dostupné jsou řádově 10 a 100 miliónů slov veliké
- je možné stáhnout paralelní texty z internetu
- vícejazyčné stránky (BBC, Wikipedie)
- problém se zarovnáním dokumentů, odstavců, . . .
- srovnatelné korpusy (comparable corpora): texty ze stejné domény, ne přímé překlady: New York Times – Le Monde
- Kapradí – korpus překladů Shakespearových dramat (FI)
- InterCorp – ručně zarovnané beletr. texty (ČNK, FFUK)

Zarovnávání vět

- věty si neodpovídají 1:1
- některé jazyky explicitně nenaznačují hranice vět (thajština)
- *It is small, but cozy. – Es is klein. Aber es ist gemütlich.*
- pro věty $e_1, \cdots e_{n_e}$ a $f_1, \cdots f_{n_f}$
- hledáme páry $s_1, \ldots s_n$
- $s_i = (\{f_{\text{start}-f(i)}, \cdots f_{\text{end}-f(i)}\}, \{e_{\text{start}-e(i)}, \cdots e_{\text{end}-e(i)}\})$

P	**typ zarovnání**
0.98	1–1
0.0099	1–0 nebo 0–1
0.089	2–1 nebo 1–2
0.011	2–2

Pravděpodobnostní rozložení

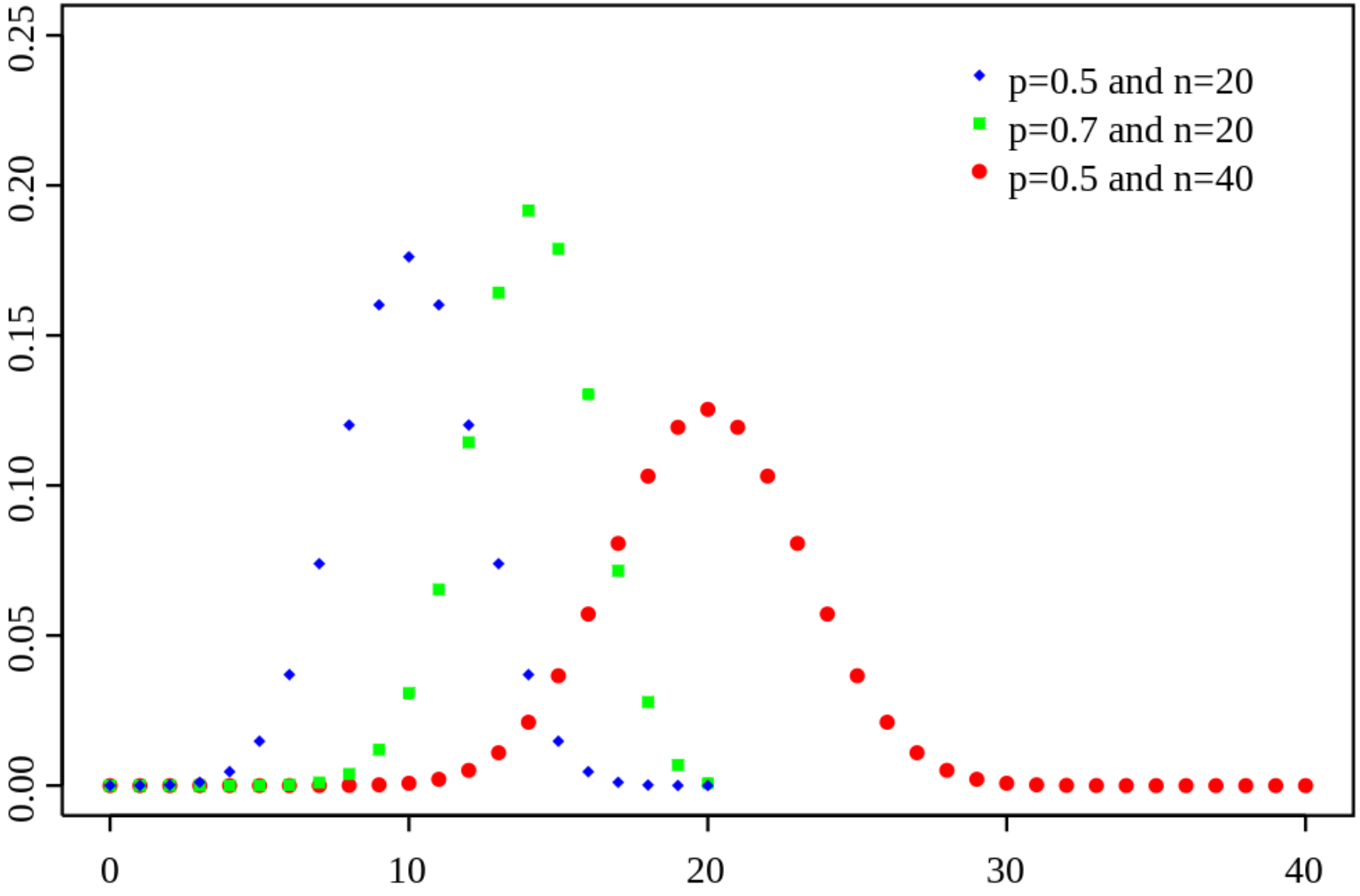
- graf hodnot pravděpodobnosti pro elementární jevy náhodné veličiny
- **rovnoměrné**: hod kostkou, mincí (diskrétní veličina)
- **binomické**: vícenásobný hod

