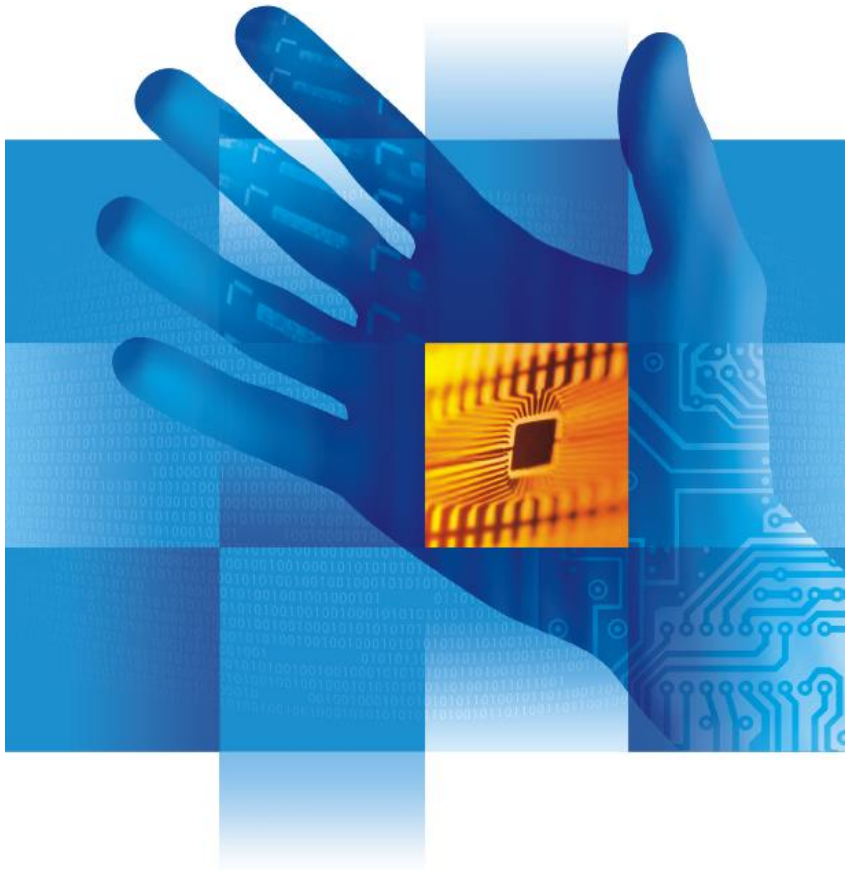
Předzpracovnání dokumentů a tvorba doménového modelu z nich

