Automated knowledge base construction

4. Taxonomy induction + entity disambiguation

Simon Razniewski

Summer term 2022

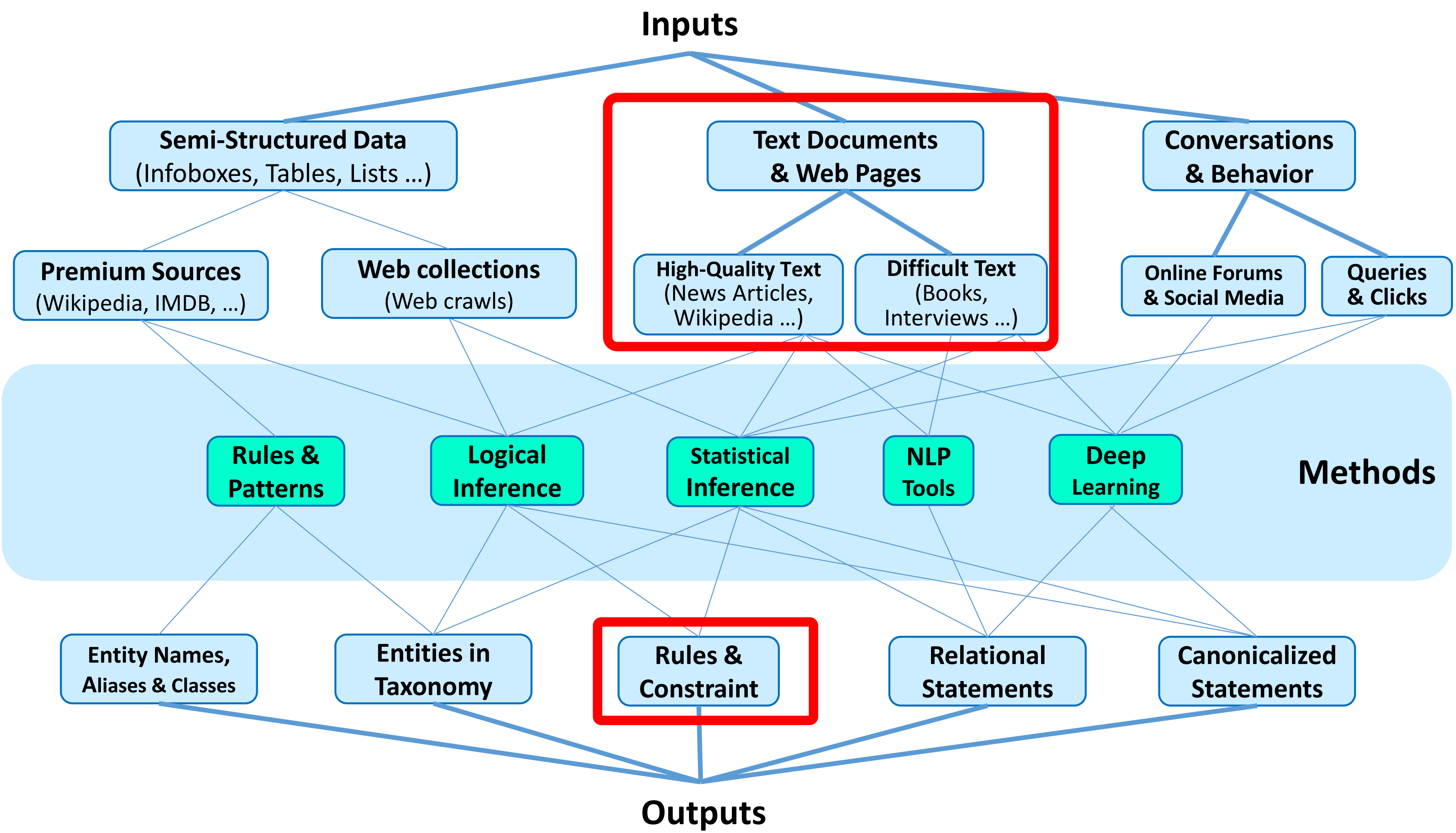
Outline

1. Taxonomy induction
2. Entity disambiguation

Recap: Entity types

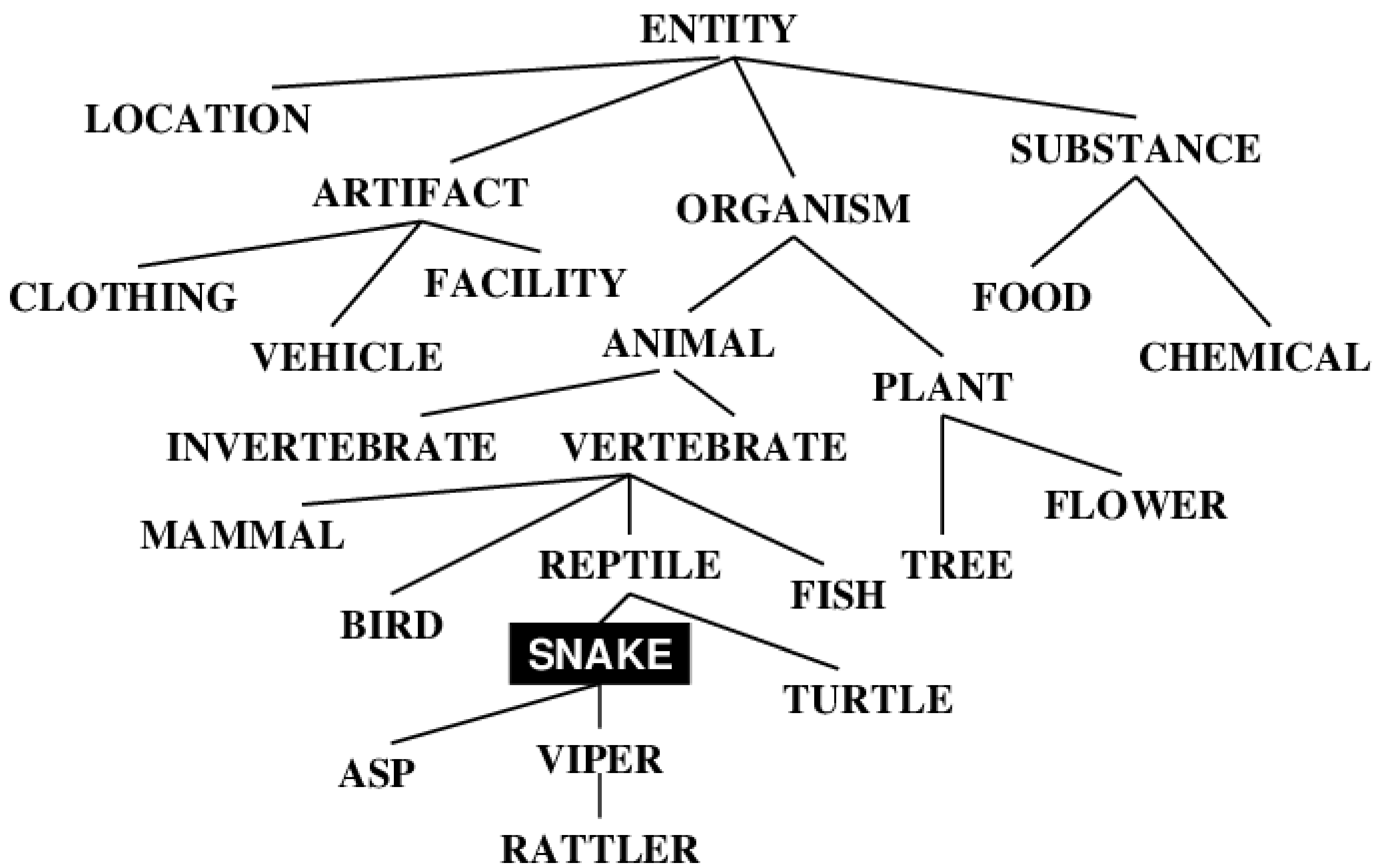
- Einstein: Physicist, Nobel prize winner, ETH alumni
- Dudweiler: Village, municipality
- RCH_2OH: chemical formula, psychoactive substance

- Why organize them?
  - Observations are usually sparse
  - Upper classes may be needed for queries:
    - German locations ending in –weiler
    - Scientists born in 1879
  - Class relations needed for constraint checking
    - graduatedFrom(Person, educationalInstitution)
    - UdS -> University -> OK?



Inputs
Semi-Structured Data
(Infoboxes, Tables, Lists …)
Text Documents
& Web Pages
Conversations
& Behavior
Premium Sources
(Wikipedia, IMDB, …)
Web collections
(Web crawls)
High-Quality Text
(News Articles,
Wikipedia …)
Difficult Text
(Books,
Interviews …)
Online Forums
& Social Media
Queries
& Clicks
Rules &
Patterns
Logical
Inference
Statistical
Inference
NLP
Tools
Deep
Learning
Methods
Entity Names,
Aliases & Classes
Entities in
Taxonomy
Rules &
Constraint
Relational
Statements
Canonicalized
Statements
Outputs

