PÁZMÁNY PÉTER KATOLIKUS EGYETEM
INFORMÁCIÓS TECHNOLÓGIAI KAR

ÚJ SZÉCHENYI TERV

Automatic structuring and correction suggestion system for Hungarian clinical records

Borbála Siklósi[1], György Orosz[1,2], Attila Novák[1,2], Gábor Prószéky[1,2]

[1]*Pázmány Péter Catholic University, Faculty of Information Technology*
siklosi.borbala@itk.ppke.hu, orosz.gyorgy@itk.ppke.hu

[2]*MTA-PPKE Language Technology Research Group*
novak.attila@itk.ppke.hu, proszeky@itk.ppke.hu

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

MAGYARORSZÁG MEGÚJUL

A projektek az Európai Unió támogatásával valósulnak meg.

Outline

- Motivation
- Where to start
- Uniform representation
  - Separation of structural units
  - **Correction of spelling errors**
- Results
- Examples
- Typical errors

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Motivation

- Processing clinical records
- Revealing deeper relations
- Exploration of hidden information
- Support searchability
- Aid doctors and researchers
- …

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Where to start

- Serious shortcomings of clinical documentation systems of Hungarian clinics
- Missing medical histories
- Flowing texts lacking any structure
- Misused, mixed language – of Hungarian and “medical Latin”

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Semmelweis Egyetem Szemészeti Klinika Tömő u.
1083 Budapest Tömő u. 25-29.
Általános Ambulancia
Intézetvezető: Prof. Németh János
Tel.: (1) 210-0280/51710

A M B U L Á N S K E Z E L Ő L A P

Státusz
2010.10.19 12:28 Székelyhidi/Füst

Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
V:0,7+0,75Dsph=1,0
1,0 +0,5 Dsph élesebb

+2.0 Dsph mko Cs IV

St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla rekciók rendben, lencse tiszta, jó vvf.
Átfecskendezés mko sikerült.

Olvasó szemüveg javasolt: +2.0 Dsph mko.
Éjszakánként műkönnygél ha szükséges.
Kontroll: panasz esetén

Diagnózis

DIAGNÓZISOK megnevezése	Kód	Dátum	Év	K	V	T
Látászavar, k.m.n.	H5390	2010.10.19				3

Beavatkozások

Kód	Megnevezés	Menny.	Pont
11041	Vizsgálat	1	750

2010.11.16

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Uniform representation

- Separation of structural units
- Tagging basic metadata
- Separation of textual from tabular data
- Spelling correction

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Separation of structural units

- XML structure based on (not uniform) formatting elements and basic surface patterns
  - Keeping the whole copy of original document
  - Content (header, diagnosis, applied treatments, status, operations, symptoms, etc)
  - Metadata (document type, institution and department IDs, tables, medical codes)
  - Named entities (dates, doctors, operations)
  - Medical history

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Separation of textual from tabular data

Semmelweis Egyetem Szemészeti Klinika Tömő u.
1083 Budapest Tömő u. 25-29.
Általános Ambulancia
Intézetvezető: Prof. Németh János
Tel.: (1) 210-0280/51710

A M B U L Á N S K E Z E L Ő L A P

Státusz
2010.10.19 12:28 Székelyhidi/Füst

Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
V:0,7+0,75Dsph=1,0
1,0 +0,5 Dsph élesebb

+2.0 Dsph mko Cs IV

St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla rekciók rendben, lencse tiszta, jó vvf.
Átfecskendezés mko sikerült.

Olvasó szemüveg javasolt: +2.0 Dsph mko.
Éjszakánként műkönnygél ha szükséges.
Kontroll: panasz esetén

Diagnózis

DIAGNÓZISOK megnevezése	Kód	Dátum	Év	K	V	T
Látászavar, k.m.n.	H5390	2010.10.19				3

Beavatkozások

Kód	Megnevezés	Menny.	Pont
11041	Vizsgálat	1	750

2010.11.16

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Separation of textual from tabular data

Semmelweis Egyetem Szemészeti Klinika Tömő u.
1083 Budapest Tömő u. 25-29.
Általános Ambulancia
Intézetvezető: Prof. Németh János
Tel.: (1) 210-0280/51710

A M B U L Á N S K E Z E L Ő L A P

Státusz
2010.10.19 12:28 Székelyhidi/Füst

Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
V:0,7+0,75Dsph=1,0
1,0 +0,5 Dsph élesebb

+2.0 Dsph mko Cs IV

St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla rekciók rendben, lencse tiszta, jó vvf.
Átfecskendezés mko sikerült.

