# Analiza comparativă a principalilor algoritmi SaaS pentru recunoașterea automată de entități în limba română

**Bogdan IANCU**
Academia de Studii Economice, Piața Romană Nr. 6, Sector 1, București, 010374, România
bogdan.iancu@ie.ase.ro

**Rezumat:** Lucrarea de față își propune analiza comparativă a principalilor algoritmi de Named Entity Recognition disponibili în cloud, aplicați pentru texte scrise în limba română. Contextul în care acești algoritmi sunt analizați este cel al web-ului semantic, în cadrul căruia încă persistă problema identificării de noi entități ce pot fi legate la ontologii existente. Sunt definite procese prin care textul este tradus într-una din limbile suportate de algoritmii furnizați de DBpedia (DBpedia Spotlight), Google (Google Cloud Natural Language API), Microsoft (modulul NER din Azure Machine Learning Studio) și IBM (IBM Watson Natural Language Understanding), pentru ca mai apoi să fie utilizat scorul $F_1$ pentru a determina procesul optim. Articolul se încheie cu o comparație între rezultatele obținute și performanța altor algoritmi NER specializați pe limba engleză sau independenți de limbă.

**Cuvinte cheie:** Web semantic, NER, LOD, SaaS.

# Comparative Analysis of the Main SaaS Algorithms for Named Entity Recognition Applied for Romanian Language

**Abstract:** This paper proposes a comparative analysis of the main Name Entity Recognition algorithms available in cloud, applied for texts written in Romanian. The context of this analysis is the one of the semantic web, where the problem of identifying new entities and linking them to existing ontologies persists. There are processes defined that allow the text written in Romanian to be translated in one of the languages supported by the algorithms provided by DBpedia (DBpedia Spotlight), Google (Google Cloud Natural Language API), Microsoft (the NER module from Azure Machine Learning Studio) and IBM (IBM Watson Natural Language Understanding), and afterwards the $F_1$ score is computed in order to identify the optimal process. The article ends with a comparison between the obtained results and the performance achieved by NER algorithms specialized for English or language independent.

**Keywords:** Semantic web, NER, LOD, SaaS

## 1. Introducere

În contextul web-ului semantic a existat încă de la început problema identificării automate de entități ce pot fi legate de ontologii existente. Chiar dacă una dintre cele mai mari ontologii interdisciplinare existente, denumită DBpedia, numără în prezent 4,22 milioane de entități pentru limba engleză [1], acestea nu sunt nici pe departe suficiente pentru a acoperi toate personalitățile, locurile, operele de creație sau organizațiile existente în acest moment sau care au existat vreodată. Dezideratul

final în domeniul web-ului semantic este legarea tuturor entităților disponibile în internet într-o ontologie a ontologiilor sau, conform creatorului world wide web, Tim Berners-Lee, conectarea la LOD – Linked Open Data [2] a tuturor entităților posibile. În acest scop, dar nu numai (algoritmii existând aprioric apariției web-ului semantic), au fost definiți diverși algoritmi de tip NER (Named Entity Recognition) ce pot înlesni identificarea automată de entități în text liber. La început, algoritmii de acest tip se concentrau în special pe indentificare de organizații, persoane și locații [12], însă în prezent mulți dintre ei oferă mai mult de atât [11]. Problema care se ridică în acest moment este dată de faptul că majoritatea acestor algoritmi funcționează pe un număr limitat de limbi [9] și, cu toate că au existat încercări de a crea unul independent de limbă cu aplicabilitate și în limba română, rezultatele nu au fost cele așteptate [3]. Scopul lucrării de față este indentificarea unui proces cu randament maxim prin care, folosind algoritmi de tip NER disponibili în cloud ca și SaaS (Software as a Service), poate fi adnotat un text oarecare scris în limba română.

