NED with two-stage coherence optimization

or

How I taught my bottle of Jack Daniel’s not to turn into a 168-years-old person with a net income of $120.000.000

Filip Ilievski

Supervisors:

Marieke van Erp
Piek Vossen
Stefan Schlobach
Wouter Beek

Kick-off

- Combinatorial explosion
- FrameNet
- NWR's Events and Situations Ontology (ESO)

- Clash of the worlds:

Roles (A0, A1, A2),
frames,
selectional restrictions

vs

types,
subjects with domains,
objects with ranges

Outline

1. Background
    a. Language processing
    b. Identity
2. Problem statement
3. Research scope
4. Contemporary approaches
5. Solution design
6. Tools and examples
7. Experimental Setup
8. Remarks

Language processing: Motivation

- “Ninety percent of all the data in the world was produced in the last two years.
  - This trend is expected to grow.”

- “80 percent of all the information in the world is unstructured information.”

- We need computers that can understand this flood of information.
  - i.e. we need tools to automatically process language

(IBM on Watson, 2012)

On the Ambiguity of language

- Language is at the base of our cognition, our ability to understand the world
- The language is subjective and relates to a discourse world

- Language is incredibly imprecise: we luv abusing and messing up
  - How can a *slim chance* and a *fat chance* be the same, but a *wise man* and a *wise guy* are opposites?
  - How can a house *burn up* as it *burns down*?
  - Why do we *fill in* a form by *filling it out*?
- And language is amazingly accurate
  - despite all its inconsistencies, irregularities and contradictions, we convey so much meaning and accomplish so much collaboration.

On the Ambiguity of language

- Language is at the base of our cognition, our ability to understand the world
- The language is subjective and relates to a discourse world

- Language is incredibly imprecise: we luv reusing and messing up
  - How can a *slim chance* and a *fat chance* be the same, but a *wise man* and a *wise guy* are opposites?
  - How can a house *burn up* as it *burns down*?
  - Why do we *fill in* a form by *filling it out*?
- And language is amazingly accurate
  - despite all its inconsistencies, irregularities and contradictions, we convey so much meaning and accomplish so much collaboration

On the Ambiguity of language

- Language is at the base of our cognition, our ability to understand the world
- The language is subjective and relates to a discourse world

- Language is incredibly imprecise: we luv reusing and messing up
  - How can a *slim chance* and a *fat chance* be the same, but a *wise man* and a *wise guy* are opposites?
  - How can a house *burn up* as it *burns down*?
  - Why do we *fill in* a form by *filling it out*?
- And language is amazingly accurate
  - despite all its inconsistencies, irregularities and contradictions, we convey so much meaning and accomplish so much collaboration

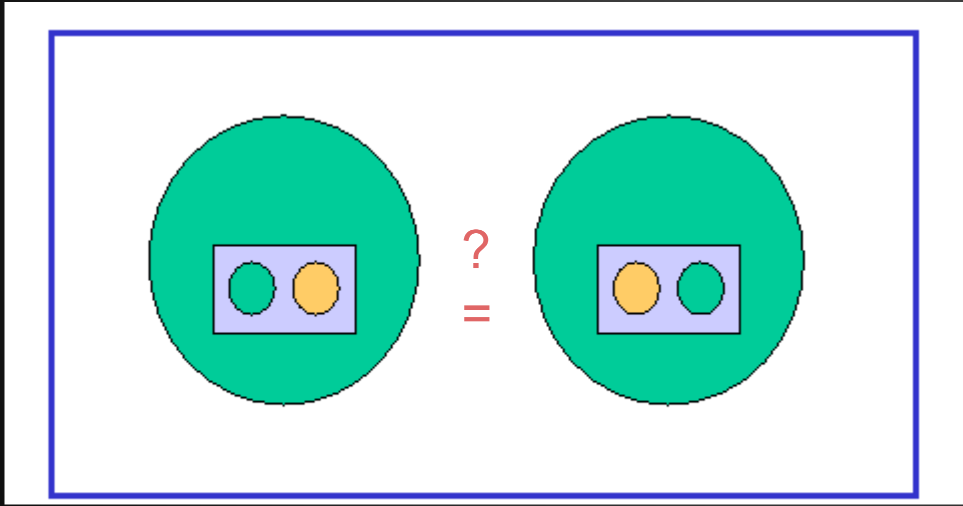
(IBM on Watson, 2012)

The burden of context in language

- The language is context-dependent
- Verbal context
  - Ford fell from a tree.
    - What is “Ford” ?
- Social context
  - What is “2+2” ?
    - In mathematics it is 4
    - In the car domain it is a car configuration: 2 front + 2 back seats
    - In psychology it is a family with 2 parents and 2 children

Identity

- Problem of identity in philosophy



$$\forall x \forall y [x = y \rightarrow \forall P (Px \leftrightarrow Py)]$$

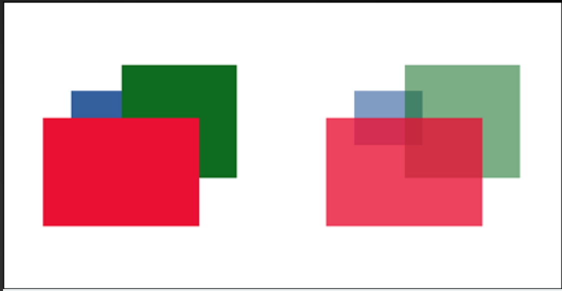
Identity in language

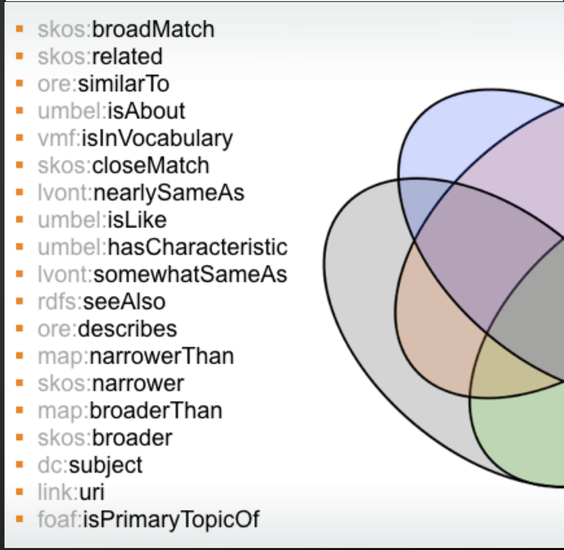
“Any two entities are both similar and dissimilar with respect to an infinite number of properties.” (Murphy & Medin, 1985)

- Entity linking becomes tricky
  - eg. Temporality: Is *Old Amsterdam* identical to *New Amsterdam* ?
- Continuum of identity
- Metonymy
  - “Plato is on the top shelf” - Who or what is Plato?

