Lingvistické předzpracování (nejen) webových textů pro web mining

Tomáš Kliegr, Vojtěch Svátek, Petr Strossa
KIZI VŠE Praha

Možné fáze předzpracování

ilustrováno na systému GATE, viz cvičení

- Tokenizace
- Identifikace vět
- Určení slovních druhů (POS)
- Určení anafor/koreferencí
- Lemmatizace a stemování
- Nalezení pojmenovaných entit
  - Gazetteery
  - Kontextová analýza
- *Extrakce relací* (už příliš nespadá do předzpracování)
- Syntaktická analýza – úplná nebo mělká
- Sémantická analýza využívající tezaury

Tokenizace

- Dokument je na začátku zpracování posloupnost znaků
- Pro účely dalšího zpracování je třeba znaky seskupit do smysluplných větších prvků – kapitol, odstavců, vět a *slov*

Tokenizací nazýváme označení slov a mezer (někdy též vět)

Tokenizer tokenu přiděluje typicky vlastnosti (features*)*

"Mr. Smith, throughout his distinguished career in government and in opposition, left a profound impression on the history of his party and his country," State Department spokesman Michael McCurry said.

Type	Set	Start	End	Id	Features
Token		390	392	1533	{kind=word, length=2, orth=upperInitial, string=Mr}
Token		392	393	1534	{kind=punctuation, length=1, string=.}
Token		394	399	1536	{kind=word, length=5, orth=upperInitial, string=Smith}
Token		399	400	1537	{kind=punctuation, length=1, string=,}
Token		401	411	1539	{kind=word, length=10, orth=lowercase, string=throughout}
Token		412	415	1541	{kind=word, length=3, orth=lowercase, string=his}
Token		416	429	1543	{kind=word, length=13, orth=lowercase, string=distinguished}
Token		430	436	1545	{kind=word, length=6, orth=lowercase, string=career}
Token		437	439	1547	{kind=word, length=2, orth=lowercase, string=in}
Token		440	450	1549	{kind=word, length=10, orth=lowercase, string=government}
Token		451	454	1551	{kind=word, length=3, orth=lowercase, string=and}
Token		455	457	1553	{kind=word, length=2, orth=lowercase, string=in}
Token		458	468	1555	{kind=word, length=10, orth=lowercase, string=opposition}

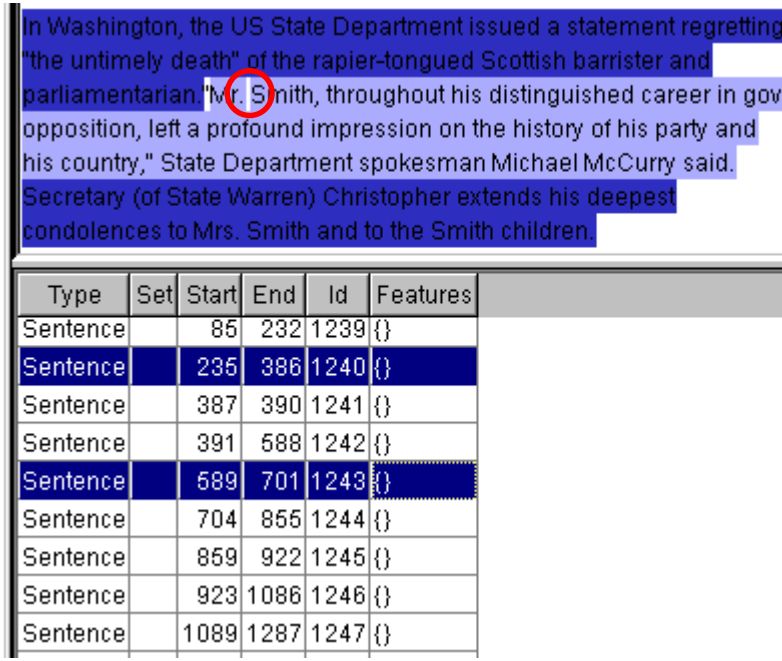
Výstup z GATE English Tokenizer

Identifikace vět

- Někdy součástí tokenizace
- Identifikace vět je obtížná, protože především tečka plní také jiné funkce, než je oddělovač vět
- Využívá se heuristik
- Seznam často používaných zkratek