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

- **normální, Gaussovo**: spojité, dobře aproximuje ostatní rozložení; zahrnuje rozptyl

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

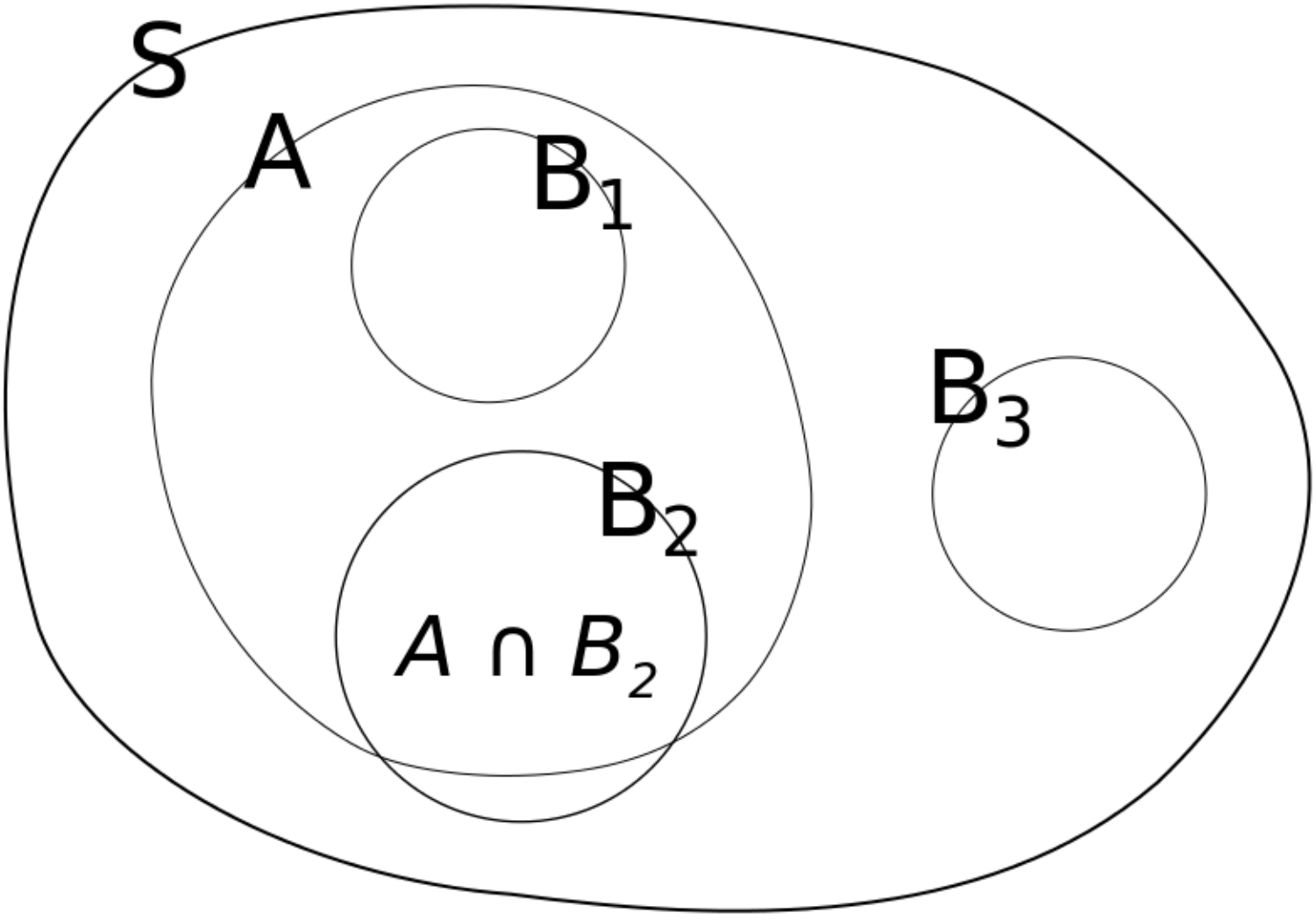
Binomické rozložení



Statistika I

- náhodná proměnná, pravděpodobnostní funkce, ...
- máme data, chceme spočítat rozložení, které nejlépe tato data vystihuje
- **zákon velkých čísel**: čím víc máme dat, tím lépe jsme schopni odhadnout pravděpodobnostní rozložení
- např.: hod falešnou kostkou; výpočet π
- nezávislé proměnné: $\forall x, y : p(x, y) = p(x).p(y)$
- **spojená (joint) pravděpodobnost**: hod mincí a kostkou
- **podmíněná pravděpodobnost**: $p(y|x) = \frac{p(x,y)}{p(x)}$
  pro nez. proměnné platí: $p(y|x) = p(y)$

Podmíněná pravděpodobnost



Shannonova hra

Pravděpodobnostní rozložení pro následující znak v textu se liší v závislosti na předchozích znacích.

Doplňujeme postupně znaky (malá abeceda a mezera).
Některé znaky nesou více informace (jsou uhádnuty později).

Bayesovo pravidlo

$$p(x|y) = \frac{p(y|x).p(x)}{p(y)}$$

- příklad s kostkou
- $p(x)$ – prior
- $p(y|x)$ – posterior

Statistika II

