PÁZMÁNY PÉTER KATOLIKUS EGYETEM - KIEMELT FELSŐOKTATÁSI INTÉZMÉNY
INFORMÁCIÓS TECHNOLÓGIAI ÉS BIONIKAI KAR

Hibrid előfeldolgozó algoritmusok morfológiailag komplex nyelvek és erőforrásszegény domainek hatékony feldolgozására

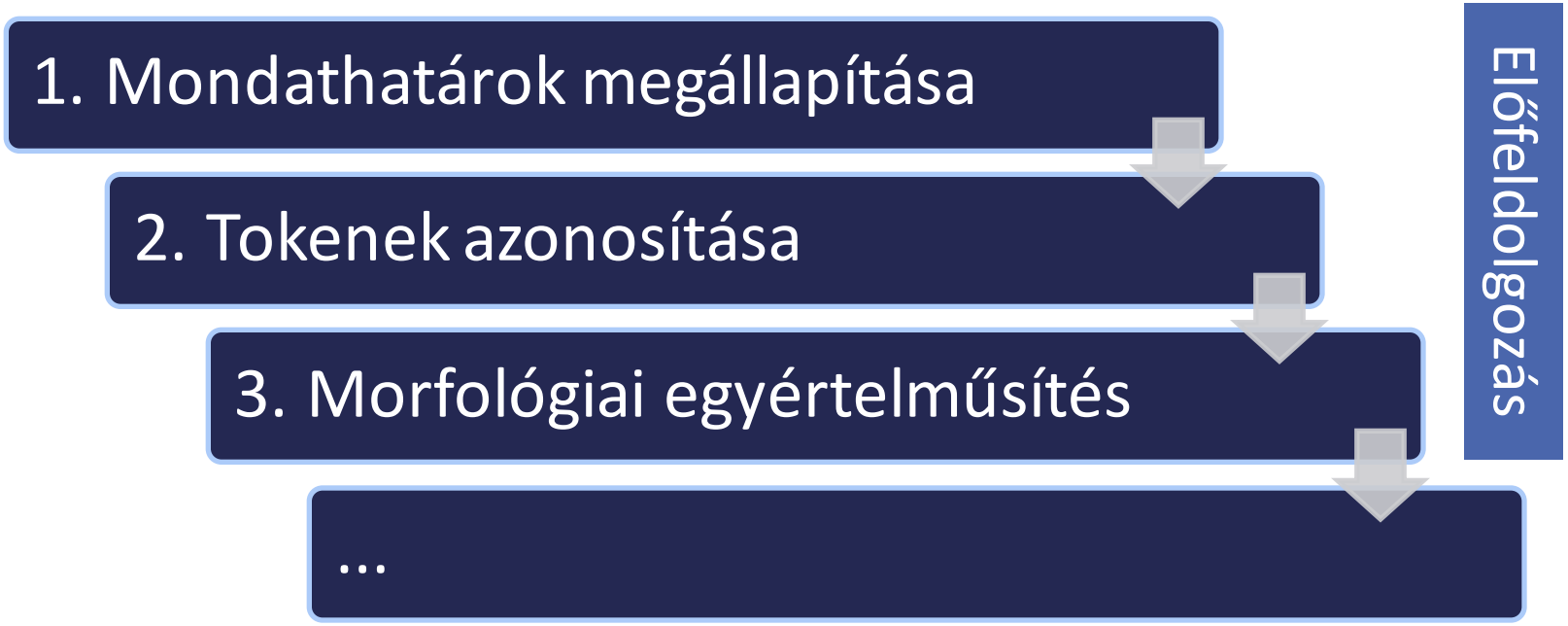
Orosz György

Témavezető: Prószéky Gábor

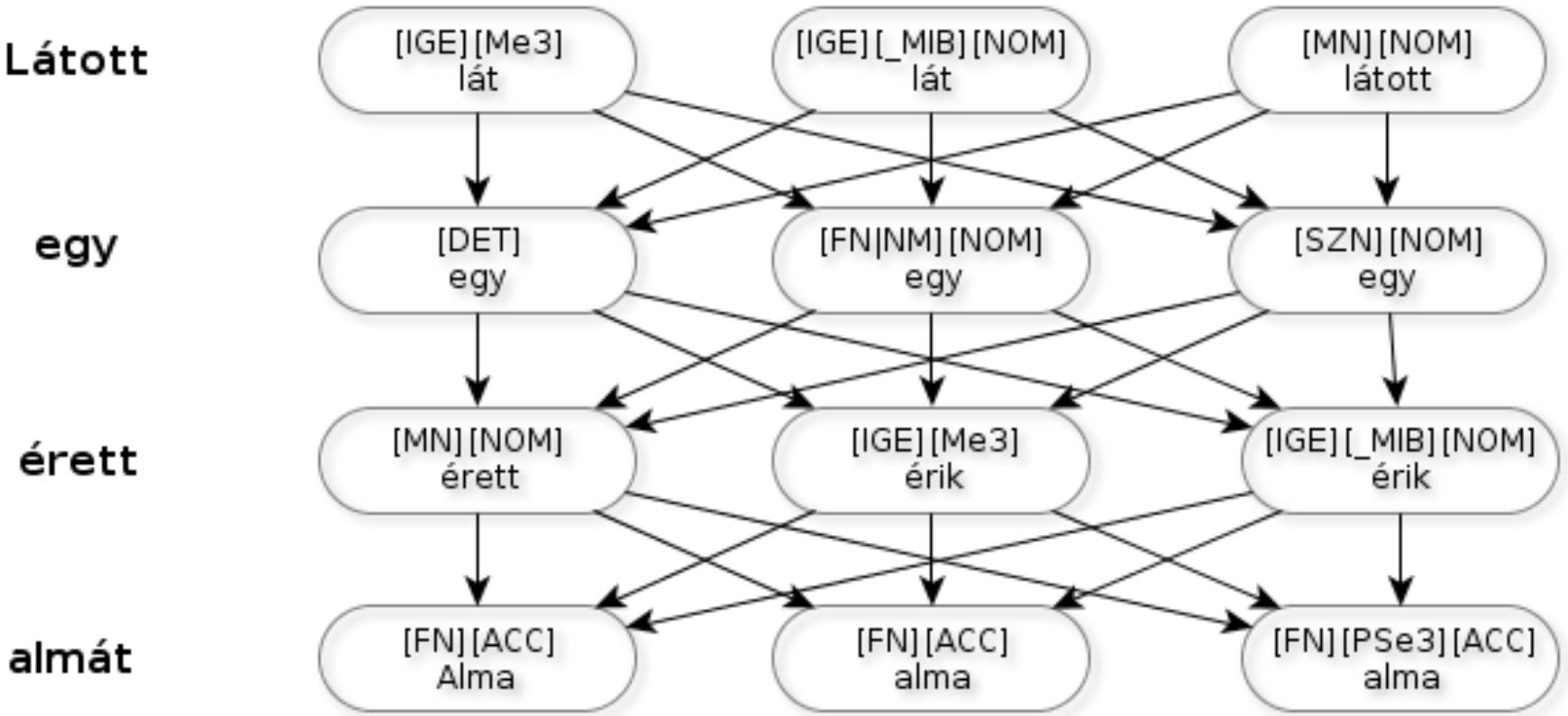
Bevezetés

Előfeldolgozó algoritmusok

Napjaink szövegfeldolgozó rendszereinek felépítése:



