

# Упражнение върху GATE

Мария Матева

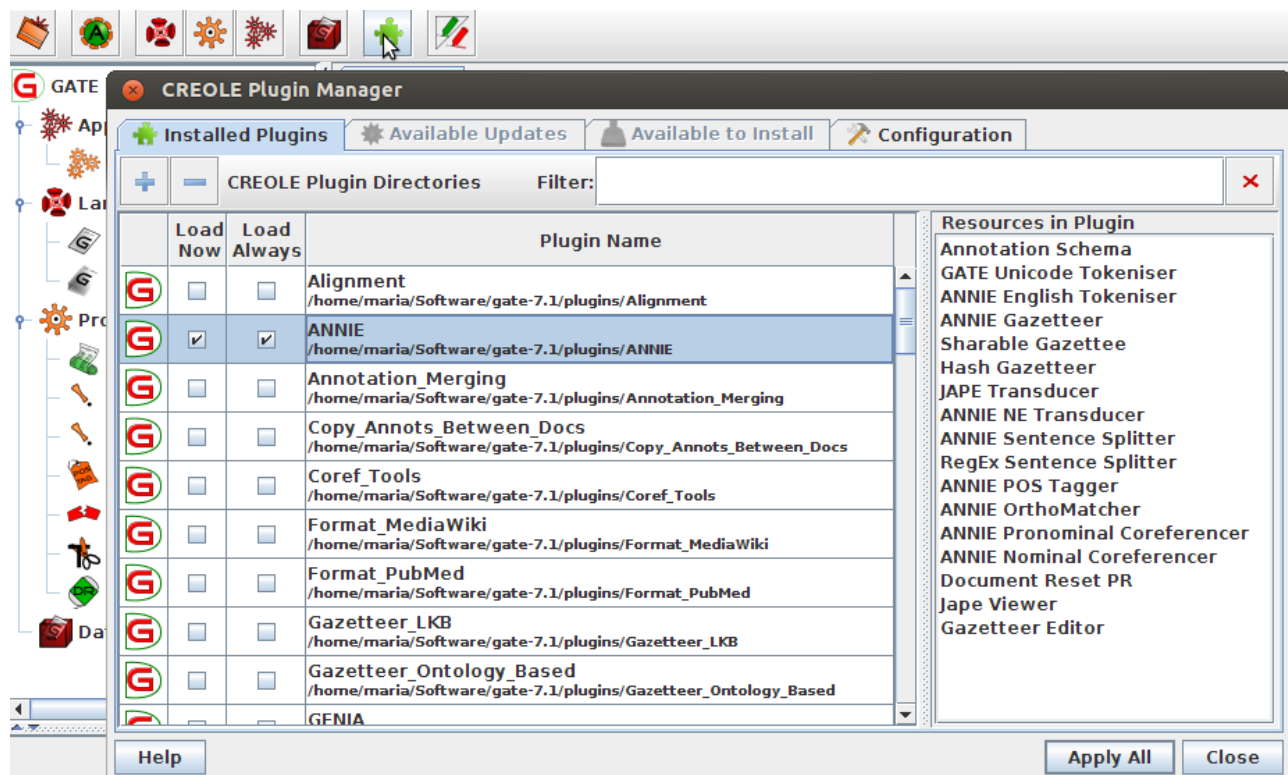
Някои понятия:

- обработка на естествен език – natural language processing
- разпознаване на именовани единици в текста – named entity recognition – намиране на обекти в текста, най-често Хора, Локации, Организации, пр.
- семантично аотиране – определяне на значещи единици в текста, асоциирайки ги с уникален идентификатор (URI – unique resource identifier) и опционално други метаданни за тях.
- GATE аотация – метаданни за значението на единица от текста (име, дума, фраза)

GATE е фреймуърк за анализ на текст и семантично аотиране.

Целта на упражнението е да се запознаем с някои основни компоненти при обработка на естествен език, да видим изхода от тяхната работа, както и да разпознаем някои значещи единици в текста.