Identity in Semantic Web

- owl:sameAs follows Leibniz’s law (maybe?)
- Approaches:
  - based on the intrinsic properties
  - weaker definitions: near-identity, intransitive, nonsymmetric, non-reflexive constructs





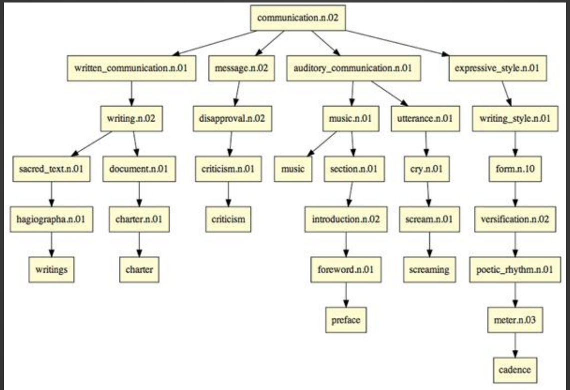
Resources

Natural Language Processing


On August 20, 2005, in a private ceremony, Thompson's ashes were fired from a cannon atop a tower of his own design (in the shape of a double-thumbed fist clutching a peyote button) to the tune of Norman Greenbaum's "Spirit in the Sky" and Bob Dylan's Mr. Tambourine Man. Red, white, blue, and green fireworks were launched along with his ashes. As the city of Aspen would not allow the cannon to remain for more than a month, the cannon has been dismantled and put into storage until a suitable permanent location can be found. According to his widow Anita, Thompson's funeral was financed by actor Johnny Depp, a close friend of Thompson. Depp told the Associated Press, "All I'm doing is trying to make sure his last wish comes true. I just want to send my pal out the way he wants to go out." Other famous attendees at the funeral included U.S. Senator John Kerry and former U.S. Senator George McGovern; 60 Minutes correspondent Ed Bradley and Charlie Rose; actors Jack Nicholson, Bill Murray, Benicio del Toro, Sean Penn, and Josh Hartnett; singers Lyle Lovett, John Oates and numerous other friends. An estimated 280 people attended the funeral.
The plans for this monument were initially drawn by Thompson and Ralph Steadman and were shown as part of an Omnibus program on the BBC entitled Fear and Loathing in Gonzovision (1978). It is included as a special feature on the second disc of the 2003 Criterion Collection DVD release of Fear and Loathing in Las Vegas (labeled on the DVD as "Fear and Loathing on the Road to Hollywood"). The video footage of Steadman and Thompson drawing the plans and outdoor footage showing where he wanted the cannon constructed were planned prior to the unveiling of his cannon at the funeral.

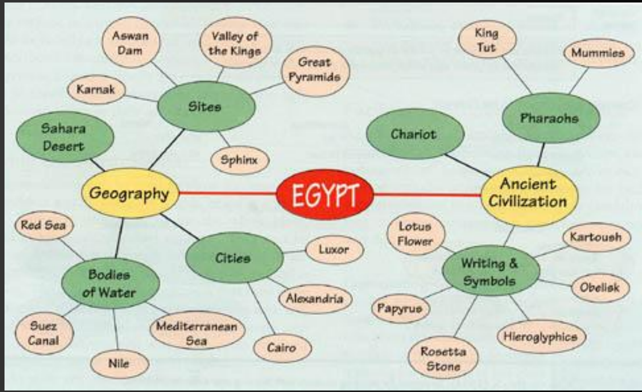
Grammatical structure and meaning of words

Lexical resources



Structured linguistic information

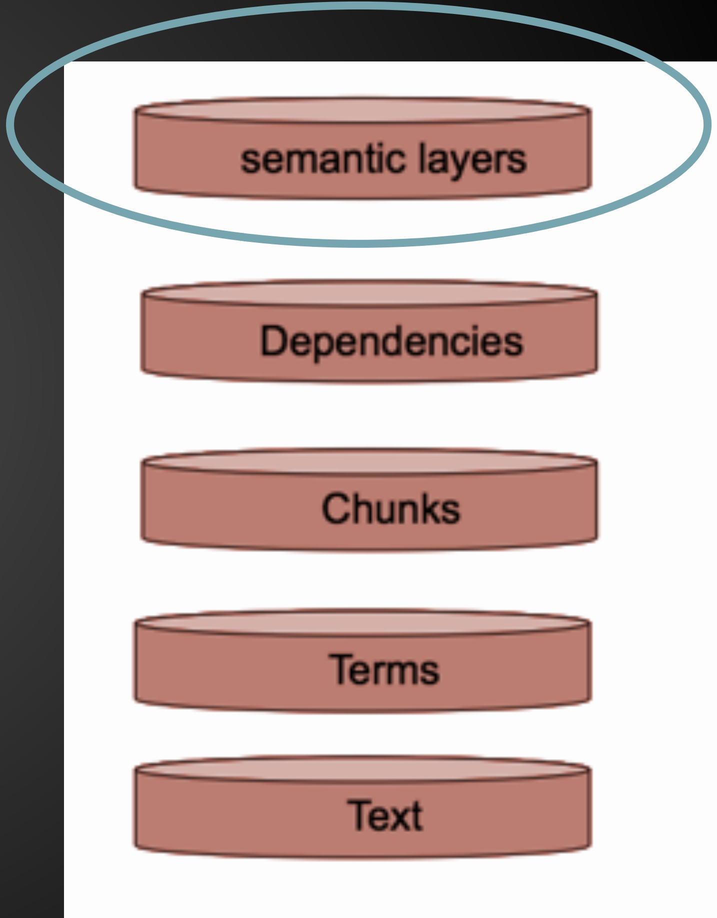
Semantic Web



Background knowledge

Research scope

- Semantic layer of the NLP pipeline
- Named Entity Disambiguation (NED) is the task of determining the identity of entities mentioned in text.
- Semantic Role Labelling (SRL) detects the semantic arguments associated with the predicate or verb of a sentence and classifies them into specific roles.



Research question

- Can the NED accuracy be improved by optimizing the coherence of the entities based on binary logic and probabilistic models?
  - What is the accuracy of each of the phases of the solution ?
    1. the binary filtering phase
    2. the probabilistic phase
  - Which components of the system are useful/useless?
  - How does this solution compare to the existing approaches (FRED, NewsReader or DBPedia Spotlight) ?

Example

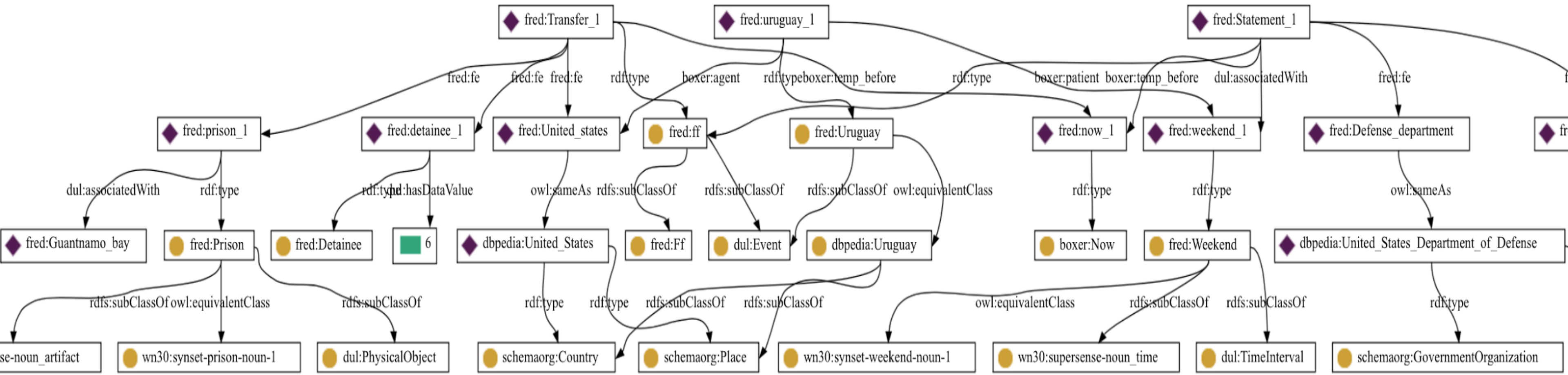
“The United States transferred six detainees from the Guantánamo Bay prison to Uruguay this weekend, the Defense Department announced early Sunday.”