## 2. Prezentarea algoritmilor NER comparați

- ***DBpedia Spotlight***

DBpedia, varianta semantică a celebrei enciclopedii electronice Wikipedia, pune la dispoziția utilizatorilor și un algoritm de NER, denumit DBpedia Spotlight [4]. Acesta permite adnotarea semantică de text scris într-una din limbile engleză, germană, flamandă, franceză, italiană, rusă, spaniolă, portugheză, maghiară sau turcă. Pe lângă textul propriu-zis, Spotlight mai permite, drept parametrii de intrare, un coeficient de încredere minim dorit și selectarea unor clase DBpedia, Freebase sau Schema.org ce sunt de interes (Figura 1). Algoritmul este open source și poate fi instalat pe un server propriu, însă este disponibil și sub formă de serviciu web (http://model.dbpedia-spotlight.org) sau prin intermediul unei interfețe demonstrative (http://demo.dbpedia-spotlight.org/).

- ***Google Cloud Natural Language***

Printre serviciile de cloud computing oferite de cei de la Google, prin platforma denumită generic Google Cloud, se numără și un API de procesare de limbaj natural [5]. Acesta oferă, pe lângă algoritmul de interes pentru lucrarea de față, cel de Entity Recognition, și alți algoritmi de analiză de sintaxă, de analiză a sentimentelor sau de clasificare de conținut. În ceea ce privește algoritmul de NER al celor de la Google, acesta este disponibil în 10 limbi (chineză simplificată, chineză tradițională, engleză, franceză, germană, italiană, japoneză, coreană, portugheză și spaniolă) și poate fi accesat prin intermediul unui serviciu web. Asemănător DBpedia, și cei de la Google oferă pe site-ul oficial o aplicație demonstrativă ce ilustrează modul de funcționare al algoritmului (Figura 2).


**Figura 1.** Interfața demonstrativă a DBpedia Spotlight

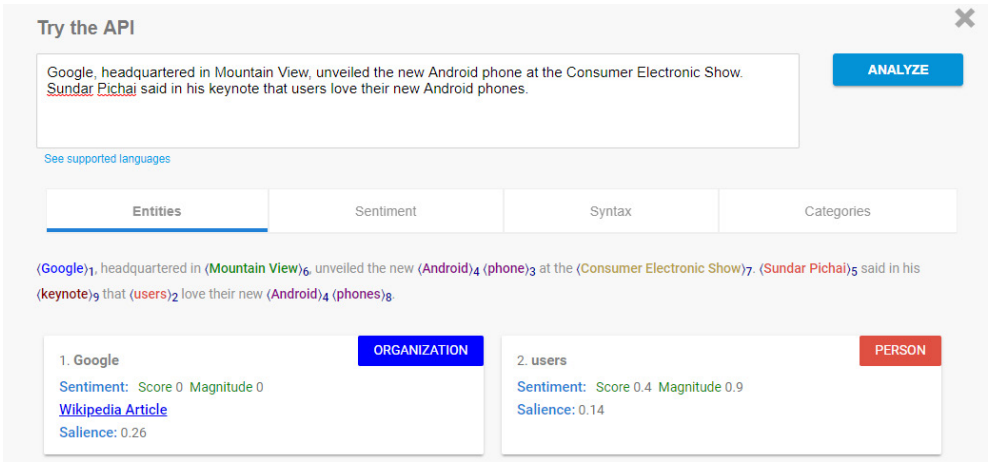


**Figura 2.** Interfața demonstrativă a Google Cloud Natural Language API

- ***Azure Machine Learning***

Și platforma de Cloud a celor de la Microsoft, denumită Azure, oferă un modul NER în cadrul componentei Azure Machine Learning Studio [10]. În comparație cu cei doi algoritmi deja prezentați, cel oferit de Microsoft poate adnota doar text în limba engleză. Rezultatele sunt și ele limitate la 3 tipuri de entități: persoane (PER), locații (LOC) și organizații (ORG). Modulul NER din Azure poate fi folosit atât prin

intermediul unei interfețe de tip WYSIWYG (What You See Is What You Get) – Figura 3, cât și exportat ca serviciu web ce poate fi apelat ulterior.