1. Изтеглете GATE, <http://gate.ac.uk/download/> Сега актуалният пакет е **gate-7.1-build4485-ALL.zip**
2. Пуснете GATE от скрипта **gate\_folder/bin/gate.bat** или **gate.sh**  
Linux users: **bash gate.sh** . Зарежда се потребителския интерфейс на GATE.
3. Заредете плъгини: ANNIE, Tools, Stemmer\_Snowball  
Дайте им "Load now", "Load always" и след това "Apply All".



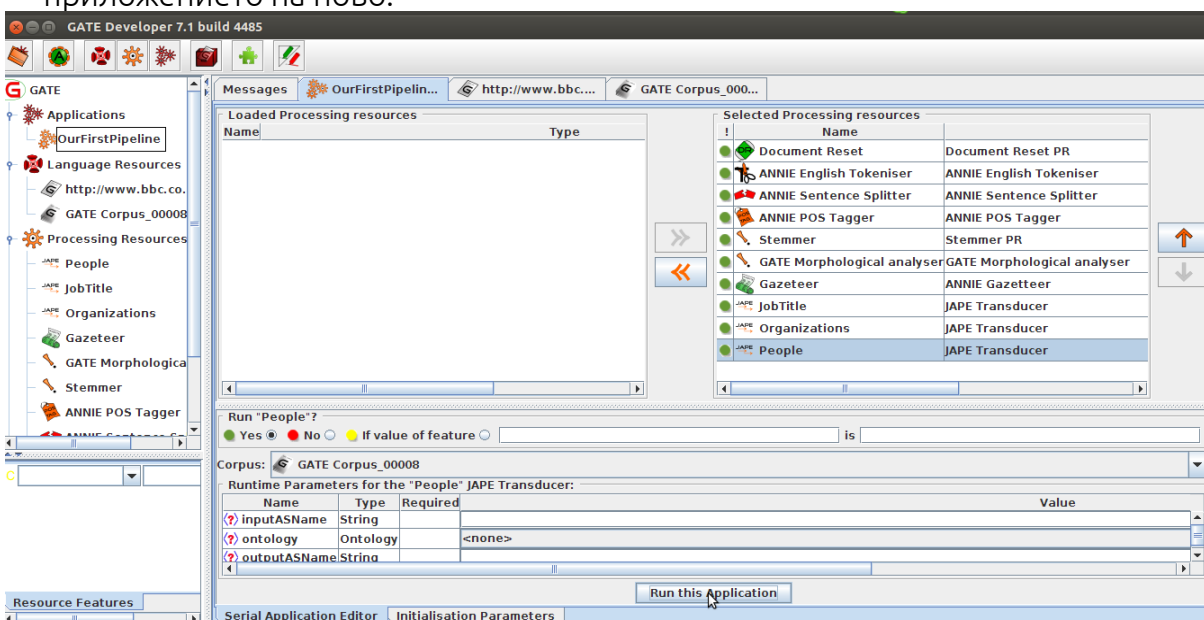
4. Създайте документ. От менюто с ресурси вляво изберете Language Resources > New > GATE Document. Тук има разнообразни възможности. Например, да

изтеглим документ от мрежата.

