Statisztikai gépi fordítás

Simon Eszter

2020. május 22.

Egy zajos csatorna modellen alapuló statisztikai gépi fordító rendszer 3 komponensre osztható:

- nyelvmodell (*language model*)
- fordítási modell (*translation model*)
- dekóder (*decoder*)

A kiinduló képlet:

$$\hat{E} = argmax P(F|E)P(E)$$

ahol $P(F|E)$ a fordítási modell (*translation model*) és $P(E)$ a nyelvmodell (*language model*).

Nyelvmodell. A nyelvmodell ugyanúgy számítódik ki, ahogy korábban láttuk az n-gram alapú nyelvmodellezés esetében.

Fordítási modell. A fordítási modell is két komponensből tevődik össze:

$$P(F|E) = \prod \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1})$$

ahol $\phi(\bar{f}_i, \bar{e}_i)$ a fordítási valószínűség (*translation probability*) és $d(a_i - b_{i-1})$ a szórendbeli különbséget büntető modell (*distortion model*).

A frázisalapú gépi fordítás (*phrase-based MT*) esetén a fordítási valószínűséget kell kitanulni frázis szinten illesztett párhuzamos szövegekből. Ha van ilyen gold standard adatunk, akkor ki tudjuk számolni:

$$\phi(\bar{f}, \bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e}))}$$

Az $(\bar{f}, \bar{e})$ párokat a hozzájuk tartozó $\phi(\bar{f}, \bar{e})$ értékekkel egy nagy frázisfordítási táblában (*phrase translation table*) tároljuk.

De: mi van, ha nincs frázis szinten kézzel illesztett adatunk? Ilyenkor szószintű illesztésből (*word alignment*) kell kinyerni a frázisszintű illesztéseket. Ezt a feladatot mondat szinten illesztett párhuzamos szövegekből tudjuk megoldani.

Dekóder. A dekóder végzi el az $\hat{E}$ maximalizálását.