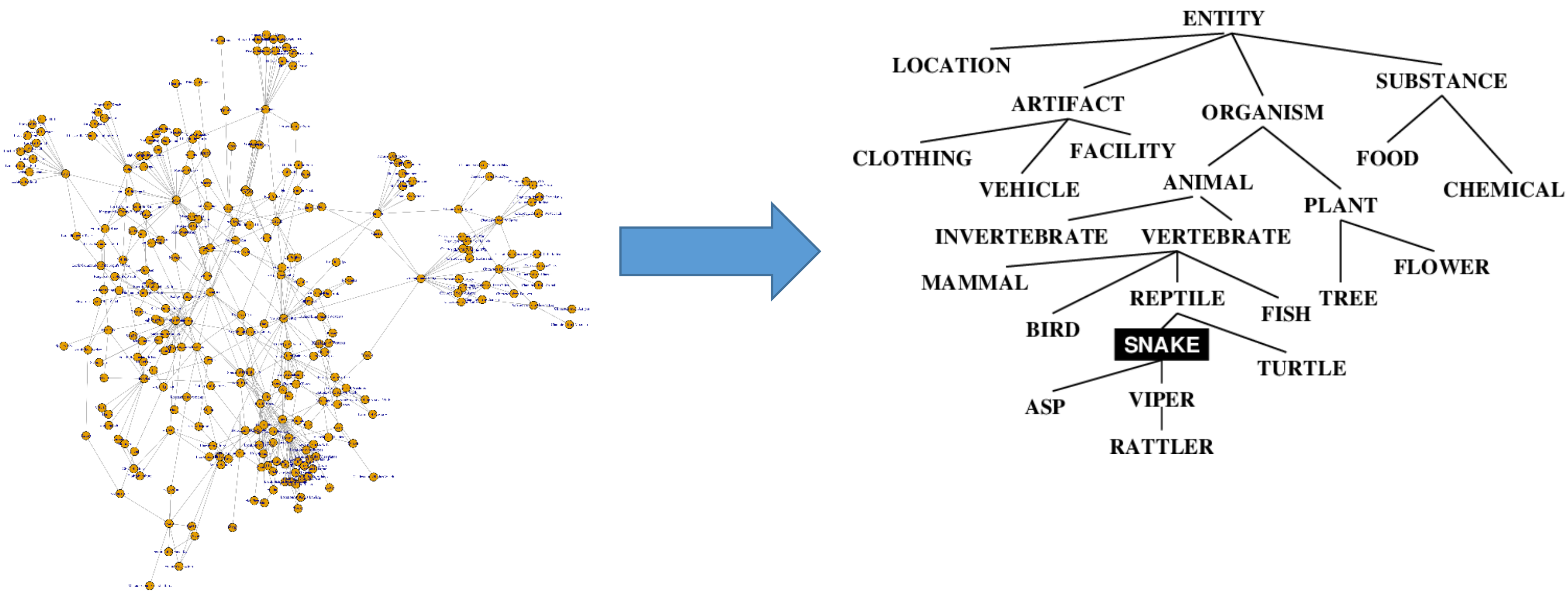
Taxonomy induction: Goal



Taxonomy induction: General approach

- Hypernymy candidates are “cheap”

→ Start with large noisy candidate graph, clean it up



Candidates #1: Hearst-patterns

- Hearst-style patterns (below: WebIsALOD for Frodo)

hyponymLabel	confidence
"hero"	0.597244
"hobbit"	0.479114
"member of the fellowship"	0.472321
"character"	0.456166
"playable character"	0.426721
"character in the lord"	0.346989
"character from the lord"	0.339778
"fellowship of the ring"	0.330798
"thing"	0.282846
"ordinary man"	0.266521
"mortal"	0.265587
"lord of the ring"	0.25944
"dog"	0.215679
"people"	0.214287

hyponymLabel	confidence
"tv show"	0.730957
"event"	0.670605
"series"	0.64273
"popular show"	0.609206
"character in the game"	0.586694
"hit tv show"	0.583963
"david bowie album"	0.578075

hyponymLabel	confidence
"creature"	0.70834
"blockbuster film"	0.611883
"thing"	0.58897
"film"	0.576852
"mythical creature"	0.560562
"anticipate film"	0.55724

hyponymLabel	confidence
"film"	0.678143
"monster"	0.622037
"horror"	0.57432
"person"	0.563758
"member"	0.547969
"word"	0.526026

Candidates #2: Sub-category relations in Wiki systems

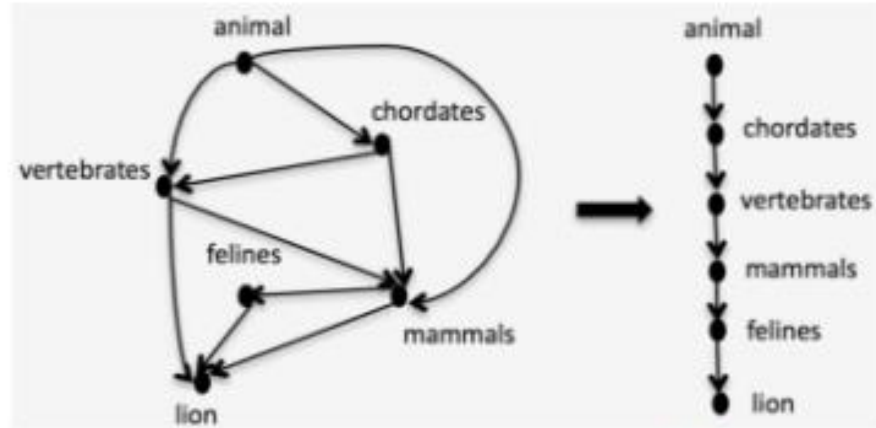
Categories: Featured articles | Characters | Cleanup | Hobbits | Baggins | Ring bearers | Elf friends | Fellowship members | Major characters (The Lord of the Rings) | The Lord of the Rings Characters | Characters that have appeared in the Hobbit and the Lord of the Rings | The Hobbit: An Unexpected Journey Characters | Bearers of the One Ring | The Lord of the Rings: The Fellowship of the Ring (film) Characters | The Lord of the Rings: The Two Towers (film) Characters | The Lord of the Rings: The Return of the King (film) Characters

