

Gépi tanulási módszerek ómagyar kori szövegek normalizálására

Oravecz Csaba, Sass Bálint, Simon Eszter

VI. Magyar Számítógépes Nyelvészeti Konferencia

Szeged, 2009. december 3-4.



- 1 **Bevezető**
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- 1 **Bevezető**
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása

Feladat

összes elektronikus formában elérhető szöveg begyűjtése
egységes formátumra alakítása (normalizált változat)

Kérdés

manuális ábrási munka kiváltható-e gépi eljárással, az
embeli erőforrás alkalmazása leszűkíthető-e a tanuló
adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

összes elektronikus formában elérhető szöveg begyűjtése egységes formátumra alakítása (normalizált változat)

Kérdés

manuális ábrási munka kiváltható-e gépi eljárással, az emberi erőforrás alkalmazása leszűkíthető-e a tartó adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegmélek
 - középmagyar korból arányos válogatás

Feladat

összes elektronikus formában elérhető szöveg begyűjtése egységes formátumra alakítása (normalizált változat)

Kérdés

manuális ábrási munka kiváltható-e gépi eljárással, az emelkedett erőforrás alkalmazása leszűkíthető-e a tartalom adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

összes elektronikus formában elérhető szöveg begyűjtése egységes formátumra alakítása (normalizált változat)

Kérdés

manuális ábrási munka kiváltható-e gépjelárással, az emiatti erőforrás alkalmazása leszűrhető-e a fenti adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

- összes elektronikus formában elérhető szöveg begyűjtése
- egységes formátumra alakítása (*normalizált változat*)

Kérdés

• manuális ábrási munka kiváltható-e gépi eljárással, az emberi ellenőrzés alkalmazása leszűkíthető-e a fájlok adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

- összes elektronikus formában elérhető szöveg begyűjtése
- egységes formátumra alakítása (*normalizált változat*)

Kérdés

• manuális ábrási munka kiváltható-e gépi eljárással, az
entfernung alkalmazása leszűrhető-e a tartalom
adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

- összes elektronikus formában elérhető szöveg begyűjtése
- egységes formátumra alakítása (*normalizált változat*)

Kérdés

A manuális ábrás műve kiváltható-e gépi eljárással az
entire corpus alkalmazása és szűrése a
adatok előállításának feladataira?



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

- összes elektronikus formában elérhető szöveg begyűjtése
- egységes formátumra alakítása (*normalizált változat*)

Kérdés

- manuális átírási munka kiváltható-e gépi eljárással, az emberi erőforrás alkalmazása leszűkíthető-e a tanuló adatok előállításának feladatára



Cél

- elektronikus nyelvtörténeti adatbázis létrehozása
 - összes ómagyar szövegemlék
 - középmagyar korból arányos válogatás

Feladat

- összes elektronikus formában elérhető szöveg begyűjtése
- egységes formátumra alakítása (*normalizált változat*)

Kérdés

- manuális átírási munka kiváltható-e gépi eljárással, az emberi erőforrás alkalmazása leszűkíthető-e a tanuló adatok előállításának feladatára



- nem egységesített írásmód, egyenetlenségek a szövegekben
- az egy hang–egy betű megfelelés ritka
- egy-egy hang jelölésmódja egy emléken belül is ingadozik (*Vylag uilaga [világ világa]*)
- kettős hangérték (*zerzete zerent [szerzete szerint]*)
- korlátozott mennyiségű specifikusan annotált tanító adat



- nem egységesített írásmód, egyenetlenségek a szövegekben
- az egy hang–egy betű megfelelés ritka
- egy-egy hang jelölésmódja egy emléken belül is ingadozik (*Vylag uilaga [világ világa]*)
- kettős hangérték (*zerzete zerent [szerzete szerint]*)
- korlátozott mennyiségű specifikusan annotált tanító adat



- nem egységesített írásmód, egyenetlenségek a szövegekben
- az egy hang–egy betű megfelelés ritka
- egy-egy hang jelölésmódja egy emléken belül is ingadozik (*Vylag uilaga [világ világa]*)
- kettős hangérték (*zerzete zerent [szerzete szerint]*)
- korlátozott mennyiségű specifikusan annotált tanító adat



