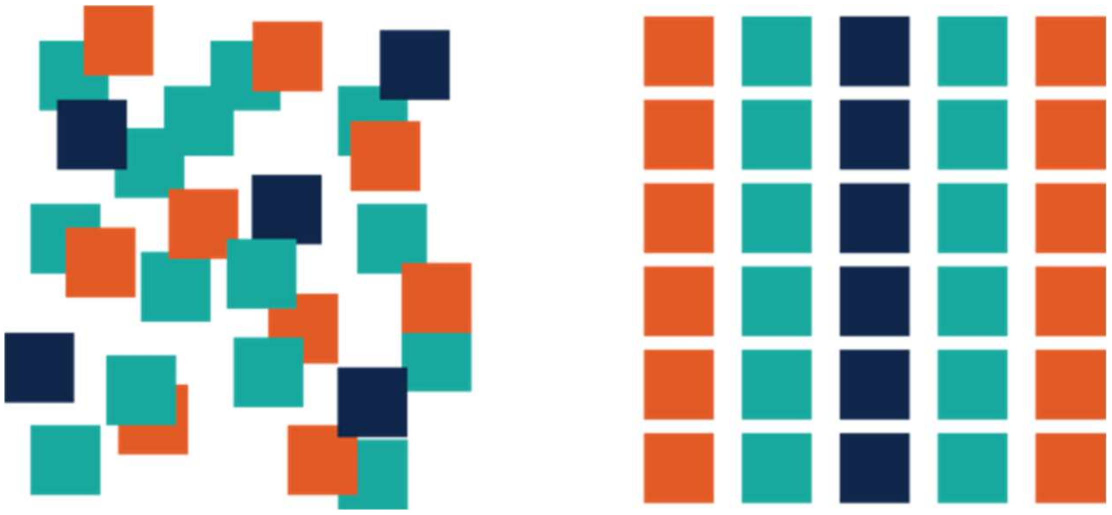
Spark NLP: A Versatile Solution for Structuring Data from Endoscopy Reports

Andrei Constantin IOANOVICI
Stefan Marius MĂRUŞTERI
Andrei Marian FEIER
Alina Dia TRÂMBIȚAS-MIRON

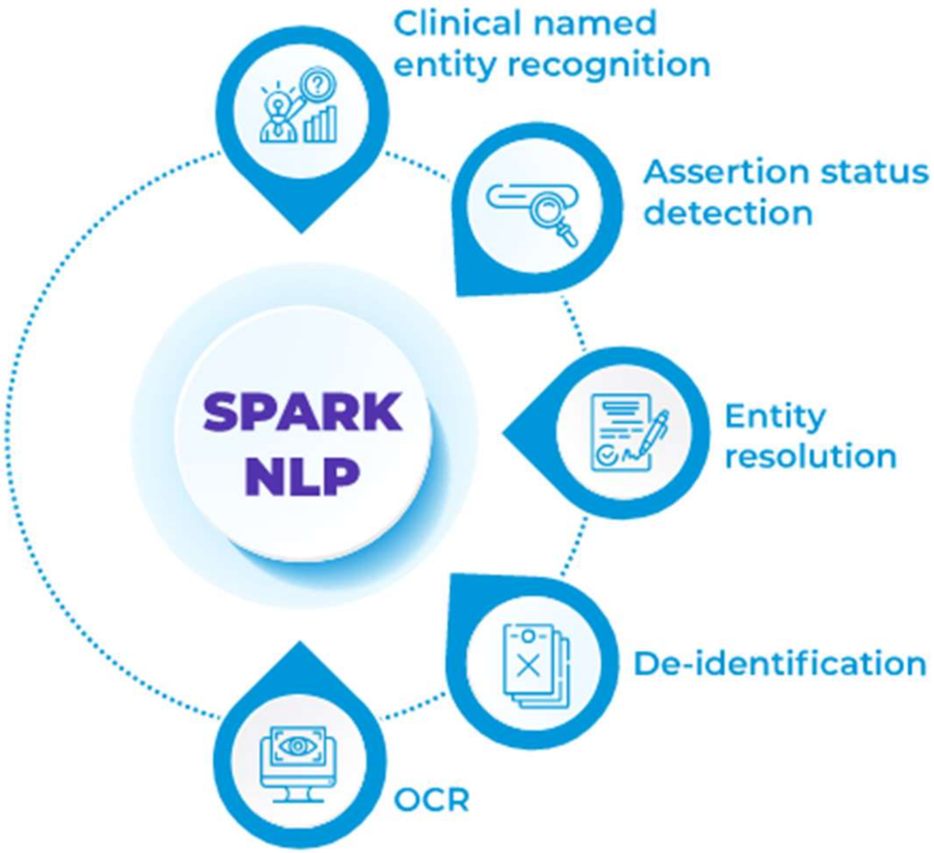
Electronic Health Records

- Electronic Health Records (EHRs) = digital storing patients healthcare events and observations
- ubiquitous yet essential to the delivery of modern healthcare and for research purposes as well.
- The data within the EHRs can be found either in a **structured** state or **unstructured**.