Categories: The Lord of the Rings characters | Middle-earth Hobbits | Adventure film characters | Fictional orphans | Bearers of the One Ring | Fictional characters who can turn invisible | Fictional characters introduced in 1954 | Fictional swordsmen | Fictional amputees | Fictional writers

Categories: Middle-earth characters | Middle-earth Men
Hidden categories: Commons category link is on Wikidata

Categories: Swordsmen | Fictional melee weapons practitioners
Hidden categories: Categories requiring diffusion

Challenges



Zornitsa Kozareva and Eduard H. Hovy:
"A semi-supervised method to learn and construct taxonomies using the web"
EMNLP 2010

- Noise
  - Meta-categories
  - Ambiguous terms
- Structural oddities
  - Cycles
  - Upward branching
  - Redundancy (transitive edges)
- Imbalance in observations and scoring
  - Score-based thresholding discards entire regions

Text-based taxonomy induction challenge [Semeval 2016, Bordea et al.]

- Input: Set of domain terms
  - Tofu, pizza, garlic
  - Computer, smartphone, printer
- Task: Induce a taxonomy over these terms
- Potential evaluation measures
  - #nodes
  - #edges
  - Acyclicity
  - Recall w.r.t. gold standard
  - Precision w.r.t. gold standard
  - Connectedness (#connected components / #c.c)
  - Categorization (#intermediate nodes / #i.i)

Taxi [Panchenko et al., 2016]

1. Crawl domain-specific text corpora in addition to WP, Commoncrawl
2. Candidate hypernymy extraction
   1. Via substrings
      - “biomedical science” isA “science”
      - “microbiology” isA “biology”
      - “toast with bacon” isA “toast”
      - Lemmatization, simple modifier processing
      - Scoring proportional to relative overlap
   2. Candidate hypernymy from 4 Hearst-Pattern extraction works
3. Supervised pruning
   1. Positive examples: gold data
   2. Negative examples: inverted hypernyms + siblings
   3. Features: Substring overlap, Hearst confidence (more features did not help)

Taxi [Panchenko et al., 2016]

4. Taxonomy induction
   - Break cycles by random edge removal
   - Fix disconnected components by attaching each node with zero outdegree to root

	Monolingual (EN)			Multilingual (NL, FR, IT)		
Measure	**Baseline**	**BestComp**	**TAXI**	**Baseline**	**BestComp**	**TAXI**
Cyclicity	*0*	**0**	**0**	*0*	**0**	**0**
Structure (F&M)	*0.005*	**0.406**	0.291	*0.009*	0.016	**0.189**
Categorisation (i.i.)	*77.67*	**377.00**	104.50	*64.28*	**178.22**	64.94
Connectivity (c.c.)	*36.83*	44.75	**1 .00**	*40.50*	34.89	**1.00**
Gold standard comparison (Fscore)	*0.330*	0.260	**0.320**	*0.009*	0.016	**0.189**
Manual Evaluation (Precision)	*n.a.*	**0.490**	0.200	*n.a.*	0.298	**0.625**

- too many hypernyms in English

Taxonomy induction using hypernym subsequences [Gupta et al., 2017]

- Looking at edges in isolation ignores important interactions
  - Hypernym candidates typically contain higher-level terms that help in predicting whole sequence
  - Crucial as abstract term hypernym extraction empirically harder (e.g., "company" → "group of friends"?)