State-of-the-art:

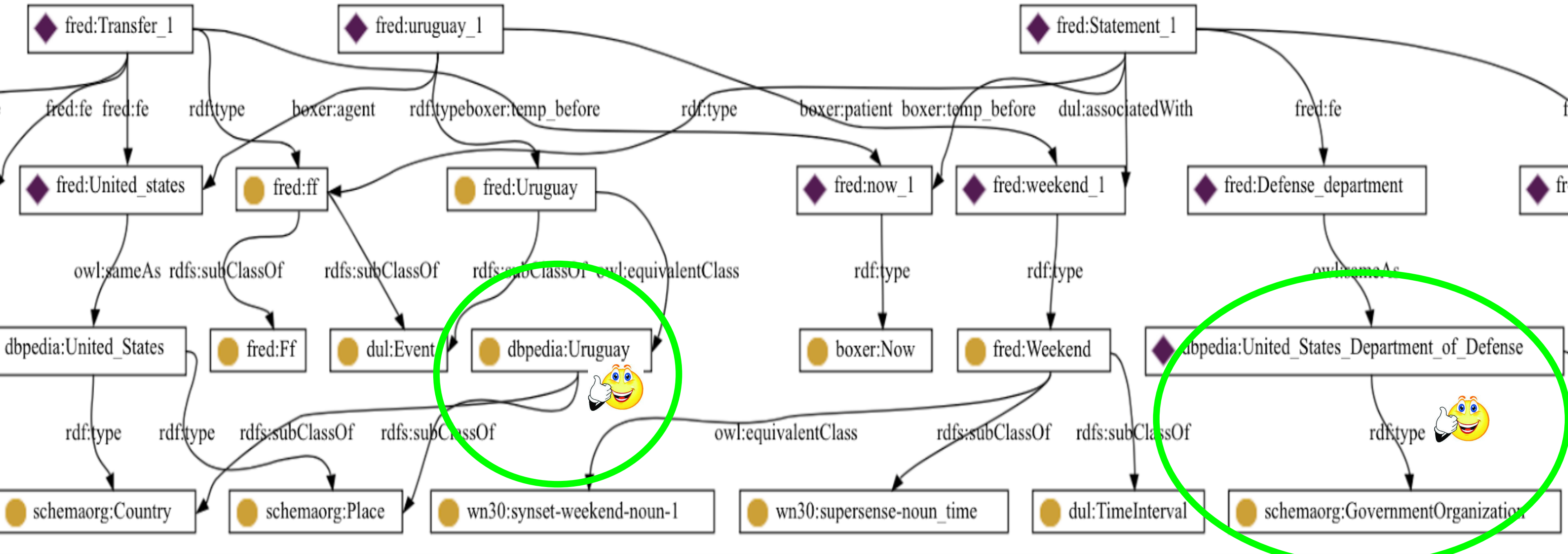
United States	Guantanamo Bay	Uruguay	Defence Department
Geographical region	GB detention camp	Geographical region	US Dept. of Defence
Fed. Government	Place	Football team	Ministry of Defence of Rep. of Korea
Men’s soccer team	The naval base	River	
Women’s soccer team	Battle of GB	Rugby union team	
Rugby union team		U20 football team	
Men’s ice hockey team		U17 football team	
Men’s basketball team			
Secondary education in US			