Morfológiai egyértelműsítés ... és alkalmazásai



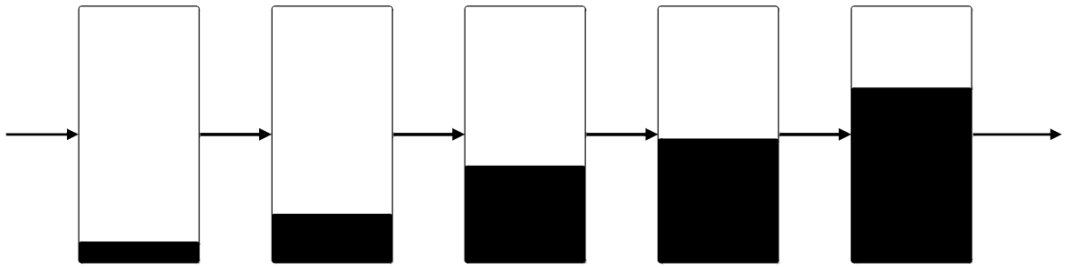
Szavak és mondatok szegmentálása

```
Mai szemészeti lelet _:
V_:_0,3-0,5 D_sph_-0,75 D_cyl 90 fok_=_0,7?
  0,8+0,5 D_sph_-0,75 D_cyl 90 fok_=_1,0?_(_felhúzott szemhéjjal _)
 add._+3,0 D_sph Csapody III. pd._69/67_mm<SE>
A saját szemüvege nem javított a látásán _.<SE>
Sü_:_-1,0 D_sph_-0,5 D_cyl 90 fok
   -1,0 D_sph_-0,5 D_cyl 90 fok<SE>
Szemüveget kap távolra és közelre is _.<SE>
ld_: változatlan<SE>
lsin_: A felső szemhéj ptosisa csökkent _, a látható polsáló terime megint
megkisebbedett _, a bulbus jobb helyzetben látható _,_mint ezelőtt_, érágas
conjunctiva_,_jó vvfény.<SE>Fundus: aedem<SE>
2006._02._21.-én_.<SE>
Amerikai úton a műtétet nem végezték el mert asok bizonytalansági tényező
miatt a beteg nem vállalta a magas rizikójú beavatkozáast _.<SE>
kontrollvizsgálatra jött <SE>
ld_:_idem<SE>
lsin_:_felső szemhéj ptotikus_, a kiemelkedő terime idem_, pulsal_,_a bulbus enyhe
exophthalmusban_,_érágas conj_,_vvfény nyerhető<SE>Fundus:_scler papilla<SE>
Tappl_: 16/24 Hgmm<SE>
Th_: 2x Azopt o._s.<SE>
```

Motiváció

Előfeldolgozás – megoldott probléma?

- Új kihívásokkal szembesülünk morfológiailag komplex nyelvek és nem sztenderd domainek esetén
- Az elérhető eszközök nehezen adaptálhatóak:
  - adatvezérelt algoritmusok – adatéhség
  - szabályalapú rendszerek – specifikusság
- Hibaterjedés a pipelineszerű architektúrákban



I. Hatékony morfológiai egyértelműsítő módszerek morfológiai komplex nyelvek elemzéséhez

Morfológiailag komplex nyelvek

- Az egyes szavakhoz (az angolhoz képest) sokkal több szóalak létezik
- Sokkal több ismeretlen szóalakkal találkoznak az adatvezérelt algoritmusok
- Nagyobb címkekészlet
- Megnövekedett többértelműséggel és adatritkasággal kell szembesülnünk