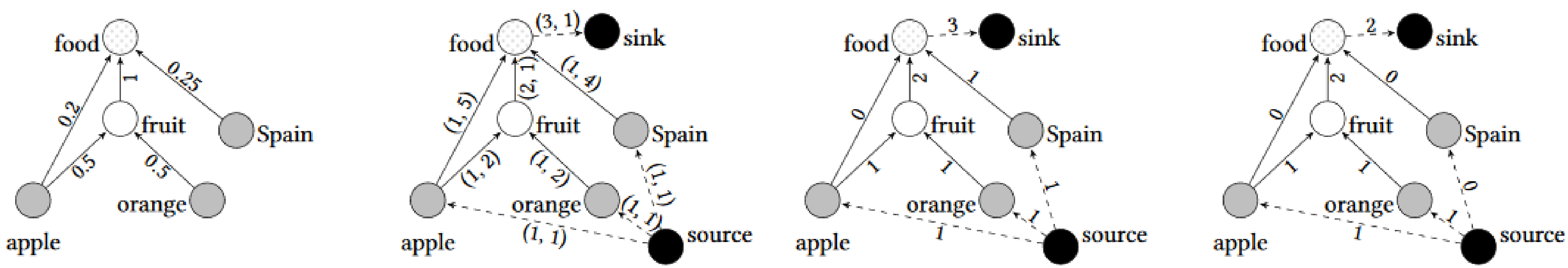
Candidate hypernym	Frequency
company	5536
fruit	3898
apple	2119
vegetable	928
orange	797
tech company	619
brand	463
hardware company	460
technology company	427
food	370

Candidate hypernyms for the term *apple*.

Taxonomy induction using hypernym subsequences [Gupta et al., 2017]

- Joint probabilistic model that estimates true hypernymy relations from skewed observations
- Break cycles by removing edges with minimal weight
- Induce tree from DAG by a min-cost-flow model

Taxonomy induction using hypernym subsequences [Gupta et al., 2017]



(a): Noisy hypernym graph (H).

(b): Flow network F with (capacity, cost) values for each edge.

(c): Flow values (f) for each edge found using demand $d = 3$.

(d): Flow values (f) for each edge found using demand $d = 2$.

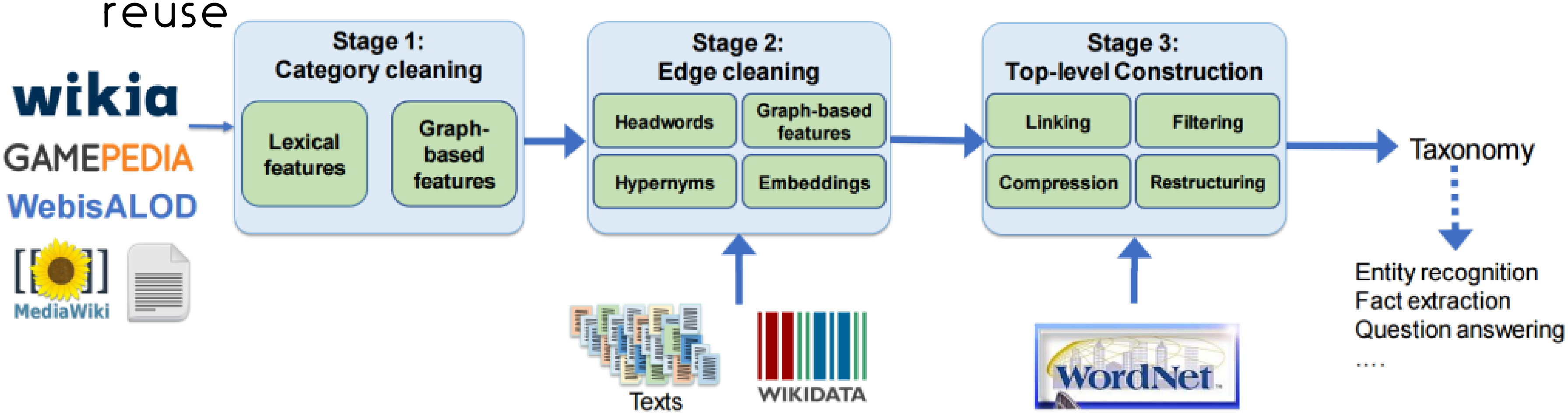
- Method: Find cheapest way to send flow from leaves to root
- Cost inversely proportional to edge weight

Wiki[pedia | a]- based taxonomy induction: TiFi [Chu et al., WWW 2019]

Observations:

- Wikia category systems are noisy
- Wikia category systems lack abstractions

Approach: Supervised filtering + WordNet reuse



TiFi: Category cleaning

- Challenge:
  - Meta-categories (Meta, Administration, Article_Templates)
  - Contextual categories (actors, awards, inspirations)
  - Instances (Arda, Mordor)
  - Extensions (Fan fiction)
- Approach: Supervised classification
  - "Featurizes" earlier rule-based category cleaning works, e.g., Marius Pasca at Google
- Features:
  - Lexical
    - Meta string dictionary (manual)
    - Headword in plural? Dark Orcs, Ring of Power
    - Capitalization? Quenya words, Ring bearers
  - Graph-based
    - #instances
    - Supercategory/subcategory count
    - Average depth
    - Connected subgraph size

Categories: Featured articles | Characters | Quenya words | Villains | Ring bearers | Major characters (The Lord of the Rings) | Servants of Morgoth | Characters in Unfinished Tales | Characters in The History of Middle-earth | The Hobbit: The Battle of the Five Armies Characters | The Hobbit: An Unexpected Journey Characters | The Hobbit: The Desolation of Smaug Characters | The Lord of the Rings: The Fellowship of the Ring (film) Characters | The Lord of the Rings: The Two Towers (film) Characters | The Lord of the Rings: The Return of the King (film) Characters | The Silmarillion Characters | Bearers of the One Ring

TiFi: Category cleaning - results

Universe	# Categories	# Edges
Lord of the Rings (LoTR)	973	1118
Game of Thrones (GoT)	672	1027
Star Wars	11012	14092
Simpsons	2275	4027
World of Warcraft	8249	11403
Greek Mythology	601	411

Table 1: Input categories from Wikia/Gamepedia.