- Často se vyskytující vzory obsahující tečky

|"Dr."|"FEB."|"Fig."|"FRI."|"GMBH."|"Gov." …



Výstup z GATE Sentence Splitter

<DOTTEDNAME: (<ALPHANUM>)+(<FULLSTOP>(<ALPHANUM>)+)+

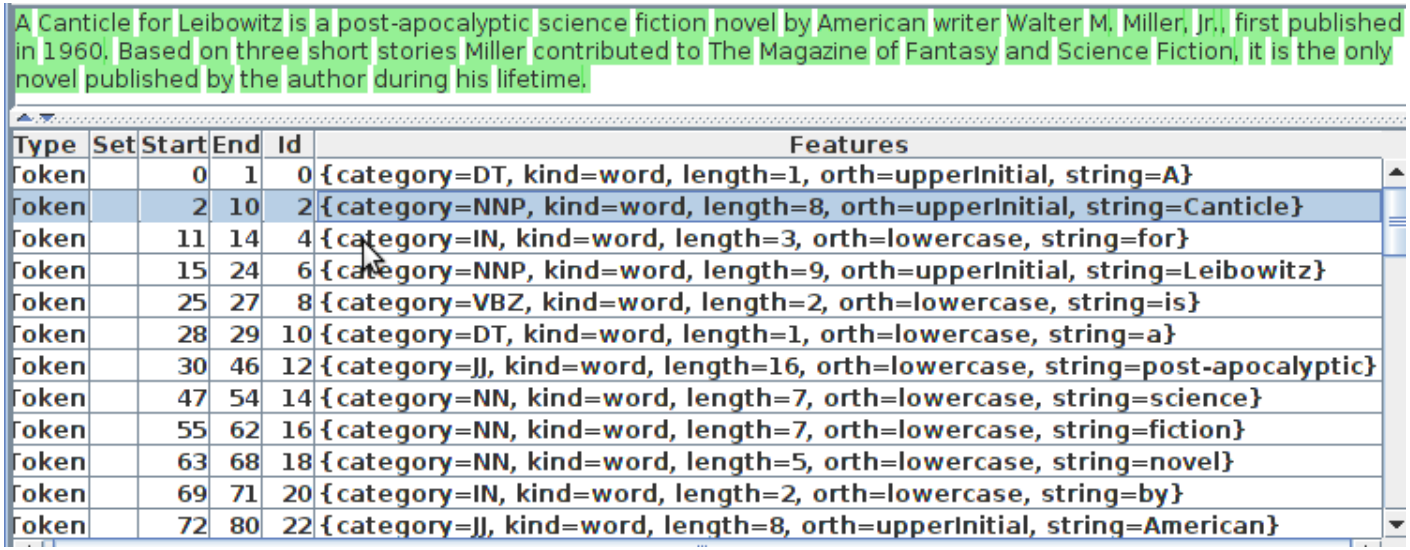
Příklady jsou z komponenty SentenceSplitter od DigitalPebble

Určení (cca) slovních druhů

- Part of Speech (POS) Tagging
- Značky vytvářené jednotlivými systémy se mírně liší
- Zde příklady značek generovaných taggerem v rámci ANNIE (GATE)

česky	anglicky	zkratka
Podstatné jméno	Noun	NN
Přídavné jméno	Adjective	JJ
Zájmeno	Pronoun	(různě)
Číslovka	Cardinal number	CD
Sloveso	Verb	VB
Příslovce	Adverb	RB
Předložky	Preposition	IN

Existují desítky dalších značek	
Osobní zájmeno	PP
Přivlastňovací zájmeno	PRP
„WH“- zájmeno	WP
Vlastní jméno	NNP
Podstatné jméno – množné číslo	NNS
Sloveso v minulém čase	VBD
Člen (určitý/neurčitý)	DT
…	…



Lemmatizace a stemování

- Snížení dimenzionality dat u statistických přístupů k WM
- (Alespoň částečně) odstranění *homonymie*
  - Situace, kdy tvar slova odpovídá více odlišným slovům, např. „left“ („This *left*-wing politician *left* us too early…“)
- Lemmatizace - slova se převádí do základního tvaru, který se nazývá *lemma* nebo slovníkový tvar
  - Lemma je, v ideálním případě, gramaticky správné a platné slovo
- Stemming - výsledkem stemování je *stem*, což může být
  - kmen slova
    - Slovo bez gramatické koncovky
  - kořen slova
    - Slovo bez předpon, přípon a koncovek
  - nebo prostě výsledek daného stemovacího algoritmu
    - Nemusí být gramaticky správné a platné slovo

Výstup z Gate STEMMER (není součástí ANNIE)

In Washington, the US State Department issued a statement regretting "the untimely death" of the rapier-tongued Scottish barrister and

{kind=word, length=10, orth=upperInitial, stem=depart, string=Department}
{kind=word, length=6, orth=lowercase, stem=issu, string=issued}
{kind=word, length=1, orth=lowercase, stem=a, string=a}
{kind=word, length=9, orth=lowercase, stem=statement, string=statement}
{kind=word, length=10, orth=lowercase, stem=regret, string=regretting}
{kind=punctuation, length=1, stem=", string="}
{kind=word, length=3, orth=lowercase, stem=the, string=the}
{kind=word, length=8, orth=lowercase, stem=untim, string=untimely}
{kind=word, length=5, orth=lowercase, stem=death, string=death}
{kind=punctuation, length=1, stem=", string="}
{kind=word, length=2, orth=lowercase, stem=of, string=of}
{kind=word, length=3, orth=lowercase, stem=the, string=the}
{kind=word, length=14, orth=lowercase, stem=rapier-tongu, string=rapier-tongu
{kind=word, length=8, orth=upperInitial, stem=scottish, string=Scottish}

Řešení anafor (koreferencí)

- Proces párování entit v textu, které odpovídají stejné entitě v reálném světě:
  - Anafora – „odkaz zpět“
  - Koreference – obecně vztah mezi dvěma jazykovými výrazy označující stejnou entitu

In Bonn, the head of the German Social Democratic Party, Rudolf Scharping, said in a statement he was "very affected by the sudden death of John Smith.

"A good friend of German social democracy has left us too early. He was very close to achieving his life's goal of making the Labour Party the largest political force in Britain" and would be "cruelly missed" in Europe, Scharping said.

