Information Extraction

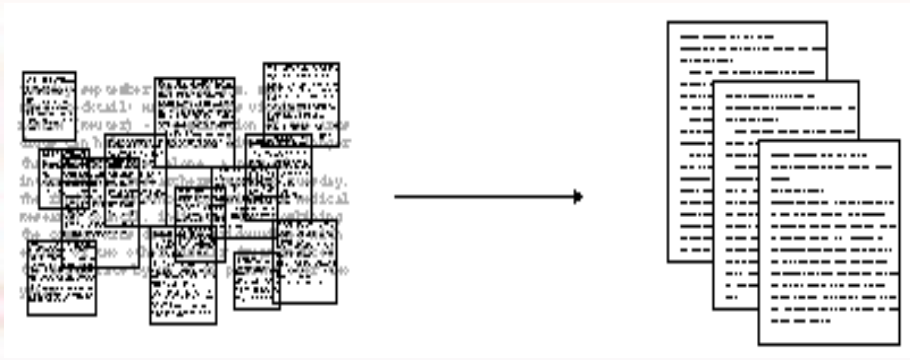
Ewa Płużek & Antoni Myłka

Co to takiego?

- technologia bazująca na analizie języka naturalnego
- ekstrakcja usystematyzowanych i ustrukturalizowanych informacji z tekstów pisanych
- szczególny rodzaj Information Retrieval

Information Retrieval

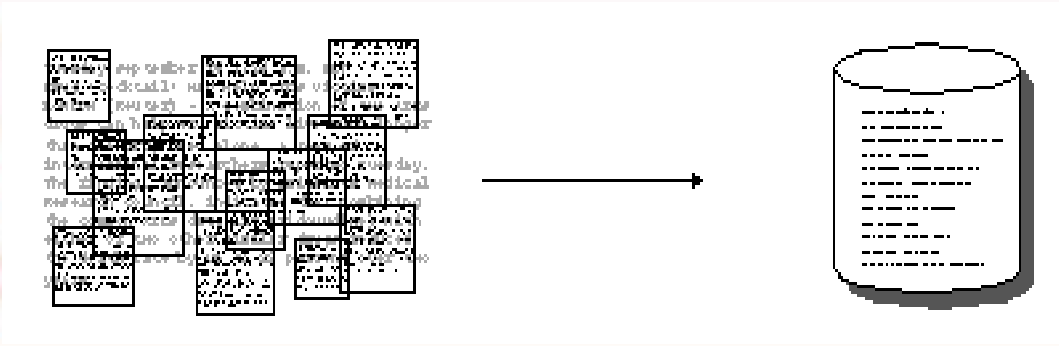
Technologie IR dostarczają po prostu listę dokumentów w których występują zadane słowa.



użytkownik sam musi analizować dokumenty

Information Extraction

Systemy IE z kolei potrafią wyłuskać odpowiednie informacje oraz przedstawić w jednym dokumencie, tylko i wyłącznie te fragmenty, które uznają za przydatne szukającemu.



użytkownik analizuje tylko fakty

Pradzieje

- Hans Peter Luhn
- “Key Words in Context” indexing
- 1958 - “A Business Intelligence System”

Burzliwy rozwój

- TAUM-METEO (lata 60-te) – tłumaczenie raportów pogodowych
- ATRANS – ekstrakcja informacji z prostych wiadomości teleksowych o przelewach bankowych
- JASPER – ekstrakcja informacji o zarobkach z krótkich zdań
- SCISOR – ekstrakcja informacji z tekstów w internecie

Message Understanding Conferences

- 7 konferencji
- lata 90-te
- sponsorowane przez DARPA
- konkurs systemów IE polegający na ekstrakcji informacji z zadanego korpusu tekstów
- systemy startowały w pięciu kategoriach

Jak działają systemy Information Extraction

Dla przykładu rozpatrzmy zdanie:

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Big Head is a staff scientist at We Build Rockets Inc.

Named Entity Recognition

odkryje, że encjami są tu:

- “**rocket**”,
- “**Tuesday**”,
- “**Dr. Big Head**”
- “**We Build Rockets Inc**”.

*The shiny red **rocket** was fired on **Tuesday**. It is the brainchild of **Dr. Big Head**. Dr. Big Head is a staff scientist at **We Build Rockets Inc**.*

Coreference Resolution

odkryje, że

- **“it”** odnosi się do encji
- **“rocket”**

*The shiny red **rocket** was fired on Tuesday. **It** is the brainchild of Dr. Big Head. Dr. Big Head is a staff scientist at We Build Rockets Inc.*

Template Element Construction

odkryje, że określeniami encji **“rocket”** są:

- **“shiny”**
- **“red”**
- **“brainchild of Dr. Big Head”**

*The **shiny red** rocket was fired on Tuesday. It is the **brainchild of Dr. Big Head.** Dr. Big Head is a staff scientist at We Build Rockets Inc.*

Template Relation Construction

odkryje, że:

- **“Dr. Big Head”**
  pracuje dla:
- **“We Build Rockets Inc.”**

*The shiny red rocket was fired on Tuesday. It is the brainchild of **Dr. Big Head.** Dr. Big Head is a staff scientist at **We Build Rockets Inc.***

Scenario Template Production

Dopasowuje wyniki TE i TR do wyspecyfikowanych scenariuszy zdarzeń.