State-of-the-art: DBpedia Spotlight

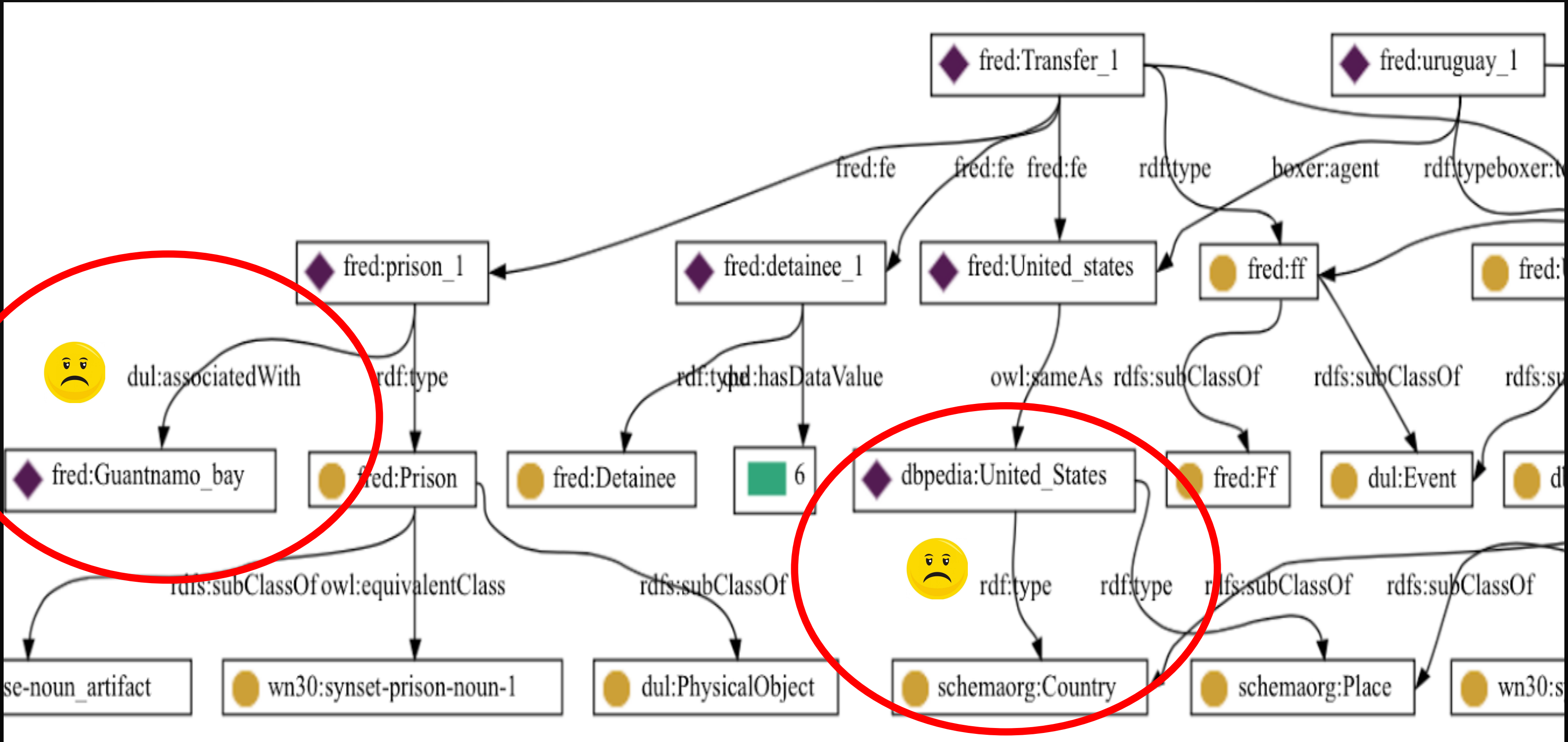
State-of-the-art: FRED



State-of-the-art: FRED



State-of-the-art: FRED



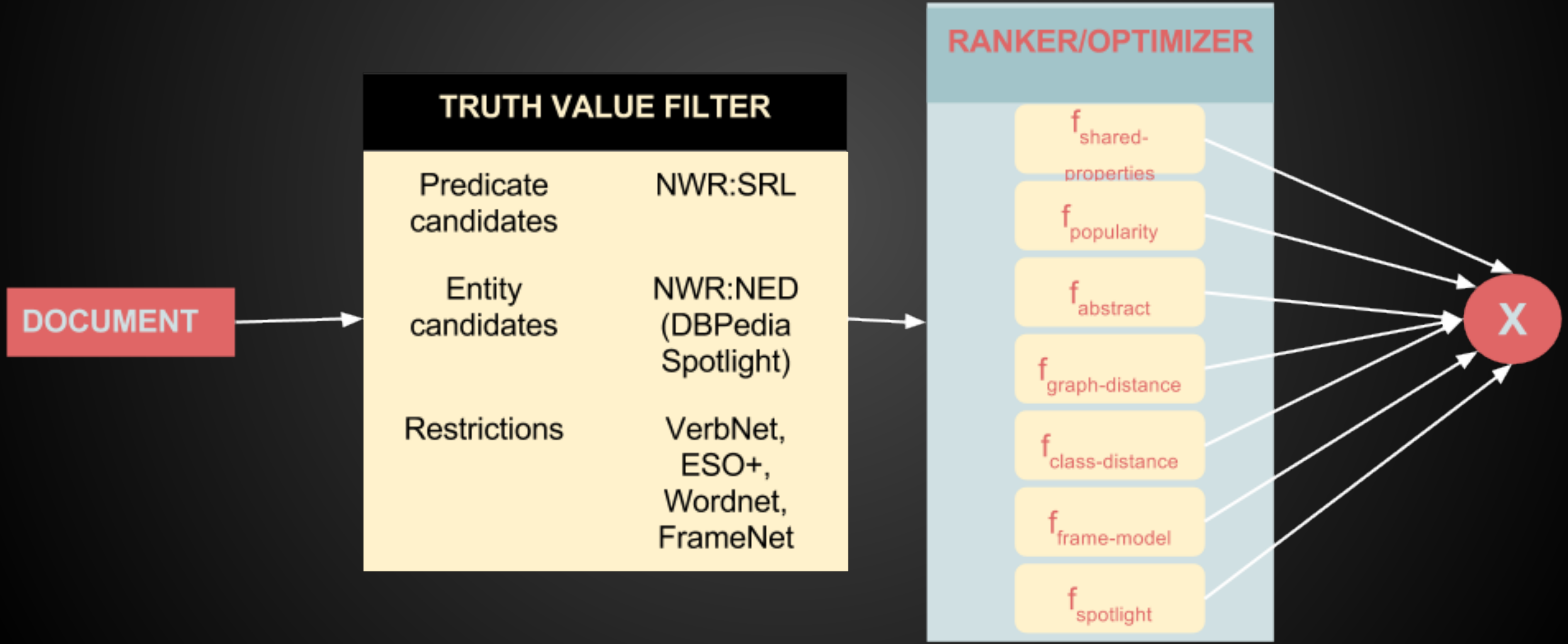
Shall we go a step further?

Proposed solution

Optimization in two phases:

1. phase excludes what is impossible
   - based on entity types and predicate restrictions
2. phase tells what is the most probable
   - based on frequency and semantic coherence of the entities

Proposed solution (II)



Phase I: Binary Filtering

DOCUMENT

TRUTH VALUE FILTER

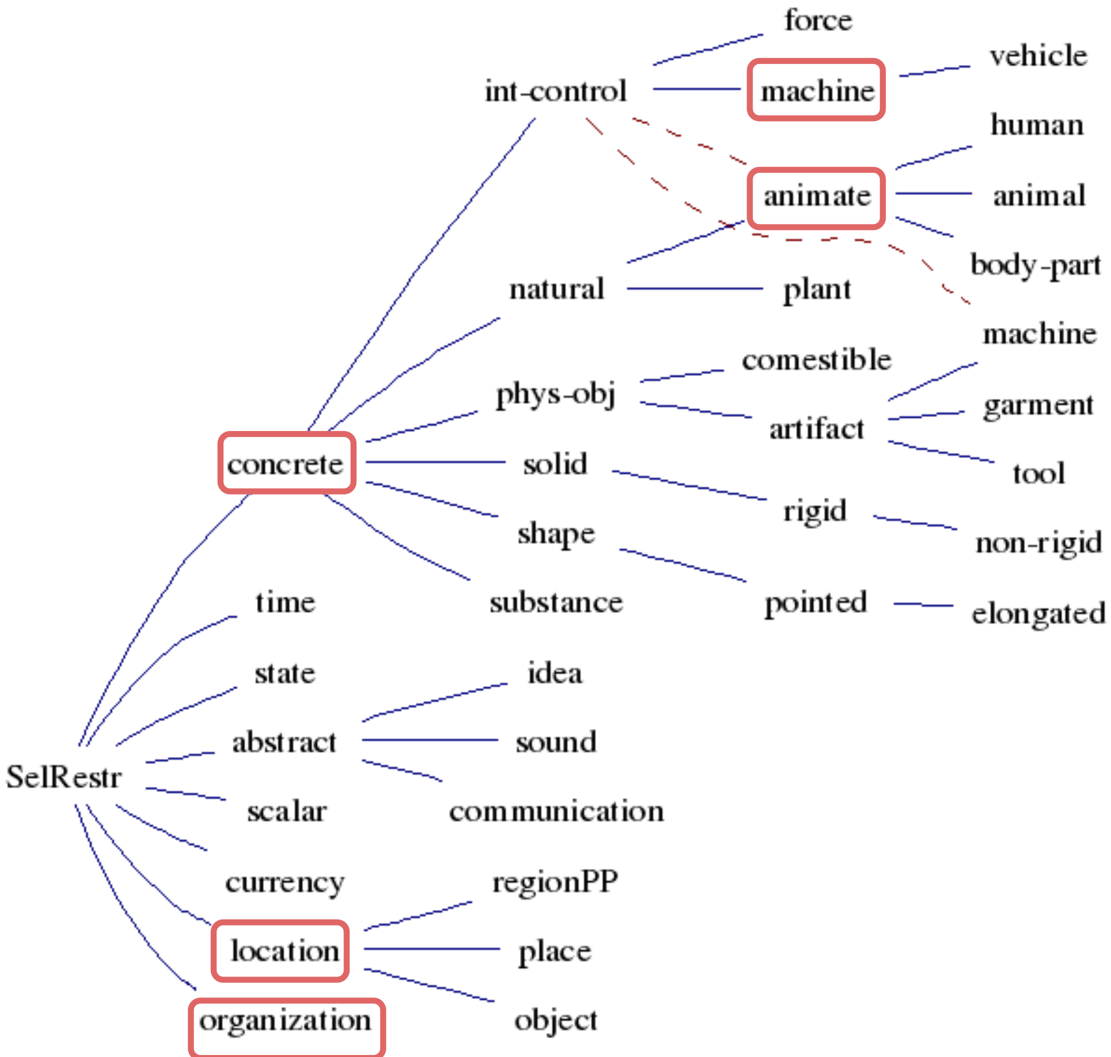
Predicate candidates	NWR:SRL
Entity candidates	NWR:NED (DBPedia Spotlight)
Restrictions	VerbNet, ESO+, Wordnet, FrameNet