- Gramatické (v angličtině zpravidla zájmenné) anafory
  - osobní zájmenná anafora (čeština: 3.osoba, angličtina: he, him, you)
  - reflexivní zájmenná anafora (čeština: se, si, svůj; angličtina: himself, herself)
  - posesivní zájmenná anafora (čeština: jeho, její, jejich; angličtina: her, his, hers)
- Jmenno-frázová anafora – jako odkaz na dříve uvedenou entitu (tzv. antecedent) slouží jmenná fráze
  - zde odkazuje příjmení na celé jméno osoby

Pro řešení anafor se používá specializovaných algoritmů (např Hobbsův algoritmus)

Anafory v ČJ viz DP *Nguy Giang Linh* . „Návrh souboru pravidel pro analýzu anafor v českém jazyce“ , MFF UK, 2006
Dále viz cvičení se systémem Phrase Detectives

Named Entity Recognition

(rozpoznávání pojmenovaných entit)

- Identifikace entit v textu a jejich klasifikace do tříd
- Nejlepší současné NER systémy jsou trénovány (pro základní druhy entit) na rozsáhlých korpusech s využitím hlubokého učení
- Detekce NE může být prováděna i na základě *pravidel* nebo *výčtů* (gazetteerů) entit
- Opírá se o obsah a/nebo kontext výskytů entit

Výsledek NER lze použít v pravidlových gramatikách nebo např. pro snížení dimenzionality textu při statistické klasifikaci

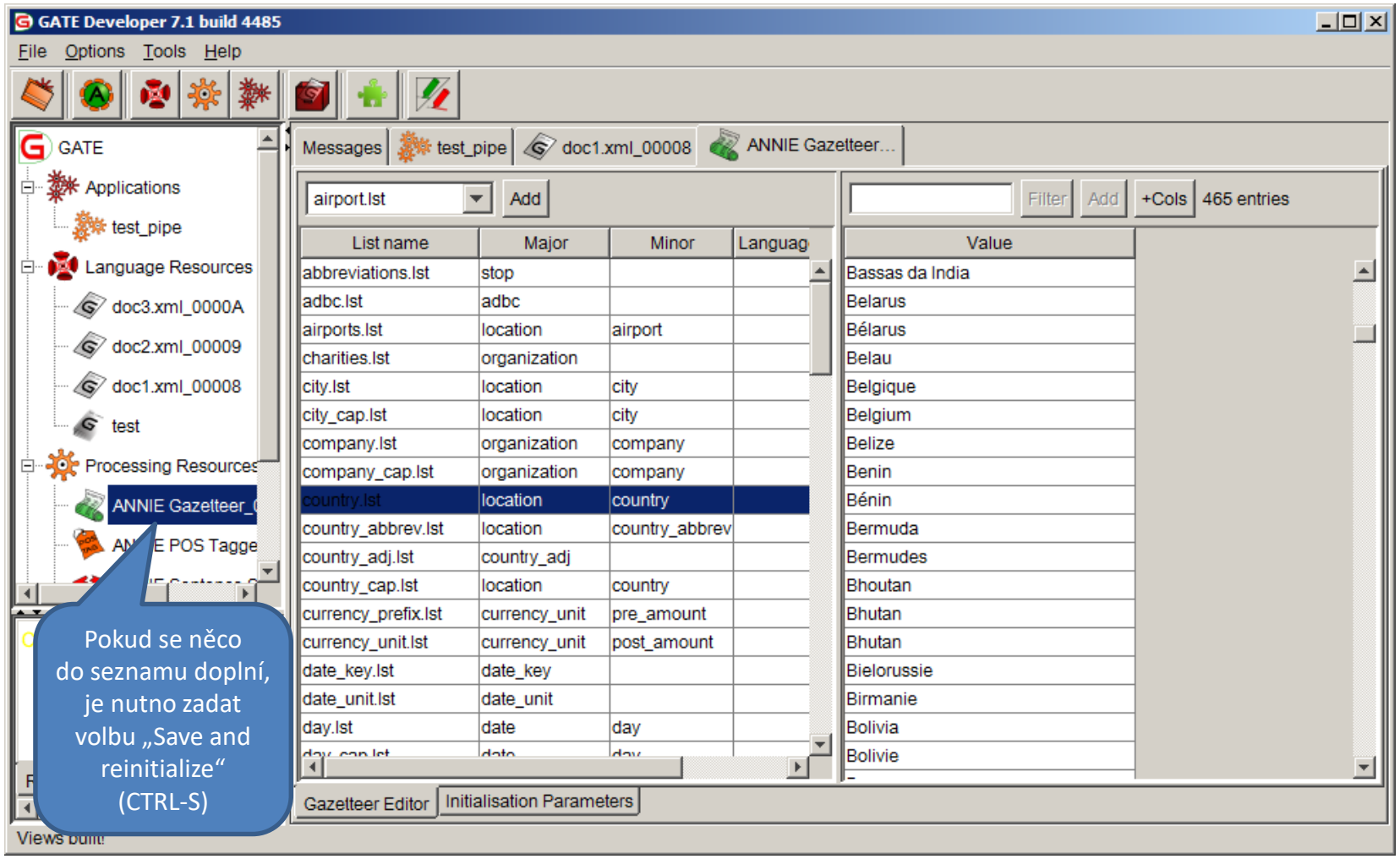
Výstup z GATE ANNIE GAZETTEER

Tributes poured in from around the world Thursday to the late Labour Party leader John Smith, who died earlier from a massive heart attack aged 55.