- nem egységesített írásmód, egyenetlenségek a szövegekben
- az egy hang–egy betű megfelelés ritka
- egy-egy hang jelölésmódja egy emléken belül is ingadozik (*Vylag uilaga [világ világa]*)
- kettős hangérték (*zerzete zerent [szerzete szerint]*)
- korlátozott mennyiségű specifikusan annotált tanító adat



- nem egységesített írásmód, egyenetlenségek a szövegekben
- az egy hang–egy betű megfelelés ritka
- egy-egy hang jelölésmódja egy emléken belül is ingadozik (*Vylag uilaga [világ világa]*)
- kettős hangérték (*zerzete zerent [szerzete szerint]*)
- korlátozott mennyiségű specifikusan annotált tanító adat



- nem egységesített írásmód, egyenetlenségek a szövegekben
- az egy hang–egy betű megfelelés ritka
- egy-egy hang jelölésmódja egy emléken belül is ingadozik (*Vylag uilaga [világ világa]*)
- kettős hangérték (*zerzete zerent [szerzete szerint]*)
- korlátozott mennyiségű specifikusan annotált tanító adat



Kutatási probléma

- az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe
- melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását adja
- a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvelmékekre



Kutatási probléma

- az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe
- melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását adja
- a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvelmékekre



Kutatási probléma

- az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe
- melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását adja
- a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvelmékekre

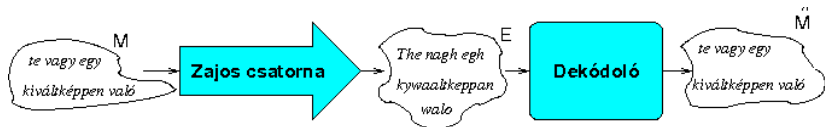


Kutatási probléma

- az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe
- melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását adja
- a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvelmékekre



- 1 Bevezető
- 2 A szövegnormalizáló modell**
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok





Eredeti szöveg

= a normalizált szöveg egy zajos kommunikációs csatornán átment „eltorzított” változata

A dekódoló feladata

M : a modem helyesírású normalizált szövegváltozat egy (rész)mondatnyi sztringje

E : ennek eredeti betűhő átirata

Azon \hat{M} karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális.



Eredeti szöveg

= a normalizált szöveg egy zajos kommunikációs csatornán átment „eltorzított” változata

A dekódoló feladata

- M : a modern helyesírású normalizált szövegváltozat egy (rész)mondatnyi sztringe
- E : ennek eredeti betűhű átírata

Azon \hat{M} karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális.



Eredeti szöveg

= a normalizált szöveg egy zajos kommunikációs csatornán átment „eltorzított” változata

A dekódoló feladata

- M : a modern helyesírású normalizált szövegváltozat egy (rész)mondatnyi sztringe
- E : ennek eredeti betűhű átírata

Azon \hat{M} karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális.



Eredeti szöveg

= a normalizált szöveg egy zajos kommunikációs csatornán átment „eltorzított” változata

A dekódoló feladata

- M : a modern helyesírású normalizált szövegváltozat egy (rész)mondatnyi sztringe
- E : ennek eredeti betűhű átírata

Azon \hat{M} karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális.



$$\hat{M} = \operatorname{argmax}_M P(M|E) \quad (1)$$

illetve béjsziesen:

$$\hat{M} = \operatorname{argmax}_M \frac{P(E|M)P(M)}{P(E)} = \operatorname{argmax}_M P(E|M)P(M) \quad (2)$$

Feladat

A $P(E|M)$ transzliterációs modell-eloszlás (csatornamodell) és a $P(M)$ normalizált szövegmodell-eloszlás (forrásmodell) meghatározása.



karakter N -gram modell

- normalizált szöveg mai magyar nyelvű \rightarrow nagy mennyiségű adat hozzáférhető és használható $\rightarrow N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



karakter N -gram modell

- normalizált szöveg mai magyar nyelvű \rightarrow nagy mennyiségű adat hozzáférhető és használható $\rightarrow N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



karakter N -gram modell

- normalizált szöveg mai magyar nyelvű \rightarrow nagy mennyiségű adat hozzáférhető és használható $\rightarrow N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmódel készlet, Good-Turing simítás



karakter N -gram modell

- normalizált szöveg mai magyar nyelvű \rightarrow nagy mennyiségű adat hozzáférhető és használható $\rightarrow N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