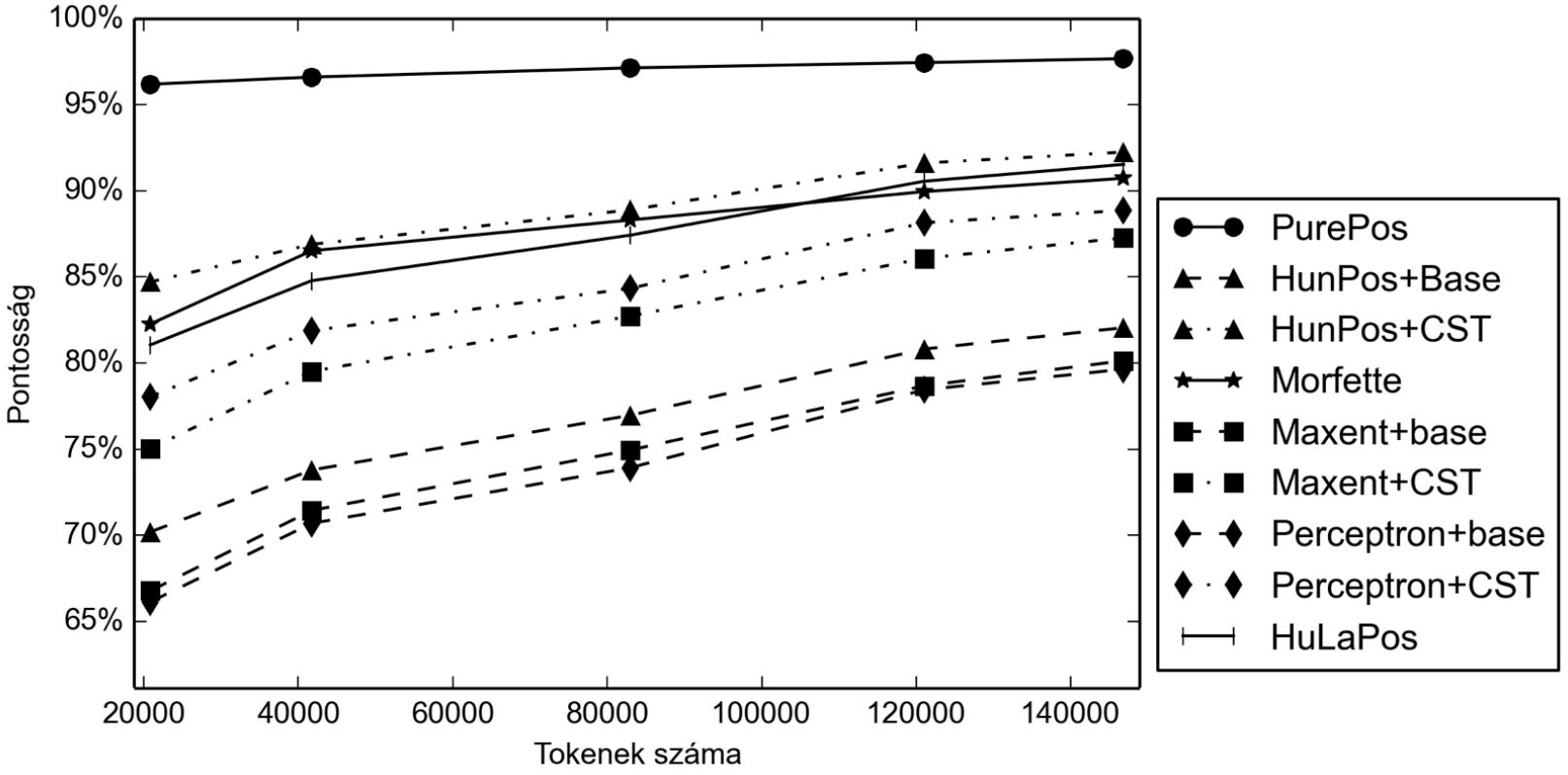
Purepos: egy hibrid morfológiai egyértelműsítő eszköz

- Szófaji egyértelműsítés:
  - rejtett Markov módszerre (főként a TnT/HunPos algoritmusaira) építve
  - a morfológiai elemző integrált használatával
- Lemmatizálás:
  - morfológiai elemző kimenetére építve
  - szövég alapú guesser + szótő modell együttes használatával
  - ismeretlen szavak hatékony kezelése

Hibrid morfológiai egyértelműsítő teljesítménye magyar nyelvű szövegeken

	PoS tagging	Morph. tagging	
		Token	Sentence
`magyarlanc`	96.50%	95.72%	54.52%
Morfette	96.94%	92.24%	38.18%
HuLaPos	96.90%	95.61%	54.57%
PurePos	96.99%	96.27%	58.06%
HunPos + BL	96.71%	92.65%	36.06%
HunPos + CST	96.71%	91.19%	35.31%
Maxent + BL	95.63%	92.21%	34.82%
Maxent + CST	95.63%	90.14%	29.70%
Perceptron + BL	95.19%	91.16%	29.42%
Perceptron + CST	95.19%	89.78%	27.91%

… kevés tanítóanyag esetén



Hibrid komponensek kiaknázása

Történeti szövegek morfológiai annotálásának feladatában jelentősen mértékben javítottam a morfológiai annotálás minőségén.