In Washington, the US State Department issued a statement regretting "the untimely death" of the rapier-tongued Scottish barrister and parliamentarian."Mr. Smith, throughout his distinguished career in government and ir

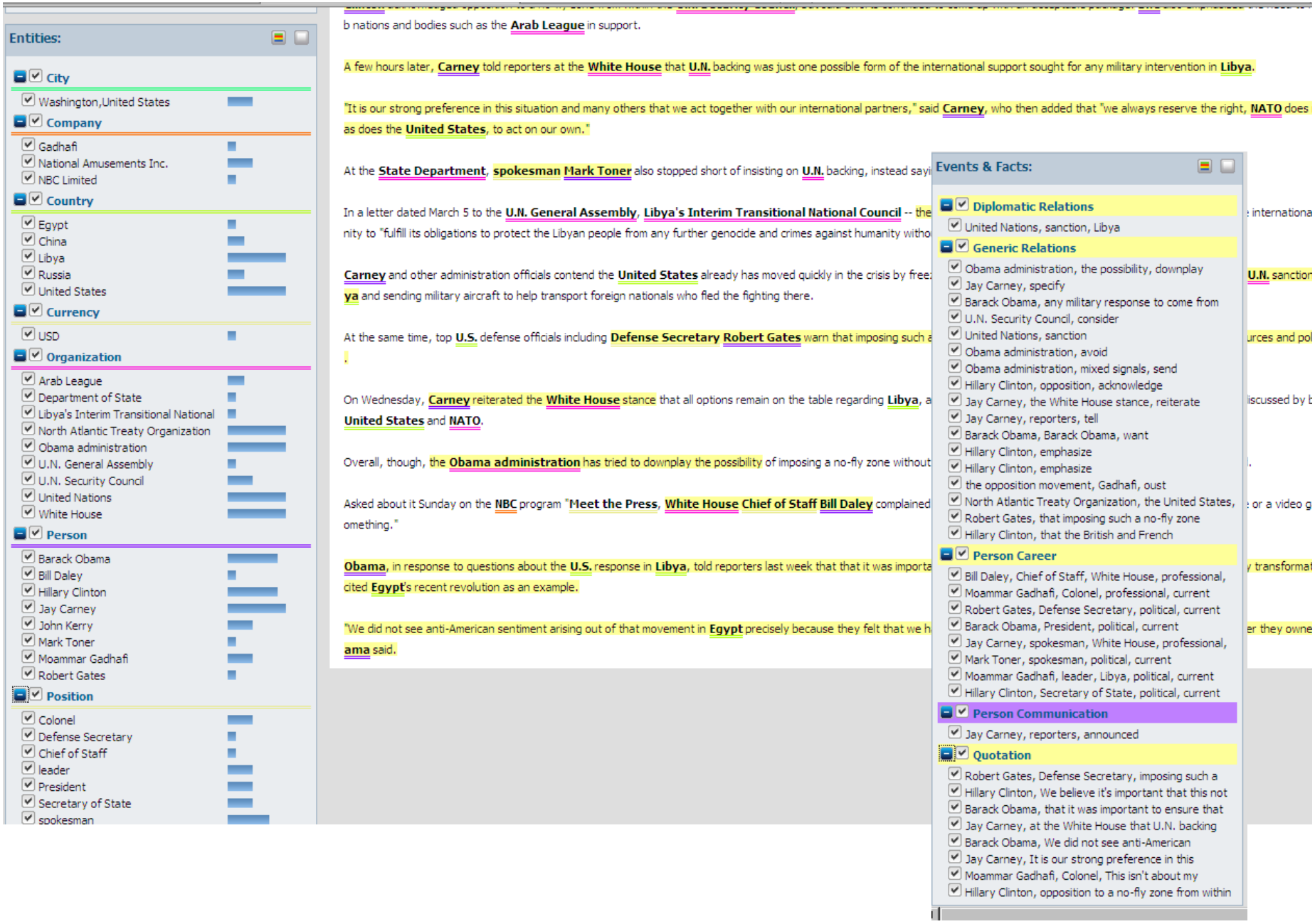
Type	Set	Start	End	Id	Features
Lookup		126	134	1283	{majorType=date, minorType=day}
Lookup		147	159	1284	{majorType=organization, minorType=government}
Lookup		154	159	1285	{majorType=org_base}
Lookup		160	166	1286	{majorType=jobtitle}
Lookup		167	171	1287	{majorType=person_first, minorType=male}
Lookup		238	248	1288	{majorType=location, minorType=city}
Lookup		254	256	1290	{majorType=location, minorType=country}
Lookup		254	256	1291	{majorType=location, minorType=country_abbrev}
Lookup		254	256	1289	{majorType=currency_unit, minorType=pre_amount}
Lookup		257	273	1292	{majorType=organization, minorType=departmen}
Lookup		263	273	1293	{majorType=org_base}
Lookup		347	355	1294	{majorType=country_adj}
Lookup		390	393	1295	{majorType=title, minorType=male}

