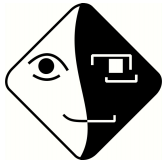
Gépi tanulási módszerek ómagyar kori szövegek normalizálására

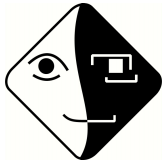
Oravecz Csaba, Sass Bálint, Simon Eszter

VI. Magyar Számítógépes Nyelvészeti Konferencia

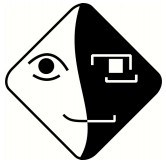
Szeged, 2009. december 3-4.



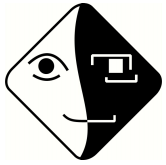
Vázlat



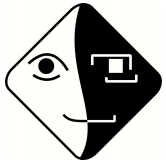
Vázlat



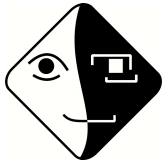
Vázlat



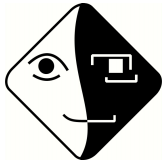
Vázlat



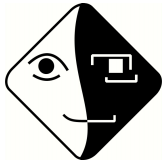
Vázlat



Vázlat



Tartalom



A projekt

Cél

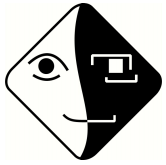
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

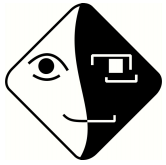
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

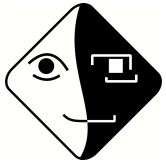
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

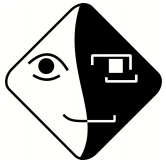
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

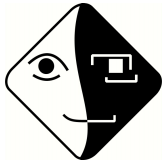
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

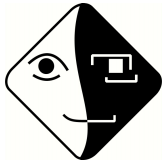
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

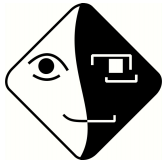
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

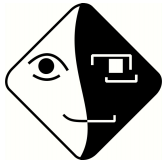
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

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



A projekt

Cél

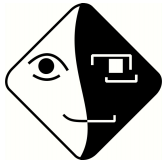
- elektronikus nyelvtörténeti adatbázis létrehozása
  - összes ómagyar szövegemlék
  - középmagyar korból arányos válogatás

Feladat

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

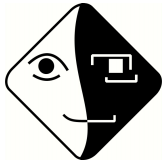
Kérdés

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



Nehézségek

- 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

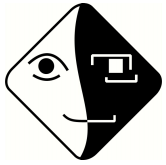


Nehézségek

- 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

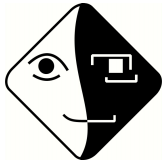
Nehézségek

- 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



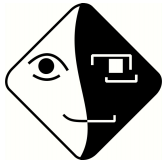
Nehézségek

- 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



Nehézségek

- 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



Nehézségek

- 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 nyelvemlékekre

Kutatási probléma

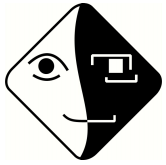
- az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe
- melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását adja
- a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre

Kutatási probléma

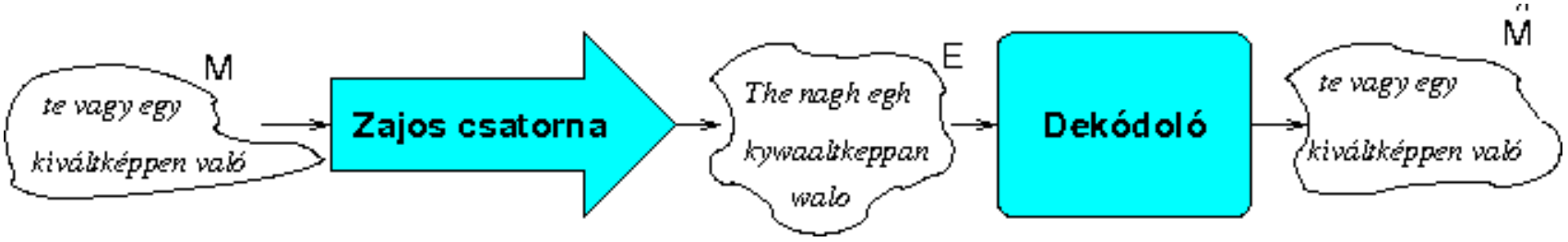
- az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe
- melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását adja
- a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre

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 nyelvemlékekre



Tartalom


M
te vagy egy
kiváltképpen való
Zajos csatorna
E
The nagh egh
kywaaltkeppan
walo
Dekódoló
M'
te vagy egy
kiváltképpen való

Eredeti szöveg

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

A dekódoló feladata

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

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

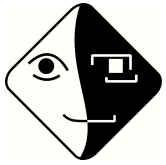
Eredeti szöveg

= 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 átirata

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



A modell

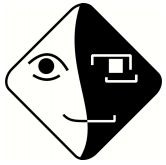
Eredeti szöveg

= 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 átirata

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



A modell

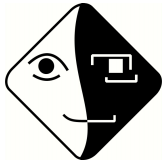
Eredeti szöveg

= 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 átirata

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



Képletesen

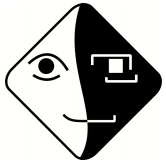
$$\hat{M} = \operatorname*{argmax}_M P(M|E) \tag{1}$$

illetve béjsziesen:

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

Feladat

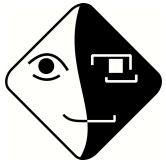
A $P(E|M)$ transzliterációs modell-eloszlás (csatornamodell) és a $P(M)$ normalizált szövegmodell-eloszlás (forrásmodell) meghatározása.



Forrásmodell

karakter N-gram modell

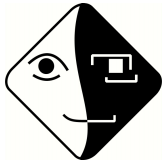
- normalizált szöveg mai magyar nyelvű $\rightarrow$ nagy mennyiségű adat hozzáférhető és használható $\rightarrow$ $N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



Forrásmodell

karakter N-gram modell

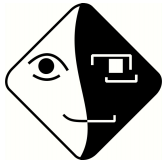
- normalizált szöveg mai magyar nyelvű $\rightarrow$ nagy mennyiségű adat hozzáférhető és használható $\rightarrow$ $N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



Forrásmodell

karakter N-gram modell

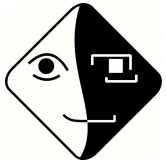
- normalizált szöveg mai magyar nyelvű $\rightarrow$ nagy mennyiségű adat hozzáférhető és használható $\rightarrow$ $N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



Forrásmodell

karakter N-gram modell

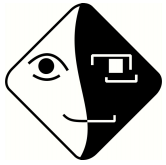
- normalizált szöveg mai magyar nyelvű $\rightarrow$ nagy mennyiségű adat hozzáférhető és használható $\rightarrow$ $N > 3$ (is lehet)
- forrás: MNSz 10 millió szavas, 65 millió karakteres részlete
- CMU nyelvmodell készlet, Good-Turing simítás



Transzliterációs modell formálisan

Előfeltétel

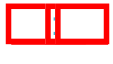
- $M_i^j \rightarrow E_k^l$ megfeleléseket tartalmazó tanító korpusz ($i < j,\ k < l$ karakterek közötti pozíciókat jelölő indexek)
- 1-nél hosszabb sztringekre definiált megfeleltetésekkel a transzliterációs modell kontextuális információt is képes reprezentálni

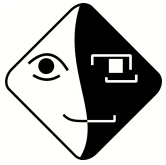


Transzliterációs modell formálisan

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

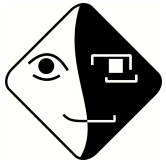




Transzliterációs modell formálisan

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



Transzliterációs modell formálisan

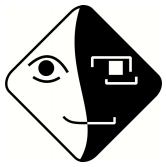
- Part(M): a *modern* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- Part(T): az *eredeti* nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza
- R_i: az i-edik szegmentum, adott $R \in \text{Part}(M)$ partícióra, ahol R j $(= |R|)$ darab szegmentumból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