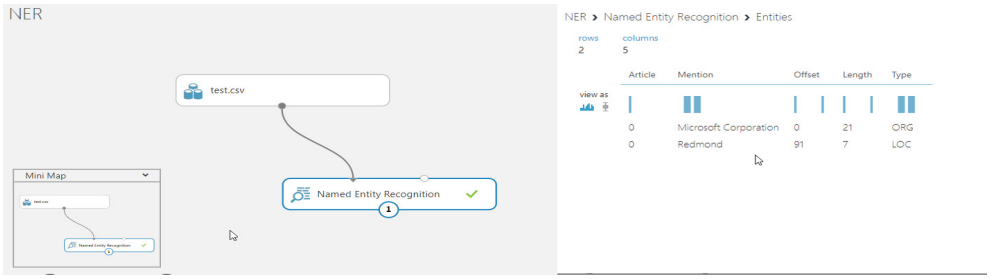


**Figura 3.** Intefața WYSIWYG a Azure Machine Learning Studio, împreună cu vizualizarea grafică a entităților indentificate

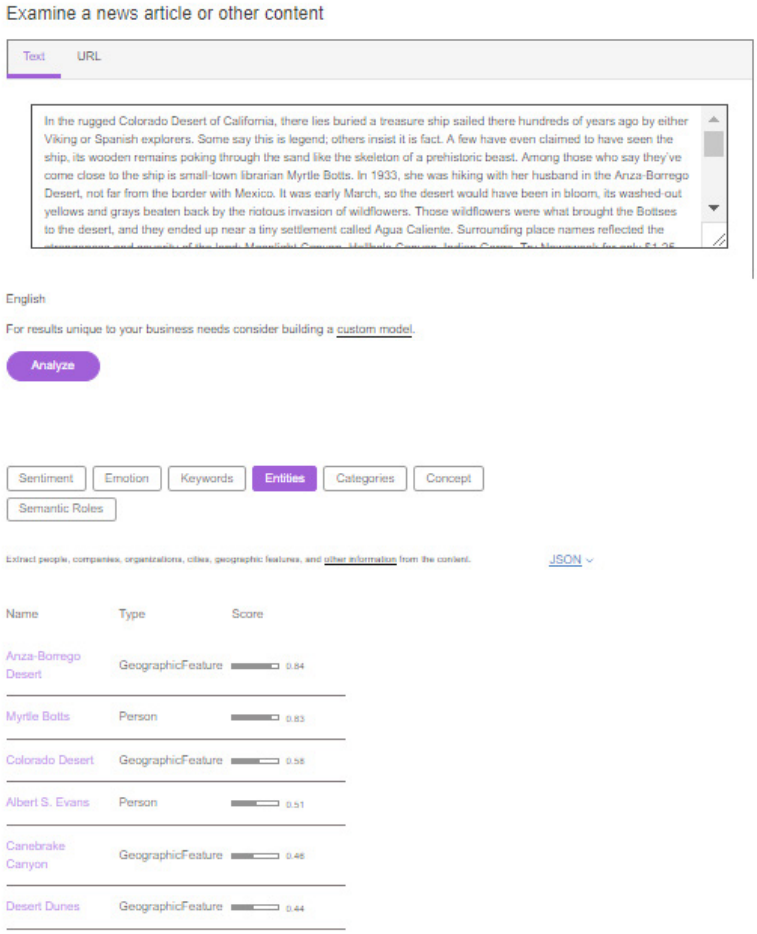


**Figura 4**. Interfața demonstrativă a IBM Watson Natural Language Understanding

- ***IBM Watson Natural Language Understanding***

Integrat în platforma de cloud computing a celor de la IBM (intitulată IBM Bluemix) găsim și serviciul de procesare de limbaj natural denumit Watson [6]. Disponibil în 6 limbi (engleză, chineză, franceză, germană, japoneză și spaniolă), acesta permite adițional algoritmului de Entity Recognition și alți algoritmi specifici procesării de limbaj natural precum: analiză a sentimentelor, a emoțiilor, a cuvintelor cheie, identificare automată de categorii sau de concepte și, nu în cele din urmă, analiza semantică a propozițiilor cu împărțirea cuvintelor componente într-una din categoriile *subiect*, *acțiune* sau *obiect*. Similar cu celelalte platforme prezentate, și algoritmul de NER al celor de la IBM este disponibil atât ca serviciu web, cât și prin intermediul unei pagini demonstrative (Figura 4).