Előfeltétel

- $M_i^j \rightarrow E_k^l$ megfeleléseket tartalmazó tanító korpusz ($i < j$, $k < l$ karakterek közötti pozíciókat jelölő indexek)
- 1-nél hosszabb sztringekre definiált megfeleltetésekkel a transzliterációs modell kontextuális információt is képes reprezentálni



Előfeltétel

- $M_i^j \rightarrow E_k^l$ megfeleléseket tartalmazó tanító korpusz ($i < j$, $k < l$ karakterek közötti pozíciókat jelölő indexek)
- 1-nél hosszabb sztringekre definiált megfeleltetésekkel a transzliterációs modell kontextuális információt is képes reprezentálni



Előfeltétel

- $M_i^j \rightarrow E_k^l$ megfeleléseket tartalmazó tanító korpusz ($i < j$, $k < l$ karakterek közötti pozíciókat jelölő indexek)
- 1-nél hosszabb sztringekre definiált megfeleltetésekkel a transzliterációs modell kontextuális információt is képes reprezentálni



- $\text{Part}(M)$: a *modern* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- $\text{Part}(T)$: az *eredeti* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- R_j : az i -edik szegmentum, adott $R \in \text{Part}(M)$ partícióra, ahol R_j ($= |R_j|$) darab szegmentumból áll
- ekkor ($|T| = |R|$ esetén, ahol $T \in \text{Part}(E)$):

$$P(E|M) = \sum_{R \in \text{Part}(M)} P(R|M) \sum_{T \in \text{Part}(E)} \prod_{i=1}^j P(T_i|R_i) \quad (3)$$



- $\text{Part}(M)$: a *modern* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- $\text{Part}(T)$: az *eredeti* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- R_j : az i -edik szegmentum, adott $R \in \text{Part}(M)$ partícióra, ahol R_j ($= |R_j|$) darab szegmentumból áll
- ekkor ($|T| = |R|$ esetén, ahol $T \in \text{Part}(E)$):

$$P(E|M) = \sum_{R \in \text{Part}(M)} P(R|M) \sum_{T \in \text{Part}(E)} \prod_{i=1}^j P(T_i|R_i) \quad (3)$$



- $\text{Part}(M)$: a *modern* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- $\text{Part}(T)$: az *eredeti* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- R_j : az i -edik szegmentum, adott $R \in \text{Part}(M)$ partícióra, ahol R_j ($= |R_j|$) darab szegmentumból áll
- ekkor ($|T| = |R|$ esetén, ahol $T \in \text{Part}(E)$):

$$P(E|M) = \sum_{R \in \text{Part}(M)} P(R|M) \sum_{T \in \text{Part}(E)} \prod_{i=1}^j P(T_i|R_i) \quad (3)$$



- $\text{Part}(M)$: a *modern* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- $\text{Part}(T)$: az *eredeti* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- R_j : az i -edik szegmentum, adott $R \in \text{Part}(M)$ partícióra, ahol R_j ($= |R_j|$) darab szegmentumból áll
- ekkor ($|T| = |R|$ esetén, ahol $T \in \text{Part}(E)$):

$$P(E|M) = \sum_{R \in \text{Part}(M)} P(R|M) \sum_{T \in \text{Part}(E)} \prod_{i=1}^j P(T_i|R_i) \quad (3)$$



- egy meghatározott illesztés megfelel adott $M_i^j \rightarrow E_k^l$ megfeleltetések halmazának
- csak a legjobb particionálást tekintve:

$$P(E|M) = \max_{R \in \text{Part}(M), T \in \text{Part}(E)} P(R|M) \prod_{i=1}^j P(T_i|R_i) \quad (4)$$

- $P(R|M)$ -et (egyelőre) nem vesszük figyelembe (mi se) (illetve a partíciók felett jobb híján jelenleg egyenletes eloszlást feltételezünk)



- egy meghatározott illesztés megfelel adott $M_i^j \rightarrow E_k^l$ megfeleltetések halmazának
- csak a legjobb particionálást tekintve:

$$P(E|M) = \max_{R \in \text{Part}(M), T \in \text{Part}(E)} P(R|M) \prod_{i=1}^j P(T_i|R_i) \quad (4)$$

- $P(R|M)$ -et (egyelőre) nem vesszük figyelembe (mi se) (illetve a partíciók felett jobb híján jelenleg egyetlen eloszlást feltételezünk)