- 88,99% → 97,58% szószintű pontosság
- 55,58% → 86,48% tagmondatszintű pontosság

I.1 tézis

Kidolgoztam egy olyan metódust, ami agglutináló nyelvek, így magyar esetén is nagy pontossággal képes szavak lemmáit azonosítani. Az eljárás a tanítóanyagban látott szavakon túl az ún. ismeretlen szóalakokat is hatékonyan kezeli, amihez a morfológiai elemző lehetséges elemzésein kívül a tanítóanyagból készített statisztikai modellekre is épít. Mérésekkel kimutattam, hogy a módszer magyar nyelv esetén kimagasló pontossággal bír.

I.2 tézis

Létrehoztam egy olyan hibrid morfológiai egyértelműsítő eszközt (PurePos), mely hatékonyan alkalmazható morfológiailag komplex és nyelvi erőforrásokban szegény nyelvek esetén. Az algoritmus statisztikai eljárásokra támaszkodva, morfológiai elemző integrált alkalmazásával és szabály alapú komponensek használatával hatékony egyértelműsítést tesz lehetővé. Az eszköz a szavak lemmáinak meghatározását az előző tézisben ismertetett módszerrel végzi. Megmutattam, hogy az eljárás magyar nyelv esetén state-of-the-art teljesítménnyel rendelkezik. Ismertettem, hogy a rendszer architektúrája lehetőséget nyújt domén specifikus szabályok hatékony alkalmazására, illetve méréseimmel alátámasztottam, hogy a létrehozott algoritmus kiemelkedő pontossággal bír kevés tanítóanyag használata esetén is.

Egyértelműsítő rendszerek diszkrepanciája

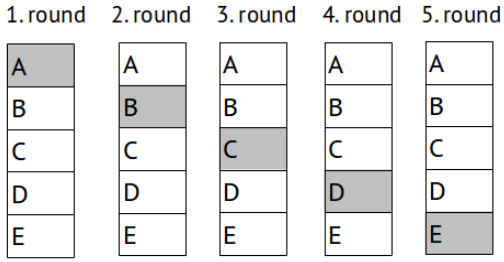
- A PurePos és a HulaPos rendszer hibái jelentősen eltérnek egymástól
- Megvizsgáltam, hogy hogyan lehetséges kombinációjukkal javítani a morf. egyértelműsítés pontosságán