Unstructured data → Structured data : Challenges

NLP



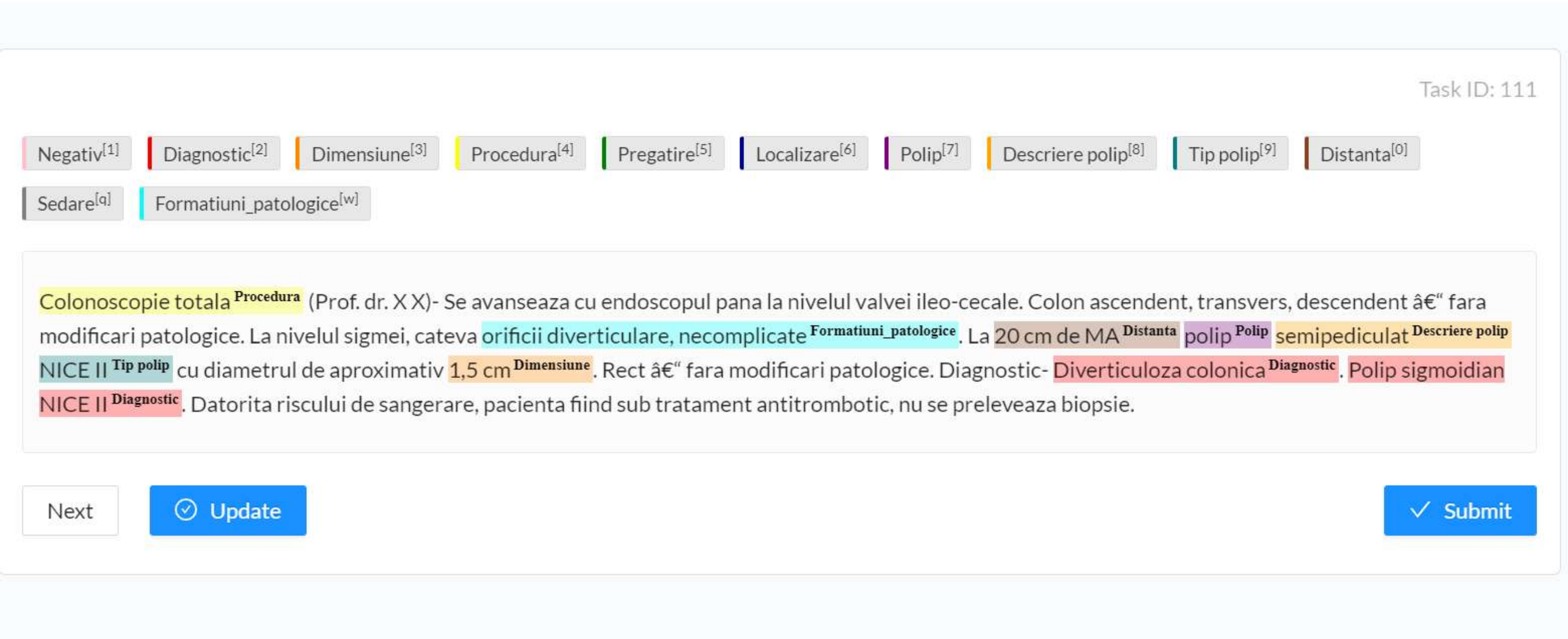
[John Snow Labs](#) - [Spark NLP for Healthcare](#)