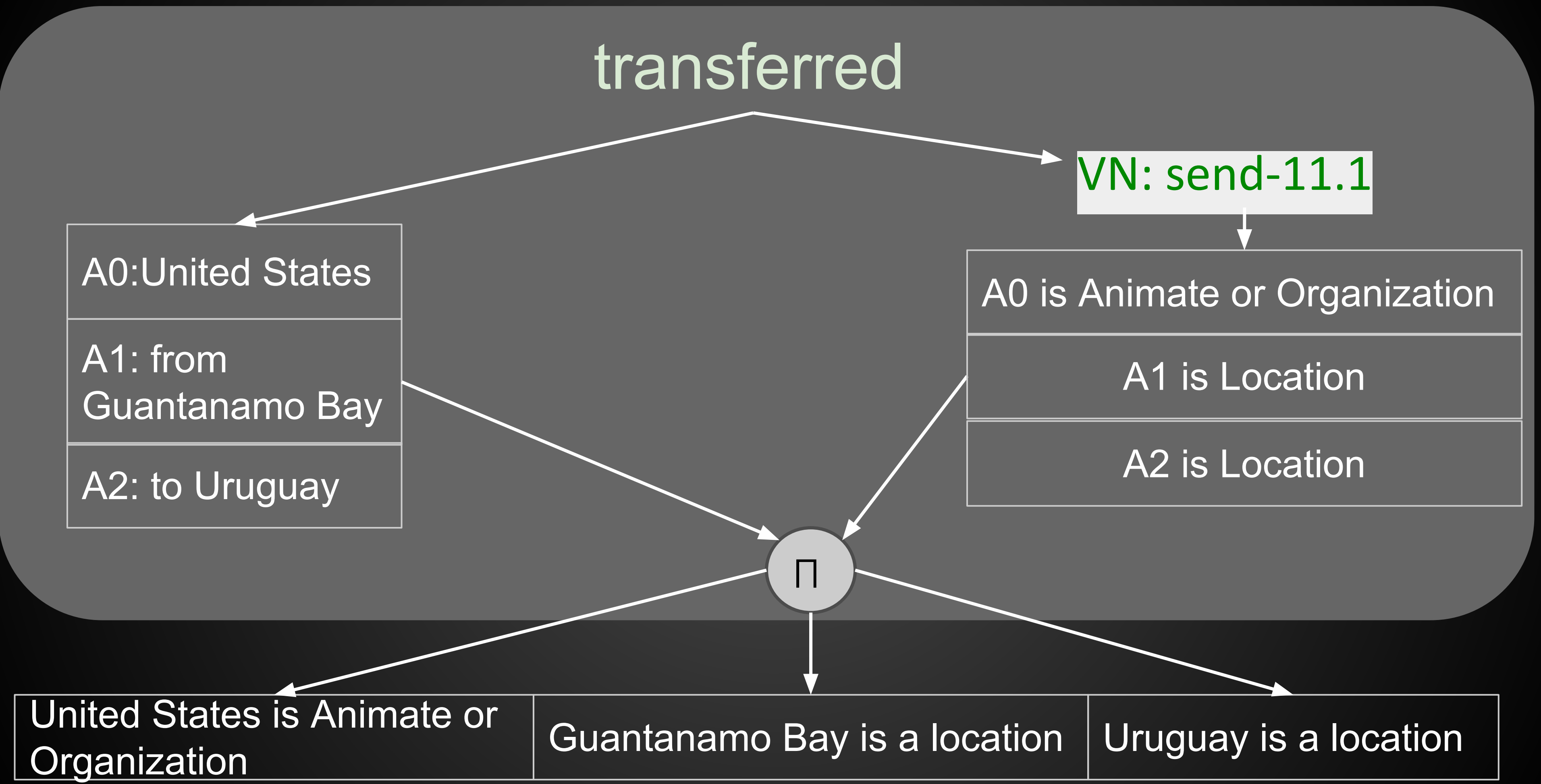
Tools: VerbNet

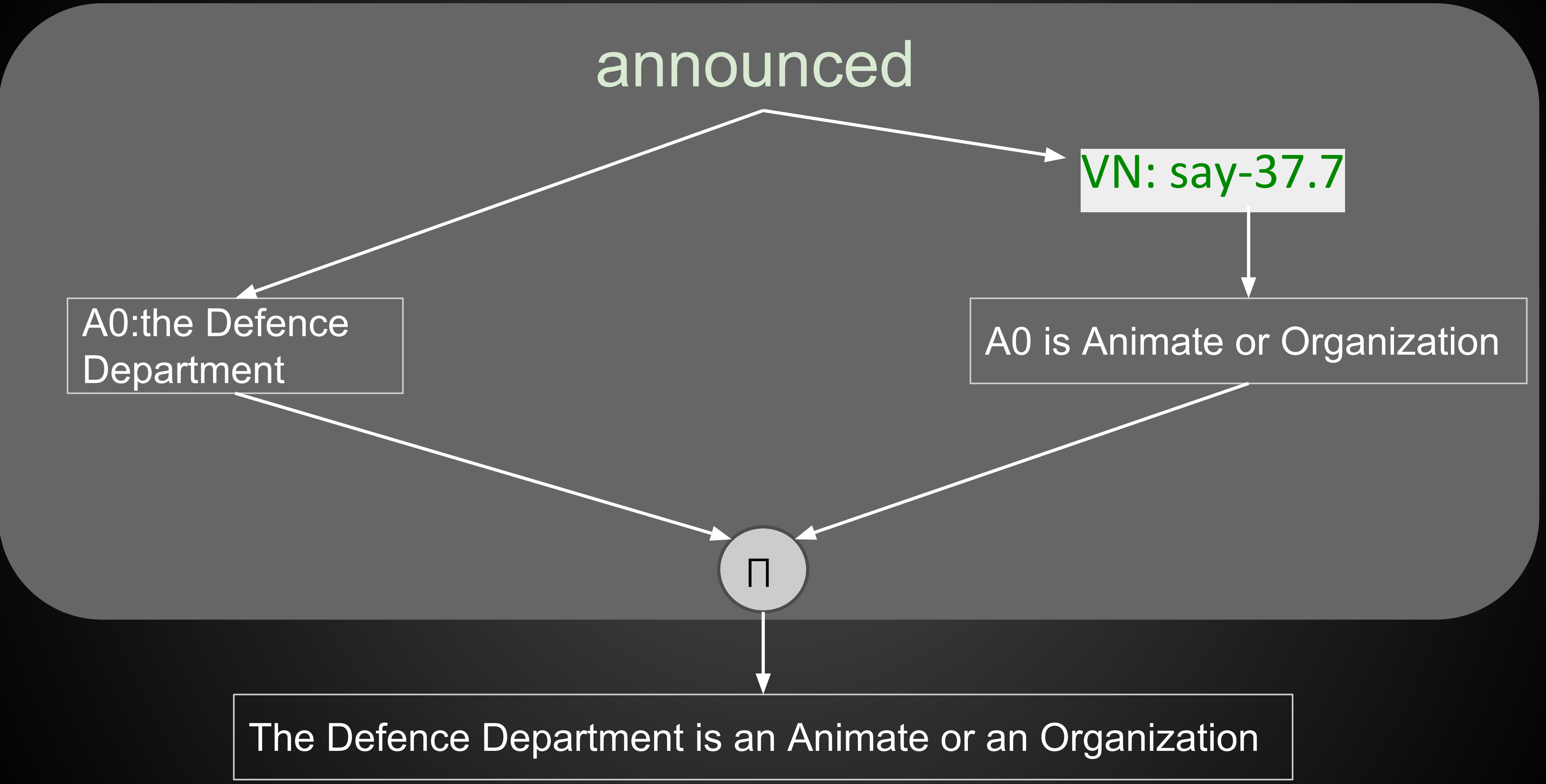
- largest online verb lexicon currently available for English
- hierarchical (not ontological though)
- domain-independent