# 3. Analiza comparativă a algoritmilor NER pentru limba română

Având în vedere că niciunul dintre algoritmii NER disponibili în cloud prezentați nu permite identificarea de entități în limba română, se va defini pentru fiecare dintre ei un proces prin care pot fi identificate entități din text liber în limba română. Pentru stabilirea procesului optim, se va folosi scorul $F_1$, calculat ca și medie armonică a ratei de precizie (P) și a celei de recuperare (R). Rata de precizie este definită ca numărul de entități corect recunoscute din totalul de entități identificate, iar cea de recuperare drept numărul de entități identificate corect din numărul total de entități ce ar putea fi identificate [8].

Pentru testarea preliminară vom folosi următoarea secvență de text, prin care este definită informatica pe pagina Wikipedia aferentă:

„Termenul informatică provine din alăturarea cuvintelor informație și matematică. Alte surse susțin că provine din combinația informație și automatică. Istoria informaticii începe înainte de momentul apariției computerului digital. Înainte de anul 1920, termenul de „computer" se referea în limba engleză la un o persoană care efectua calcule (un funcționar). Primii cercetători în ceea ce avea să se numească informatică, cum sunt Kurt Gödel, Alonzo Church și Alan Turing, au fost interesați de problema computațională: ce informații ar putea un funcționar uman să calculeze având hârtie și creion, prin urmărirea pur și simplu a unei liste de instrucțiuni, atât timp cât este necesar, fără să fie nevoie ca el să fie inteligent sau să presupună capacități intuitive. Una din motivațiile acestui proiect a fost dorința de a proiecta și realiza „mașini computaționale" care să automatizeze munca, deseori plictisitoare și nu lipsită de erori, a unui calculator sau computer uman." (*sursă text: https://ro.wikipedia.org/wiki/Informatic%C4%83*).

Astfel, pentru DBpedia Spotlight se va traduce textul în limba engleză folosind Google Translate, se va aplica NER pe acesta, iar mai apoi se vor lua din DBpedia entitățile corespondente din limba română pentru fiecare din cele din limba engleză identificate. Acest lucru este posibil prin intermediul relației de tip *owl:sameAs* din ontologia DBpedia ce leagă entitățile similare din limbi diferite.

La o primă rulare a algoritmului folosindu-se un coeficient de încredere de 50% și traducerea în engleză oferită de Google, au fost identificate următoarele entități: *http://dbpedia.org/resource/Automation*, *http://dbpedia.org/resource/Computer_science*, *http://dbpedia.org/resource/English_language*, *http://dbpedia.org/resource/Kurt_G%C3%B6de*, *http://dbpedia.org/resource/Alonzo_Church*, *http://dbpedia.org/resource/Alan_Turing*, toate corect identificate.

La o nouă rulare, folosind de această dată un coeficient de încredere de 40% și tot o traducere în limba engleza, am obținut următoarele entități în plus: *http://dbpedia.org/resource/Computer*, *http://dbpedia.org/resource/Pencil*, ambele corect identificate.

Cea de-a treia rulare cu un coeficient de încredere scăzut la 30% a oferit un total de 37 de entități, din care 10 identificate greșit.

Toate rezultatele obținute sunt disponibile în Tabelul I, cu precizarea că au fost folosite și alte limbi de origine latină disponibile atât în Goolge Translate, cât și în DBpedia Spotlight (italiană și franceză), pentru a testa dacă pot fi obținute rezultate mai bune în acest fel.

**Tabelul I.** Scorul $F_1$ pentru entitățile identificate de către DBpedia Spotlight

| Nr. crt. | Limba în care a fost tradus textul | Coeficient de încredere | P (%) | R (%) | $F_1$ (%) |
|---|---|---|---|---|---|
| 1 | engleză | 50% | 100 | 16,21 | 27,89 |
| 2 | engleză | 40% | 100 | 21,62 | 35,55 |
| 3 | engleză | 35% | 100 | 37,83 | 54,89 |
| 4 | engleză | 30% | 72,97 | 72,97 | 72,97 |
| 5 | italiană | 50% | 100 | 18,91 | 31,8 |
| 6 | italiană | 40% | 100 | 35,13 | 51,99 |
| 7 | italiană | 35% | 100 | 37,83 | 54,89 |
| 8 | italiană | 30% | 100 | 40,54 | 57,69 |
| 9 | franceză | 50% | 100 | 10,81 | 19,51 |
| 10 | franceză | 40% | 100 | 16,21 | 27,89 |
| 11 | franceză | 35% | 100 | 18,91 | 31,8 |
| 12 | franceză | 30% | 88,88 | 21,62 | 34,77 |

