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.

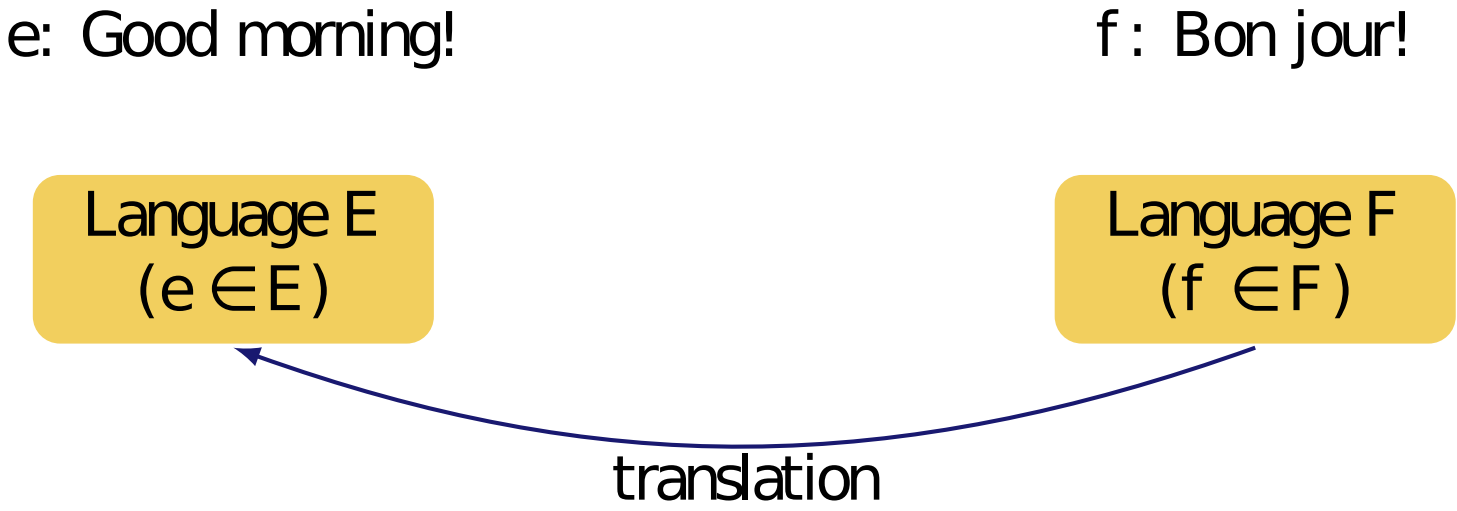
Tartalom

Motiváció

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

Létező megoldások

Statistical machine translation



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

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

Létező megoldások

Statistical machine translation

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

- 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 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 [**?**]

Létező megoldások

Aktuális próbálkozások a javításra

- WSD beépítése az SMT-be: [?] [?]

