Többszavas kifejezések használata kétnyelvű prediktív szótárakban

Szerző: Szabó Gábor

Konzulens: Dr. Juhász Sándor

AAIT

MSc, Mérnök Informatikus

2014

Változtat-e az elérhető gépelés megtakarításon többszavas kifejezések használata egy prediktív szótárban?

Miről lesz szó?

- Feladat, kontextus
- Szótár alapú módszer
- Mérési cél, paraméterek
- Eredmények
- Összegzés

Feladat

- Fordítás (translation) prediktív gépeléssel
- Szótár alapú módszerrel
- Egynyelvűhöz képest több információ
  - Aktuálisan fordított mondat

PREDICTIVE TEXT

It's like Marmite.
You either love it or hate it.
There is no in between!

I'll be flying to Toronto next

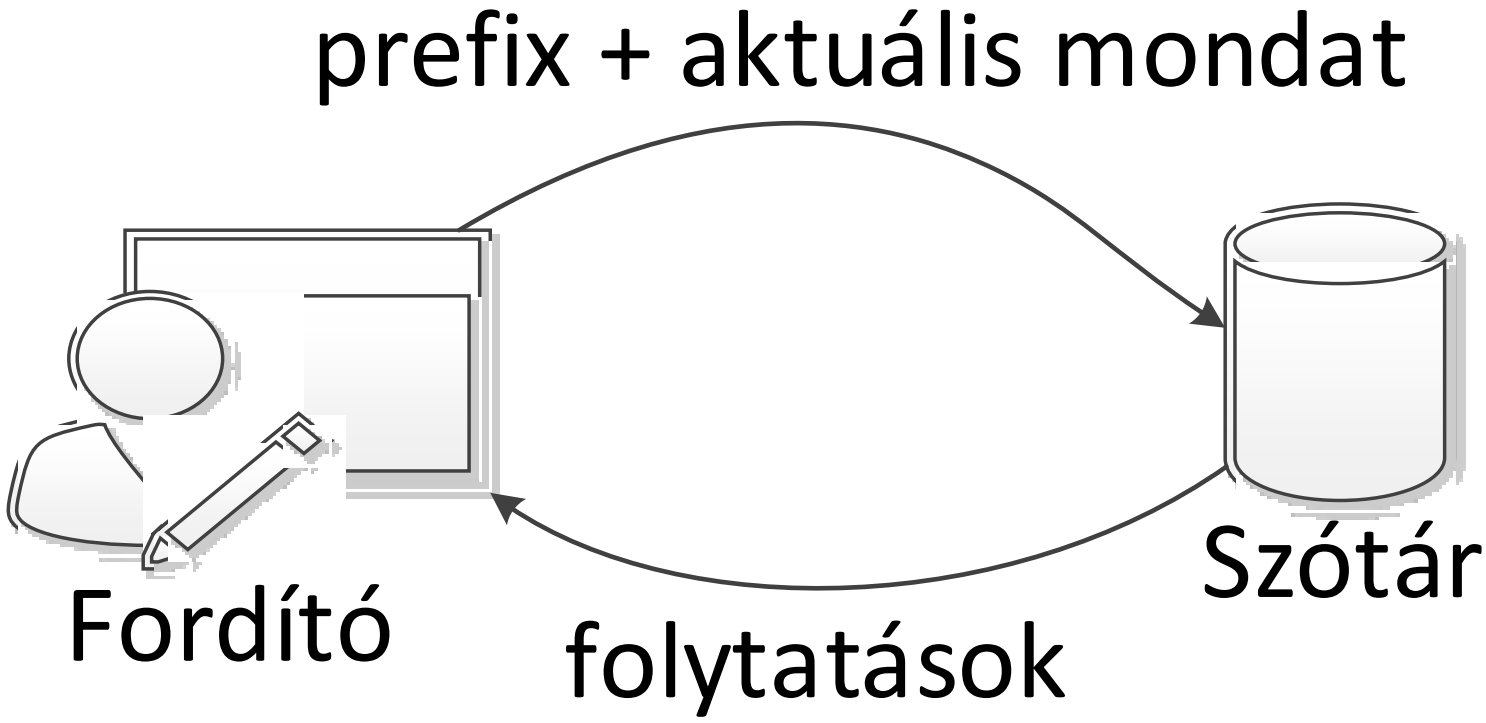
Sent from my Windows Phone

Szótár felépítése

- Forrás nyelvű kifejezésekhez egy lista a lehetséges fordításokról és azok „valószínűségéről”

Forrás kif.	Fordítások
I would	(möchte; 0.45), (möchte ich; 0.31)
new	(neuen; 0.55), (neue; 0.49)

Szótár használata



Szótár készítése

Kétnyelvű párhuzamos szövegtestből

1. Gyakori **kifejezések** kigyűjtése nyelvenként (küszöb! 1‰)
2. Gyakori **kifejezések** együttes előfordulásának megszámolása
3. Együttes és külön-külön előfordulás alapján score
4. Score alapján bekerül a szótárba (küszöb!)