ANNIE Gazetteer - seznamy



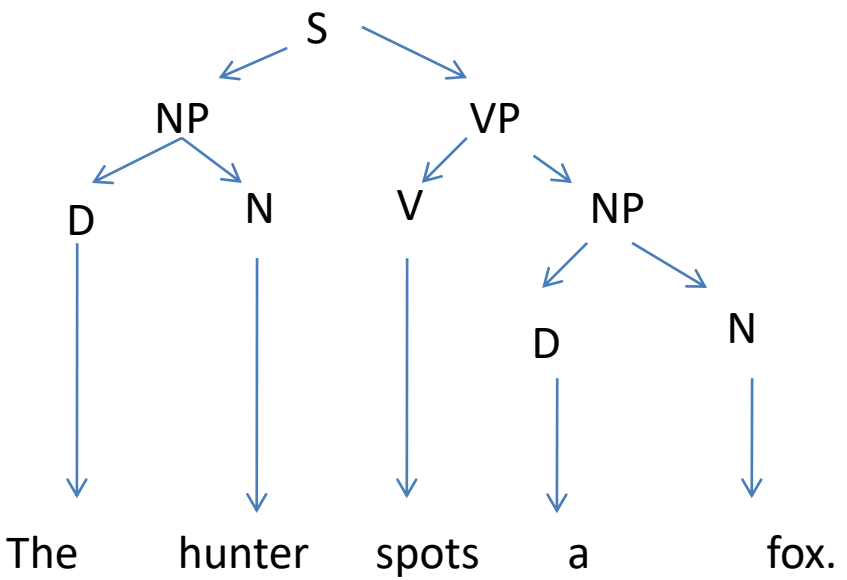
Od NER k extrakci relací

- Příklad: systém OpenCalais - nyní https://permid.org/tagging
- Vyvinutý Thomson Reuters, využívání v praxi
  - Webová služba, možnost exportu do RDF, i importu RDF dat
  - Rychlé, nevyužívá všechny dostupné možnosti NLP
- Extrakce entit
  - základních pojmenovaných entit (osoba, organizace, místo)
  - zjemněných základních entit (pozice, firma, stát, oblast)
  - dalších entit (program TV, svátek, …)
- Extrakce relací
  - Specifické relace, např. „citace“, „diplomatické vztahy“, „osobní kariéra“ (kdo, čím je, kde, politik/odborník, aktuálně?)
  - (Dříve?) i „generické relace“ založené na doslovném využití sloves


Entities:
City
Washington,United States
Company
Gadhafi
National Amusements Inc.
NBC Limited
Country
Egypt
China
Libya
Russia
United States
Currency
USD
Organization
Arab League
Department of State
Libya's Interim Transitional National
North Atlantic Treaty Organization
Obama administration
U.N. General Assembly
U.N. Security Council
United Nations
White House
Person
Barack Obama
Bill Daley
Hillary Clinton
Jay Carney
John Kerry
Mark Toner
Moammar Gadhafi
Robert Gates
Position
Colonel
Defense Secretary
Chief of Staff
leader
President
Secretary of State
spokesman
b nations and bodies such as the Arab League in support.
A few hours later, Carney told reporters at the White House that U.N. backing was just one possible form of the international support sought for any military intervention in Libya.
"It is our strong preference in this situation and many others that we act together with our international partners," said Carney, who then added that "we always reserve the right, NATO does as does the United States, to act on our own."
At the State Department, spokesman Mark Toner also stopped short of insisting on U.N. backing, instead sayi
In a letter dated March 5 to the U.N. General Assembly, Libya's Interim Transitional National Council -- the
nity to "fulfill its obligations to protect the Libyan people from any further genocide and crimes against humanity witho
Carney and other administration officials contend the United States already has moved quickly in the crisis by free
ya and sending military aircraft to help transport foreign nationals who fled the fighting there.
At the same time, top U.S. defense officials including Defense Secretary Robert Gates warn that imposing such a
.
On Wednesday, Carney reiterated the White House stance that all options remain on the table regarding Libya, a
United States and NATO.
Overall, though, the Obama administration has tried to downplay the possibility of imposing a no-fly zone without
Asked about it Sunday on the NBC program "Meet the Press, White House Chief of Staff Bill Daley complained
omething."
Obama, in response to questions about the U.S. response in Libya, told reporters last week that that it was importa
cited Egypt's recent revolution as an example.
"We did not see anti-American sentiment arising out of that movement in Egypt precisely because they felt that we h
ama said.
Events & Facts:
Diplomatic Relations
United Nations, sanction, Libya
Generic Relations
Obama administration, the possibility, downplay
Jay Carney, specify
Barack Obama, any military response to come from
U.N. Security Council, consider
United Nations, sanction
Obama administration, avoid
Obama administration, mixed signals, send
Hillary Clinton, opposition, acknowledge
Jay Carney, the White House stance, reiterate
Jay Carney, reporters, tell
Barack Obama, Barack Obama, want
Hillary Clinton, emphasize
Hillary Clinton, emphasize
the opposition movement, Gadhafi, oust
North Atlantic Treaty Organization, the United States,
Robert Gates, that imposing such a no-fly zone
Hillary Clinton, that the British and French
Person Career
Bill Daley, Chief of Staff, White House, professional,
Moammar Gadhafi, Colonel, professional, current
Robert Gates, Defense Secretary, political, current
Barack Obama, President, political, current
Jay Carney, spokesman, White House, professional,
Mark Toner, spokesman, political, current
Moammar Gadhafi, leader, Libya, political, current
Hillary Clinton, Secretary of State, political, current
Person Communication
Jay Carney, reporters, announced
Quotation
Robert Gates, Defense Secretary, imposing such a
Hillary Clinton, We believe it's important that this not
Barack Obama, that it was important to ensure that
Jay Carney, at the White House that U.N. backing
Barack Obama, We did not see anti-American
Jay Carney, It is our strong preference in this
Moammar Gadhafi, Colonel, This isn't about my
Hillary Clinton, opposition to a no-fly zone from within