UMFST Targu Mures – JSL License for research purposes

Materials and methods

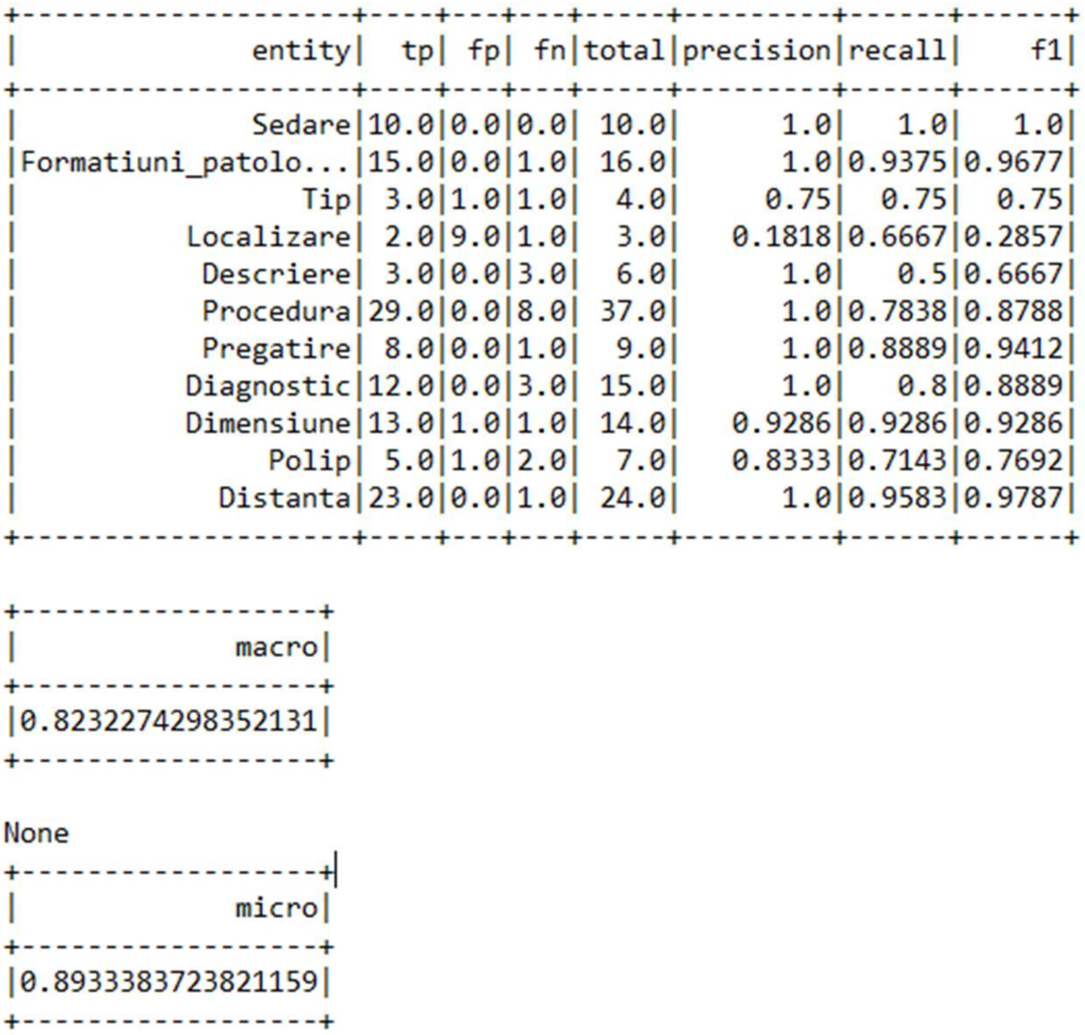
- Endoscopy reports (colon polyps):
  - Gastroenterology Dept. Mures County Clinical Hospital
  - Reports in Romanian language
- Extracted features pertaining to colon polyps:
  - Size, Type, etc.
  - Endoscopy findings, procedure, diagnosis;
- Annotated 100 documents
- 80 documents used for training set
- 20 documents used for test set

Data Annotation



Colonoscopie totala Procedura (Prof. dr. X X)- Se avanseaza cu endoscopul pana la nivelul valvei ileo-cecale. Colon ascendent, transvers, descendent â€“ fara modificari patologice. La nivelul sigmei, cateva orificii diverticulare, necomplicate Formatiuni_patologice. La 20 cm de MA Distanta polip Polip semipediculat Descriere polip NICE II Tip polip cu diametrul de aproximativ 1,5 cm Dimensiune. Rect â€“ fara modificari patologice. Diagnostic- Diverticuloza colonica Diagnostic. Polip sigmoidian NICE II Diagnostic. Datorita riscului de sangerare, pacienta fiind sub tratament antitrombotic, nu se preleveaza biopsie.