*The shiny red **rocket was fired** on Tuesday. It is the brainchild of Dr. Big Head. Dr. Big Head is a staff scientist at We Build Rockets Inc.*

Jeżeli kazaliśmy mu szukać zdarzeń odpalenia rakiety to zwróci nam informacje z tej notatki.

Podsumowanie

Proces ekstrakcji informacji składa się z pięciu etapów:

- **NE** – Named Entity Recognition
- **CO** – Coreference Resolution
- **TE** – Template Element Construction
- **TR** – Template Relation Construction
- **ST** – Scenario Template Production

Jak się tworzy systemy IE?

Wprowadzanie wiedzy:

- Knowledge Engineering – wiedza wprowadzana przez eksperta w danej dziedzinie
- Learning Approach – system uczy się “sam” na podstawie ręcznie opisywanych dokumentów i interakcji z użytkownikiem

Anatomia systemu IE

Dwa najważniejsze elementy:

- Procesor tekstów
- Generator wzorców

Procesor tekstów

- analiza leksykalna
- segmentacja tekstu (podział na zdania)
- interpretacja skrótów
- analiza podstaw słowotwórczych
- wychwytywanie z tekstu struktur istotnych z punktu widzenia dziedziny problemu

Generator wzorców

- łączy wyniki działania procesora tekstów z wiedzą dziedzinową
- wyraża informacje z tekstu w formie pewnych wzorców

Skuteczność systemów IE

	Współczesne systemy IE	Człowiek
NE	95 %	< 100 %
CO	50 – 60 %	< 100 %
TE	80 %	95 %
TR	75 %	< 100 %
ST	60 %	80 %

Zastosowania

- obszary, gdzie nie jest niezbędna informacja dokładna
- wyławianie wiedzy z dużego zbioru tekstów (np. sieci WWW)

Analiza finansowa

- “Jak często w sieci pojawiają się wzmianki wskazujące na dobre prognozy nt. przyszłości firmy?”
- “Jak w ostatnim roku zmieniały się oczekiwania wobec firmy?”
- “Jak dużo opinii pozytywnych i negatywnych na temat nowego zarządu opublikowano w zeszłym roku?”

Marketing

- Jak przyjmowany jest nasz program lojalnościowy?
- Jak duży odzew spowodowało wprowadzenie nowego produktu, czy faktycznie wywołało sensację?
- Czy potencjalna grupa docelowa komentuje na forach internetowych nowy produkt i jakie opinie przeważają?

Public Relations

- “Podaj 12 najbardziej zjadliwych komentarzy na temat ostatniej wypowiedzi szefa, opublikowanych w prasie w ciągu ostatnich dwóch dni”
- “Jaką szkodę dla wizerunku przyniosła ostatnia afera, jaka grupa klientów najbardziej się nią przejęła?”

Analiza Mediów

- “Jaka jest odległość medialna między nazwą naszej firmy a pojęciem 'XML'?”
- “Jaka jest linia polityczna gazety X?. Które partie są w niej częściej opisywane pozytywnie?”