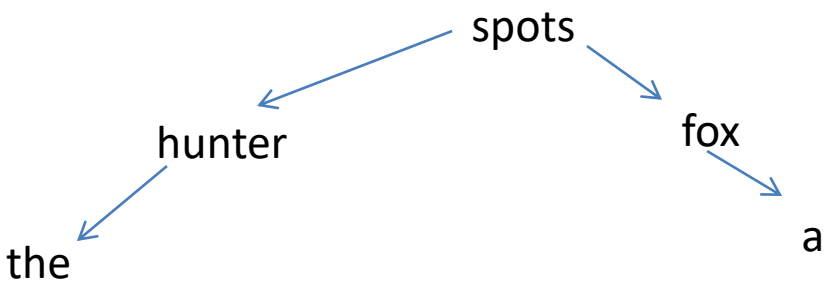
Syntaktická analýza (parsing)

- Složkové gramatiky (constituency grammars)
  - Stromový rozklad věty na složky, v daném pořadí
    - Dává smysl jen pro jazyky s pevným slovosledem
  - Na vyšších úrovních jde především o jmenné a slovesné fráze
    - Mohou být rekurzivní, v příkladu je jmenná fráze součástí slovesné fráze
  - Listy stromu jsou konkrétní slova
  - Fráze mohou být opatřeny i rolemi (podmět, předmět, doplněk…)



Syntaktická analýza (parsing)

- Závislostní gramatiky (dependency grammars)
  - Opět stromová struktura; hrany však nevyjadřují rozklad věty na složky, ale vztah mezi řídícím a závislým členem ve větě
    - viz „analýza větné skladby“ na českých ZŠ
  - Za hlavní kritérium pro rozlišení členu řídícího a závislého je možné považovat syntaktickou vypustitelnost závislého členu



Používá např.
Prague Dependency Corpus (PDT)

Více: http://www.kb-old.upol.cz/data/soubor_kb_804.pdf
http://www.ling.helsinki.fi/kit/2008s/clt231/nltk-0.9.5/doc/en/ch07.html

Syntaktická analýza (parsing)

- Demo https://corenlp.run/
  - Zobrazuje závislostní strukturu na podkladě POS

Mělká syntaktická analýza

- Komplexní analýza textu je časově náročná
- Pro praktické účely se provádí mělká (shallow) synt. analýza
- Mělké parsování se považuje za robustnější
- Výstupem jsou jmenné a slovesné fráze nebo skupiny

Slovesná fráze
sloveso + specifický typ jeho rozvití
Slovesná skupina
sloveso + pomocné/modální sloveso
Jmenná fráze
jméno + čím je rozvíjeno

Jmenné fráze (GATE Noun Phrase Chunker)

In Washington, the US State Department issued a statement regretting "the untimely death" of the rapier-tongued Scottish barrister and parliamentarian."Mr. Smith, throughout his distinguished career in government and in opposition, left a profound impression on the history of his party and his country," State Department spokesman Michael McCurry said. Secretary (of State Warren) Christopher extends his deepest condolences to Mrs. Smith and to the Smith children.

Slovesné skupiny (GATE VB Chunker)

'A good friend of German social democracy has left us too early. He was very close to achieving his life's goal of making the Labour Party the largest political force in Britain" and would be "cruelly missed" in Europe, he said.

Type	Set	Start	End	Id	Features
VG		900	908	3756	{tense=PrePer, type=FVG, voice=active}
VG		926	929	3757	{tense=SimPas, type=FVG, voice=active}
VG		944	953	3758	{tense=Pre, type=PART, voice=active}
VG		973	979	3759	{tense=Pre, type=PART, voice=active}
VG		1041	1049	3760	{tense=none, type=MODAL, voice=none}
VG		1059	1065	3761	{tense=SimPas, type=FVG, voice=active}
VG		1081	1085	3762	{tense=SimPas, type=FVG, voice=active}

SimPas = simple past, PrePer = present perfect

Sémantická analýza

využití tezaurů, ev. ontologií

- Zatímco lemmatizace sdružuje různé tvary stejného slova a analýza POS slova stejného jazykového druhu, sémantická analýza **sdružuje výrazy se stejným nebo souvisejícím významem**
- Používá se
  - při předzpracování (redukce dimenzionality, vážení rysů)
  - při vlastním dolování, nebo následné vizualizaci
- Tezaury (Wordnet): obecné „tezaurové“ vztahy výrazů
- Ontologie: obsahuje různé doménově-specifické vztahy

Synsety seřazeny dle pravděpodobnosti výskytu

S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*

- *direct hyponym* / *full hyponym*
- ***part meronym***
  - S: (n) accelerator, accelerator pedal, gas pedal, gas, throttle, gun (a pedal that controls the throttle valve) *"he stepped on the gas"*
  - S: (n) air bag (a safety restraint in an automobile; the bag inflates on collision and prevents the driver or passenger from being thrown forward)
  - S: (n) auto accessory (an accessory for an automobile)
  - S: (n) automobile engine (the engine that propels an automobile)
  - S: (n) automobile horn, car horn, motor horn, horn, hooter (a device on an automobile for making a warning noise)
  - S: (n) buffer, fender (a cushion-like device that reduces shock due to an impact)