Method	Universe	Precision	Recall	F1-score
Pasca 2018 [34]	LoTR	0.33	0.75	0.46
	GoT	0.57	0.85	0.68
Ponzetto & Strube 2011 [38]	LoTR	0.44	**1.0**	0.61
	GoT	0.45	**1.0**	0.62
Pasca + Ponzetto & Strube	LoTR	0.41	0.75	0.53
	GoT	0.64	0.85	0.73
TiFi	LoTR	**0.84**	0.82	**0.83**
	GoT	**0.85**	0.85	**0.85**

Table 2: Step 1 - In-domain category cleaning.

- Most important feature: Plural
  - Occasional errors (Food)

TiFi: Edge cleaning

- Challenge:
  - Type mismatches
    - Frodo → The Shire
    - Boromir → Death in Battle
    - Chieftains of the Dúnedain → Dúnedain of the North
- Approach: Supervised classification
  - Combination of lexical, semantic and graph-based features

TiFi: Edge cleaning - features

- Lexical
  - Head word generalization (c $\text{subclass}Of$ d?)
    - $head(c) + post(c) = head(d) + post(d)$ and $pre(d)\ in\ pre(c)$ — *Dwarven Realms → Realms*
    - $pre(c) + head(c) = pre(d) + head(d)$ and $post(d)\ in\ post(c)$ — *Elves of Gondolin → Elves*
  - Only plural parents?
- Semantic
  - WordNet hypernym relations
  - Wikidata hypernym relations
  - Text matches
    - Wikia first sentence Hearst
      - Haradrim: The Haradrim, known in Westron as the Southrons, were a race of Men from Harad in the region of Middle-earth.
    - WordNet synset headword
      - Ex: Werewolves: a monster able to change appearance from human to wolf and back again
  - Distributional similarity
    - WordNet graph distance (Wu-Palmer score)
    - Diretional embedding scores (HyperVec – directional interpretation of embeddings)
      - Distributional inclusion hypothesis: flap is more similar to bird than to animal
      - Hypernyms occur in more general contexts
- Graph-based
  - #common children
  - Parent.#children/parent.avg-depth

TiFi - WordNet synset headword

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: castle Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) palace, **castle** (a large and stately mansion)
- S: (n) **castle** (a large building formerly occupied by a ruler and fortified against attack)
- S: (n) **castle**, rook ((chess) the piece that can move any number of unoccupied squares in a direction parallel to the sides of the chessboard)
- S: (n) **castle**, castling (interchanging the positions of the king and a rook)

TiFi – WordNet synset linking

Algorithm 1: WordNet Synset Linking

Data: A category c
Result: WordNet synset s of c
c = *pre* + *head* + *pos*, l = null;
l = list of WordNet synset candidate for c;
if $l = null$ **then**
  l = list of WordNet synset candidates for *pre* + *head*;
  if $l = null$ **then**
    l = list of WordNet synset candidates for *head*;
if $l = null$ **then**
  return null;
max = 0, s = null;
for *all WordNet synset* s_i *in* l **do**
  $sim(s_i, c) = cosine(V_{s_i}, V_c)$ with V: context vector;
  $sim(s_i, c = sim(s_i, c) + 1/(2R_{s_i})$ where R: rank in WordNet;
  if $sim(s_i, c) > max$ **then**
    max = $sim(s_i, c)$;
    $s = s_i$;
return s;

TiFi: Edge cleaning - results

Method	Universe	Precision	Recall	F1-score
HyperVec [31]	LoTR	0.82	0.8	0.81
	GoT	**0.83**	0.81	0.82
HEAD [16]	LoTR	**0.85**	0.83	0.84
	GoT	0.81	0.78	0.79
TiFi	LoTR	0.83	**0.98**	**0.90**
	GoT	**0.83**	**0.91**	**0.87**

Table 4: Step 2 - In-domain edge cleaning.