Olvasó szemüveg javasolt: +2.0 Dsph mko.
Éjszakánként műkönnygél ha szükséges.
Kontroll: panasz esetén

Diagnózis

DIAGNÓZISOK megnevezése	Kód	Dátum	Év	K	V	T
Látászavar, k.m.n.	H5390	2010.10.19				3

Beavatkozások

Kód	Megnevezés	Menny.	Pont
11041	Vizsgálat	1	750

2010.11.16

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Separation of textual from tabular data

- Rules and pattern matching do not work
- Clustering (k-means)
  - Input: „concatenated” lines
  - Output: manual selection from several clusters
- Classification (Naive-Bayes)
  - Applied to new documents
  - Trained on output of clustering
  - 98%

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Spelling correction

- Domain and language specific difficulties
- Standardized corpus ⟶ „0th” goal
- Approaching an error model with language models
  - Stopword list
  - Abbreviations list – automatic generation
  - Judgment of morphological analyzer (spell checker)
    - Licensed
    - Non licensed ⟶ if frequent, then “correct”
  - General and domain specific word lists

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Spelling correction

- Tokenization (abbreviations, punctuation, imperfect syntactic structures)
- Generation of candidate corrections:
  - One edit distance from original form
  - Suggestions of the speller
- Scoring:
  - Weighted language models
  - Weighted edit distance generation (e.g. accents)
  - Features of the original form

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Results

- First five elements of the ranked candidate list

túvel -> tűvel

'tűvel' : 0.15000030725

'tevel' : 0.15

'tővel' : 0.15

'túvel' : 2.59336099585e-05

'túmel' : 0.0

implatatumot -> implantatumot

'implantatumot' : 0.150008644537

'implatatumot' : 1.72890733057e-05

'impaltatumot' : 0.0

'implatatemot' : 0.0

'óimplatatumot' : 0.0

telefonnegbeszélés -> telefonmegbeszélés

'telefonmegbeszélés' : 0.15

'telefonnegbeszélés' : 1.72890733057e-05

'telefonnqgbeszélés' : 0.0

'telefonnegbeszélpés' : 0.0

'teleifonnegbeszélés' : 0.0

Meibm -> Meibom

'meibom' : 0.15016748598

'meibm' : 2.59336099585e-05

'meilbm' : 0.0

'meicm' : 0.0

'mheibm' : 0.0

mirgy -> mirigy

'mirigy' : 0.150101293933

'miragy' : 0.15

'mirgy' : 2.59336099585e-05

'mitrgy' : 0.0

'miagy' : 0.0

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Results

- Manually corrected test set (~3500 tokens)
- Linear model with different weighting schemes
- Precision, recall, F-measure
- Correct suggestion in first 5 ⟶ 99%

OOV	VOC	SZEGED	BNO	HUMOR	ORIG	PRECISION	RECALL	F
0.05	0.25	0.15	0.2	0.15	0.1	0.70	0.75	0.72

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Results

- Best combination:
  - The clinical records corpus has highest weight
  - Other models, morphology
  - Original word form

OOV	VOC	SZEGED	BNO	HUMOR	ORIG	PRECISION	RECALL	F
0.05	0.25	0.15	0.2	0.15	0.1	**0.70**	**0.75**	**0.72**

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Examples

EREDETI SZÖVEG:

Meibm mirgy nyílások helyenként sárgás kupakszeráűen elzáródtak, ezeket megint túvel megnyitom

JAVÍTOTT SZÖVEG:

Meibom mirigy nyílások helyenként sárgás kupakszervűen elzáródtak , ezeket megint tűvel megnyitom

EREDETI SZÖVEG:

A beteg intraorbitalis implatatumot is kapott ezért klinikánkon szeptember végén, október elején előzetes telefonnegbeszélés után kontrollvizsgálat javasolt.

JAVÍTOTT SZÖVEG:

A beteg intraorbitalis implantatumot is kapott ezért klinikánkon szeptember végén , október elején előzetes telefonmegbeszélés után kontrollvizsgálat javasolt .

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Typical errors

- Unintentional typing errors
  - Weighted edit distance can handle these (except for non neighboring letters)
- „Intentional" deviation from standard orthography (a mismatch of actual usage and official standard)
  - Multiword vs. one word expressions, hyphenation
  - Vowel length
  - Spelling of foreign words, affixes
  - Abbreviations
  - Lower/uppercase forms

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Usage vs. standard

- zöldhályog, bentfekvés, kézbeadva, éleshatású
- ugy, leirt, degenerativ
- degeneratioja, progredial, fluorometholon, szemhéjtoilettet
- lsin

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Usage vs. standard

- zöld·hályog, bent·fekvés, kézbe·adva, éles·hatású
- úgy, leírt, degeneratív
- degeneratiója, progrediál, fluorometholone, szemhéjtoilette-et/szemhéjtoalettet
- l.·sin.