$$score = 2 \cdot \frac{N_{együtt}}{N_{forráskif} + N_{célkif}}$$

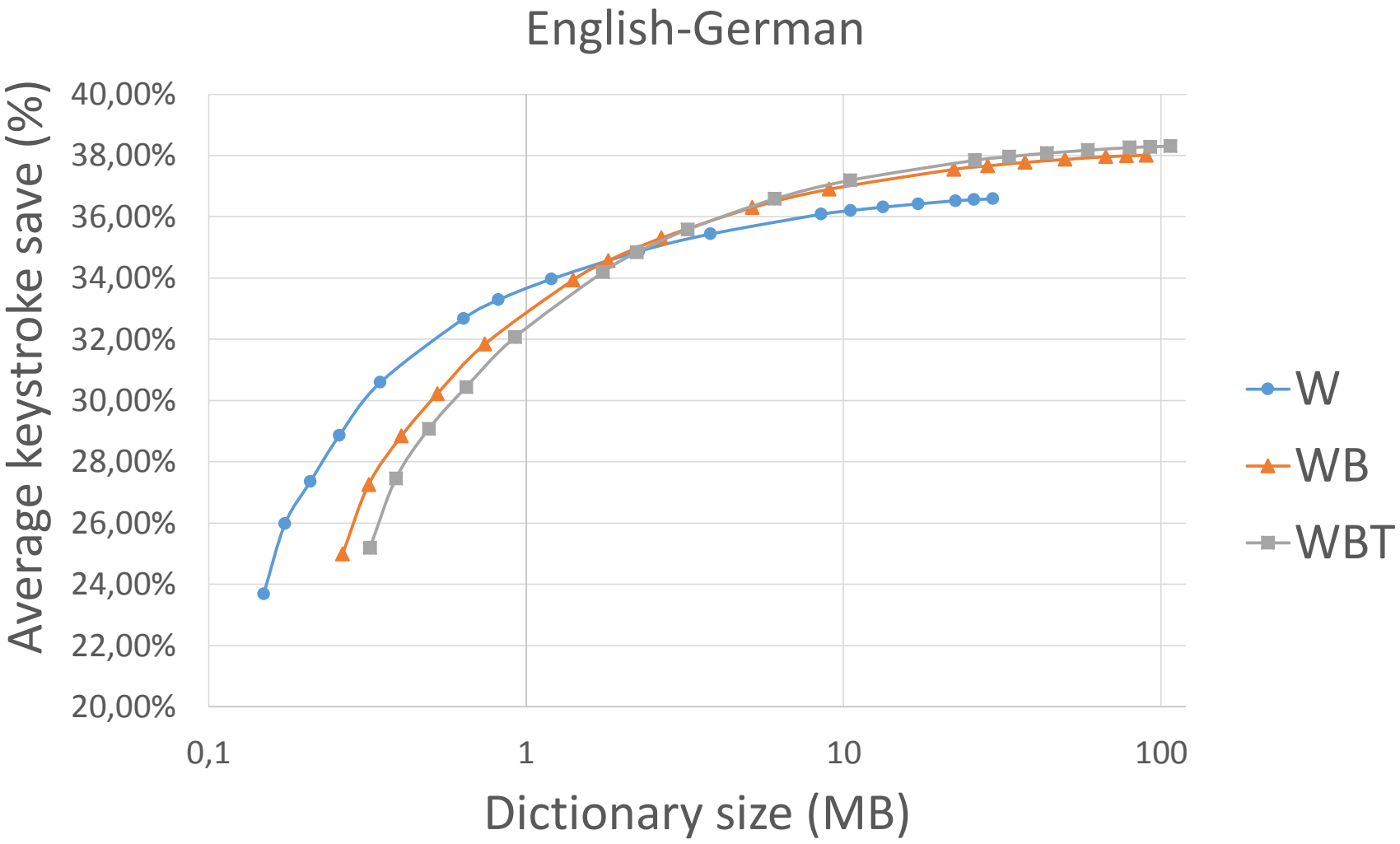
Mérési cél

- Különböző **kifejezés képzési módok** összehasonlítása
  - **csak szavak (W)**
  - **szavak és bigramok (WB)**
  - **szavak, bigramok és trigramok (WBT)**
- Jóság mértéke: **gépelés spórolás szótárméret függvényében**
  - méret befolyásolása küszöb változtatással