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

Transzliterációs modell formálisan

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

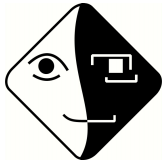
$$P(E|M) = \sum_{R \in \text{Part}(M)} P(R|M) \sum_{T \in \text{Part}(E)} \prod_{i=1}^{j} P(T_i|R_i) \quad (3)$$



Transzliterációs modell formálisan

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

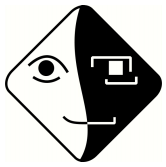
$$P(E|M) = \sum_{R \in \text{Part}(M)} P(R|M) \sum_{T \in \text{Part}(E)} \prod_{i=1}^{j} P(T_i|R_i) \quad (3)$$



Transzliterációs modell formálisan

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

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

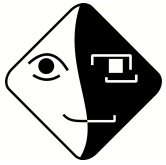


Transzliterációs modell formálisan

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

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

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

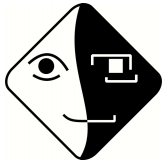


Transzliterációs modell formálisan

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

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

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

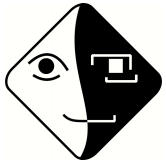


Transzliterációs modell formálisan

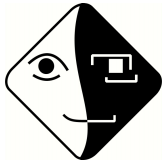
- egy meghatározott illesztés megfelel adott $M_i^j \rightarrow E_k^l$ megfeleltetések halmazának
- csak a legjobb particionálást tekintve:

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

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

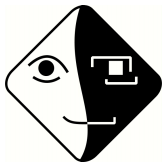


Tartalom



A TM tanító korpuszának előállítása

- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 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

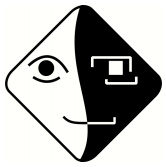


A TM tanító korpuszának előállítása

- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 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



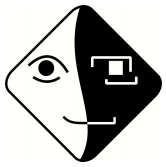
A TM tanító korpuszának előállítása

- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 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



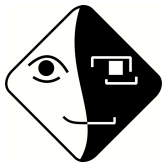
A TM tanító korpuszának előállítása

- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 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



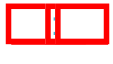
A TM tanító korpuszának előállítása

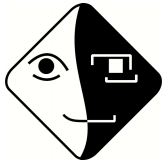
- két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából
- 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 megfeleltetés

\+ 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





Példa leképezési szabályokra

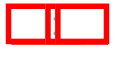
$N = 3, M = \text{te}, E = \text{the}$

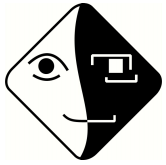
- kiinduló:

$$\begin{aligned} \text{t} &\rightarrow \text{t} \\ \epsilon &\rightarrow \text{h} \\ \text{e} &\rightarrow \text{e} \end{aligned}$$

- generált:

$$\begin{aligned} \text{t} &\rightarrow \text{th} \\ \text{e} &\rightarrow \text{he} \\ \text{te} &\rightarrow \text{the} \end{aligned}$$





Példa leképezési szabályokra

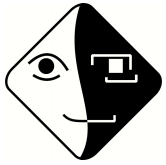
$N = 3, M = \text{te}, E = \text{the}$

- kiinduló:

$$\begin{array}{ccc} \text{t} & \rightarrow & \text{t} \\ \epsilon & \rightarrow & \text{h} \\ \text{e} & \rightarrow & \text{e} \end{array}$$

- generált:

$$\begin{array}{ccc} \text{t} & \rightarrow & \text{th} \\ \text{e} & \rightarrow & \text{he} \\ \text{te} & \rightarrow & \text{the} \end{array}$$



Példa leképezési szabályokra

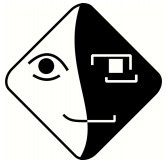
$N = 3, M = \text{te}, E = \text{the}$

- kiinduló:

$$\begin{aligned} \text{t} &\rightarrow \text{t} \\ \epsilon &\rightarrow \text{h} \\ \text{e} &\rightarrow \text{e} \end{aligned}$$