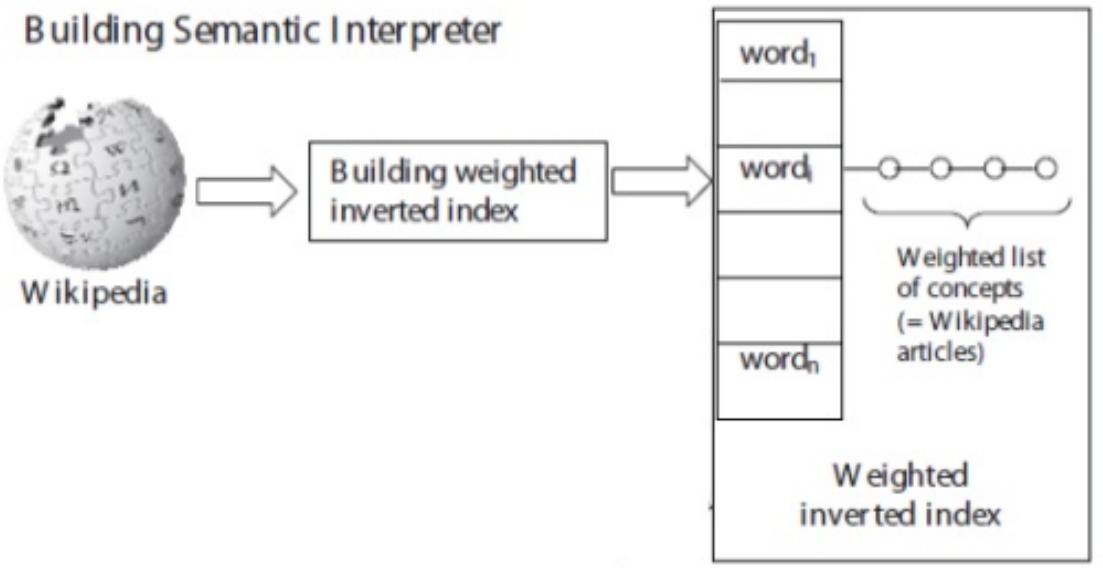
Megtenni szándékozott lépések

- 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
  1. Statisztikai modellek nem veszik figyelembe többértelmű szavakat
  2. A fordítás pontossága nagyban növelhető lenne ezen problémák megoldásával

Megtenni szándékozott lépések

Wikipedia alapú WSD rendszer

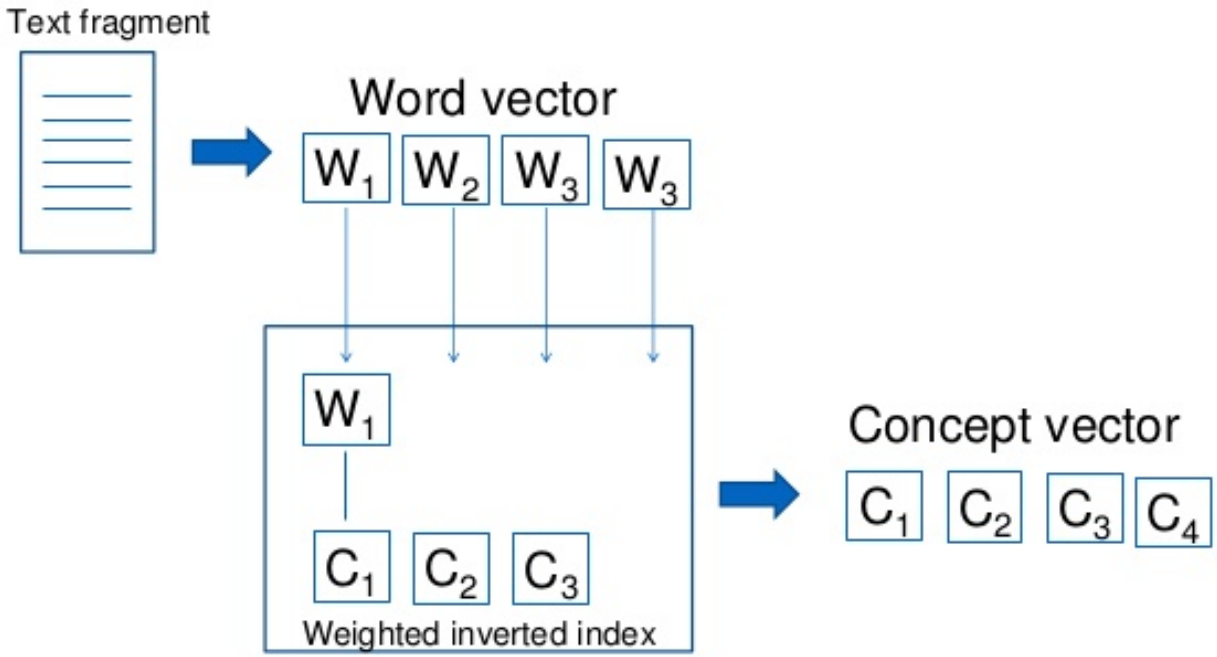
1. Invertált index felépítése minden nyelv számára
   - Minden szóhoz hozzárendelünk concepteket és a hozzájuk tartozó súlyt
   - A conceptek az angol Wikipedia címek
   - Súlyozáshoz használhatjuk például a tf-idf súlyozást



