E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Motiváci

Létező megoldások

Létező megoldáso

Kutatáci tan

Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok

Kell ide egy cím

A statisztikus gépi fordítás hatékonyságának jelentés egyértelműsítéssel történő javítása

E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Babeș-Bolyai Tudományegyetem, Matematika és Informatika Kar, Kolozsvár

2016 április 15.

Létező megoldások

Kutatási terv Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok

Tartalom

- Motiváció
- 2 Létező megoldások
- 3 Létező megoldások
- 4 Kutatási terv

Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok Munkabeosztás és ütemtery

Létező megoldások

Létező megoldáso

Kutatási ten

Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok

Munkabeosztás

Motiváció

- Szövegek fordítása a számítógép által
- Statisztikus gépi fordítás (SMT)
- Filmfeliratok használata tanulási adatként
- Hatékonyságon javítani
- Jelentés egyértelműsítése (WSD)

E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Motivácio

Létező megoldások

Létező megoldások

Kutatási terv

Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó Létező megoldások

e: Good morning!

f: Bon jour!



$$P(e|f) = \frac{P(e)P(f|e)}{P(f)} \tag{1}$$

$$T(e) = \hat{e} = argmax_e P(e|f) = argmax_e P(e)P(f|e)$$
 (2)

Motivácio

Létező megoldásol

Létező megoldások

Kutatási terv Megtenni

szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok

Megvalósítano anyagok Munkabeoszta és ütemterv

Létező megoldások

$$T(e) = \hat{e} = argmax_e P(e)P(f|e)$$

- nyelvi modell: P(e)
 - folytonosságot biztosít a célnyelvben
- fordítási modell: P(f|e)
 - lexikális megfeleltetés a nyelvek között
 - szó alapú (word based) modellek [?] [?]
 - kifejezés alapú (phrase based) modellek [?] [?]
- argmax: argmax_e
 - keresés

Létező megoldáso

Létező megoldások

Kutatási ten

Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok

Létező megoldások Előnyök, hátrányok

Szó alapú vs kifejezés alapú modellek:

- szó alapú modell: nehéz a tokeninzálás [?]
- kifejezés alapú modell pontosabb komplex nyelveknél [?]

Hátrányai ezeknek a rendszereknek:

 Még mindig nem elég pontosak az SMT rendszerek a WSD-hez képest [?]

E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Motivácio

Létező megoldásol

Létező megoldások

Kutatási ten

Megtenni szándékozott lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok Munkabeosztás és ütemterv

Létező megoldások

Aktuális próbálkozások a javításra

• WSD beépítése az SMT-be: [?] [?]

Létező megoldások

Létező megoldásol

Kutatási ter

Megtenni szándékozott lépések

Csapat tagjainak hozzájárulása Megvalósítandó anyagok Munkabeosztás

Megtenni szándékozott lépések

- Wikipedia alapú WSD rendszer felépítése
 - 1 Egyik legnagyobb létező adathalmaz
 - 2 Több nyelven elérhető
 - 3 A cikkek címei felhasználhatóak, mint conceptek
- WSD rendszer integrálása a létező SMT rendszerekbe
 - Statisztikai modellek nem veszik figyelembe többértelmű szavakat
 - 2 A fordítás pontossága nagyban növelhető lenne ezen problémák megoldásával

E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A Szász

Motiváci

Létező megoldások

Létező megoldások

Kutatási terv

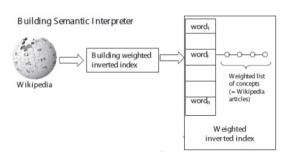
Megtenni szándékozott

lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok Munkabeosztás

Megtenni szándékozott lépések

Wikipedia alapú WSD rendszer

- 1 Invertált index felépítése minden nyelv számára
 - Minden szóhoz hozzárendelünk concepteket és a hozzájuk tartozó súlyt
 - A conceptek az angol Wikipedia címek
 - Súlyozáshoz használhatjuk például a tf-idf súlyozást



E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A Szász

Motivácio

Létező megoldások

Létező megoldásol

Kutatási terv

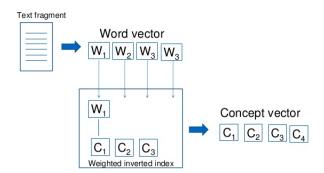
Megtenni szándékozott

lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok Munkabeosztás

Megtenni szándékozott lépések

Wikipedia alapú WSD rendszer

- Súlyozott concept vektor hozzárendelése a fordítandó és SMT által fordított szöveghez
 - Szöveg ábrázolása, mint szavak vektora
 - Conceptek megfelelő súllyal történő hozzárendelése minden szóhoz
 - Ősszesített súlyozott concept lista felépítése a szöveghez



cím
E. Bordi,
T. Both,
Sz. Pável,
Cs. Sándor.

A Szász

Kell ide egy

Motiváció

Létező megoldások

Létező megoldásol

Kutatási terv

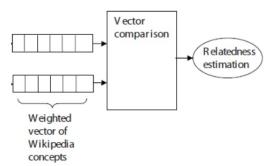
Megtenni szándékozott lépések

Csapat tagjainak hozzájárulása Megvalósítandó anyagok Munkabeosztás

Megtenni szándékozott lépések

Wikipedia alapú WSD rendszer

- 3 Szemantikai hasonlóság meghatározza a fordítandó és célszöveg között
 - A fordítandó és célszöveghez felépítjük a concept vektorokat
 - A hasonlóság vizsgálatát a vektorok összehasonlítása jelenti
 - A hasonlóság metrika lehet például a gyakran használt cos távolság



