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.

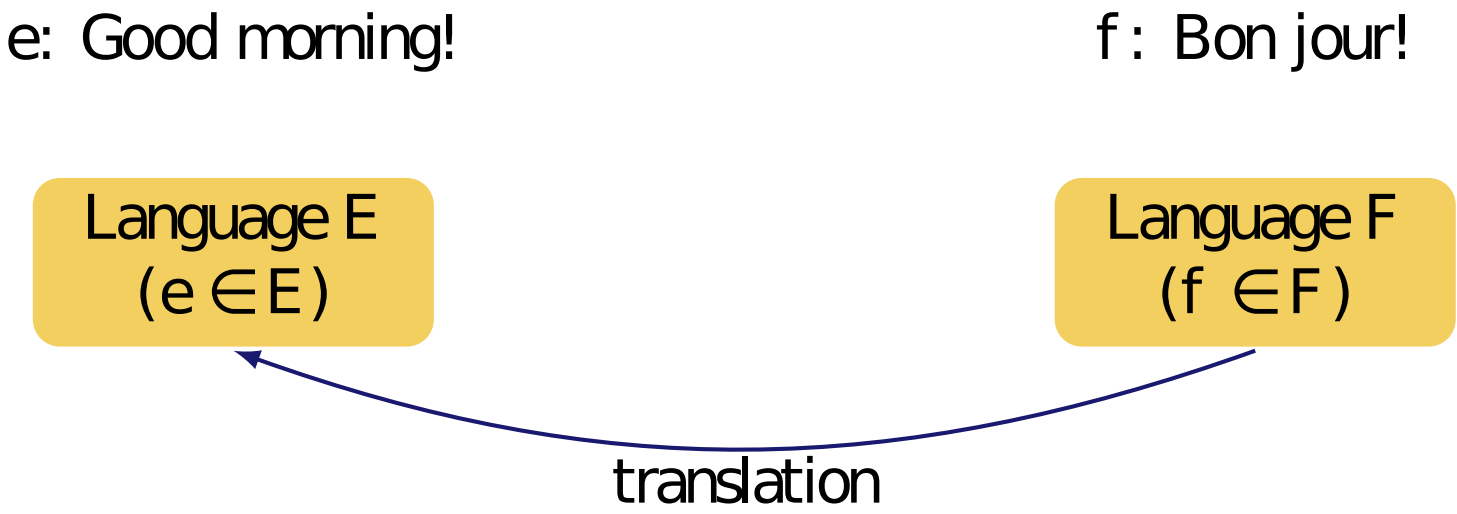
Tartalom

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 [BCP+90] [BBDP+94]
  - kifejezés alapú (phrase based) modellek [OTNI99] [MW02]
- argmax: $argmax_e$
  - keresés

Létező megoldások

Előnyök, hátrányok

Szó alapú vs kifejezés alapú modellek:

- szó alapú modell: nehéz a tokeninzálás [Lop07]
- kifejezés alapú modell pontosabb komplex nyelveknél [Lop07]

Hátrányai ezeknek a rendszereknek:

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

Létező megoldások

Aktuális próbálkozások a javításra

- WSD beépítése az SMT-be: [CW05] [CW07]

Megtenni szándékozott lépések

- Jelenleg legjobban teljesítő algoritmusok megtalálása
  1. Statistical Machine Translation módszerek keresése
  2. SMT algoritmusok telesítményének feltérképezése
- Jelentés egyérelműsítés alkalmazása
  1. Jelentésbeli közelséget vizsgáló módszerek felkutatása
  2. Modellek beépítése SMT módszerekbe
- Tanuló és teszt adathalmaz összeállítása
  1. Párosított mondatok adatbázisának felépítése
  2. Lehetséges források: filmfeliratok, TED talk feliratok
  3. Szemantikai közelség vizsgálatához Wikipedia cikkek indexelése

Megtenni szándékozott lépések

- Kiértékelési módszertan meghatározása
  1. Használt metrikák meghatározása
  2. Automata kiértékelés szerkesztése
  3. Módszer validálása manuális kiértékelés által
- Prototípus elkészítése
  1. Algoritmus implementálása
  2. Algoritmus alkalmazása valós környezetben

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:

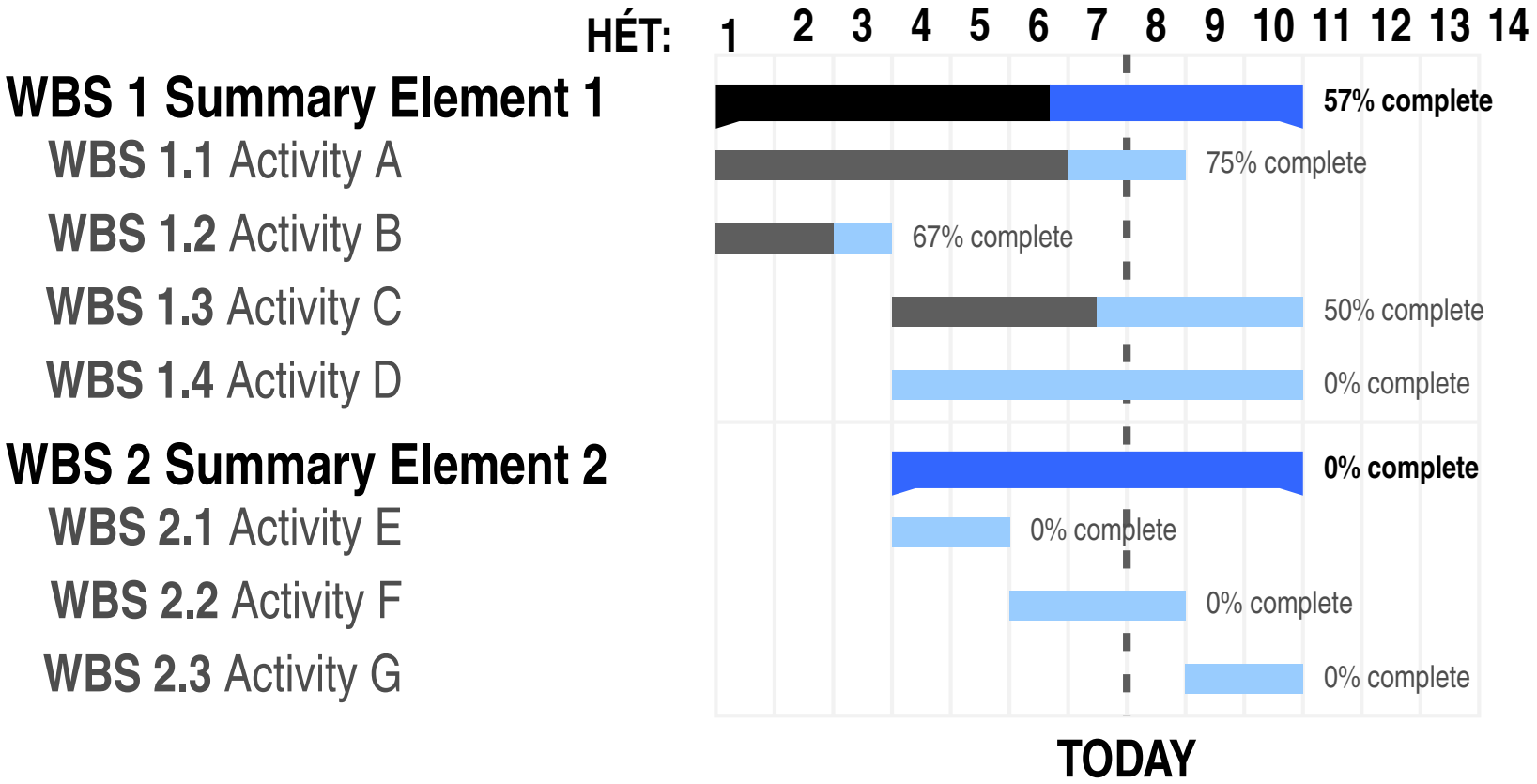
- Az új és régi módszerek összevetése
  1. Tesztfuttatások különböző algoritmusokkal és adatokkal, a statisztikus gépi fordítás hatékonyságával kapcsolatos hatások vizsgálata
  2. Az eltérés statisztikus validációja
  3. Szintézis
- Az algoritmus optimalizálása, prototípus-fejlesztés
  1. Performanciaoptimalizált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



Könyvészet I

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Luboš Ureš, *The candide system for machine translation*, Proceedings of the Workshop on Human Language Technology (Stroudsburg, PA, USA), HLT '94, Association for Computational Linguistics, 1994, pp. 157–162.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, *A statistical approach to machine translation*, Comput. Linguist. **16** (1990), no. 2, 79–85.

Könyvészet II

Marine Carpuat and Dekai Wu, *Evaluating the word sense disambiguation performance of statistical machine translation*, Second International Joint Conference on Natural Language Processing (IJCNLP-2005, p. 2005.

Marine Carpuat and Dekai Wu, *Word sense disambiguation vs. statistical machine translation*, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (Stroudsburg, PA, USA), ACL '05, Association for Computational Linguistics, 2005, pp. 387–394.

Könyvészet III

Marine Carpuat and Dekai Wu, *Improving statistical machine translation using word sense disambiguation*, In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 61–72.

Adam Lopez, *A survey of statistical machine translation*, Tech. report, 2007.

Daniel Marcu and William Wong, *A phrase-based, joint probability model for statistical machine translation*, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (Stroudsburg, PA, USA), EMNLP '02, Association for Computational Linguistics, 2002, pp. 133–139.

Könyvészet IV

Franz Josef Och, Christoph Tillmann, Hermann Ney, and Lehrstuhl Fiir Informatik, *Improved alignment models for statistical machine translation*, University of Maryland, College Park, MD, 1999, pp. 20–28.