Vom repeta experimentul folosind Google Cloud Natural Language API de această dată, cu precizarea că acesta nu suportă modificarea coeficientului de încredere. Vom considera drept entități corect identificate doar pe acelea ce nu sunt incluse în clasa OTHER sau conțin o trimitere către o resursă Wikipedia. Procedăm astfel deoarece fiecărei clase dintre celelalte (PERSON, ORGANIZATION, CONSUMER GOOD, LOCATION, EVENT) îi putem asocia o clasă schema.org și astfel poate fi conectată în LOD. La fel putem proceda și pentru entitățile ce fac trimitere la Wikipedia, pe baza

URL-ului putându-se determina clasa DBpedia aferentă (majoritatea paginilor Wikipedia au drept corespondent o entitate DBpedia). Tabelul II prezintă rezultatele obținute:

**Tabelul II.** Scorul $F_1$ pentru entitățile identificate de către Google Cloud Natural Language API

| Nr. crt. | Limba în care a fost tradus textul | P (%) | R (%) | $F_1$ (%) |
|---|---|---|---|---|
| 1 | engleză | 85,71 | 32,43 | 47,05 |
| 2 | italiană | 100 | 37,83 | 54,89 |
| 3 | franceză | 87,5 | 37,83 | 52,82 |

În continuare vom testa modulul NER din Azure Machine Learning Studio. Având în vedere că acesta suportă doar limba engleză, vom proceda la traducerea textului folosind atât Google Translate, cât și Bing Translator (serviciul de traduceri oferit de cei de la Micosoft). În Tabelul III sunt prezentate rezultatele ce arată că, indiferent de platforma ce a furnizat traducerea, modulul de NER din Azure ML a reușit identificarea doar a celor 3 entități de tip persoană (Kurt Gödel, Alonzo Church și Alan Turing).

**Tabelul III.** Scorul $F_1$ pentru entitățile identificate de către modulul NER din Azure ML

| Nr. crt. | Platforma de traducere utilizată | P (%) | R (%) | $F_1$ (%) |
|---|---|---|---|---|
| 1 | Google | 100 | 8,1 | 14,98 |
| 2 | Bing | 100 | 8,1 | 14,98 |

În cele din urmă vom proceda la testarea algoritmului de NER din cadrului serviciului IBM Watson Natural Language Understanding. Vom folosi ca și date de intrare textul tradus folosind Google Translate și vom obține drept date de ieșire toate entitățile identificate. Conectarea entităților la LOD în cazul IBM Watson este foarte simplă, deoarece, în cadrul fișierului JSON returnat de serviciu, se regăsește și adresa DBpedia a resursei identificate (dacă aceasta există, în caz contrar conectarea se va face prin intermediul clasei schema.org aferente).Cu toate că în cadrul IBM Watson există și serviciu cloud ce furnizează traduceri, acesta nu a putut fi utilizat deoarece limba română nu se regăsește printre cele disponibile. Rezultatele obținute în urma testării algoritmului pentru text tradus din limba română în limbile engleză, spaniolă și italiană sunt disponibile în Tabelul IV.

**Tabelul IV.** Scorul $F_1$ pentru entitățile identificate de către IBM Watson Natural Language Understanding

| Nr. crt. | Limba în care a fost tradus textul | P (%) | R (%) | $F_1$ (%) |
|---|---|---|---|---|
| 1 | engleză | 100 | 5,4 | 10,24 |
| 2 | spaniolă | 100 | 8,1 | 14,98 |
| 3 | franceză | 100 | 8,1 | 14,98 |

Vom repeta experimentul pentru un text cu un volum mai mare de cuvinte de această dată (2.854 de cuvinte), care conține un număr de 597 de entități ce pot fi adnotate. Față de experimentul precedent, am renunțat la analiza traducerilor în limba italiană și franceză folosind DBpedia Spotlight pentru coeficienți de încredere mai mari de 30%, rezultatele optime înregistrându-se pentru aceasta valoare. Tabelul V prezintă coeficienții calculați pentru noul text analizat, iar Figura 5 reprezentarea grafică a rezultatelor.