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

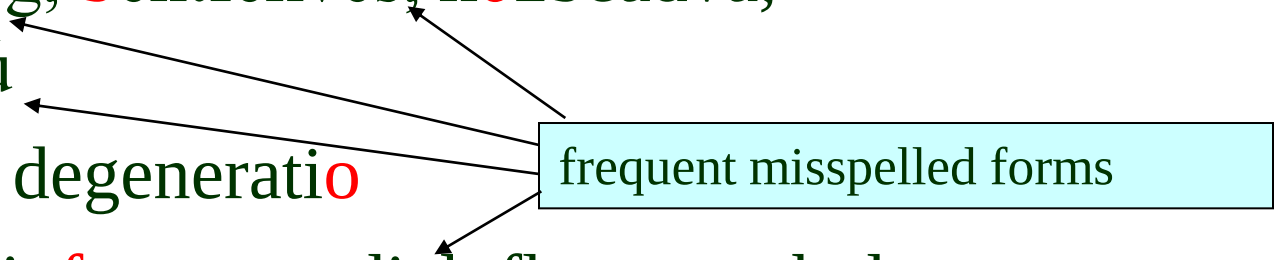
Actual output

- zöldhályog, bentfekvés, közbeadva, éleshatárú
- egy, leirt, degeneratio
- degeneratiofa, progredial, fluorometholone, szemhéjtoilette
- sin

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Actual output

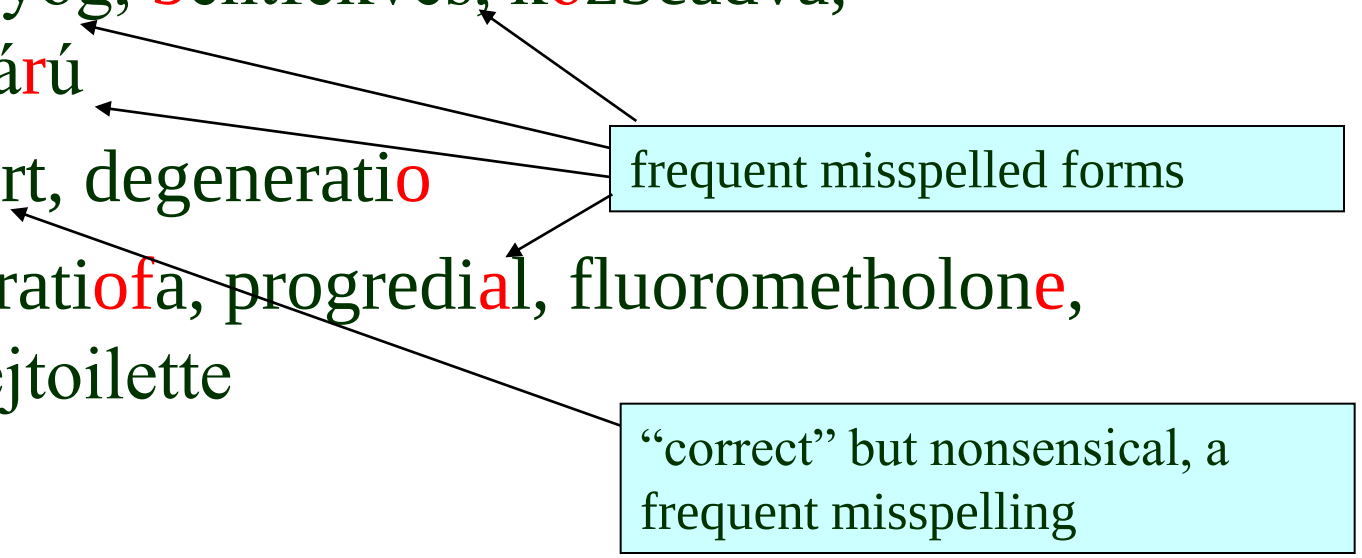
- zöldhályog, bentfekvés, közbeadva, éleshatárú
- egy, leirt, degeneratio
- degeneratiofa, progredial, fluorometholone, szemhéjtoilette
- sin



