

Named Entity Recognition with NooJ

Aleksandar Petrovski
International Slavic University, Sveti Nikole
North Macedonia



Zagreb, NooJ 2020

What are Named Entities?



Named entities, in a narrower sense, are concrete or abstract elements of the real world, including people, organizations, companies, places, etc., e.g. *Bill Gates*, *Google*, *Zagreb*.

In a broader sense, the expressions related to time, space, quantity, such as *September 11*, *80 EUR*, are considered as named entities too.

Why is Named Entity Recognition (NER) Important?



- Understanding documents
- Information retrieval - IR
- Machine Translation – MT
- Question Answering - QA
- Speech Recognition

NER Techniques



- Grammars (higher precision, but at the cost of lower recall, a team of computer engineers and computer linguists required, expensive)
- Statistical models (requires a large amount of manually annotated machine learning data, expensive, often only for a specific domain)

Suggested Solution



- NooJ linguistic environment
- NooJ morphological lexicons
- NooJ grammars (morphological and syntactic)

Macedonian Morphological Lexicons

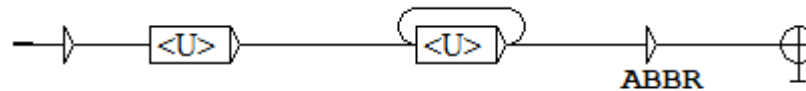


- Basic morphological lexicon (89,026 lemmas, 1,480.201 word forms)
- Lexicon of toponyms (1,398 lemmas, 40,246 word forms)
- Lexicon of names of persons and companies (5,422 lemmas, 157,321 word forms)
- Lexicon of derived adjectives from verbs, with suffixes -chki and -bilon (12,073 lemmas, 281,488 word forms)
- Lexicon of participles derived from verbs (19,552 lemmas, 1,252,328 word forms)
- Lexicon of compound units consisting of multiple words (784 lemmas, 6,289 word forms)

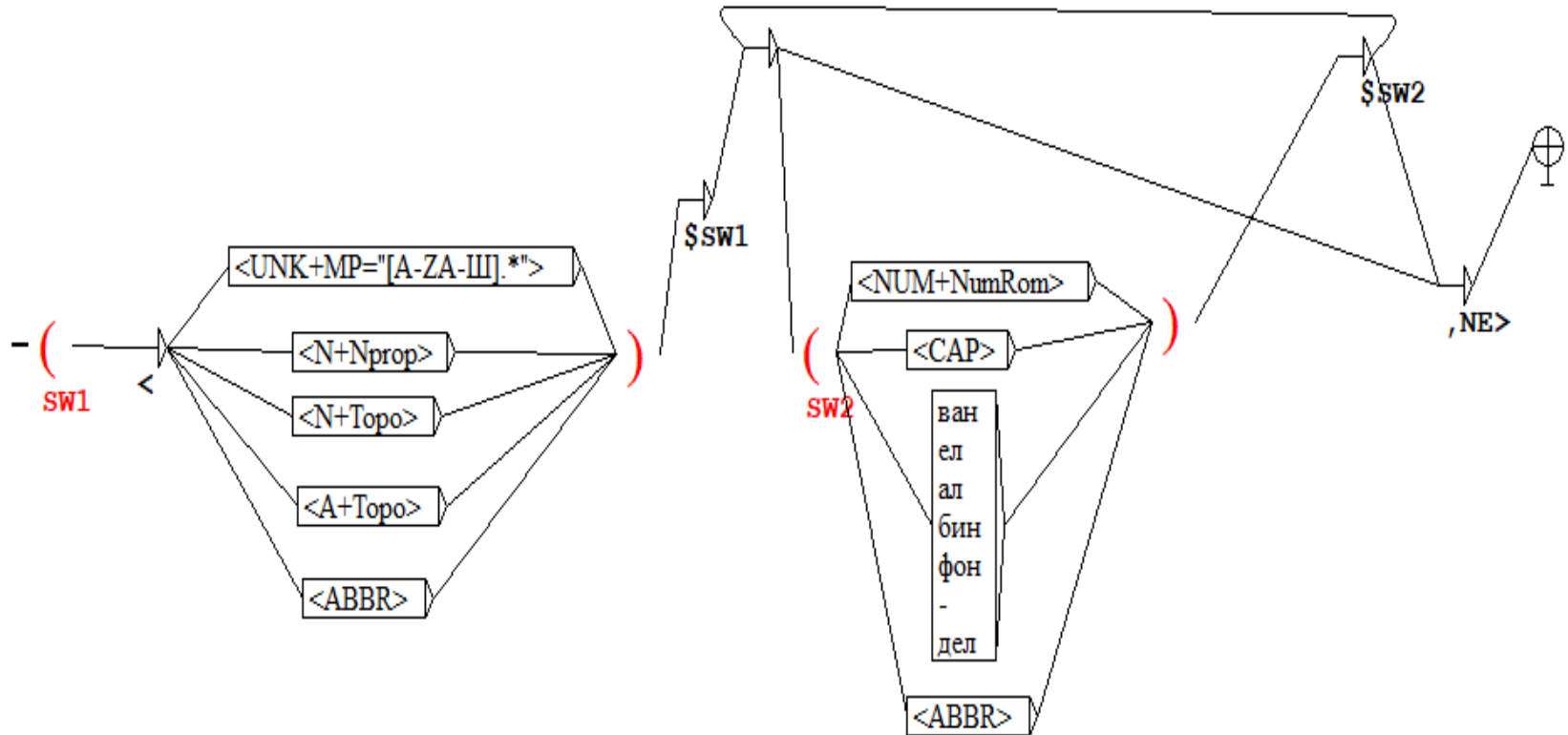


Morphological Grammars

- Negative adjectives and adverbs
- Prefixed verbs
- Diminutives
- Augmentatives
- Roman numbers
- Abbreviations