**Tabelul V.** Analiza comparativă a scorului $F_1$ pentru algoritmii analizați

| **Nr. crt.** | **Platforma SaaS utilizată pentru NER** | **Limba în care a fost tradus textul** | **Coeficient de încredere** | **P (%)** | **R (%)** | **$F_1$ (%)** |
|---|---|---|---|---|---|---|
| 1 | DBpedia Spotlight | engleză | 50% | 95,45 | 28,14 | 43,46 |
| 2 | DBpedia Spotlight | engleză | 40% | 97,58 | 40,53 | 57,27 |
| 3 | DBpedia Spotlight | engleză | 35% | 95,36 | 61,39 | 74,69 |
| 4 | DBpedia Spotlight | engleză | 30% | 68,68 | 83,41 | 75,33 |
| 5 | DBpedia Spotlight | italiană | 30% | 97,18 | 52,09 | 67,82 |
| 6 | DBpedia Spotlight | franceză | 30% | 96,36 | 44,38 | 60,77 |
| 7 | Google Cloud | engleză | N/A | 94,31 | 47,23 | 62,93 |
| 8 | Google Cloud | franceză | N/A | 94,16 | 56,78 | 70,84 |
| 9 | Google Cloud | italiană | N/A | 92,78 | 49,58 | 64,62 |
| 10 | Microsoft Azure | engleză (Google Translate) | N/A | 96,11 | 16,58 | 28,28 |
| 11 | Microsoft Azure | engleză (Bing Translator) | N/A | 93,26 | 16,24 | 27,66 |
| 12 | IBM Watson | engleză | N/A | 93,61 | 7,87 | 14,51 |
| 13 | IBM Watson | franceză | N/A | 94 | 8,37 | 15,37 |
| 14 | IBM Watson | spaniolă | N/A | 95,91 | 8,2 | 15,10 |

