Počítačové zpracování češtiny

Kontrola pravopisu

Daniel Zeman

http://ufal.mff.cuni.cz/daniel-zeman/

Úloha

- Rozpoznat slovo, které není ve slovníku
  - Triviální
  - Těžší je rozpoznat slovo, které ve slovníku je, ale v daném kontextu je nepravděpodobné
- Navrhnout opravy
  - Nejpravděpodobnější nejdříve
  - Seznam 100 oprav nemá smysl

Jak navrhnout opravu

- Hledáme ve slovníku podobná slova
- Které slovo je podobné?
- Editační vzdálenost: počet operací
  - Vynechání písmene (deletion)
  - Přidání písmene (insertion)
  - Záměna písmene za jiné (replacement)
  - Prohození dvou sousedních písmen
  - Záměnu a prohození lze simulovat posloupností vynechání a přidání, ale psychologicky jde o samostatný jev

Pravděpodobnost opravy

- Pravděpodobnost, že slovo *w* má být opraveno na slovo *c* (correction)
- Bayesův vzorec:

$$P(c|w) = \frac{P(w|c) \cdot P(c)}{P(w)}$$

Pravděpodobnost opravy

$$P(c|w) = \frac{P(w|c) \cdot P(c)}{P(w)}$$

$$\arg\max_c P(c|w)$$
$$= \arg\max_c \left(P(w|c) \cdot P(c)\right)$$

- P(c|w) … pravděpodobnost, že *c* je opravou *w*
- P(w) … pravděpodobnost opravovaného slova. Nemá vliv na nalezení nejlepšího *c*.
- P(c) … pravděpodobnost náhradního slova. Je dána jazykovým modelem.
- P(w|c) … pravděpodobnost zkomolení správného slova *c* právě na tvar *w*. Je dána chybovým modelem.

Model kanálu s šumem

- *Noisy channel* (např. rádiové spojení, kvůli šumu jsme některá slova dostali chybná)
- Bayesův vzorec se na rozklad používá často
- Jazykový model: n-gram, viz značkování
- Chybový model je složitější

Jazykový model

- Pravděpodobnosti se naučíme z korpusu
- N-gramový model, ale na rozdíl od značkování
  - Neodhadujeme pravděpodobnost značky
  - Odhadujeme pravděpodobnost slova
  - Na základě kontextu předcházejících slov
- Podobně slouží jazykový model jako pomocný model ve strojovém překladu, při rozpoznávání řeči a jinde.

Chybový model

- Potřebovali bychom korpus chyb a jejich oprav
  - Např. pro angličtinu existuje
- Nemáme-li, pomůžeme si heuristikami

Heuristiky pro chybový model

- Pravděpodobnost chyby klesá s editační vzdáleností mezi správným a chybným slovem
- Editační vzdálenost je počet editačních operací
  - Vynechání písmene
  - Přidání písmene
  - Záměna jednoho písmene za jiné
  - Prohození dvou sousedních písmen
- Není triviální zjistit editační vzdálenost mezi dvěma slovy
  - Není snadné najít nejkratší posloupnost editací (minimální cesta v grafu, graf je ale obrovský)

Heuristiky pro chybový model

- Některé chybné editace jsou pravděpodobnější než jiné
  - Zdvojení písmene není totéž co vložení libovolného znaku (autor si nebyl jist pravopisem)
  - Záměna *a* za *e* je asi pravděpodobnější než *a* za *t* (blízká výslovnost)
  - Písmeno s diakritikou je podobné témuž písmenu bez diakritiky
  - Písmena, která jsou vedle sebe na klávesnici, se zamění nebo přidají snadněji
  - Vynechání nebo přidání mezery je stejně snadné jako úhoz na jinou klávesu, ale změní nám počet slov a zkomplikuje celou úlohu

Trénování

- Slovník
  - Nemáme-li, lze naučit z korpusu
    - Ale pozor, v korpusu bývají překlepy!
- Jazykový model
  - Z korpusu, překlepy potlačí statistika
- Chybový model
  - Speciální korpus nebo heuristiky
- Chtělo by to alespoň malý korpus chyb kvůli vyhodnocení