+ good coverage of predicates
+ defines syntactic-semantic relations

- thematic roles are too generic


SelRestr
concrete
int-control
force
machine
vehicle
human
animate
animal
body-part
natural
plant
machine
phys-obj
comestible
artifact
garment
tool
solid
rigid
non-rigid
shape
pointed
elongated
substance
time
state
abstract
idea
sound
communication
scalar
currency
location
regionPP
place
object
organization


transferred
VN: send-11.1
A0:United States
A1: from Guantanamo Bay
A2: to Uruguay
A0 is Animate or Organization
A1 is Location
A2 is Location
⊓
United States is Animate or Organization
Guantanamo Bay is a location
Uruguay is a location


announced
VN: say-37.7
A0:the Defence Department
A0 is Animate or Organization
⊓
The Defence Department is an Animate or an Organization

After VerbNet

United States	Guantanamo Bay	Uruguay	Defence Department
~~Geographical region~~	**GB detention camp**	**Geographical region**	**US Dept. of Defence**
Fed. Government	Place	~~Football team~~	Ministry of Defence of Rep. of Korea
Men's soccer team	The naval base	River	
Women's soccer team	~~Battle of GB~~	~~Rugby union team~~	
Rugby union team		~~U20 football team~~	
Men's ice hockey team		~~U17 football team~~	
Men's basketball team			
~~Secondary education in US~~			

Tools: WordNet & FrameNet

WordNet

- lexical database for English
- groups words into synsets
- provides definitions and semantic relations between the synsets.

+ very good coverage of the verbs hierarchy

- does not capture the syntactic nor semantic behaviour.

FrameNet

- large-scale lexical resource with information on semantic frames (situations) and semantic roles.

+ Good generalization across predicates

- Does not define selectional restrictions for semantic roles
- Has limited coverage

Tools: NWR Events & Situations Ontology

- Reuse of existing ESO ontology (current version: 0.6)
- Global automotive industry
- Manually constructed, hence (hopefully) trustworthy
- Previous experiment on 1.3 million car industry articles demonstrated the 59 ESO classes with FrameNet and SUMO mappings cover 23% of the predicates
- Will contain domain and range information for frequently used classes linked to FN frames
- Complements VerbNet when no (enough) restrictions or granularity

ESO+ example

"Detroit Lions fired Mornhinweg."

- VN comes up with two predicates:
  - fire-10.10
  - throw-17.1

- Mornhinweg can be a: Concrete, Animate or Organization

- But in the car domain they usually dismiss an employee :-)

ESO+ example

firing

FN: Firing

FN: Shoot_projectiles

FN: Use_firearm

correspondsToFNFrame

ESO+: LeaveAnOrganization

Tools: DBpedia/Schema.org

- Semantic web resources with domain-independent factual information
  - DBpedia has 685 classes, over 4 million instances
- But how to map the VerbNet / ESO+ restrictions to DBpedia classes?
  - Easy and manually
  - ESO+ is natively mapped. As for VN:

VerbNet	DBPedia	Schema.org
Animate	Person	Person
Organization	Organization	Organization
Location	Place	Place

