The Polish Cyc lexicon as a bridge between Polish language and the Semantic Web

Aleksander Pohl

Computational Linguistics Departament
Jagiellonian University

18-20, September 2010

Agenda

Goals & Motivation

Methodology

Results

Agenda

Goals & Motivation

Methodology

Results

Strategic goals

- knowledge based **information extraction**:
  „Google Inc. is road-testing cars that steer, stop and start without a human driver, the company says.”

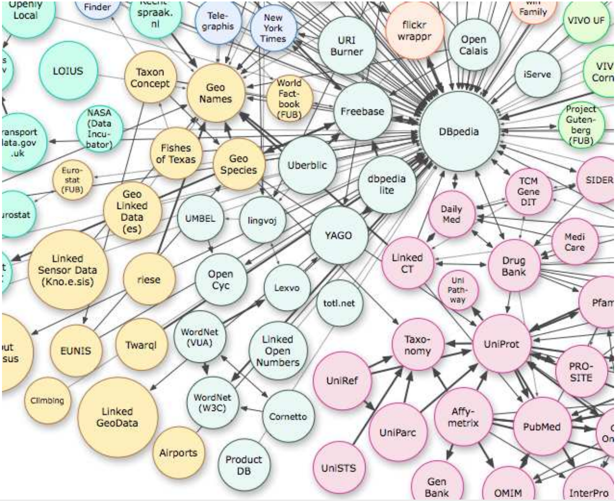
```
(#$and
(#$isa #$Test-123 #$PerformenceTesting)
(#$performedBy #$Test-123 #$GoogleInc)
(#$objectOfEvaluation #$Test-123
(#$InstanceFn
#$TransportationDevice-Unmanned)))
```

- knowledge based **text generation**:
  „Google testuje pojazdy autonomiczne.”

Tactical goals

- Polish mappings for Cyc concepts which form semantic restrictions of Cyc relations arguments
  - `(#$arg1Isa #$hasMembers #$Organization)`
  - `(#$arg2Isa #$hasMembers #$Agent-Generic)`
  - `(#$interArgIsa1-2 #$hasMembers #$SoccerTeam #$SoccerPlayer)`
- Mapping of (Polish) Wikipedia semantic categories to Cyc concepts
  - polski piłkarz → `#$SoccerPlayer`, `#$PolishPerson`
    - Andrzej Salach
    - Józef Kałuża
    - . . . 772 more
  - polski klub piłkarski → `#$SoccerTeam`
    - Wisła Kraków
    - Jarota Jarocin
    - . . . 210 more

Motivation – **Data + Knowledge**



Agenda

Goals & Motivation

Methodology

Results

Lexicon creation

Previous results (IIS 2009):

- overall accuracy: 54%
- concepts translated: approx. 20K

Current attempt:

- computer-aided translation
- transfer-base approach
- multi-word expressions
- preservation of full lexical information
- several thousands of concepts

The Algorithm

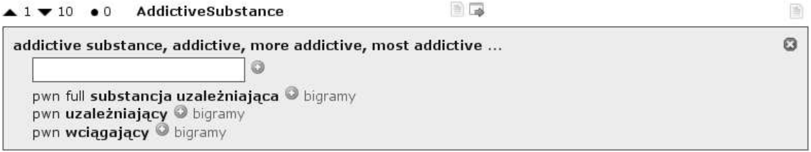
- *translate* the English mapping of the concept into Polish (many results might be produced)
- *map* the words of each translation to the entries of Polish inflectional dictionary
- *transform* the translations to match syntax constraints
- *rank* the translations
- *present* the results to the human operator
- *store* the selected result in the database
- *search* for semantic categories extracted from the Polish Wikipedia, corresponding to the translation
- *merge* or *link* the selected categories with the Cyc concept