- generált:

$$\begin{aligned} \text{t} &\rightarrow \text{th} \\ \text{e} &\rightarrow \text{he} \\ \text{te} &\rightarrow \text{the} \end{aligned}$$

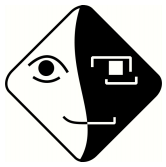


A TM tanító korpuszának előállítása

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

$$P(\alpha \to \beta) = \frac{C(\alpha \to \beta)}{C(\alpha)} \tag{5}$$

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

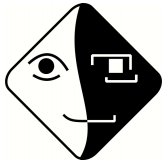


A TM tanító korpuszának előállítása

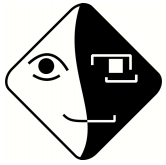
- manuális előállítás különféle heurisztikákkal történő gépi támogatása
- *modellparaméterek*: az egyes helyettesítések valószínűsége a következőképpen számítható

$$P(\alpha \to \beta) = \frac{C(\alpha \to \beta)}{C(\alpha)} \tag{5}$$

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



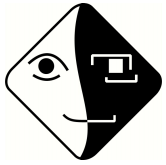
Tartalom



Dekódolás

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

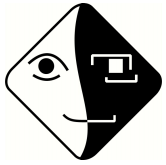
- eredeti szöveg minden partíciójából a transzliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrarendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



Dekódolás

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

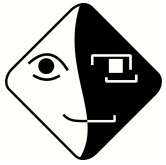
- eredeti szöveg minden partíciójából a transzliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrarendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



Dekódolás

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

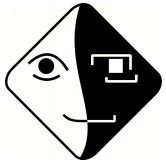
- eredeti szöveg minden partíciójából a transzliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrarendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



Dekódolás

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

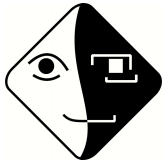
- eredeti szöveg minden partíciójából a transzliterációs modell helyettesítései alapján a lehetséges modern változatok generálása
- valószínűségük hozzárendelése a modell alapján
- a változatokra kapott rangsor újrarendezése a nyelvmodell szerint
- optimalizálás: ritkítás (pruning), beam-keresés



Dekódolás

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

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



Tartalom

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

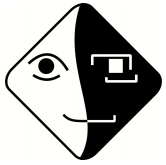
- használhatósági kritérium: a manuális annotáció redukálható a felkínált alakok közötti választásra
- az alapmodell kiegészíthető az egyes tokenek feletti szóalapú *n*-gram nyelvmodellel
- a kimenet szűrhető illetve átrangsorolható morfológiai elemzés segítségével

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



Példák

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) ⇒

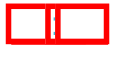
-10,87	igén
-11,32	igen
-11,60	igény
-13,42	igyen
-14,36	igin
-14,48	igyén

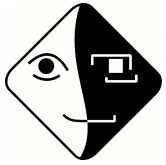
honneg (honnét) ⇒

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

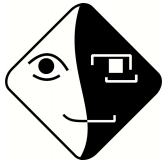
sabach (szabács) ⇒

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



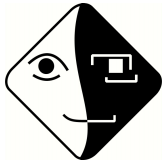


Tartalom



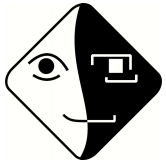
További feladatok, kilátások

- 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



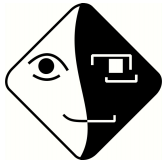
További feladatok, kilátások

- 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



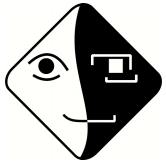
További feladatok, kilátások

- 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



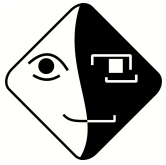
További feladatok, kilátások

- 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



További feladatok, kilátások

- 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



Kǫzǫnyuk a fyġelmeth!