Syntactic Grammar



Results



Pattern is:

- a string of characters:
- a PERL regular expression:
- a NooJ regular expression:

<NE>

a NooJ grammar:

Syntactic Analysis

Index

- Shortest matches
- Longest matches
- All matches

Limitation

- All occurrences
- Only: occ.
- 1 occ. per match

Reset Concordance

Concordance

Reset Display: characters word forms before, and after. Display: Matches Outputs

Text	Before	Seq.	After
tekst3.not	на пријателот, каде се наоѓа	Кочабамба	? Професорот по географија молчеше за
tekst3.not	шега. – Не знаеш, праша нервозно	Зоран	– Во Јужна Америка, рече професорот
tekst3.not	знаеш, праша нервозно Зоран. – Во	Јужна Америка	, рече професорот. Чекај... Да, во
tekst3.not	рече професорот. Чекај... Да, во	Боливија	Во Андите. Некоје рударско место
tekst3.not	Чекај... Да, во Боливија. Во	Андите	. Некоје рударско место. Зошто ти
tekst3.not	ти е потребно? Да не...	Зоран	го затвори телефонот. Стоеше еден
tekst3.not	Не разпознаваше ниту една буква. ...	Кочабамба	", промрмори за себе. Сфаќаше дека
tekst3.not	од железничката станица. „Возот за	Кочабамба	стои на седмиот колосек. Поаѓа
tekst3.not	седмиот колосек. Поаѓа веднаш.“ И	Зоран	, наместо кон телефонот, потрча кон
tekst3.not	се оддалечуваше со голема брзина.	Зоран Туртул	не стаса да помисли на
tekst3.not	се трие едно од друго.	Зоран	се обидуваше да се фати
tekst3.not	вагонот се веднаш кон земјата,	Зоран Туртул	виде низ прозорецот дека пејзажот
tekst3.not	исчезна. 2. Утредента, на 6 јули, во	скопските	весници можеше да се прочита
tekst3.not	часот, во непосредна близина на	скопската	железничка станица се случи полесна
tekst3.not	Патничкиот воз кој одеше за	Кочани	и Бања, и кој непосредно
tekst3.not	на другата што води од	Ла Паз	до Пуната. Веста прв ја
tekst3.not	води од Ла Паз до	Пуната	. Веста прв ја донесе болиvisкиот
tekst3.not	Пуната. Веста прв ја донесе	болиvisкиот	весник „Ултима ора“, а од
tekst3.not	прв ја донесе болиvisкиот весник „	Ултима	ора“, а од него, со
tekst3.not	измени, ја пренесоа и „Ла	Пресенсиа	и „Ел Диарио“. Веста завршуваше
tekst3.not	пренесоа и „Ла Пресенсиа“ и ..	Ел Диарио	“. Веста завршуваше со следниов детаљ
tekst3.not	случена во непосредна близина на	Кочабамба	, во пет часот и седумнаесет
tekst3.not	се судејќи излегува некаде во	Европа	. Најинтересно во сето тоа е
tekst3.not	ден кога тој излегува во	Европа	е неизводливо.“
tekst4.not	Моите први сеќавања за	Емилија	поврзани се со појавата на
tekst4.not	зелена мува. Се сеќавам дека	Емилија	клекна и ја избрка мувата
tekst4.not	каде и како се појави	Емилија	? Моето сеќавање не е наполно
tekst4.not	албум и записите на дедо	Симон	се во најголем број случаи
tekst4.not	нејасни; понекогаш дури мислам дека	Емилија	успеала – со некаква само нејзе
tekst4.not	војната сликата на мојата родина	Емилија	станува несомнено појасна и поизострена
tekst4.not	некаква стара рамка. Тоа се	африкански	слонови, реков јас поучно. Се
tekst4.not	Се разликуваат од оние од	Индија	, додадов, горд на своето знаење
tekst4.not	артилерија. Слушај, и' реков на	Емилија	која се уште ги гледаше
tekst4.not	душек се беше истурила волна.	Емилија	ја одбра главата на куклата

Query 467/467

Conclusion



In order to recognize the named entities, the program environment NooJ can be used, with the help of several language resources for Macedonian: morphological lexicons, morphological and syntactic grammars. A syntactic grammar that recognizes named entities in texts exploiting the capital letters rule is the basis on which their recognition is based. All of these language resources are applied to a small corpus, and the system detects 2.18% of named entities from the total number of word forms.

The system does not detect multi-word named entities, in which only the first word starts with a capital letter. This problem can be solved by developing new syntactic grammars.