- střední hodnota (diskrétní): $\mathrm{E}\,X = \sum_I s_i.p_i$
- rozptyl: $\sigma^2 = \sum_{i=1}^{n} [x_i - \mathrm{E}(X)]^2 p_i$
- očekávaná hodnota: $\mathrm{E}[X] = \sum_{x \in X} x.p(x)$

SMT – princip noisy channel

Vyvinut Shannonem (1948) pro potřeby samoopravujících se kódů, pro korekce kódovaných signálů přenášených po zašuměných kanálech na základě informace o původní zprávě a typu chyb vznikajících v kanálu.

Příklad s OCR. Rozpoznávání textu z obrázků je chybové, ale dokážeme odhadnout, co by mohlo být v textu (jazykový model) a jaké chyby často vznikají: záměna l-1-I, rn-m apod.

$$\begin{aligned} e^* &= \arg\max_e p(e|f) \\ &= \arg\max_e \frac{p(e)p(f|e)}{p(f)} \\ &= \arg\max_e p(e)p(f|e). \end{aligned}$$

SMT – komponenty noisy channel principu

- jazykový model:
  - jak zjistit $p(e)$ pro libovolný řetěz e
  - čím víc vypadá e správně utvořené, tím vyší je $p(e)$
  - problém: co přiřadit řetězci, který nebyl v trénovacích datech?
- překladový model:
  - pro e a f vypočítej $p(f|e)$
  - čím víc vypadá e jako správný překlad f, tím vyšší p
- dekódovací algoritmus
  - na základě předchozího najdi pro větu f nejlepší překlad e
  - co nejrychleji, za použití co nejmenší paměti