Megtenni szándékozott lépések

Wikipedia alapú WSD rendszer

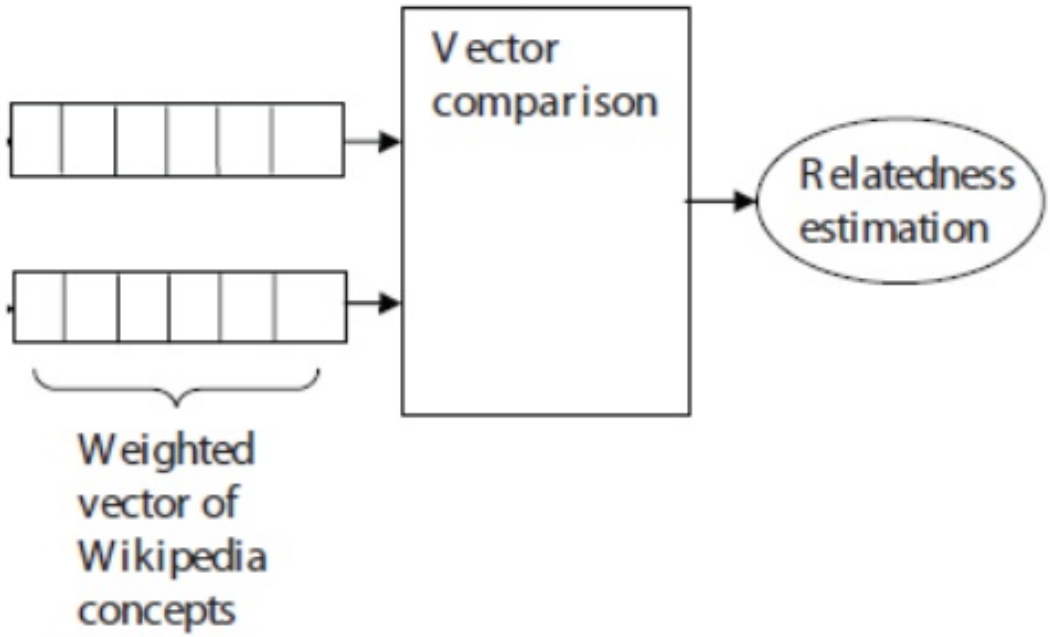
2. Súlyozott concept vektor hozzárendelése a fordítandó és SMT által fordított szöveghez
   - Szöveg ábrázolása, mint szavak vektora
   - Conceptek megfelelő súllyal történő hozzárendelése minden szóhoz
   - Összesített súlyozott concept lista felépítése a szöveghez



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



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

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

Megvalósítandó anyagok

Munkabeosztás és ütemterv

Munkaegységek:

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

Munkabeosztás és ütemterv

Munkaegységek és részfeladatai:

- 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
  1. 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

Munkabeosztás és ütemterv

Munkaegységek és részfeladatai:

- 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
  1. 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

Munkabeosztás és ütemterv

Munkaegységek és részfeladatai:

- 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
  1. 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

Munkabeosztás és ütemterv

Munkaegységek és részfeladatai:

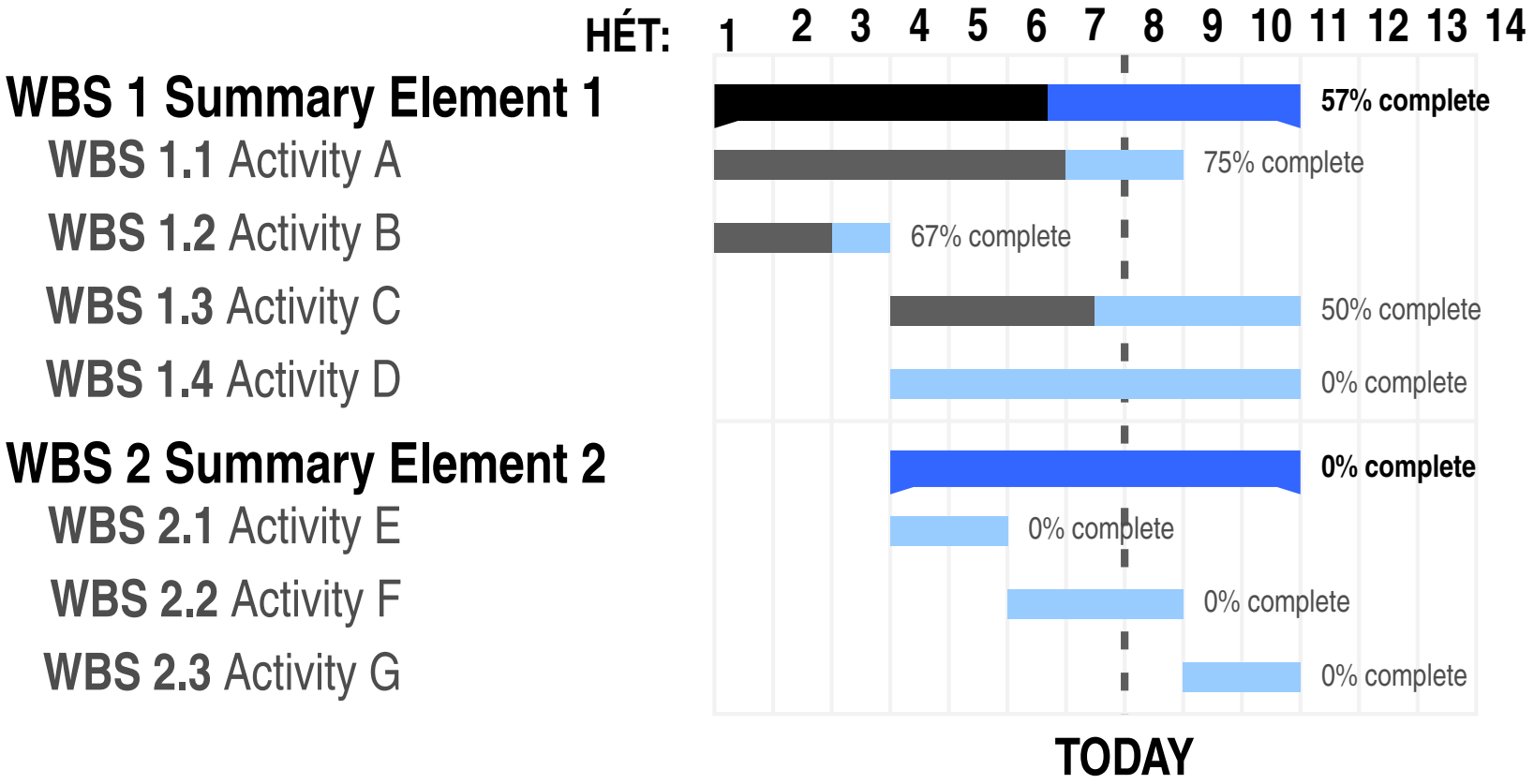
- Az új és régi módszerek összevetése
  1. Tesztfuttatások különböző algoritmusokkal és adatokkal, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata
  2. Az eltérés statisztikus validációja
  3. Szintézis
- Az algoritmus optimalizálása, prototípus-fejlesztés
  1. Performanciaoptimalizált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

Munkabeosztás és ütemterv

Munkaegységek és részfeladatai:

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

Munkabeosztás és ütemterv



Irodalomjegyzék I

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal of This and That, 2(1):50–100, 2000.