Petr Šaloun
VŠB-Technická univerzita Ostrava
FEI, katedra informatiky



Textové dokumenty

- Obsah, struktura, forma
- Značkovací jazyky
- XML
  - DOM
  - SAX
- Ontologie

Nástroje pro analýzu textu

- ✓ SimMetrics
- ✓ Vzdálenost – Matching coefficient

$$distance = |X \,\&\, Y|$$

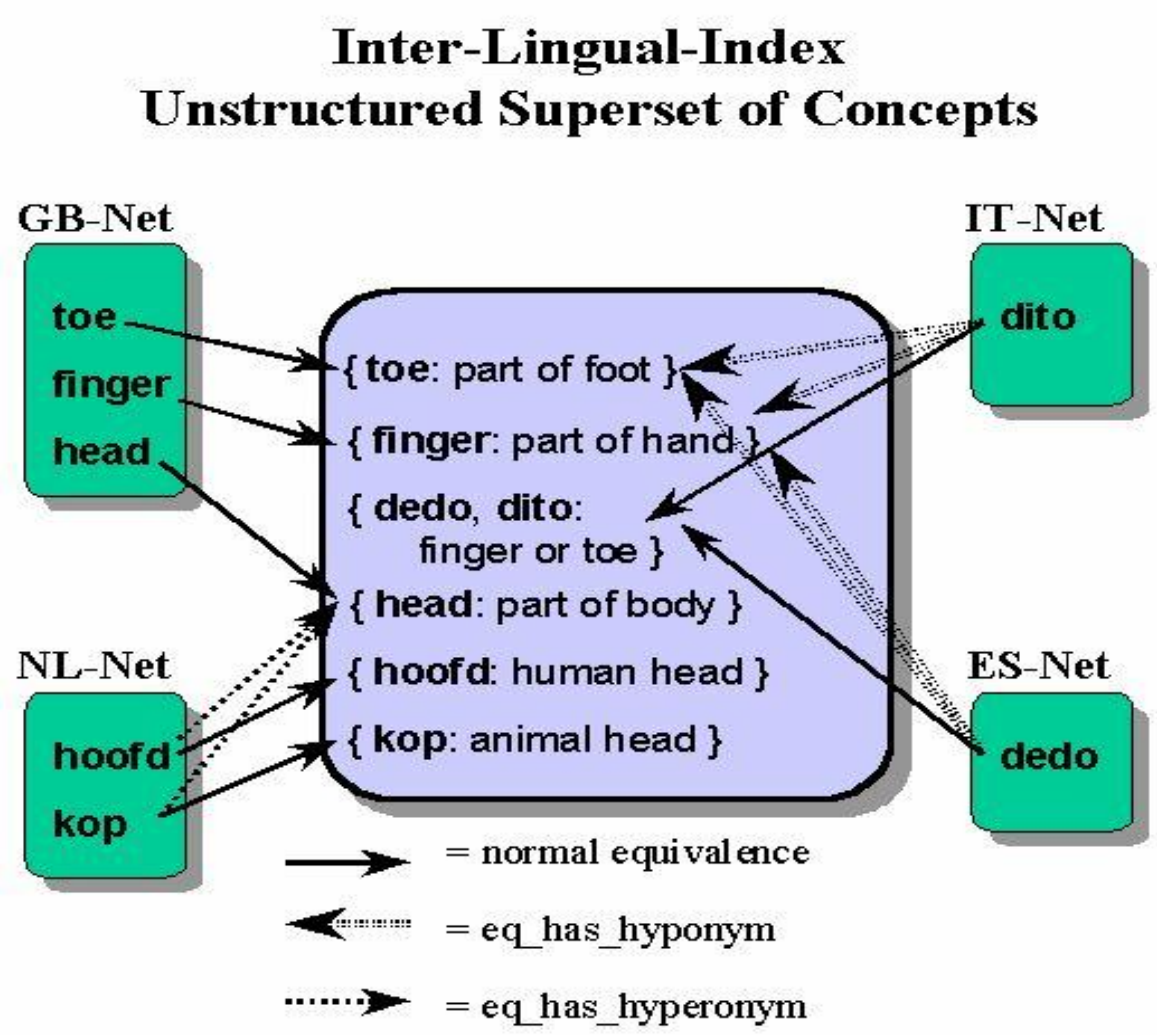
- ✓ Podobnost – Dice´s coefficient

$$DicesCoefficient = \frac{2 * CommonTerms}{NumberOfTermsInString1 + NumberOfTermsInString2}$$