Trained Models: Metrics

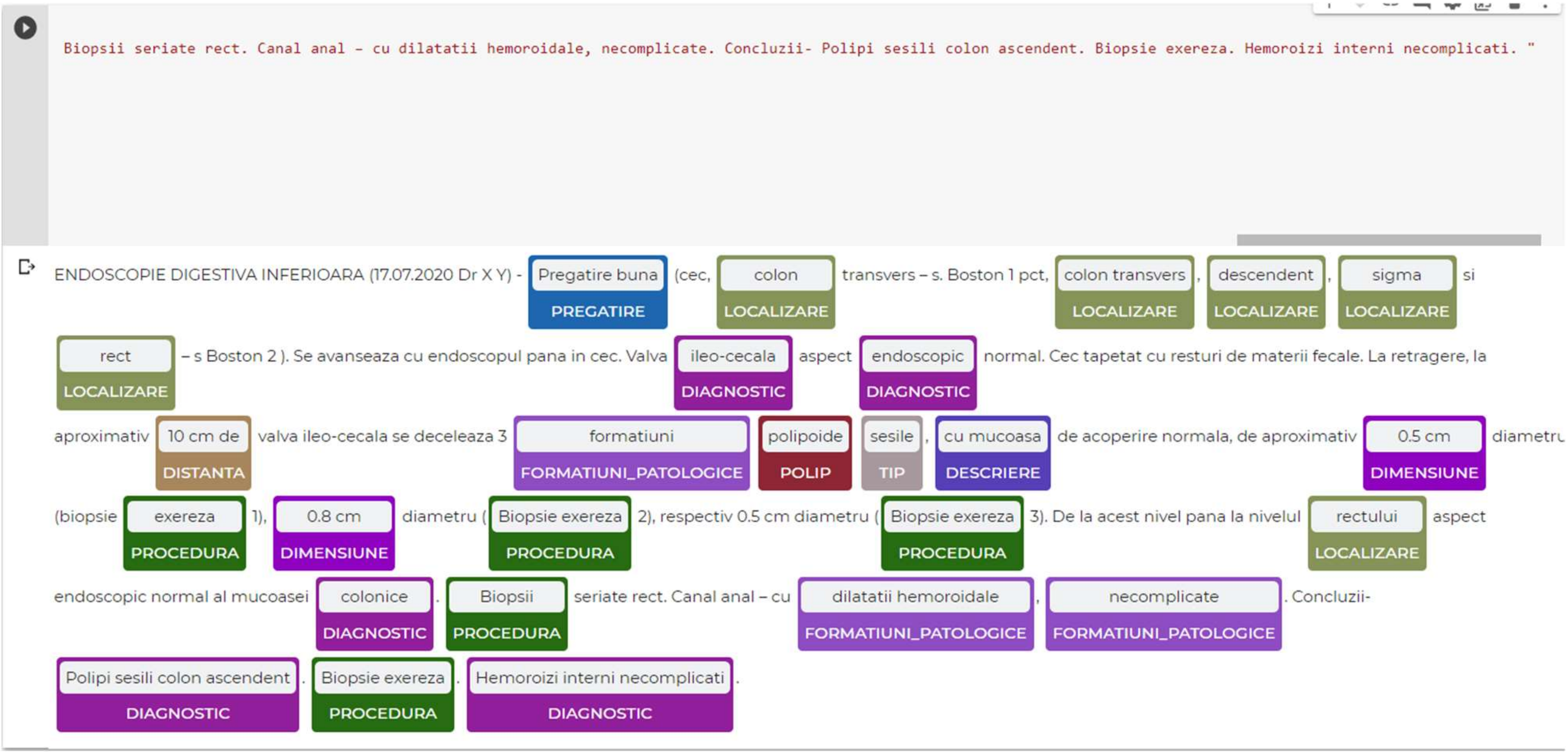


```
+--------------------+----+---+---+-----+---------+------+------+
|              entity|  tp| fp| fn|total|precision|recall|    f1|
+--------------------+----+---+---+-----+---------+------+------+
|              Sedare|10.0|0.0|0.0| 10.0|      1.0|   1.0|   1.0|
|Formatiuni_patolo...|15.0|0.0|1.0| 16.0|      1.0|0.9375|0.9677|
|                 Tip| 3.0|1.0|1.0|  4.0|     0.75|  0.75|  0.75|
|         Localizare| 2.0|9.0|1.0|  3.0|   0.1818|0.6667|0.2857|
|           Descriere| 3.0|0.0|3.0|  6.0|      1.0|   0.5|0.6667|
|           Procedura|29.0|0.0|8.0| 37.0|      1.0|0.7838|0.8788|
|           Pregatire| 8.0|0.0|1.0|  9.0|      1.0|0.8889|0.9412|
|          Diagnostic|12.0|0.0|3.0| 15.0|      1.0|   0.8|0.8889|
|          Dimensiune|13.0|1.0|1.0| 14.0|   0.9286|0.9286|0.9286|
|               Polip| 5.0|1.0|2.0|  7.0|   0.8333|0.7143|0.7692|
|            Distanta|23.0|0.0|1.0| 24.0|      1.0|0.9583|0.9787|
+--------------------+----+---+---+-----+---------+------+------+

+------------------+
|             macro|
+------------------+
|0.8232274298352131|
+------------------+

None
+------------------+
|             micro|
+------------------+
|0.8933383723821159|
+------------------+
```