- egy meghatározott illesztés megfelel adott $M_i^j \rightarrow E_k^l$ megfeleltetések halmazának
- csak a legjobb particionálást tekintve:

$$P(E|M) = \max_{R \in \text{Part}(M), T \in \text{Part}(E)} P(R|M) \prod_{i=1}^j P(T_i|R_i) \quad (4)$$

- $P(R|M)$ -et (egyelőre) nem vesszük figyelembe (mi se) (illetve a partíciók felett jobb híján jelenleg egyetlen eloszlást feltételezünk)



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása**
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok



- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- $10000 M_i^j \rightarrow E_k^l, j = i + 1, l = k + 1, j = l - 1$ megfeleltetés
- + nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezések, ahol a leképezés megfelelő oldalán üres szimbólum áll
- + kiterjesztés: olyan leképezések, ahol a két oldalhoz adott N szomszédos leképezésből származó szimbólumok konkatenálódnak
- üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek



- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- $10000 M_i^j \rightarrow E_k^l, j = i + 1, l = k + 1, j = l - 1$ megfeleltetés
- + nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezések, ahol a leképezés megfelelő oldalán üres szimbólum áll
- + kiterjesztés: olyan leképezések, ahol a két oldalhoz adott N szomszédos leképezésből származó szimbólumok konkatenálódnak
- üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek



- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- $10000 M_i^j \rightarrow E_k^l, j = i + 1, l = k + 1, j = l - 1$ megfeleltetés
- + nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezések, ahol a leképezés megfelelő oldalán üres szimbólum áll
- + kiterjesztés: olyan leképezések, ahol a két oldalhoz adott N szomszédos leképezésből származó szimbólumok konkatenálódnak
- üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek



- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- $10000 M_i^j \rightarrow E_k^l, j = i + 1, l = k + 1, j = l - 1$ megfeleltetés
- + nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezések, ahol a leképezés megfelelő oldalán üres szimbólum áll
- + kiterjesztés: olyan leképezések, ahol a két oldalhoz adott N szomszédos leképezésből származó szimbólumok konkatenálódnak
- üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek



- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- $10000 M_i^j \rightarrow E_k^l, j = i + 1, l = k + 1, j = l - 1$ megfeleltetés
- + nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezések, ahol a leképezés megfelelő oldalán üres szimbólum áll
- + kiterjesztés: olyan leképezések, ahol a két oldalhoz adott N szomszédos leképezésből származó szimbólumok konkatenálódnak
- üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek



$N = 3, M = te, E = the$

- kiinduló:

t	→	t
ε	→	h
e	→	e

- generált:

t	→	th
e	→	he
te	→	the



$N = 3, M = te, E = the$

- kiinduló:

t	→	t
ε	→	h
e	→	e

- generált:

t	→	th
e	→	he
te	→	the



$N = 3, M = te, E = the$

- kiinduló:

t → t

ε → h

e → e

- generált:

t → th

e → he

te → the



- manuális előállítás különféle heurisztikákkal történő gépi támogatása
- *modellparaméterek*: az egyes helyettesítések valószínűsége a következőképpen számítható

$$P(\alpha \rightarrow \beta) = \frac{C(\alpha \rightarrow \beta)}{C(\alpha)} \quad (5)$$

$C(\alpha \rightarrow \beta)$: a tanítókorpuszban látott $\alpha \rightarrow \beta$ helyettesítések
 $C(\alpha)$: az α sztring előfordulásainak száma.



- manuális előállítás különféle heurisztikákkal történő gépi támogatása
- *modellparaméterek*: az egyes helyettesítések valószínűsége a következőképpen számítható

$$P(\alpha \rightarrow \beta) = \frac{C(\alpha \rightarrow \beta)}{C(\alpha)} \quad (5)$$

$C(\alpha \rightarrow \beta)$: a tanítókorpuszban látott $\alpha \rightarrow \beta$ helyettesítések
 $C(\alpha)$: az α sztring előfordulásainak száma.