Nástroje pro analýzu textu

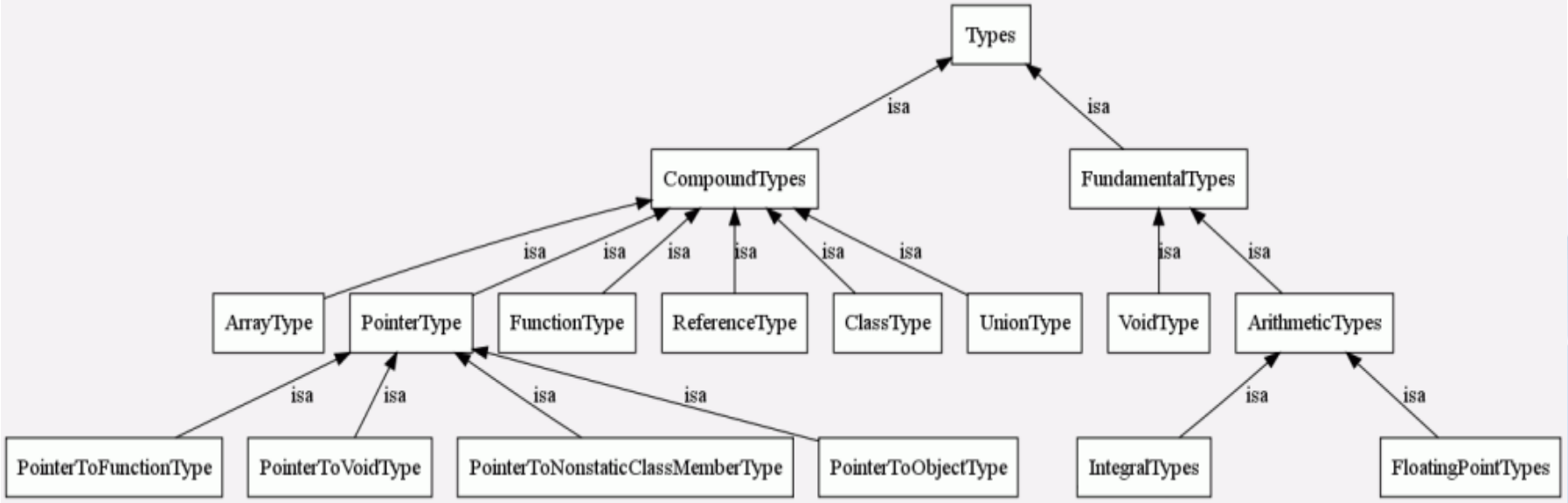
- WordNet
- Anglický WordNet
  - Volně ke stažení
- Český WordNet
  - Pouze vzdálený přístup
- EuroWordNet
  - Placený

Princip ILI


Inter-Lingual-Index
Unstructured Superset of Concepts
GB-Net
toe
finger
head
IT-Net
dito
{ toe: part of foot }
{ finger: part of hand }
{ dedo, dito:
finger or toe }
{ head: part of body }
{ hoofd: human head }
{ kop: animal head }
NL-Net
hoofd
kop
ES-Net
dedo
= normal equivalence
= eq_has_hyponym
= eq_has_hyperonym

Prostředí - data

- ✓ Zaměření na programovací jazyk C++
- ✓ Kniha Thinking in C++, Bruce Eckel
- ✓ Ontologie C++, Zdeněk Velart



Prostředí - knihovny

- ✓ Implementační jazyk C++
- ✓ WordNet API
- ✓ TinyXML
- ✓ Zkompilovány do statických .lib

Předzpracování dokumentů

- ✓ Zjednodušení
- ✓ Převod HTML do XHTML
- ✓ Sjednocení znakových entit
- ✓ Odstranění nadbytečných informací

```
<FONT FACE = "Verdana"><H1 ALIGN="LEFT">
6: Initialization &amp; Cleanup</H1></FONT>
<DIV ALIGN="LEFT"><P><FONT FACE="Verdana" SIZE=4>Chapter 4 made a
significant improvement in library </FONT><BR><FONT FACE="Verdana"
SIZE=4>use by taking all the scattered components of a typical
</FONT><BR><FONT FACE="Verdana" SIZE=4>C library and encapsulating
them into a structure (an abstract data type, called a <I>class</I>
from now on). </FONT><BR></P></DIV>
```