Wordnet – hlavní tezaurové vztahy
Hyponymie/hyperonymie: obecnější vs. specifičtější výraz
Holonymie/meronymie: celek vs. část
Synonymie: stejný nebo blízký význam
Antonymie: opačný význam

Extrakce tezaurových vztahů z textu

- Tezaurové vztahy mohou být nejen na vstupu, ale i na výstupu lingvistického zpracování
- Příkladem je pravidlová gramatika (JAPE) pro extrakci výskytů **vzorů Hearstové**

Vzor Hearstové („Hearst pattern“) je lexiko-syntaktický vzor pro extrakci dvojic hyperonymum – hyponymum.

- V textu
  „Maradona (born October 30, 1960) is a former Argentine football player“
  je hyperonymem „player“
  a hyponymem „Maradona“
- Jde o specifický typ extrakce relace

```
// pouze v rámci vět
Rule: HearstRuleExactMatch Priority:1000
//nalézt řetězec odpovídající dotazu (Query) 'Maradona'
// libovolný počet libovolných tokenů
({Token})*
//'(born October 30, 1960)'
// tvar "to be", zde 'is'
{Token.string == "is"}|{Token.string == "are"}|
{Token.string == "were"}|{Token.string == "was"}
// následuje člen, zde 'a'
({Token.string == "a"}|{Token.string == "an"}|
{Token.string == "the"})
//makro následované povolenými slovy
//které mohou předcházet vlastní hyperonymum
(NounChunkBody)
// 'former Argentine football'
//Makro zastupnující NN, NNS nebo NNP
(Head)
//'player'
:hearstPattern
--> //odděluje levou a pravou stranu pravidla
//přidáme novou anotaci 'hearst' nad 'player'
//na tuto anotaci přidáme rys "rule" s hodnotou
"ExactMatch"
:hearstPattern.hearst = {rule = "ExactMatch"}
```

Extrakce tezaurových vztahů z textu

- Tezaurové vztahy mohou být nejen na vstupu, ale i na výstupu lingvistického zpracování
- Příkladem je pravidlová gramatika (JAPE) pro extrakci výskytů **vzorů Hearstové**

Vzor Hearstové („Hearst pattern“) je lexiko-syntaktický vzor pro extrakci dvojic hyperonymum – hyponymum.

- V textu
  „Maradona (born October 30, 1960) is a former Argentine football player“
  je hyperonymem „player“
  a hyponymem „Maradona“
- Jde o specifický typ extrakce relace

Abstraktní vyjádření tohoto vzoru:
<Hypo> is a <Hyper>

Jiné příklady vzorů Hearstové, s různou mírou spolehlivosti:
<Hyper> such as <Hypo1>, <Hypo2>, …
<Hyper>: <Hypo1>, <Hypo2>, …
the <Hypo> <Hyper>
(např. „the Maracana arena“)

'aradona'

