Resolving abbreviations in clinical texts without pre-existing structured resources

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

{siklosi.borbala, novak.attila, proszeky.gabor}@itk.ppke.hu

[1]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics

[2]MTA-PPKE Hungarian Language Technology Research Group

1083 Budapest, Práter u. 50/a

Resolving abbreviations in clinical texts without pre-existing structured resources

Borbála Siklósi[1], Attila Novák[1,2], Gábor Prószéky[1,2]
{siklosi.borbala, novak.attila, proszeky.gabor}@itk.ppke.hu
[1]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
[2]MTA-PPKE Hungarian Language Technology Research Group
1083 Budapest, Práter u. 50/a

Clinical abbreviations

Domain	Abbr.	Resolution	Hungarian	English
Standard	o.d.	oculus dexter	jobb szem	right eye
	med. gr.	mediocris gradus	közepes fokú	medium grade
Domain specific	o.	oculus	szem	eye
	o.	os	csont	bone
Domain specific common	sü	saját szemüveg	saját szemüveg	own glasses
	fén	fényérzés nélkül	fényérzés nélkül	no sense of light
	n	normál	normál	normal
Common	köv	következő	következő	next
	lsd	lásd	lásd	see

Results

Length	Lexicon size	Precision	Recall	F2	Baseline (F)
ALL	136	93.23%	78.29%	80.88%	60.52%
	44	88.37%	68.73%	71.92%	75.71%
>1	136	96.22%	86.23%	88.05%	
	44	90.63%	79.46%	81.46%	

Resolution process

1. **Find each possible partitioning of the tokens into non-overlapping spans**

```
| Exstirp. | tu. | et | reconstr. | pp. | inf. | l. | d. |
  | Exstirp. tu. | et | reconstr. | pp. | inf. | l. d. |
   | Exstirp. tu. | et  | reconstr. pp. | inf. | l. d. |
     | Exstirp. tu. et | reconstr. pp. inf. | l. d. |
       | Exstirp. tu. et reconstr. pp. inf. l. d. |
                                  etc.
```

2. **Generate regular expressions for each span**

```
o. s.   →  o[^ ]* s[^ ]*   →  oculus sinister
os      →  os[^ ]*         →  osteoporosis
        →  o[^ ]* s[^ ]*   →  oculus sinister
```

3. **Search for matches in the corpus (with context)**
4. **Regenerate patterns based on (partially) resolved fragments**
   - Department of ophthalmology (600,792 tokens)
   - Single token fragments are not considered
   - Pruning based on
     - corpus frequency of the matched result
     - length of the covered span
5. **Match regular expression against the lexicons**
   - Ophthalmology sections of the descriptions of the official coding system for diseases, anatomical structures, medical procedures,
   - a medical dictionary
   - domain-specific lexicon created manually with the help of a medical expert → 97 entries
6. **Rank resolution candidates**
   - Three components of the score:
     1. the number of all tokens in the sequence covered by a resolved form
     2. the size of the longest span covered
     3. the size of the shortest span covered

```
| myop. maj. | gr. | o. u. |  →
|myopia major | gradus | oculi utrisque| 5,2,1
```