Automatic Construction of the Polish Nominal Lexicon for the OpenCyc Ontology

Aleksander Pohl

Intelligent Information Systems 2009

16^{th} of June 2009

Goal & Motivation

What?

- OpenCyc – formal representation of common sense knowledge, eg. `(#$genls #$Dog #$CanisGenus)`
- OpenCyc Lexicon – mapping between Cyc symbols and English words, eg. `#$Dog` – „dog”, „doggie”, „hound”
- **General goal**: Polish Lexicon – mapping between Cyc symbols and Polish words, eg. `#$Dog` – „pies”
- **First step**: Polish *Nominal* Lexicon – only nouns and proper names

Why?

- Cyc ontology as a foundation for Polish Semantic Dictionary
- Ontology-based information extraction and translation

Goal & Motivation

What?

- OpenCyc – formal representation of common sense knowledge, eg. `(#$genls #$Dog #$CanisGenus)`
- OpenCyc Lexicon – mapping between Cyc symbols and English words, eg. `#$Dog` – „dog”, „doggie”, „hound”
- **General goal**: Polish Lexicon – mapping between Cyc symbols and Polish words, eg. `#$Dog` – „pies”
- **First step**: Polish *Nominal* Lexicon – only nouns and proper names

Why?

- Cyc ontology as a foundation for Polish Semantic Dictionary
- Ontology-based information extraction and translation

Why not WordNet?

WordNet and OpenCyc contents is overlapping:

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

but:

- „sense density” is different – fine-grained WordNet synsets vs. coarse-grained Cyc concepts
- Cyc was designed in a language-agnostic manner
- CycL expressiveness is higher (rules, functions, microtheories, arbitrary arity relations):
  - `(#$distalTo (#$The(#$LeftObjectOfPairFn #$GonadalVein)) (#$The #$AorticArch))`
- Cyc is shipped with sophisticated *inferencing engine*

Why not WordNet?

WordNet and OpenCyc contents is overlapping:

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

but:

- „sense density” is different – fine-grained WordNet synsets vs. coarse-grained Cyc concepts
- Cyc was designed in a language-agnostic manner
- CycL expressiveness is higher (rules, functions, microtheories, arbitrary arity relations):
  - `(#$distalTo (#$The(#$LeftObjectOfPairFn #$GonadalVein)) (#$The #$AorticArch))`
- Cyc is shipped with sophisticated *inferencing engine*

Tools

- OpenCyc (`http://opencyc.org`):
  - 300 thousands concepts
  - 3 millions assertions
  - 15 thousands relations
- The Great English-Polish Polish-English Multimedia Dictionary Oxford/PWN 2004:
  - designed for humans, not computers
  - uses SGML
  - approx. 78 thousands entries in English-Polish part
  - besides simple translations, contains grammatical, lexical and domain qualifications, as well as examples

Tools

- OpenCyc (`http://opencyc.org`):
  - 300 thousands concepts
  - 3 millions assertions
  - 15 thousands relations
- The Great English-Polish Polish-English Multimedia Dictionary Oxford/PWN 2004:
  - designed for humans, not computers
  - uses SGML
  - approx. 78 thousands entries in English-Polish part
  - besides simple translations, contains grammatical, lexical and domain qualifications, as well as examples

The algorithm

Iterate over all the entries in the dictionary, trying to find best matchings between Cyc symbols and Polish words corresponding to given English word.

The problem – homonymy

- English-Polish Dictionary **grain**:
  1. (commodity) **zboże**; (different kinds) **zboża**
  2. (seed) **ziarno**
  3. (small piece) (of sand) **ziarnko**; (of salt) **kryształek**
  4. *fig* (of hope, comfort) **odrobina**; ...
  5. (pattern) (in wood) **słoje**; (in paper, fabric, flesh) **włókna**; ...
  6. (roughness) Phot **ziarno**
  7. Meas (weight) **gran** (= 0,0648 g)
- OpenCyc **grain**:
  1. `(#$FruitFn #$CerealPlant)` → 1(?), 2
  2. `#$GrainOfCereal` → 2
  3. `#$Grain-UnitOfMass` → 7

The problem – homonymy

- English-Polish Dictionary **grain**:
  1. (commodity) **zboże**; (different kinds) **zboża**
  2. (seed) **ziarno**
  3. (small piece) (of sand) **ziarnko**; (of salt) **kryształek**
  4. *fig* (of hope, comfort) **odrobina**; ...
  5. (pattern) (in wood) **słoje**; (in paper, fabric, flesh) **włókna**; ...
  6. (roughness) Phot **ziarno**
  7. Meas (weight) **gran** (= 0,0648 g)
- OpenCyc **grain**:
  1. **(#$FruitFn #$CerealPlant)** → 1(?), 2
  2. **#$GrainOfCereal** → 2
  3. **#$Grain-UnitOfMass** → 7

Grouping heuristics

Semantic groups vs. Cyc concepts:

- **1-to-1** – map with strong confidence
- **1-to-n** – apply semantic h., then map to all with medium confidence
- **n-to-1** – apply semantic h., then map to the first with medium, and rest with weak confidence
- **n-to-n** – apply semantic h., then map Cartesian product of sets with weak confidence

Grouping heuristics

Semantic groups vs. Cyc concepts:

- **1-to-1** – map with strong confidence
- **1-to-n** – apply semantic h., then map to all with medium confidence
- **n-to-1** – apply semantic h., then map to the first with medium, and rest with weak confidence
- **n-to-n** – apply semantic h., then map Cartesian product of sets with weak confidence

Grouping heuristics

Semantic groups vs. Cyc concepts:

- **1-to-1** – map with strong confidence
- **1-to-n** – apply semantic h., then map to all with medium confidence
- **n-to-1** – apply semantic h., then map to the first with medium, and rest with weak confidence
- **n-to-n** – apply semantic h., then map Cartesian product of sets with weak confidence

Grouping heuristics

Semantic groups vs. Cyc concepts:

- **1-to-1** – map with strong confidence
- **1-to-n** – apply semantic h., then map to all with medium confidence
- **n-to-1** – apply semantic h., then map to the first with medium, and rest with weak confidence
- **n-to-n** – apply semantic h., then map Cartesian product of sets with weak confidence

Semantic heuristics

- paradigmatic qualification – search the concept hierarchy
- syntagmatic qualification – use mapping between pre-defined categories (Animal, BodyPart, etc.) taken from Polish Semantic Dictionary and Cyc concepts related to them by means of syntagmatic relations
  - BodyPart – `#$BiologicalLivingObject`, just.: **foot** of cat, dog
- domain qualification – use mapping between domains and some general Cyc concepts closely related to given domain
  - Botany – `#$Plant`, `#$NaturalTangibleStuff`, `#$OrganismPart`

Semantic heuristics

- paradigmatic qualification – search the concept hierarchy
- syntagmatic qualification – use mapping between pre-defined categories (Animal, BodyPart, etc.) taken from Polish Semantic Dictionary and Cyc concepts related to them by means of syntagmatic relations
  - BodyPart – `#$BiologicalLivingObject`, just.: **foot** of cat, dog
- domain qualification – use mapping between domains and some general Cyc concepts closely related to given domain
  - Botany – `#$Plant`, `#$NaturalTangibleStuff`, `#$OrganismPart`

Semantic heuristics

- paradigmatic qualification – search the concept hierarchy
- syntagmatic qualification – use mapping between pre-defined categories (Animal, BodyPart, etc.) taken from Polish Semantic Dictionary and Cyc concepts related to them by means of syntagmatic relations
  - BodyPart – `#$BiologicalLivingObject`, just.: **foot** of cat, dog
- domain qualification – use mapping between domains and some general Cyc concepts closely related to given domain
  - Botany – `#$Plant`, `#$NaturalTangibleStuff`, `#$OrganismPart`

Results

- Only nouns and proper names were mapped (grammatical qualifier *n*, *npl*, *prn*)
- ∼27 thousands mappings were created for ∼16 thousands lexemes
- ∼3,5 thousands mappings were verified (∼ 12%)
- General precision: **54%**

confidence	strong	medium	weak
precision	64,7%	49,8%	23,1%

Results details (1)

	Abstract-Obj	Animal	Artifact	BodyPart	Event
# of map.	4652	878	4807	758	6957
strong	48.29%	87.5%	44.86%	70.42%	54.69%
medium	38.97%	61.64%	32.69%	84.21%	32.30%
weak	22.22%	18.75%	31.11%	15.38%	16.26%
overall	**42.39%**	**76.42%**	**40.22%**	**66.01%**	**35.29%**

	Human	Instrument	Location	Meter	NaturalObj
# of map.	2551	3486	2373	110	1432
strong	80.23%	57.26%	62.42%	91.89%	76.92%
medium	79.10%	54.90%	63.43%	80.95%	60.52%
weak	29.62%	12.0%	14.28%	100.0%	69.23%
overall	**74.71%**	**54.37%**	**59.61%**	**88.33%**	**72.61%**

Results details (2)

	Proper	Self	Set	State	Structure
# of map.	168	659	592	1590	358
strong	79.31%	53.84%	51.61%	82.95%	60.60%
medium	73.03%	62.5%	30.0%	69.23%	31.81%
weak	54.54%	46.42%	10.0%	37.14%	0.0%
overall	**72.86%**	**54.65%**	**39.34%**	**69.71%**	**41.53%**

	Food	Plant			
# of map.	489	208			
strong	84.31%	97.5%			
medium	60.0%	83.33%			
weak	21.42%	25.0%			
overall	**67.77%**	**89.28%**			

Conclusions

- **Completely automatic construction of the lexicon is not feasible – the result has to be reviewed manually.**
- The smaller the semantic category, the better the result.
- The notion of mapping confidence proved to be useful – the results might be ordered according to the confidence.
- The lack of grammatical categories in OpenCyc significantly influenced the result of event category – the ResearchCyc should give better much results.

Thank you!