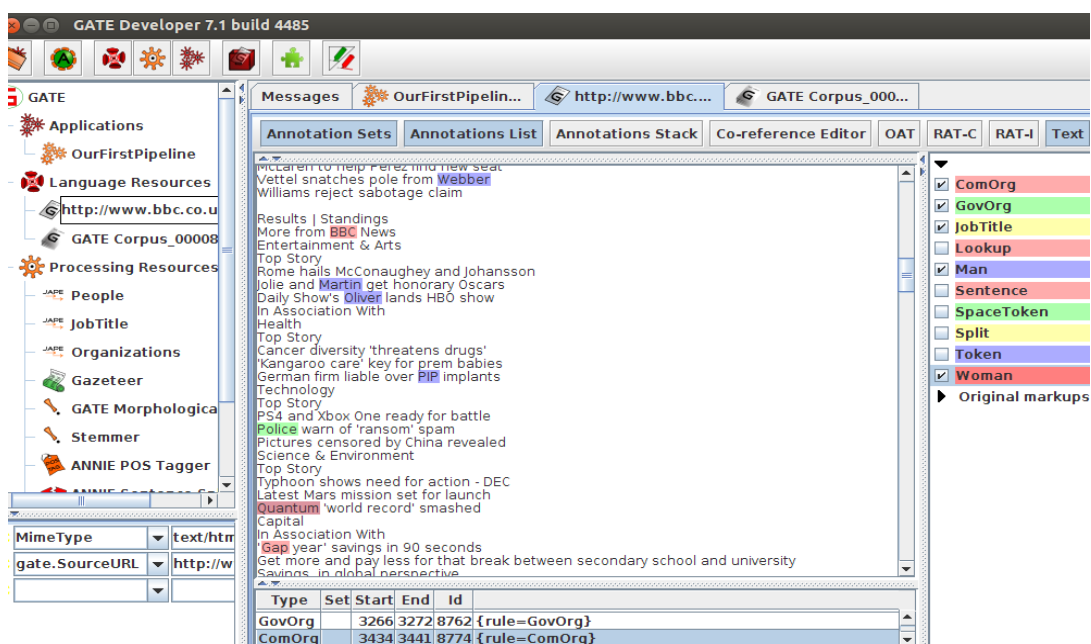
1. В полето sourceURL въведете адреса на любим новинарски сайт на английски, например <http://bbc.co.uk>
2. Потвърдете с OK
3. Създайте още няколко документа по същия начин
5. Създайте корпус за документите: Language Resources > New > GATE Corpus
6. Сложете създадените в 4. документи в корпуса
7. Създайте следните компоненти за обработка на естествен език  
Processing Resources > New >...
  1. Document Reset(изчиства GATE анотации)
  2. ANNIE English Tokenizer – токънайзър, разделя текста на думи
  3. ANNIE Sentence Splitter – определя изречения в текста
  4. ANNIE POS tagger – добавя етикет за част на речта
  5. Stemmer – стемер, определя основната част от думата(работи на евристичен принцип)
  6. GATE Morphological Analyzer – добавя лема(лема е основна дума, дума от която произхожда текущата), работи на принципа на морфологичен анализ
  7. ANNIE Gazeteer – открива смислови обекти в текста, от предефинирани речници с имена на хора, държави, организации, популярни заглавия и т.н.
8. Създайте ново приложение пайплайн:  
Applications > Create New Applications > Conditional Corpus Pipeline
9. Подредете компонентите от 7. в този ред в пайплайна и го пуснете върху корпуса от 6. с бутона “Run this application”
10. Кликнете на някой от документите за да разгледате резултатите. От менюто изберете “Annotation Sets” и “Annotation List”. Ще се появят списък с анотациите и списък с типовете анотации – Sentence, SpaceToken, Split, Token. Изберете Token.
11. Разгледайте конкретни анотации – техните category(част на речта според <http://gate.ac.uk/sale/tao/splitap7.html>), stem – стем, root – лема и др.

Type	Set	Start	End	Id
Token		774	778	1423 {category=NN, kind=word, length=4, orth=lowercase, r
Token		778	779	1424 {category=., kind=punctuation, length=1, root=?, stem
Token		780	783	1426 {category=WRB, kind=word, length=3, orth=upperinitia
Token		784	786	1428 {category=VBP, kind=word, length=2, orth=lowercase,
Token		787	794	1430 {affix=s, category=VBZ, kind=word, length=7, orth=low
Token		795	799	1432 {category=JJ, kind=word, length=4, orth=lowercase, r
Token		800	806	1434 {affix=s, category=NNS, kind=word, length=6, orth=low
Token		807	819	1436 {category=RB, kind=word, length=12, orth=lowercase,

12. Добавете газетиър – речник разпознаващ значещи единици(Lookups) в текста:  
Processing Resources > New > ANNIE Gazetteer
13. Добавете следните JAPE правила:  
- *person.jape*, *orgs.jape*, *jobtitle.jape* по следния начин  
Processing Resources > New > JAPE Transducer;  
URL: път до всеки от горните файлове  
Разгледайте правилата.
14. JAPE правилата се използват за пряка работа върху различни типове анотации, които GATE добавя. С тях можете да се запознаете на:  
<http://gate.ac.uk/sale/tao/splitch8.html#x12-2120008> и още няколко примера за правила са: <http://gate.ac.uk/gate/plugins/ANNIE/resources/NE/>
15. Добавете 4-те компонента, създадени в 12. и 13. към пайплайна и пуснете приложението на ново.