E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Motivácio

Létező megoldásol

Létező megoldáso

Kutatási ten

Megtenni szándékozott

lépések Csapat tagjainak hozzájárulása Megvalósítandó anyagok

Megtenni szándékozott lépések WSD súlyok integrálása

- 4 Legjobb N találat újrarangsorolása
 - SMT rendszer alapján meghatározni a legjobb N fordítást
 - WSD rendszer alapján súly hozzárendelése a találatokhoz
 - A találatok újrarangsorolása az SMT és WSD együttes eredményei alapján

E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Motivácio

Létező megoldásol

Létező megoldáso

Kutatási terv

Megtenni

Csapat tagjainak hozzájárulása

hozzájárulása Megvalósítandó

Munkabeosztás és ütemterv

Csapat tagjainak hozzájárulása

Feladatkör	Csapattag
Megvalósítandó anyagok	Bordi Eszter
Munkabeosztás	
Gantt diagram	
Motiváció	Both Tibor
Megtenni szándékozott lépések	Pável Szabolcs
Létező megoldások	Sándor Csanád
Csapat tagjainak hozzájárulása	Szász Adorján

E. Bordi, T. Both. Sz. Pável, Cs. Sándor, A. Szász

Csapat tagjainak

Megvalósítandó anyagok Munkabeosztás és ütemterv

Megvalósítandó anyagok

Motivácio

megoldás

megoldás

Kutatası ter

lépések
Csapat tagjainak

hozzájárulása Megvalósítandó anyago

Munkabeosztás és ütemterv

Munkabeosztás és ütemterv

Munkaegységek:

- A kutatás beindítása, adatgyűjtés
- Az adathalmaz összeállítása
- Algoritmus kidolgozása
- Az új és régi módszerek összevetése
- Az algoritmus optimalizálása, prototípus-fejlesztés

megoldáso

megoldás

Kutatasi terv

lépések Csanat tanjainak

Csapat tagjainak hozzájárulása

Megvalósítandó any

Munkabeosztás és ütemterv

- A kutatás beindítása, adatgyűjtés
 - Kutatási módszertani lehetőségek felvázolása
 - 2 Kutatási terv kidolgozása
 - Meglévő módszerek feltérképezése
 - 4 Prototípus első koncepciójának kidolgozása
- Az adathalmaz összeállítása
 - Adatok gyűjtése
 - Standard formátumra való alakítás
 - Adathalmaz validálása
- Algoritmus kidolgozása
 - 1 Algoritmus leírása
 - 2 Teszt jegyzőkönyvek, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata

megoldáso

Létező megoldás

Kutatası terv

lépések Csapat tagjainak

Csapat tagjainak hozzájárulása

Munkabeosztás és ütemterv

Munkabeosztás és ütemterv

- A kutatás beindítása, adatgyűjtés
 - Kutatási módszertani lehetőségek felvázolása
 - 2 Kutatási terv kidolgozása
 - 3 Meglévő módszerek feltérképezése
 - Prototípus első koncepciójának kidolgozása
- Az adathalmaz összeállítása
 - Adatok gyűjtése
 - 2 Standard formátumra való alakítás
 - 3 Adathalmaz validálása
- Algoritmus kidolgozása
 - 1 Algoritmus leírása
 - 2 Teszt jegyzőkönyvek, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata

megoldaso Létező

megoldás

Kutatası terv

lépések Csapat tagjainak

hozzájárulása Megvalósítandó anya

Munkabeosztás és ütemterv

Munkabeosztás és ütemterv

- A kutatás beindítása, adatgyűjtés
 - 1 Kutatási módszertani lehetőségek felvázolása
 - 2 Kutatási terv kidolgozása
 - 3 Meglévő módszerek feltérképezése
 - 4 Prototípus első koncepciójának kidolgozása
- Az adathalmaz összeállítása
 - Adatok gyűjtése
 - 2 Standard formátumra való alakítás
 - 3 Adathalmaz validálása
- Algoritmus kidolgozása
 - 1 Algoritmus leírása
 - 2 Teszt jegyzőkönyvek, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata

Munkahoneztás és

Munkabeosztás és ütemterv

- Az új és régi módszerek összevetése
 - Tesztfuttatások különböző algoritmusokkal és adatokkal, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata
 - Az eltérés statisztikus validációja
 - Szintézis
- Az algoritmus optimalizálása, prototípus-fejlesztés

Munkabeosztás és ütemterv

Munkaegységek és részfeladatai:

Az új és régi módszerek összevetése

- Tesztfuttatások különböző algoritmusokkal és adatokkal, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata
- Az eltérés statisztikus validációja
- Szintézis
- Az algoritmus optimalizálása, prototípus-fejlesztés
 - Performanciaoptimalizált algoritmus
 - 2 A gépi fordító prototípusának tesztelése és tesztjegyzőkönyvek
 - Nyilvánosságra hozandó eredmények dokumentációja
 - A projekt lezárása, dokumentáció

Munkahoneztás és

E. Bordi, T. Both, Sz. Pável, Cs. Sándor, A. Szász

Motiváció

Motivacio

Létező

megoldás

Kutatási ter

lépések

Csapat tagjainak

hozzájárulása Menyalósítandó ar

Munkabeosztás és űtemterv

Munkabeosztás és ütemterv

WBS 1 Summary Element 1

WBS 1.1 Activity A

WBS 1.2 Activity B

WBS 1.3 Activity C

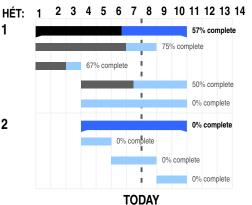
WBS 1.4 Activity D

WBS 2 Summary Element 2

WBS 2.1 Activity E

WBS 2.2 Activity F

WBS 2.3 Activity G







A. Author.

Handbook of Everything.

Some Press, 1990.



S. Someone.

On this and that.

Journal of This and That, 2(1):50-100, 2000.