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása**
- 5 Kiértékelés
- 6 További feladatok



$\operatorname{argmax}_M P(E|M)P(M)$ érték kiszámítása

- eredeti szöveg minden partíciójából a transliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



$\operatorname{argmax}_M P(E|M)P(M)$ érték kiszámítása

- eredeti szöveg minden partíciójából a transliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



$\operatorname{argmax}_M P(E|M)P(M)$ érték kiszámítása

- eredeti szöveg minden partíciójából a transliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



$\operatorname{argmax}_M P(E|M)P(M)$ érték kiszámítása

- eredeti szöveg minden partíciójából a transliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



$\operatorname{argmax}_M P(E|M)P(M)$ érték kiszámítása

- eredeti szöveg minden partíciójából a transliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés**
- 6 További feladatok



Legjobb n -es (n -best) listák

- használhatósági kritérium: a manuális annotáció redukálható a felkínált alakok közötti választásra
- az alapmodell kiegészíthető az egyes tokenek feletti szóalapú n -gram nyelvmodellel
- a kimenet szűrhető illetve átrangsorolható morfológiai elemzés segítségével



Legjobb n -es (n -best) listák

- használhatósági kritérium: a manuális annotáció redukálható a felkínált alakok közötti választásra
- az alapmodell kiegészíthető az egyes tokenek feletti szóalapú n -gram nyelvmodellel
- a kimenet szűrhető illetve átrangsorolható morfológiai elemzés segítségével



Legjobb n -es (n -best) listák

- használhatósági kritérium: a manuális annotáció redukálható a felkínált alakok közötti választásra
- az alapmodell kiegészíthető az egyes tokenek feletti szóalapú n -gram nyelvmodellel
- a kimenet szűrhető illetve átrangsorolható morfológiai elemzés segítségével



Legjobb n -es (n -best) listák

- használhatósági kritérium: a manuális annotáció redukálható a felkínált alakok közötti választásra
- az alapmodell kiegészíthető az egyes tokenek feletti szóalapú n -gram nyelvmodellel
- a kimenet szűrhető illetve átrangsorolható morfológiai elemzés segítségével



fwl (fül) ⇒

-8,81	föl
-10,72	fel
-11,06	fül
-11,28	fől
-12,46	fol
-12,79	ful
-13,52	fely

ygen (igen) ⇒

-10,87	igén
-11,32	igen
-11,60	igény
-13,42	igyen
-14,36	igin
-14,48	igyén

honneg (honnét) ⇒

-19,11	honneg
-19,52	honnég
-20,84	honnét
-21,85	honyneg
-22,21	honynég
-22,56	hónneg

sabach (szabács) ⇒

-17,26	szabács
-18,12	sabács
-18,68	szabacs
-19,18	sábacs
-19,55	szabach
-19,97	szabách



- 1 Bevezető
- 2 A szövegnormalizáló modell
- 3 A modell tanítása
- 4 A modell alkalmazása
- 5 Kiértékelés
- 6 További feladatok**



- újabb tanulási módszerek alkalmazása és kiértékelése → (kiváltható|támogatható)-e a manuális átírás gépi eljárással
- modell kidolgozása:
 - a szóhatárok kezelésére
 - a lehetséges partíciók feletti eloszlásra
- implementáció hatékonyságának növelése



- újabb tanulási módszerek alkalmazása és kiértékelése → (kiváltható|támogatható)-e a manuális átírás gépi eljárással
- modell kidolgozása:
 - a szóhatárok kezelésére
 - a lehetséges partíciók feletti eloszlásra
- implementáció hatékonyságának növelése



- újabb tanulási módszerek alkalmazása és kiértékelése → (kiváltható|támogatható)-e a manuális átírás gépi eljárással
- modell kidolgozása:
 - a szóhatárok kezelésére
 - a lehetséges partíciók feletti eloszlásra
- implementáció hatékonyságának növelése



- újabb tanulási módszerek alkalmazása és kiértékelése → (kiváltható|támogatható)-e a manuális átírás gépi eljárással
- modell kidolgozása:
 - a szóhatárok kezelésére
 - a lehetséges partíciók feletti eloszlásra
- implementáció hatékonyságának növelése



- újabb tanulási módszerek alkalmazása és kiértékelése → (kiváltható|támogatható)-e a manuális átírás gépi eljárással
- modell kidolgozása:
 - a szóhatárok kezelésére
 - a lehetséges partíciók feletti eloszlásra
- implementáció hatékonyságának növelése



Közönnyuk a figyelmeeth!