Předzpracování dokumentů

- ✓ Zjednodušení
- ✓ Převod HTML do XHTML
- ✓ Sjednocení znakových entit
- ✓ Odstranění nadbytečných informací

```
<FONT FACE = "Verdana"><H1 ALIGN="LEFT">
6: Initialization &amp; Cleanup</H1></FONT>
<DIV ALIGN="LEFT"><P><FONT FACE="Verdana" SIZE=4>Chapter 4 made a
significant improvement in library </FONT><BR><FONT FACE="Verdana"
SIZE=4>use by taking all the scattered components of a typical
</FONT><BR><FONT FACE="Verdana" SIZE=4>C library and encapsulating
them into a structure (an abstract data type, called a <I>class</I>
from now on). </FONT><BR></P></DIV>
```

Předzpracování dokumentů

- ✓ Lemmatizace textu
- ✓ Označení pojmů WordNetu
- ✓ Zachování původního obsahu
- ✓ Hledání klíčových slov ontologie

Výsledky

- Zjednodušené a normalizované dokumenty

```
<?xml version="1.0" encoding="UTF-8" ?>
<root>
  <head>
    <title>
      <content type="original">6: Initialization…
      <content type="wordnet">(6) (initialization)…
    </title>
  </head>
  <body>
    <content type="original">[ Viewing Hints ]…
    <content type="wordnet">(viewing) (hint) (exercise)…
    <h2>
      <content type="original">Thinking in C++, 2nd ed. Volume…
      <content type="wordnet">(thinking) (in) (c) (2nd) (ed) (volume)…
    </h2>
```

Výsledky

- Předzpracované vyhledávání klíčových slov

```
<found where="orig" num="607">
    <keyword num="3">&amp;</keyword>
    <keyword num="1">access control</keyword>
    ...
    <keyword num="3">try</keyword>
    <keyword num="10">using</keyword>
</found>
<found where="wn" num="709">
    <keyword num="2">access control</keyword>
    <keyword num="2">addition</keyword>
    ...
    <keyword num="3">try</keyword>
    <keyword num="11">using</keyword>
</found>
```

Problémy

- metody pro poloautomatickou tvorbu doménového modelu
- různé formáty dokumentů
- různé světové jazyky
- implementace nástroje
- ověření funkcionality
- interpretace výsledků, návrhy vylepšení

Doménový model

- doménová ontologie
- koncepty, vazby, hierarchie
- tvorba
  - manuálně
  - automaticky
  - poloautomaticky

Potřeba, současný stav

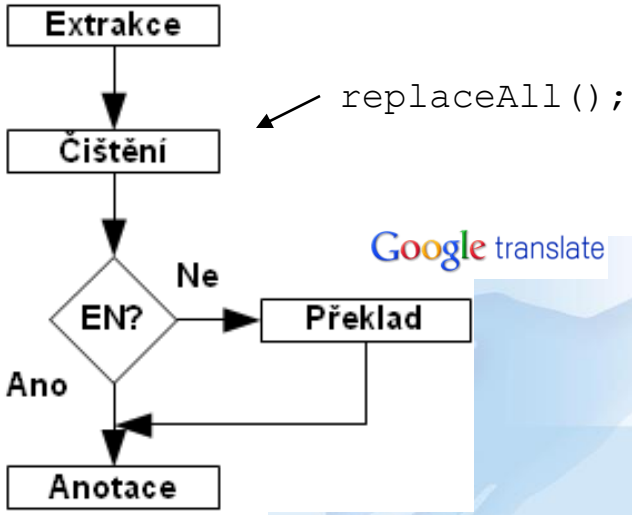
- XAPOS
- algoritmy nad doménovou ontologií
- tvorba doménové ontologie
  je časově náročná
- různé formáty dokumentů
- různé světové jazyky

Nástroje, metody

- zpracování přirozeného jazyka - NLP
- vícejazyčné prostředí
  Google Translate
- WordNet (synsety)
- Stanford NLP
- Java, SWING, XML, jTidy, JAWS, SNLP, JUNG

Příprava dat

- načtení textu, TXT, (X)HTML, PDF
- odstranění nežádoucích prvků
- překlad
- anotace




```
<html><body><p>An integer character
constant has type int.</p></body></html>
```

```
An/DT integer/NN character/NN
constant/NN has/VBZ type/NN int/NN ./.
```