Narzędzia

Darmowe, open source

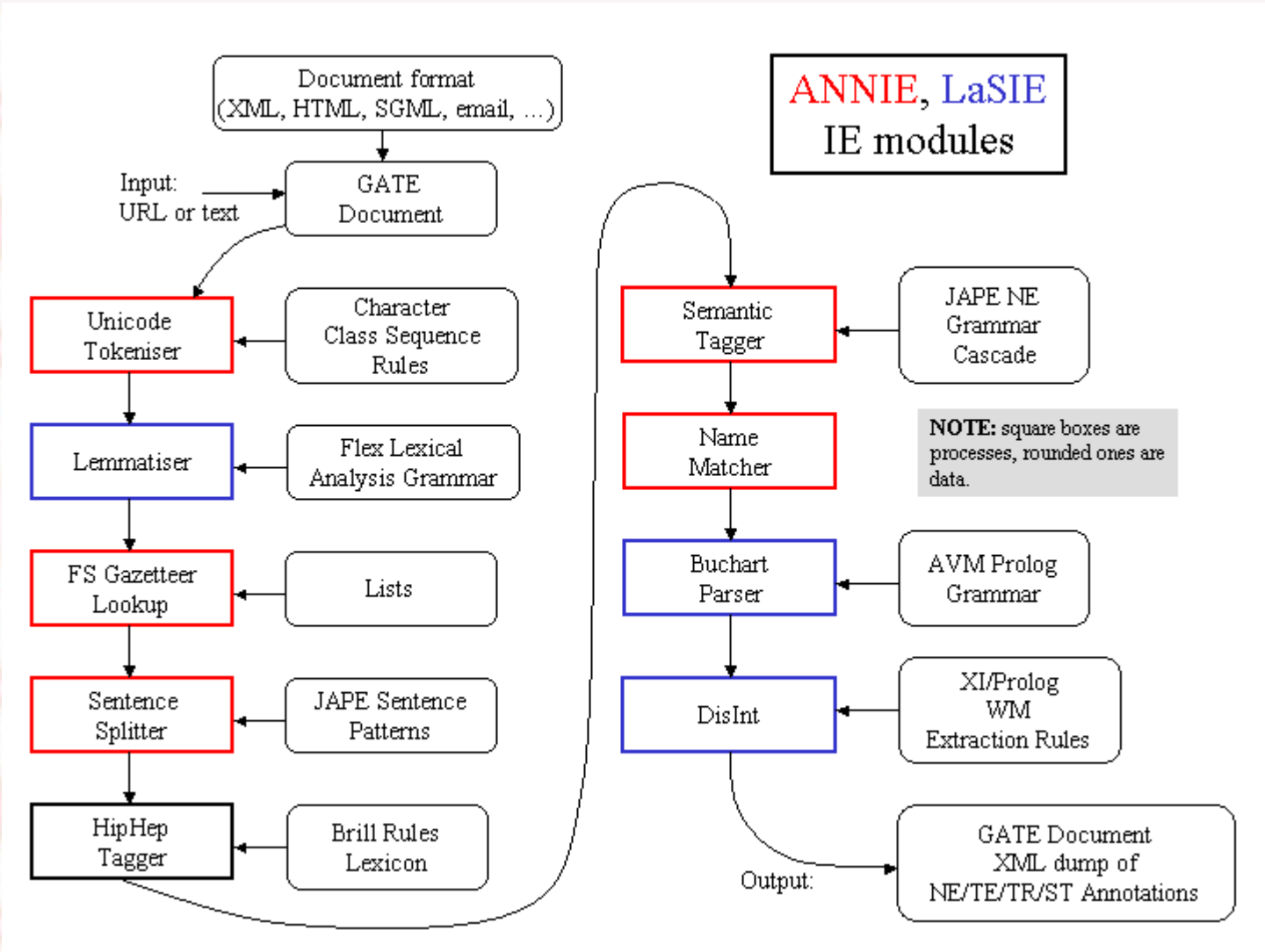
- GATE + ANNIE (http://gate.ac.uk)
- MAGPIE (http://kmi.open.ac.uk/projects/magpie/)
- BADGER (http://www-nlp.cs.umass.edu/software/badger.html)

Komercyjne

- KIM (www.ontotext.com/kim)
- SPSS (www.spss.com)

General Architecture for Text Engineering

- tworzone w University of Sheffield
- środowisko do tworzenia aplikacji operujących na tekście, m. in. do ekstrakcji informacji
- zestaw wyspecjalizowanych komponentów, których można używać w różnych sytuacjach


ANNIE, LaSIE
IE modules
Document format
(XML, HTML, SGML, email, ...)
Input:
URL or text
GATE
Document
Unicode
Tokeniser
Character
Class Sequence
Rules
Lemmatiser
Flex Lexical
Analysis Grammar
FS Gazetteer
Lookup
Lists
Sentence
Splitter
JAPE Sentence
Patterns
HipHep
Tagger
Brill Rules
Lexicon
Semantic
Tagger
JAPE NE
Grammar
Cascade
Name
Matcher
Buchart
Parser
AVM Prolog
Grammar
DisInt
XI/Prolog
WM
Extraction Rules
Output:
GATE Document
XML dump of
NE/TE/TR/ST Annotations
NOTE: square boxes are processes, rounded ones are data.

Tokeniser

- zamienia ciąg znaków wejściowych na tokeny:
  - słowa,
  - liczby,
  - znaki przestankowe
  - symbole
  - białe znaki

Gazeteer

- przechowuje listy wyrazów, zazwyczaj nazw własnych (miast, organizacji, walut)
- wykrywa wystąpienia tych wyrazów w przetwarzanym tekście i odpowiednio je oznacza

Sentence Splitter

- dzieli ciąg wejściowy na poszczególne zdania
- potrafi odróżnić kropkę występującą po jakimś skrócie od kropki kończącej zdanie.

Part of speech tagger

- każdemu wyrazowi z pliku wejściowego przydziela część mowy
- korzysta ze słownika i zespołu reguł (niezbędnych by odróżnić wystąpienia tego samego słowa jako różne części mowy)

Coreference

- Ortographical Coreference – moduł odpowiedzialny za wynajdowanie dwóch odwołań do tego samego pojęcia za pomocą różnych określeń: np: “Coca Cola” i “Coke”
- Pronominal Coreference – rozstrzyga do czego odnoszą się zaimki

JAPE – Java Annotations Pattern Engine

- Wszystkie moduły GATE operują na wiedzy wyrazonej w języku JAPE
- Język ten operuje wyrażeniami regularnymi tworzonymi na łańcuchach anotacji w tekście