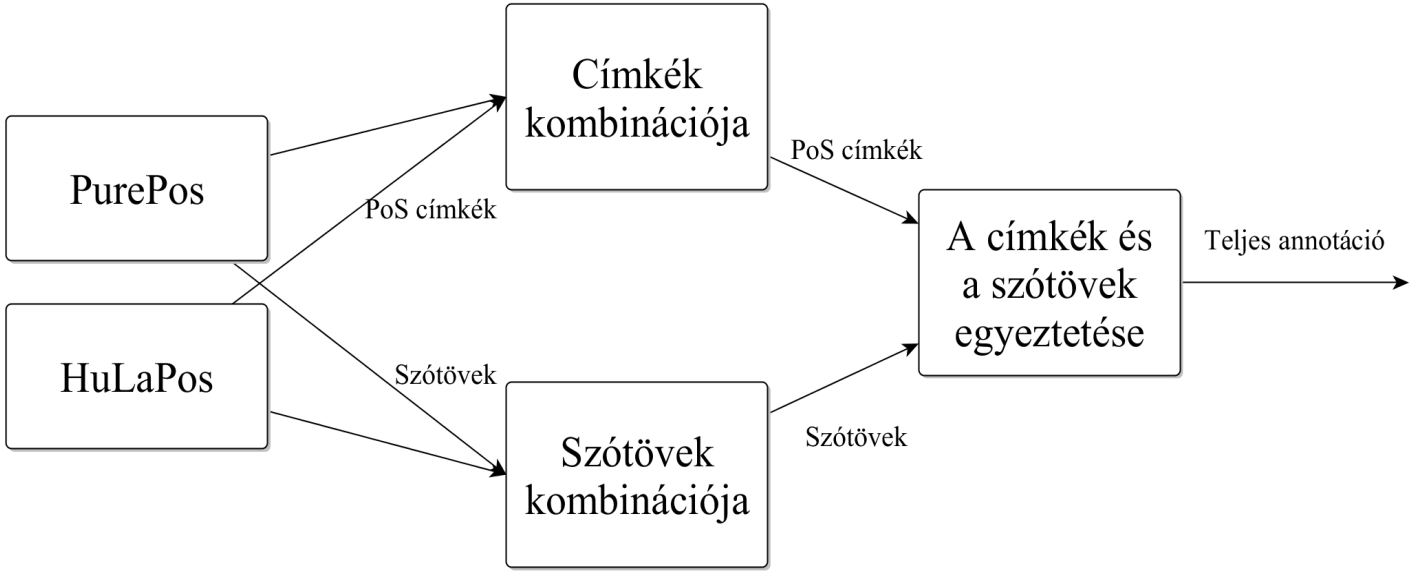
$$\mathrm{OER}(A,B) = \frac{\#\text{errors of } A \text{ only}}{\#\text{errors of either } A \text{ or } B}$$

	Tagging	Lemmatization	Full disambig.
Agreement rate	97.60%	98.02%	96.92%
They are right when they agree	99.30%	99.85%	99.29%
One is right when they disagree	97.53%	98.89%	97.14%
OER(PP, HL)	22.41%	11.66%	21.16%
OER(HLP, PP)	53.58%	80.21%	58.24%

Egyértelműsítő rendszerek kombinációja

- Metaosztályzó tanítása stacking módszerrel
- Példány alapú tanulás
- Morfológiai elemző integrált használata





28% hiba-csökkenési ráta

I.3 tézis

Létrehoztam egy olyan módszert, mely morfológiai egyértelműsítő rendszerek kombinációjával hatékonyan növeli a címkézés pontosságát magyar nyelv esetén. A kidolgozott eljárás újdonsága, hogy külön modulban végzi a lemmák és morfoszintaktikai címkék azonosítását, majd azok kimenetét egyesítve határozza meg a teljes morfológiai annotációt. A módszer példány alapú tanulásra épül és az egyes alrendszereket keresztvalidáción keresztül tanítja. Méréseimmel alátámasztottam, hogy az ismertetett módszer jelentős mértékben képes növelni a címkézési feladat pontosságán.

II. Morfoszintaktikai komplexitás automatikus becslése egyértelműsítő algoritmusok alkalmazásával

Morfoszintaktikai komplexitás mérése

- Az átlagos megnyilatkozáshossz fontos metrika a nyelvészeti kutatásokban
  - Korrelál a beszélő nyelvi fejlettségével
- Magyar esetén (angollal ellentétben) morfémák számában mérjük
- Magyar nyelvre nem létezett automatikus módszer

Gyermeknyelvi beszédátiratok morfológiai címkézése

Lexikon bővítés

- Tipikus beszéltnyelvi szavak felvétele
- Kicsinyítőképzős alakok

Egyértelműsítő adaptációja

- "... akkor ... amikor ..."
- "... azért ... mert ..."
- "...utána..."
- "...meg..."