Translation, Mapping & Transformation

- Translation
  - English-Polish large transfer dictionary Oxford/PWN
  - ignoring additional information (gender, categorization, domain, etc.)
  - each word of a multi-word expression translated separately
- Mapping to inflectional dictionary
  - lexemes indexed by base form and inflectional paradigm, e.g. `<uzależniający,CAA>`
- Transformation
  1. noun + adjective
  2. noun + noun
  3. noun + verb
  4. noun + other
  5. other

Ranking & Selection

- Ranking
  - IPI PAN corpus (250 mil. segments), not balanced
  - Google considered, but too slow for on-line translation
- Selection

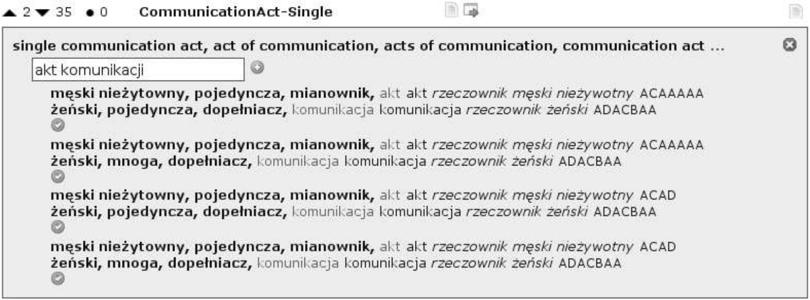


▲ 1 ▼ 10 ● 0 **AddictiveSubstance**

addictive substance, addictive, more addictive, most addictive ...

pwn full **substancja uzależniająca** bigramy
pwn **uzależniający** bigramy
pwn **wciągający** bigramy

Interpretation



▲ 2 ▼ 35 ● 0 **CommunicationAct-Single**

single communication act, act of communication, acts of communication, communication act ...

akt komunikacji

męski nieżytowny, pojedyncza, mianownik, akt akt *rzeczownik męski nieżywotny* ACAAAAA
żeński, pojedyncza, dopełniacz, komunikacja komunikacja *rzeczownik żeński* ADACBAA

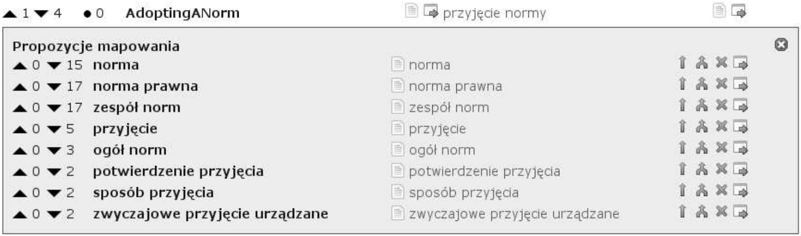
męski nieżytowny, pojedyncza, mianownik, akt akt *rzeczownik męski nieżywotny* ACAAAAA
żeński, mnoga, dopełniacz, komunikacja komunikacja *rzeczownik żeński* ADACBAA

męski nieżytowny, pojedyncza, mianownik, akt akt *rzeczownik męski nieżywotny* ACAD
żeński, pojedyncza, dopełniacz, komunikacja komunikacja *rzeczownik żeński* ADACBAA

męski nieżytowny, pojedyncza, mianownik, akt akt *rzeczownik męski nieżywotny* ACAD
żeński, mnoga, dopełniacz, komunikacja komunikacja *rzeczownik żeński* ADACBAA