Results



```
Biopsii seriate rect. Canal anal – cu dilatatii hemoroidale, necomplicate. Concluzii- Polipi sesili colon ascendent. Biopsie exereza. Hemoroizi interni necomplicati. "
```

ENDOSCOPIE DIGESTIVA INFERIOARA (17.07.2020 Dr X Y) - Pregatire buna PREGATIRE (cec, colon LOCALIZARE transvers – s. Boston 1 pct, colon transvers LOCALIZARE , descendent LOCALIZARE , sigma LOCALIZARE si
rect LOCALIZARE – s Boston 2). Se avanseaza cu endoscopul pana in cec. Valva ileo-cecala DIAGNOSTIC aspect endoscopic DIAGNOSTIC normal. Cec tapetat cu resturi de materii fecale. La retragere, la
aproximativ 10 cm de DISTANTA valva ileo-cecala se deceleaza 3 formatiuni FORMATIUNI_PATOLOGICE polipoide POLIP sesile TIP , cu mucoasa DESCRIERE de acoperire normala, de aproximativ 0.5 cm DIMENSIUNE diametru
(biopsie exereza PROCEDURA 1), 0.8 cm DIMENSIUNE diametru (Biopsie exereza PROCEDURA 2), respectiv 0.5 cm diametru (Biopsie exereza PROCEDURA 3). De la acest nivel pana la nivelul rectului LOCALIZARE aspect
endoscopic normal al mucoasei colonice DIAGNOSTIC . Biopsii PROCEDURA seriate rect. Canal anal – cu dilatatii hemoroidale FORMATIUNI_PATOLOGICE , necomplicate FORMATIUNI_PATOLOGICE . Concluzii-
Polipi sesili colon ascendent DIAGNOSTIC . Biopsie exereza PROCEDURA . Hemoroizi interni necomplicati DIAGNOSTIC .

Discussion

- This is one of the first experiments in Romanian language using NLP for extracting structured data from unstructured clinical notes
- Given the small dataset, the model performed well, with an overall precision of 0.823. Because there was a certain amount of heterogeneity in the labeled documents, a bigger dataset is required to improve the metrics. In the future, we aim to increase the dataset to at least 400 documents.
- The solution can be used in combination with other structured data such as laboratory tests, imaging, images/videos from endoscopy procedures in order to create an optimal patient profiling.

Conclusion

- This paper has presented a solution for obtaining structured data from unstructured endoscopy reports regarding colon polyps.
- Because it used reports in Romanian language, it paves the way for future work for developing optimal solutions that can be used in real life in Romanian Hospitals
- This can be integrated into an information system to assist physicians, as the implementation can be a web or mobile application for hospital and clinic use.

References

- Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. Int J Med Inform. 2018 Jun;114:57-65. doi: 10.1016/j.ijmedinf.2018.03.013. Epub 2018 Mar 26. PMID: 29673604.
- https://www.researchgate.net/publication/342878934_Machine_Learning_Models_for_Cancer_Type_Classification_with_Unstructured_Data [accessed Sep 12 2021].
- Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42:760–772.
- Loui, Ronald P., and Ashley Hollinshead. "Efficient Population of Structured Data Forms for Medical Records Using Syntactic Constraints and Intermediate Text." *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2016
- Kong HJ. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res. 2019 Jan;25(1):1-2. doi: 10.4258/hir.2019.25.1.1. Epub 2019 Jan 31. PMID: 30788175; PMCID: PMC6372467.
- Veysel Kocaman, David Talby, Spark NLP: Natural Language Understanding at Scale, Software Impacts, Volume 8, 2021, 100058, ISSN 2665-9638, https://doi.org/10.1016/j.simpa.2021.100058.