Phase II: Probabilistic optimization

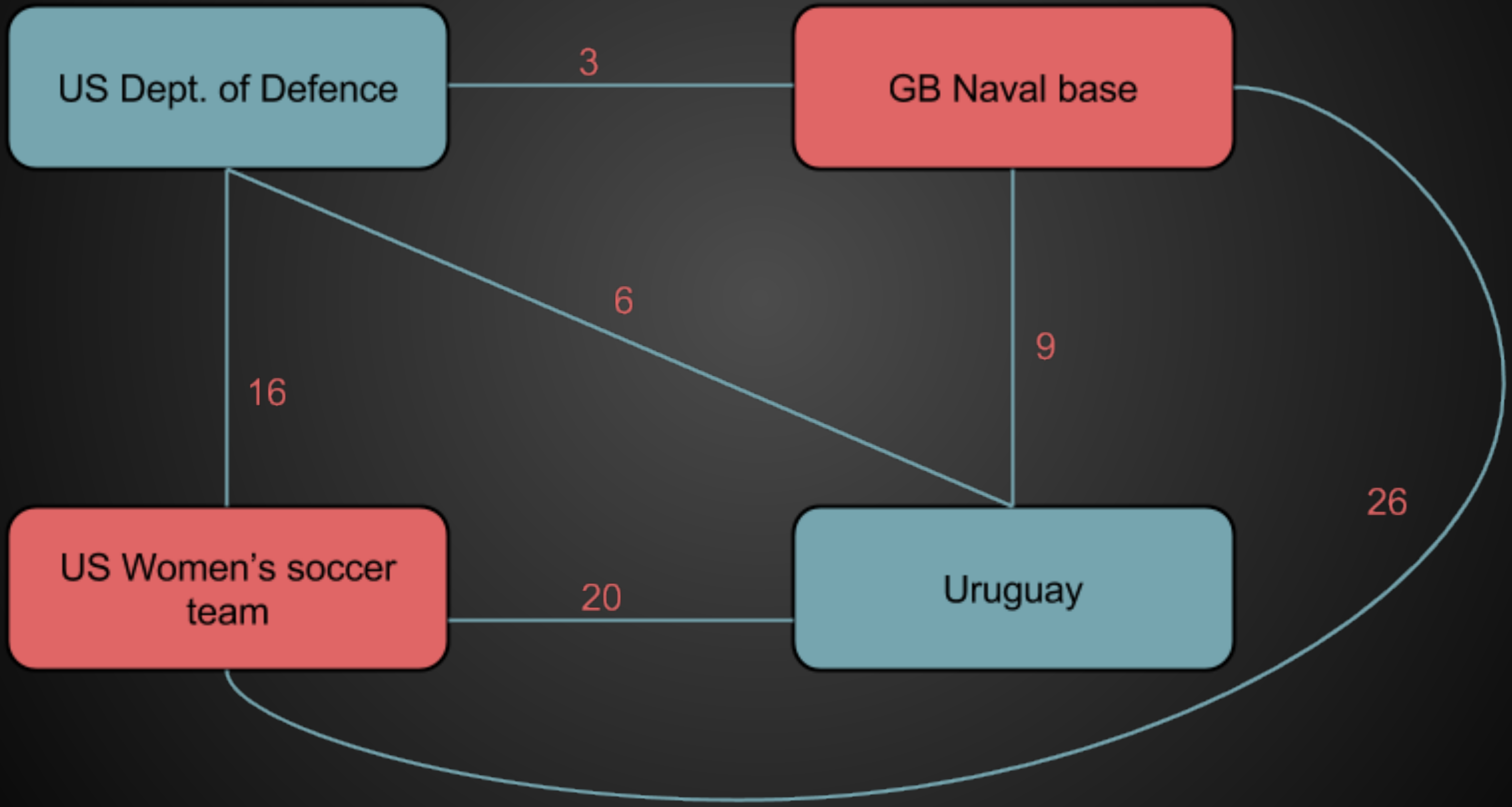
Module: DBpedia Spotlight

United States	Guantanamo Bay	Uruguay	Defence Department
~~Geographical region~~	**GB detention camp**	**Geographical region**	**US Dept. of Defence**
Fed. Government	Place	~~Football team~~	Ministry of Defence of Rep. of Korea
Men's soccer team	The naval base	River	
Women's soccer team	~~Battle of GB~~	~~Rugby union team~~	
Rugby union team		~~U20 football team~~	
Men's ice hockey team		~~U17 football team~~	
Men's basketball team			
~~Secondary education in US~~			

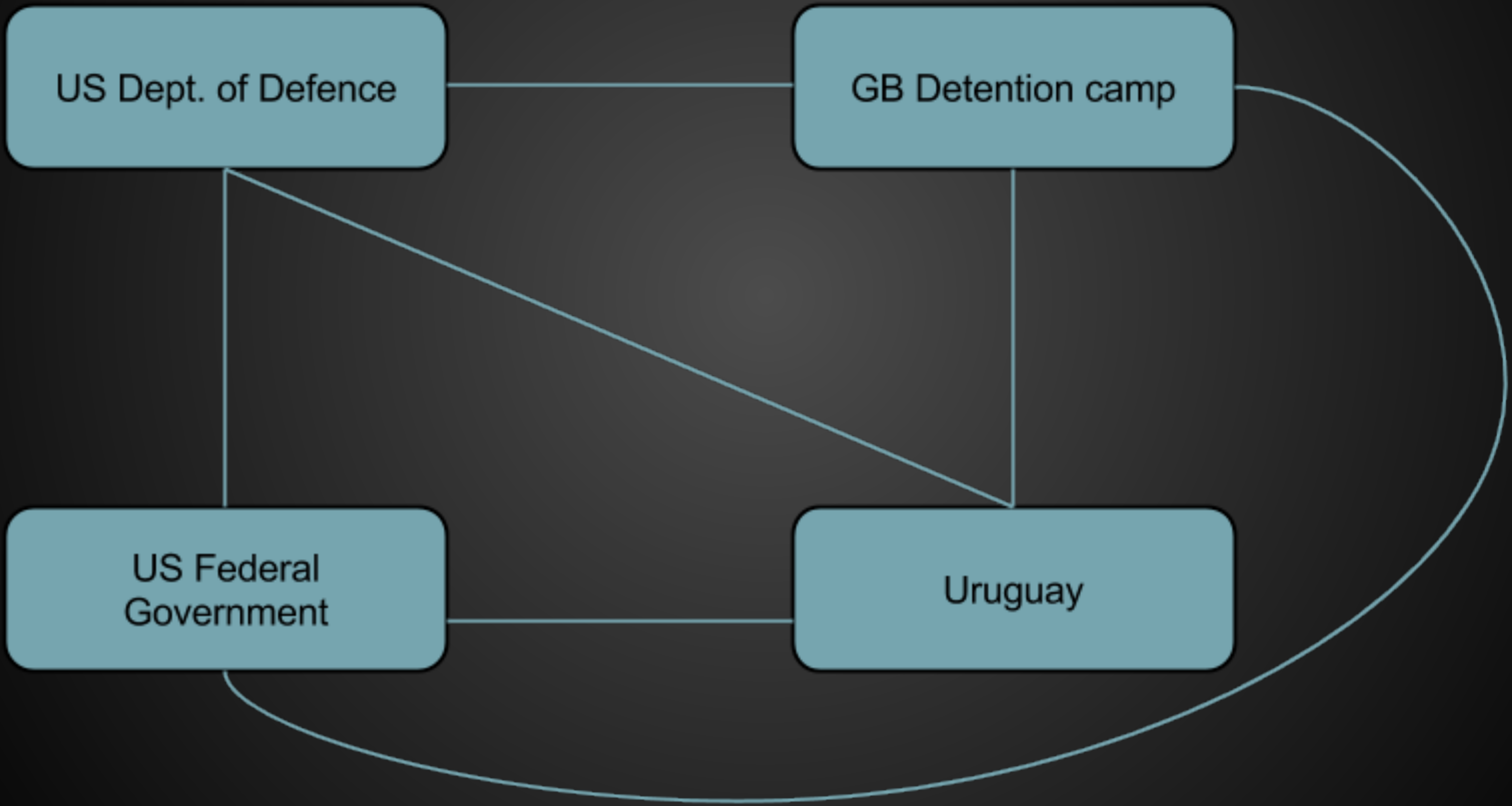
Module: Popularity = #ins / #outs

United States	Guantanamo Bay	Uruguay	Defence Department
~~817.71~~	**1.74**	**10.95**	**1.60**
6.69	0.89	~~2.73~~	0.26
4.31	0.46	0.48	
0.72	~~0.34~~	~~0.14~~	
0.29		~~1.82~~	
0.51		~~0.51~~	
0.32			
2.55			

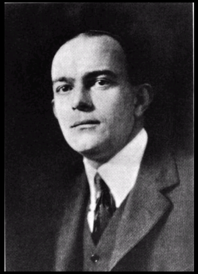
Optimization: Graph proximity

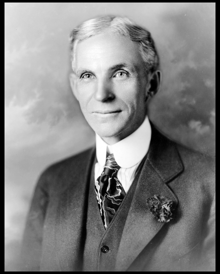


Optimization: Graph proximity



Counter - example





"With the help of C. Harold Wills, Ford designed, built, and successfully raced a 26-horsepower automobile in October 1901."