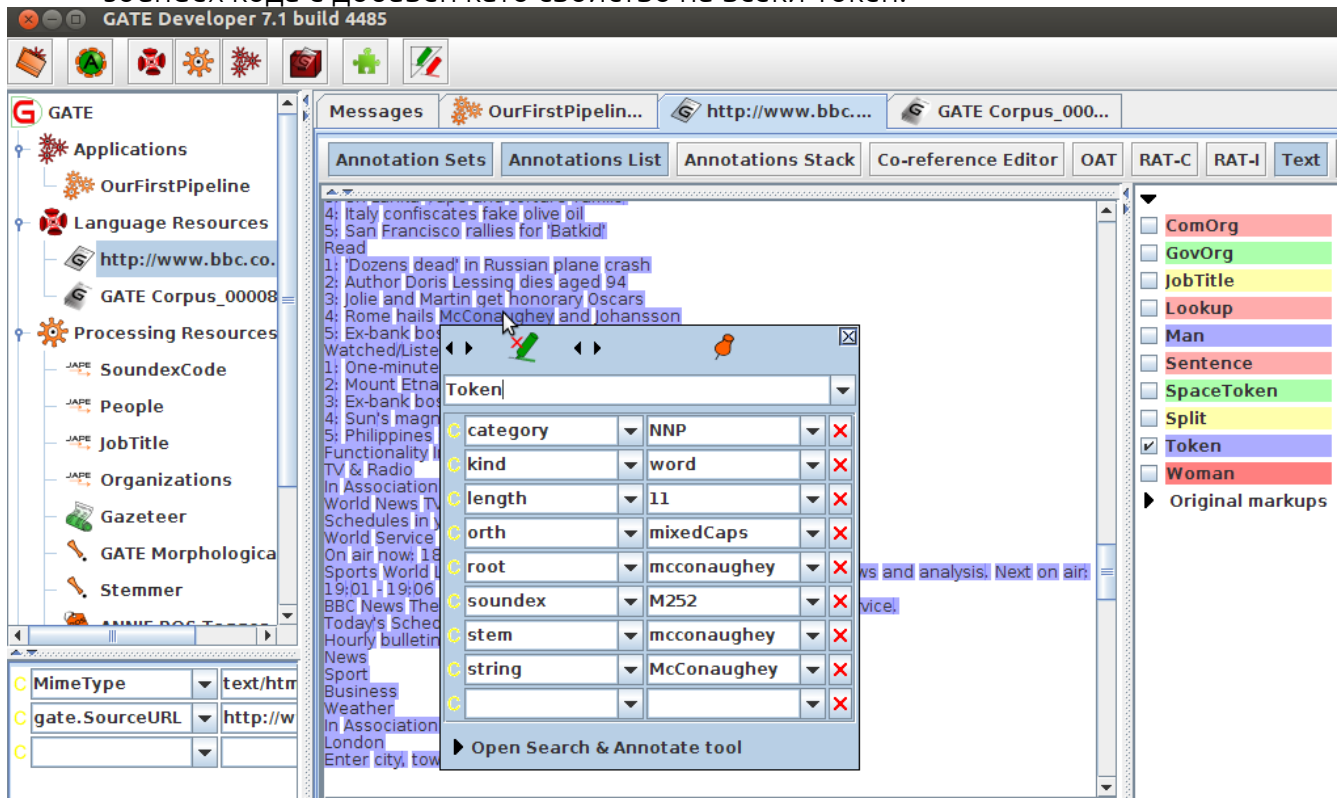


16. Първо газетиъра е открил често-срещани единици в текста(Lookups), после правилата, които включихме са разпознали обществени и частни организации, имена на мъже и жени, както и професии. Това е най-проста форма на



разпознаване на смислови единици в текст.

17. За демонстрация на възможностите на JAPE правилата добавяме и още едно правило, имплементиращо soundex алгоритъма (<http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm#Top>) и добавящо soundex код на всяка дума. Създайте JAPE Transducer със soundex.jape, добавете го в приложението, изпълнете и вижте резултата – soundex кода е добавен като свойство на всеки Token.



Soundex хеш-кода е подходящ при намиране на объркани близко звучащи имена(напр. "Кошлуков"/"Кушлуков").