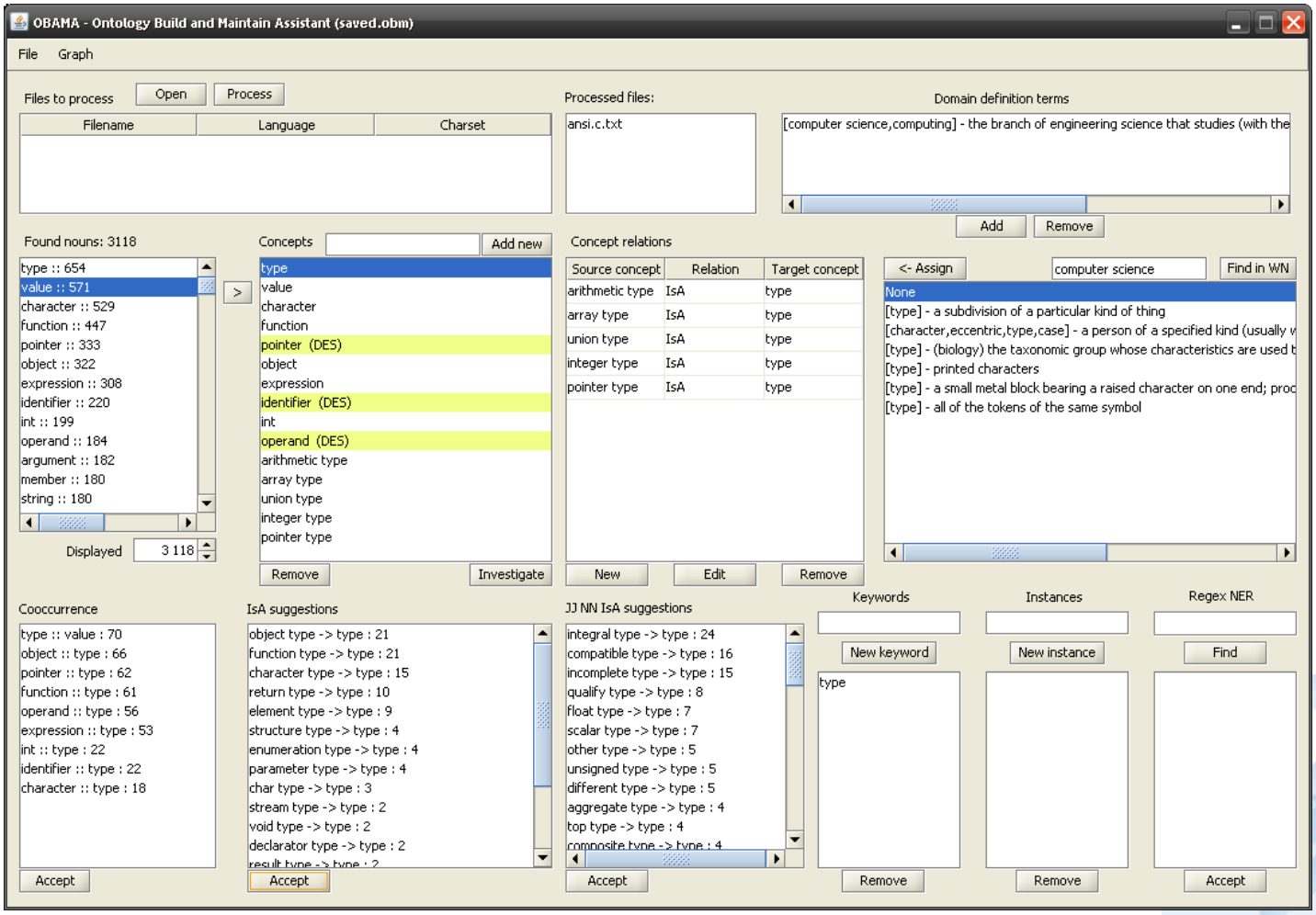
Vytěžování konceptů

- souvislé úseky podstatných jmen
- seřazení podle četností
- asistovaný výběr uživatelem
- desambiguace
- definice doménových termů
- instance konceptů - NER