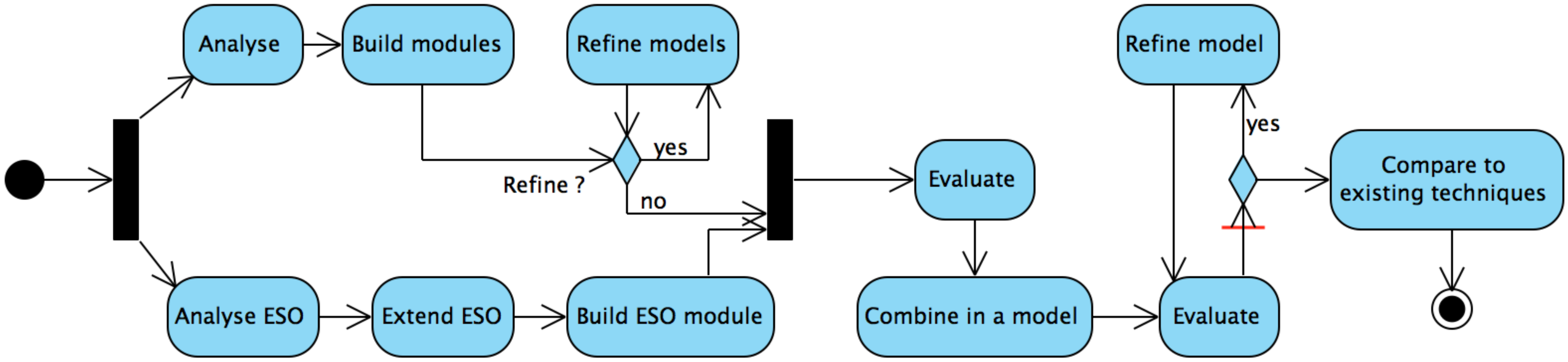
Using DBpedia

	(Henry Ford, C. Wills)	(Gerald Ford, C. Wills)
Shared property values	3	0
Graph distance	direct connection	no direct connection
Abstract search	keywords found	no keywords found
Class distance	0	0
Popularity	/	/
Frame model	/	/
DBPedia spotlight rank	> 10	5

Points to make

- No module is perfect
    - but >1 module will be helpful
- Popularity Bias
- Topic Bias
    - old movies, the western culture, entertainment, etc.
- Computational complexity (reduced implicitly)
- Incompleteness and maintenance of datasets
- The lexical resources are not ontological

Experimental setup



Data

- 120 news articles
  - 30 GM, Ford, Chrysler car data
  - 30 Apple corpus articles
  - 30 Boeing Airbus
  - 30 stock market
- Training data
  - 1.3 M car articles

++Awesome things I won’t do

Cross-sentence check

Build a model over more ontologies

FrameNet situations reasoning

Questions?

Appendices

Identity in language

“Any two entities are both similar and dissimilar with respect to an infinite number of properties.” (Murphy & Medin, 1985)

- Entity coreference becomes tricky
  - Temporality: Is *Old Amsterdam* identical to *New Amsterdam* ?
  - Pragmatism: Is *Lord Lipton* identical to *the wealthiest tea importer* ?
  - Granularization: Is *passengers* and *people* identical?
- Continuum of identity

Identity in NLP

- Traditional techniques interpret identify in a shallow way, based on:
    - popularity
    - TF-IDF score
    - bag-of-words similarity
- Contemporary techniques start looking into semantic coherence
    - pair-wise interpretation
    - collective interpretation

Problem statement

- Entity linking ambiguity in the NLP world
- Also present within the Semantic Web
- Lexical resources are not exploited enough

State-of-the-art:

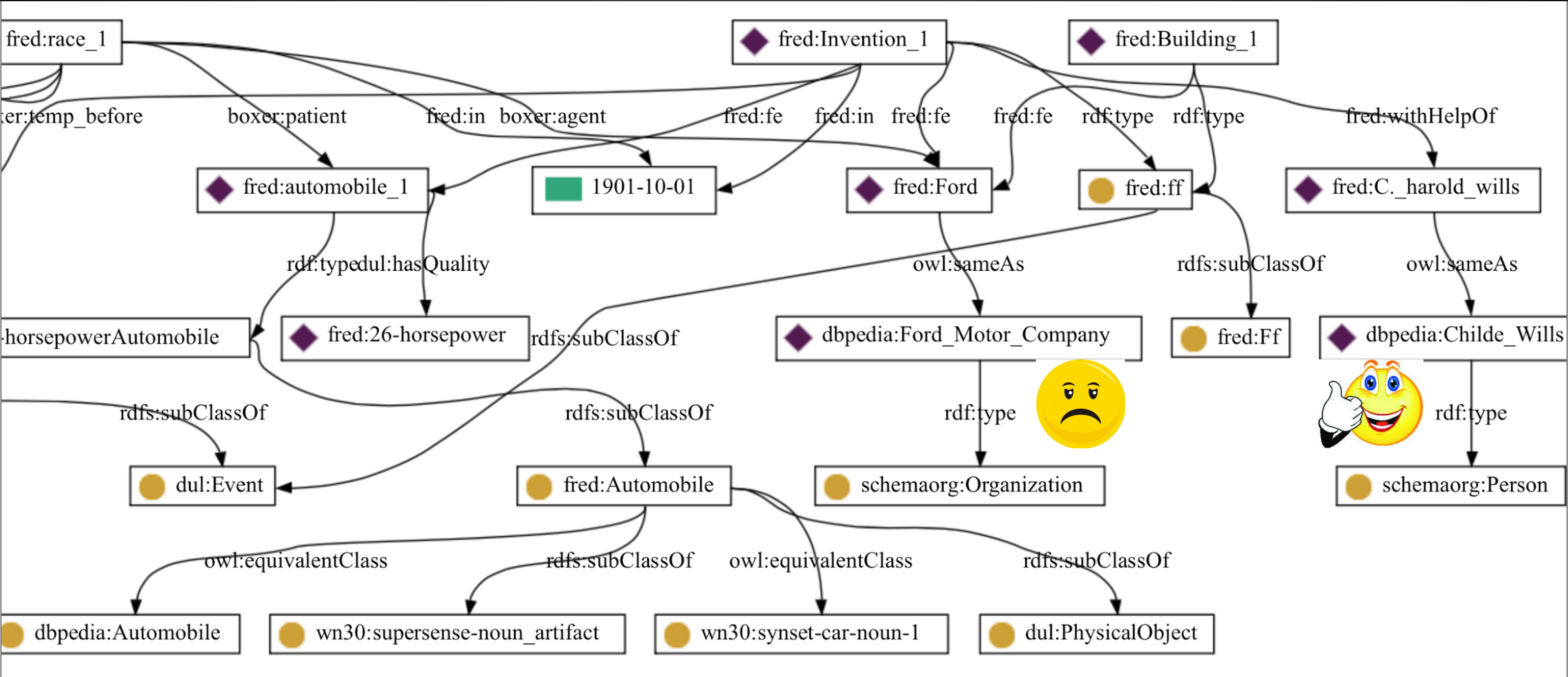
C. Harold Wills:


```
<externalReferences>
  <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Childe_Wills"
confidence="1.0" reftype="en" />
</externalReferences>
```


Ford:


```
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ford_Motor_Company"
confidence="0.99966425" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ford_of_Britain"
confidence="2.6139864E-4" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ford_Germany"
confidence="5.3583182E-5" reftype="en" />
    <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/Ford_Motor_Company_of_Australia" confidence="1.929829E-5"
reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Gerald_Ford"
confidence="1.4137138E-6" reftype="en" />
    <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/Ford_World_Rally_Team" confidence="8.657681E-8" reftype="en" />
    <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/List_of_Ford_vehicles" confidence="7.0210043E-10" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ford_of_Europe"
confidence="4.9917383E-11" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ford_Brasil"
confidence="9.321995E-13" reftype="en" />
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/Ford,_West_Sussex"
confidence="1.699912E-13" reftype="en" />
```

State-of-the-art: FRED



Tools: ConceptNet 5

- because common sense knowledge is lacking often in DBpedia (and alike) resources