Categories selection

$$R_i = \frac{cm_{i,j}}{cl_i} * \frac{cm_{i,j}}{cl_j} * children_i \tag{1}$$

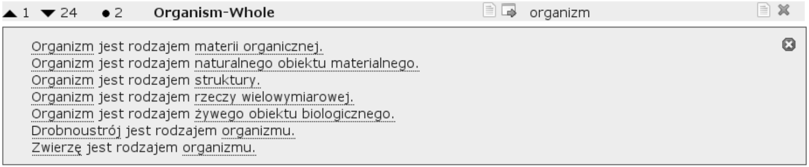


▲ 1 ▼ 4 ● 0 AdoptingANorm przyjęcie normy

Propozycje mapowania

▲ 0	▼ 15	norma	norma
▲ 0	▼ 17	norma prawna	norma prawna
▲ 0	▼ 17	zespół norm	zespół norm
▲ 0	▼ 5	przyjęcie	przyjęcie
▲ 0	▼ 3	ogół norm	ogół norm
▲ 0	▼ 2	potwierdzenie przyjęcia	potwierdzenie przyjęcia
▲ 0	▼ 2	sposób przyjęcia	sposób przyjęcia
▲ 0	▼ 2	zwyczajowe przyjęcie urządzane	zwyczajowe przyjęcie urządzane

Verification



Organizm jest rodzajem materii organicznej.
Organizm jest rodzajem naturalnego obiektu materialnego.
Organizm jest rodzajem struktury.
Organizm jest rodzajem rzeczy wielowymiarowej.
Organizm jest rodzajem żywego obiektu biologicznego.
Drobnoustrój jest rodzajem organizmu.
Zwierzę jest rodzajem organizmu.

Example of a translation

- #$AddictiveSubstance → addictive substance
- → [(uzależniający, wciągający)$_1$,(substancja, istota, ciężar, waga, podstawa, treść, realność, majątek)$_2$]
- → [(<uzależniać,BDA>, <uzależniać się,BDA>, <uzależniający,CAA>, ...)$_1$, (<substancja,ADACBAA>, <istota,ADAAA>, <ciężar,ACAAAAA>, <Ciężar,AAAAD>,...)$_2$]
- → [uzależniająca substancja, uzależniająca się substancja, wciągająca substancja, wciągająca się substancja,...]
- → „substancja uzależniająca”

Agenda

Goals & Motivation

Methodology

Results

Results

- Precision: 37% (560 translations)
- Baseline: 19% (Google Translate)
- Recall: 88%
- Inter-translator agreement: 56%
- Concepts translated : **1128** (target approx. 4000)
- Concepts mapped : 493
- Total number of concepts covered: **218942**

Thank You!

Why not WordNet?

WordNet and OpenCyc contents is overlapping:

- dog *direct hypernym* canine
- `(#$genls #$Dog #$CanineAnimal)`

but:

- Cyc relations have formal semantics
- CycL expressiveness is higher (rules, functions, microtheories, arbitrary arity relations):
  - `(#$relationAllExists #$bodyPartsUsed #$AnimalWalkingProcess #$Leg)`
  - `(#$relationAllExists #$properPhysicalParts #$CanineAnimal #$Leg)`
- Cyc is shipped with sophisticated *inferencing engine*

Why not WordNet?

WordNet and OpenCyc contents is overlapping:

- dog *direct hypernym* canine
- `(#$genls #$Dog #$CanineAnimal)`

but:

- Cyc relations have formal semantics
- CycL expressiveness is higher (rules, functions, microtheories, arbitrary arity relations):
    - `(#$relationAllExists #$bodyPartsUsed #$AnimalWalkingProcess #$Leg)`
    - `(#$relationAllExists #$properPhysicalParts #$CanineAnimal #$Leg)`
- Cyc is shipped with sophisticated *inferencing engine*

Why not DBpedia (YAGO, SUMO)?

- DBpedia
  - 259 classes
  - 1200 relations
  - no rules
  - SPARQL end-point
  - no inflectional data for Polish labels
- Cyc is the biggest ontology
  - 500K symbols (70K collections)
  - 17K (26K) relations
  - 5M assertions (mostly *defining* the terms)
  - rules
  - sophisticated inferencing engine

Why not DBpedia (YAGO, SUMO)?

- DBpedia
  - 259 classes
  - 1200 relations
  - no rules
  - SPARQL end-point
  - no inflectional data for Polish labels
- Cyc is the biggest ontology
  - 500K symbols (70K collections)
  - 17K (26K) relations
  - 5M assertions (mostly *defining* the terms)
  - rules
  - sophisticated inferencing engine