This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Actual output

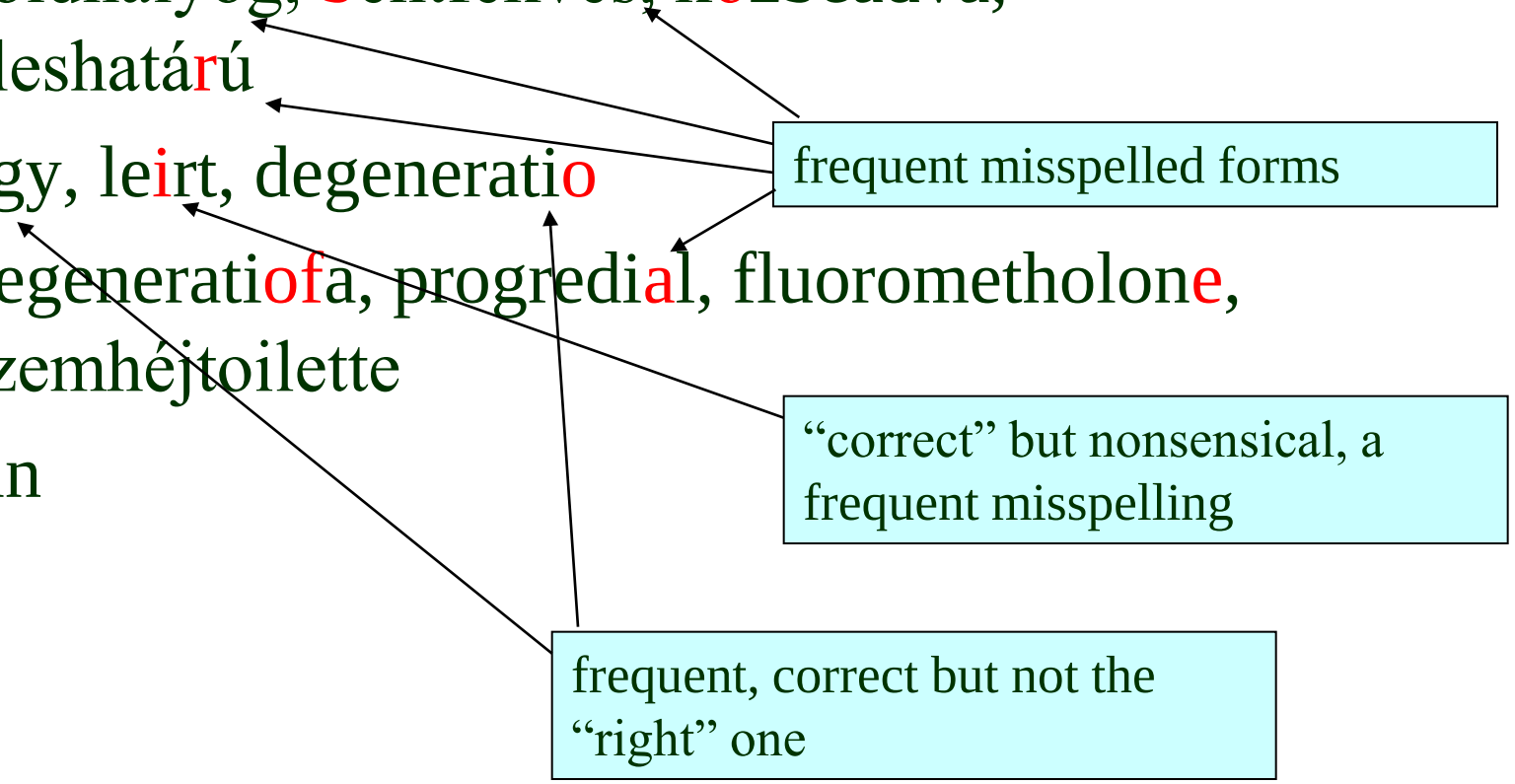
- zöldhályog, bentfekvés, közbeadva, éleshatárú
- egy, leirt, degeneratio
- degeneratiofa, progredial, fluorometholone, szemhéjtoilette
- sin



This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Actual output

- zöldhályog, bentfekvés, közbeadva, éleshatárú
- egy, leirt, degeneratio
- degeneratiofa, progredial, fluorometholone, szemhéjtoilette
- sin



This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Because

- We do not handle insertion or deletion of space
- Differences in frequency might overweight differences of forms
- Edit distance of correct form is greater than 1

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Or because

- Morphology does not recognize word and it is not frequent either
- The morphology accepts the misspelled form (we do not check word context)
- Our word lists and corpus are not big enough – they do not compensate for the overgeneration of morphology

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Some other problematic cases

- oismret ↛ ismert, szövődméynmetes ↛ szövődménymentes
- kórelőzméynébenidőskori ↛kórelőzménnyében·időskori
- Alcon → Arcon, exophthalmusban → enophthalmusban
- Neomycin → Neomucin, PolyLens → MolyLens

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

- keeping frequent “intentional” misspelled forms is better than “correcting” them to an orthographically correct but not intended (possibly nonsensical) form
- results are subjectively better than the F score suggests

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Further plans

- Typical spelling errors might be corrected systematically
  - Adj+N, N+Dir+V
- Instead of using simple edit distance, a better error model is to be built from the corpus
  - once we have a corpus created
- Using larger corpora and word lists

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

Teh Edn

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002

The End

This work was partially supported by TÁMOP-4.2.1.B-11/2/KMR-2011-0002