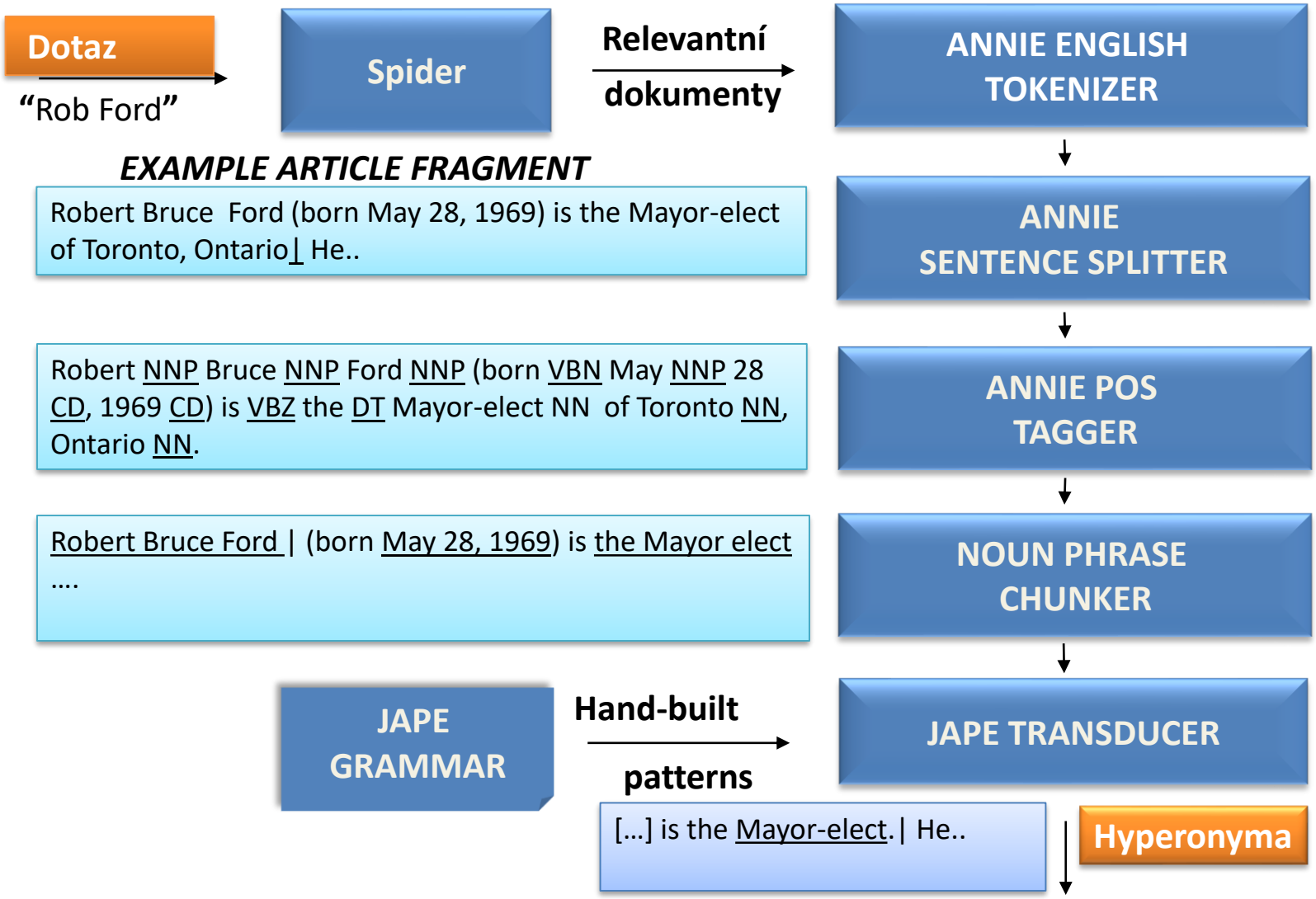
ıs"}

m

//Makro zastupující NN, NNS nebo NNP
(Head)
//'player'
:hearstPattern
--> *//odděluje levou a pravou stranu pravidla*
//přidáme novou anotaci 'hearst' nad 'player'
//na tuto anotaci přidáme rys "rule" s hodnotou "ExactMatch"
:hearstPattern.hearst = {rule = "ExactMatch"}

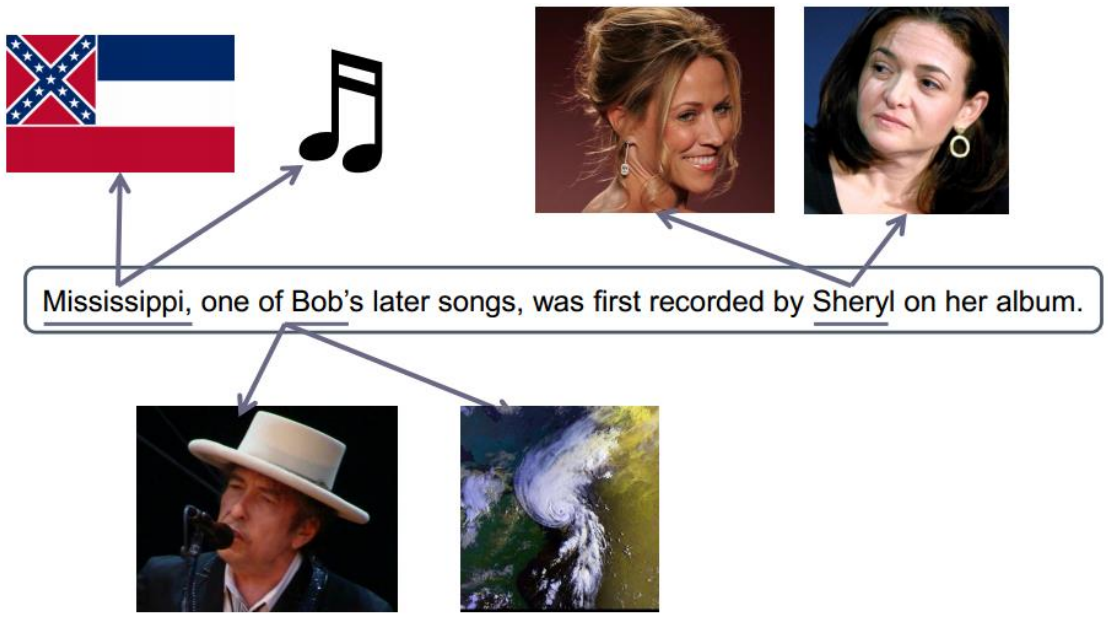
Cílená extrakce hyperonym

Systém THD z KIZI, http://entityclassifier.eu/



Entity linking

- Obdoba NER, ale cílem není jen určit typ entity, ale i konkrétní entitu, o které je záznam v databázi / znalostní bázi / znalostním grafu (dnes nejčastěji Wikidata, DBpedia, npod.)
  - Nalezení možných zmínek o dané entitě v jejím textovém popisku
  - V případě více možných kandidátů na základě kontextu ve větě se odvodí nejpravděpodobnější kombinace konkrétních entit



Shrnutí

- Lingvistické předzpracování umožňuje získat příznaky (features) pro následné dolování a současně omezit jejich počet na přijatelný
- Na techniky pracující s povrchovou strukturou jazyka postupně navazují techniky zaměřené na věcný obsah textu (NER, případně extrakce relací) – ty už patří i do oblasti extrakce informací jako samostatné disciplíny v rámci dolování z textů