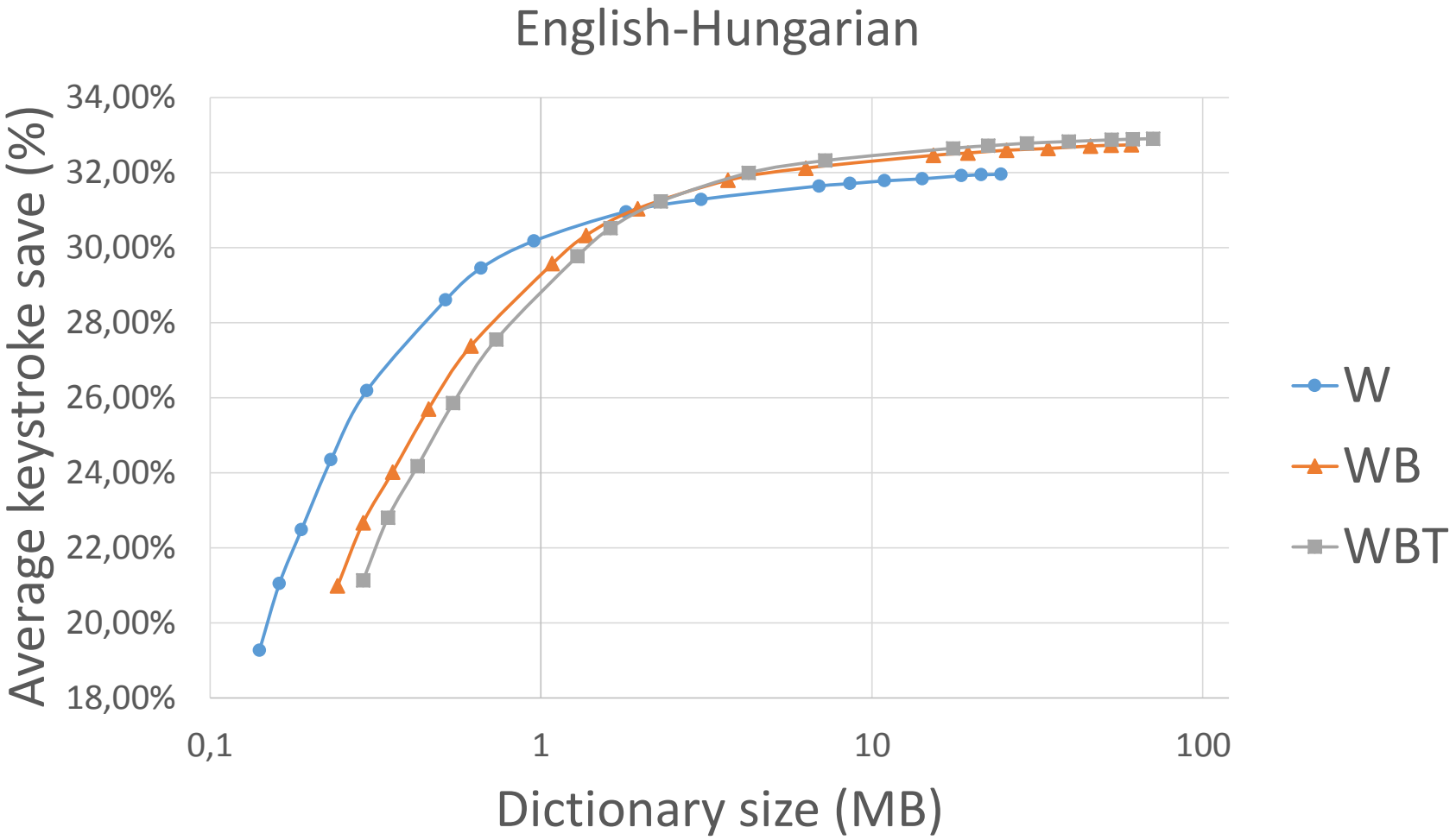
Mérési paraméterek

- Szövegtest:
  Proceedings of the European Parliament
- Vizsgált nyelvpárok:
  angol-német, angol-magyar
- 200k mondatpár, több különböző min. score
- 20k mondatpáron gépelés spórolás mérése

Eredmények


English-German
Average keystroke save (%)
40,00%
38,00%
36,00%
34,00%
32,00%
30,00%
28,00%
26,00%
24,00%
22,00%
20,00%
0,1
1
10
100
Dictionary size (MB)
W
WB
WBT

Eredmények


English-Hungarian
Average keystroke save (%)
34,00%
32,00%
30,00%
28,00%
26,00%
24,00%
22,00%
20,00%
18,00%
0,1
1
10
100
Dictionary size (MB)
W
WB
WBT

Összegzés

- Kifejezés képzési mód képes változtatni a szótár prediktív jóságán (és méretén is)
- Nyelvfüggő (nyelvcsalád?), hogy mennyit javít a bigramok és trigramok bevétele
- Lehetséges irányok:
  - További kifejezés képzési módszerek vizsgálata
  - Több nyelvpárra kiterjedően
  - Score számító képletek összehasonlítása

Hivatkozások

- OPUS - http://opus.lingfil.uu.se/
- Google Image Search