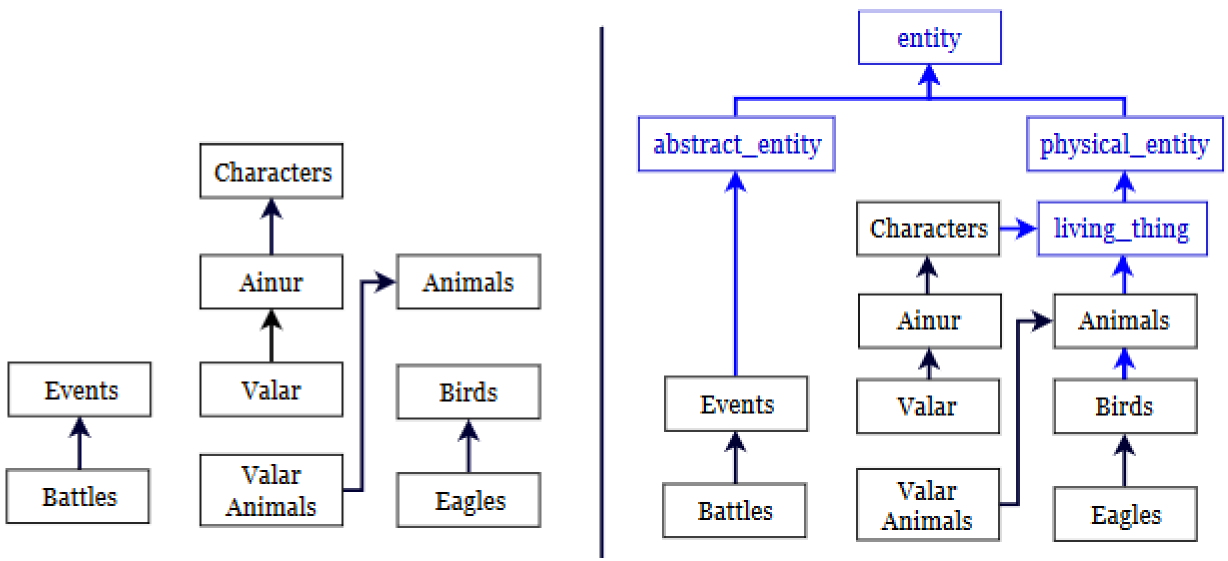
← Embedding only

← Rules only

- Most important features:
  - Only plural parent
  - Lexical generalization
  - Common child support
  - Page type matching

TiFi: Top-level construction

- Problem: Wikia categories represent many disconnected components
- Solution: Link sinks to WordNet taxonomy and import further top level



TiFi – Top-level construction

- Using same algorithm as for linking in edge cleaning
  - Birds is mapped to bird%1:05:00::
    Subsequent hypernyms: wn_vertebrate → wn_chordate → wn_animal → wn_organism → wn_living_thing → wn_whole → wn_object → wn_physical_entity → wn_entity
  - Removal of long paths (nodes with only one child and one parent)
  - Dictionary-based filtering of ~100 too abstract classes (whole, sphere, imagination, …)

TiFi: Top-level construction - results

Universe	#New Types	#New Edges	Precision
LoTR	43	171	0.84
GoT	39	179	0.84
Starwars	373	3387	0.84
Simpsons	115	439	0.92
World of Warcraft	257	2248	0.84
Greek Mythology	22	76	0.84

Table 7: Step 3 - WordNet integration.