Vytěžování vazeb

- neorientované / orientované
- nepojmenované / pojmenované
- z WordNetu
  - IsA, Has part, Part of
- z textu
  - lexico-syntactic patterns
  - IsA z textu (JJ NN IsA NN)
  - nepojmenované (co-occurrence)

Aplikace


OBAMA - Ontology Build and Maintain Assistant (saved.obm)
File
Graph
Files to process
Open
Process
Filename
Language
Charset
Processed files:
ansi.c.txt
Domain definition terms
[computer science,computing] - the branch of engineering science that studies (with the
Add
Remove
Found nouns: 3118
type :: 654
value :: 571
character :: 529
function :: 447
pointer :: 333
object :: 322
expression :: 308
identifier :: 220
int :: 199
operand :: 184
argument :: 182
member :: 180
string :: 180
Displayed
3 118
>
Concepts
Add new
type
value
character
function
pointer (DES)
object
expression
identifier (DES)
int
operand (DES)
arithmetic type
array type
union type
integer type
pointer type
Remove
Investigate
Concept relations
Source concept
Relation
Target concept
arithmetic type
IsA
type
array type
IsA
type
union type
IsA
type
integer type
IsA
type
pointer type
IsA
type
New
Edit
Remove
<- Assign
computer science
Find in WN
None
[type] - a subdivision of a particular kind of thing
[character,eccentric,type,case] - a person of a specified kind (usually
[type] - (biology) the taxonomic group whose characteristics are used
[type] - printed characters
[type] - a small metal block bearing a raised character on one end; proc
[type] - all of the tokens of the same symbol
Cooccurrence
type :: value : 70
object :: type : 66
pointer :: type : 62
function :: type : 61
operand :: type : 56
expression :: type : 53
int :: type : 22
identifier :: type : 22
character :: type : 18
Accept
IsA suggestions
object type -> type : 21
function type -> type : 21
character type -> type : 15
return type -> type : 10
element type -> type : 9
structure type -> type : 4
enumeration type -> type : 4
parameter type -> type : 4
char type -> type : 3
stream type -> type : 2
void type -> type : 2
declarator type -> type : 2
Accept
JJ NN IsA suggestions
integral type -> type : 24
compatible type -> type : 16
incomplete type -> type : 15
qualify type -> type : 8
float type -> type : 7
scalar type -> type : 7
other type -> type : 5
unsigned type -> type : 5
different type -> type : 5
aggregate type -> type : 4
top type -> type : 4
Accept
Keywords
New keyword
type
Remove
Instances
New instance
Remove
Regex NER
Find
Accept

Data, experiment

- text návrhu ANSI/ISO normy jazyka C
- porovnání s existující ontologií

Varianta	Přidáno	Položek v modelu	Nalezeno konceptů	Nalezeno/ položek (v %)	Nalezeno/ celkem v ontologii (v %)	Nalezeno/ možno nalézt (v %)
Všechny	-	3137	395	12,6	38,1	73,3
	IsA	4519	450	10,0	43,4	83,5
	IsA + NER	4558	465	10,2	44,8	86,3
200	-	200	98	49,0	9,4	18,2
	IsA	1802	152	8,4	14,6	28,2
	IsA + NER	1962	318	16,2	30,6	59,0

Výsledky

Varianta	Položek v modelu	Nalezených konceptů (v %)	Konceptů/ položek (v %)	Konceptů/ celkem v ontologii (v %)	Konceptů/ možno nalézt (v %)
Všechny + NER	3204	444	13,9	42,8	82,4
200 + NER	360	265	73,6	25,5	49,2

- IsA návrhy nepřijímat bezhlavě
- velký význam NER