96,15% szószintű pontosság

Morfoszintaktikai komplexitás mérése

- Az adaptált egyértelműsítő kimenetére építve
- Nyelvészetileg releváns szabályok implementálásával

0,99 korrelációs érték
0,04 átlagos eltérés

II.1 tézis

Létrehoztam egy hibrid morfológiai egyértelműsítő láncot magyar gyermeknyelvi beszédátiratok nagy pontosságú elemzésére. Az algoritmus alapját az I.2 tézisben ismertetett rendszer képezi, amelyet a beszélt nyelv címkézéséhez szükséges szabályokkal adaptáltam. Méréseimmel igazoltam, hogy a létrejött elemzési lánc teljesítménye megközelíti az általános nyelvi címkézők eredményességét.

II.2 tézis

Kifejlesztettem egy olyan új eljárást, amely magyar nyelvű beszédátiratok morfoszintaktikai összetettségét képes automatikusan becsülni. Az algoritmus a II.1 tézisben bemutatott elemző-láncra épülve számolja a megnyilatkozások morfémában mért hosszát. Méréseimmel ki-mutattam, hogy a módszer megfelelően képes helyettesíteni az időigényes manuális számolást.

III. Előfeldolgozó algoritmusok egy erőforrásszegény és zajos domainhez

Klinikai rekordok jellemzői

- Latin és magyar nyelv együttes használata
- Hibás illetve nem sztenderd szóalakok
- A mondathatárokat jelző írásjelek és kapitalizáció gyakori hiánya
- Nagy számú, változatos rövidítések

Klinikai rekordok mondatokra és szavakra bontása

- Skálázott log λ módszert használata a (szó, •) párok egyértelműsítésére
  - Felszíni és morfoszintaktikai tulajdonságok
- Morfológiai elemző alkalmazása rövidítések és tulajdonnevek megkülönböztetésére

Tokenizálás eredményessége

	Pontosság (P)	Fedés (R)	F_1
Baseline	99,74%	74,94%	85,58%
A teljes lánc	98,54%	95,32%	96,90%

Mondatrabontás eredményessége

	Pontosság (P)	Fedés (R)	$F_{0,5}$
`magyarlanc`	72,59%	77,68%	73,55%
HTG	44,73%	49,23%	45,56%
HTM	43,19%	42,09%	42,97%
Punkt	58,78%	45,66%	55,59%
OpenNLP	52,10%	96,30%	57,37%
A hibrid lánc	93,28%	86,73%	91,89%

Klinikai rekordok morfológiai egyértelműsítése

Sztenderd egyértelműsítő rendszer leggyakoribb hibái:

1. Rövidítések és betűszavak 49,17%
2. Ismeretlen szavak 27,27%
3. Domainspecifikus szóhasználat 14,88%

Domainadaptációs kísérletek

1. Rövidítések és betűszavak megkülönböztetett kezelése
2. Adaptált morfológiai lexikon használata
3. *Tanítóanyag választás*

93,73% szószintű pontosság

III.1 tézis

Létrehoztam egy olyan hibrid eljárást, mely magyar nyelvű klinikai rekordokat képes magas pontossággal mondatokra és szavakra bontani. A módszer alapját egy szabály-alapú szegmentáló algoritmus képezi, amelyet felügyelet nélküli gépi tanulással egészítettem ki. Méréseimmel alátámasztottam, hogy a hibrid rendszer által azonosított mondat- és szóhatárok kellően pontosak a gyakorlati alkalmazhatósághoz. Ezen túl kimutattam még, hogy a magyar nyelvre elérhető algoritmusok közül sem a szabályalapú, sem a gépi tanulást használó rendszerek nem alkalmasak orvosbiológiai szövegek tokenizálására és mondatokra bontására.

III.2 tézis

Megmutattam, hogy az I.2 tézisben ismertetett rendszer, megfelelő adaptációs technikákkal kombinálva alkalmas orvosbiológiai szövegek elfogadható minőségű morfológiai egyértelműsítésére. Méréseimmel kimutattam, hogy az ismertetett szabály-alapú és statisztikai domén-adaptációs módszerek jelentős mértékben javítanak a teljes elemzési lánc pontosságán.

Köszönöm!