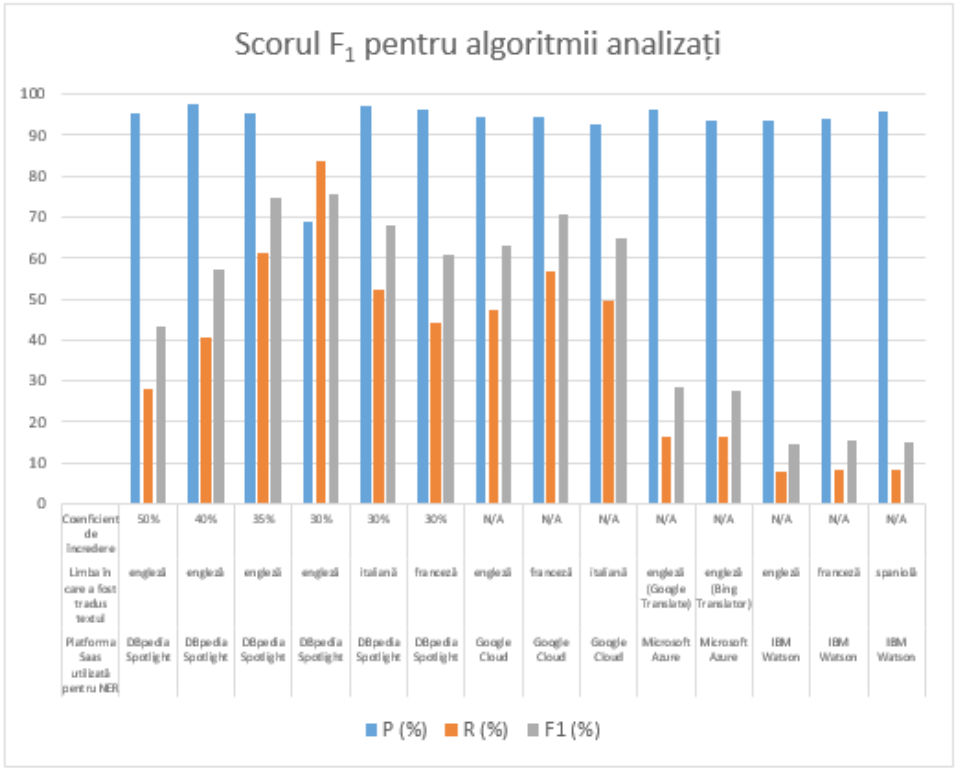


**Figura 5.** Reprezentarea grafică a rezultatelor obținute

# 4. Concluzii

Scopul lucrării de față a fost identificarea unui proces prin care, utilizând un algoritm de NER disponibil ca SaaS în cloud, putem adnota semantic text scris în limba română. Astfel, s-a procedat la utilizarea DBpedia Spotlight pe text tradus într-una din limbile engleză, italiană și franceză, a Google Cloud Natural Language API pe același set de limbi, a modului NER din Azure Machine Learning Studio pe limba engleză și a IBM Watson Natural Language Understanding pentru limbile engleză, franceză și spaniolă. Rezultatele cele mai bune, atât pe un text de mici dimensiuni, cât și pe unul de 2.854 de cuvinte, au fost obținute de algoritmul DBpedia Spotlight pe text tradus în limba engleză, cu un coeficient de încredere setat la 30%. Conform datelor din Figura 5, putem observa că acest algoritm a obținut cea mai bună rată de recuperare (R), chiar dacă cea de precizie (P) a fost cea mai slabă. Toți ceilalți alogoritmi au avut valori bune pentru rata de precizie, însă numărul mic de valori identificate a dus la o rată scăzută a recuperării, acest lucru afectând scorul $F_1$ final.

Dacă ar fi să comparăm scorul final obținut cu alți algoritmi de NER cum sunt cei prezentați în [3] sau [12], am observa că procesul definit peste modulul de NER al DBpedia

Spotlight ($F_1$ = 75,33%) are o eficiență asemănătoare cu cu cel propus de S. Cucerzan și D. Yarowsky ($F_1 \subset$ [65,69%, 75,43%]) și chiar mai bună decât cea obținută de algoritmii CoNLL-2003 pentru limba germană ($F_1 \subset$ [47,74%, 72,41%]). Toate acestea pot face din procesul propus de lucrarea de față alegerea potrivită pentru procedeele de NER pe texte scrise în limba română, în contextul web-ului semantic, în lipsa implementării complete [7] a unui instrument de Prelucrare a Limbajului Natural din limba română.

# BIBLIOGRAFIE


1. About DBpedia, *DBpedia.org*. Accessed 25 January 2018. <http://wiki.dbpedia.org/about>.
2. Bizer, C., Heath, T., Idehen, K. & Berners-Lee, T. (2008). Linked data on the web (LDOW2008). In *Proceedings of the 17th international conference on World Wide Web* (pp. 1265-1266).
3. Cucerzan, S. & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
4. Daiber, J., Jakob, M., Hokamp, C. & Mendes, P. N. (2013, September). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems* (pp. 121-124).
5. Google Cloud Natual Language API Documentation, *Google Cloud Platform*. Accessed 25 January 2018. <https://cloud.google.com/natural-language/docs/>.
6. IBM Knowledge Center - Named Entity Recognition annotator, *IBM Knowledge Center.* Accessed 25 January 2018. <https://www.ibm.com/support/knowledgecenter/en/SS8NLW_10.0.0/com.ibm.watson.wex.aac.doc/aac-tasystemt.html>.
7. Irimia, E. (2015). Accelerarea dezvoltării unui corpus digital adnotat cu relaţii de dependenţă pentru limba română utilizând resurse şi instrumente construite pentru alte limbi, *Revista Română de Informatică* şi *Automatică*, *25*(3), 5-16.
8. Mohit, B. (2014). Named entity recognition, *Natural language processing of semitic languages*, 221-245.
9. Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification, *Lingvisticae Investigationes*, *30*(1), 3-26.
10. Named Entity Recognition – Azure Machine Learning Studio, *Microsoft Docs*. Accessed 25 January 2018. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/named-entity-recognition>.
11. Ritter, A., Clark, S. & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1524-1534).
12. Tjong Kim Sang, E. F. & De Meulder, F. (2003, May). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 142-147).