TiFi – Relevance for entity search

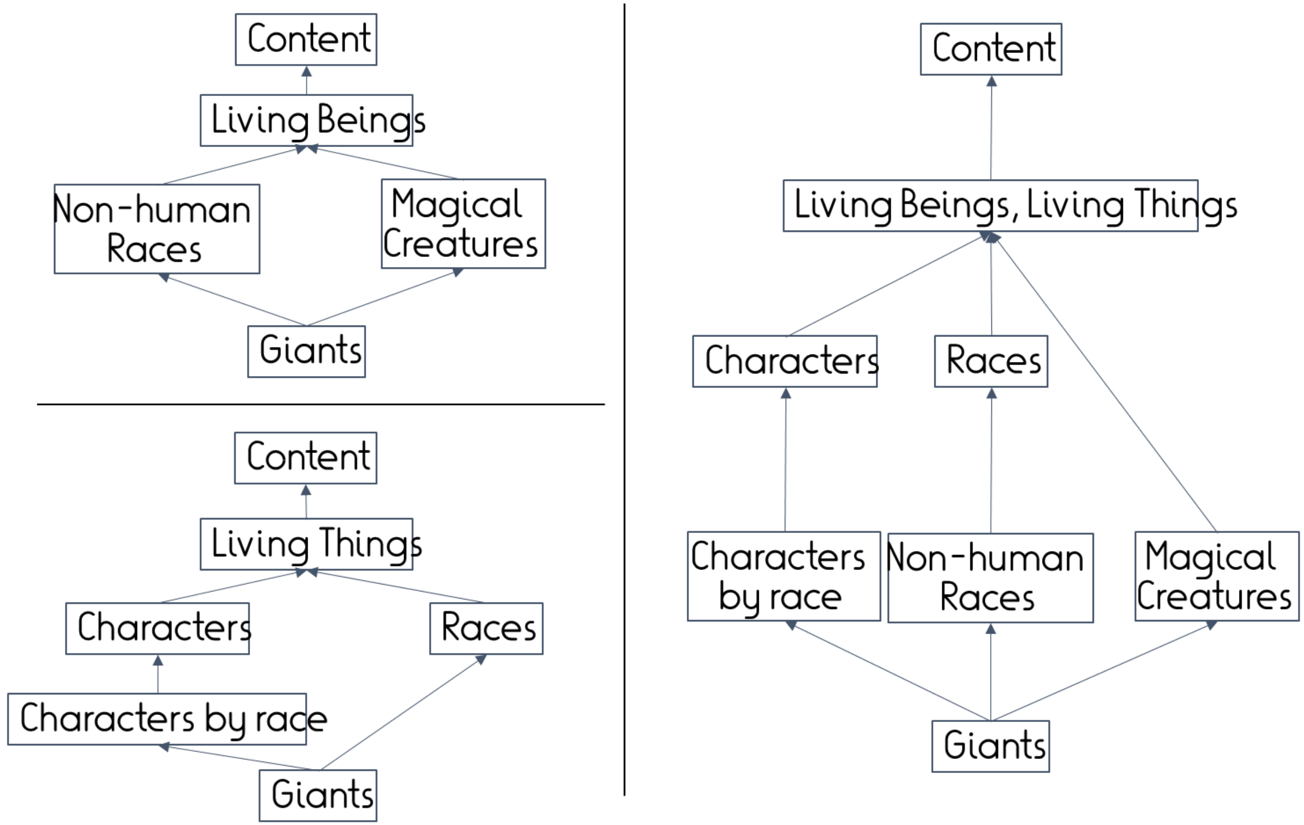
	Text		Structured Sources	
Query	Google	Wikia	Wikia-categories	TiFi
Dragons in LOTR	Glaurung, Túrin, Turambar, Eärendil, Smaug, Ancalagon	Dragons, ~~Summoned Dragon~~, Spark-dragons	~~Urgost~~,Long-worms,Gostir,~~Drogoth the Dragon Lord~~,~~Cave-Drake~~, ~~War of the Dwarves and Dragons~~, ~~Dragon-spell~~,Stone Dragons, Fire-drake of Gondolin,Spark-dragons, Were-worms, ~~Summoned Dragon~~, Fire-drakes, Glaurung,Ancalagon,Dragons,Cold-drakes, Sea-serpents, ~~User blog:Alex Lioce/Kaltdrache the Dragon~~, Smaug, ~~Dragon (Games~~, ~~Workshop)~~, ~~Drake~~, Scatha, ~~The Fall of Erebor~~	Long-worms, ~~War of the Dwarves and Dragons~~, ~~Dragon-spell~~,Stone Dragons, Fire-drake of Gondolin, Spark-dragons, Were-worms, Fire-drakes, Glaurung, Ancalagon, Dragons, Cold-drakes, Sea-serpents, Smaug, Scatha ,~~The Fall of Erebor~~, Gostir
Which Black Numenoreans are servants of Morgoth	-	Black Númenórean	Men of Carn Dûm,Corsairs of Umbar,Witch-king of Angmar, ~~Thrall Master~~,Mouth of Sauron,Black Númenórean,Fuinur	Men of Carn Dûm,Corsairs of Umbar,Witch-king of Angmar, Mouth of Sauron, Black Númenórean, Fuinur
Which spiders are not agents of Saruman?	-	-	Shelob, ~~Spider Queen and Swarm~~,~~Saenathra~~, ~~Spiderling~~, Great Spiders, ~~Wicked, Wild, and Wrath~~	Shelob, Great Spiders

Table 12. Example queries and results for the entity search evaluation.

	Text		Structured Sources	
Query	**Google**	**Wikia**	**Wikia-categories**	**TiFi**
t	2 (52%)	7 (65%)	10 (62%)	8 (87%)
$t_1 \cap t_2$	1 (23%)	2 (11%)	8 (40%)	3 (70%)
$t_1 \setminus t_2$	1 (20%)	4 (36%)	8 (63%)	6 (79%)
Average	1 (32%)	4 (37%)	9 (55%)	$\boxed{6 (79\%)}$

Table 11: Avg. #Answers and precision of entity search.

Open: Taxonomy Merging



~Complex alignment problem requiring joint optimization

Summary: Taxonomy induction

- Usually a filtering process on larger candidate set
- Structure matters for local decisions
- Local-only decision OK but not optimal
- Top-level situation
  - Sparse observations
  - Generality makes reuse easier
- Relevance for AKBC:
  - Queries for type conditions not explicitly observed
  - Constraints on relation arguments

Outline

1. Taxonomy induction
2. Entity disambiguation

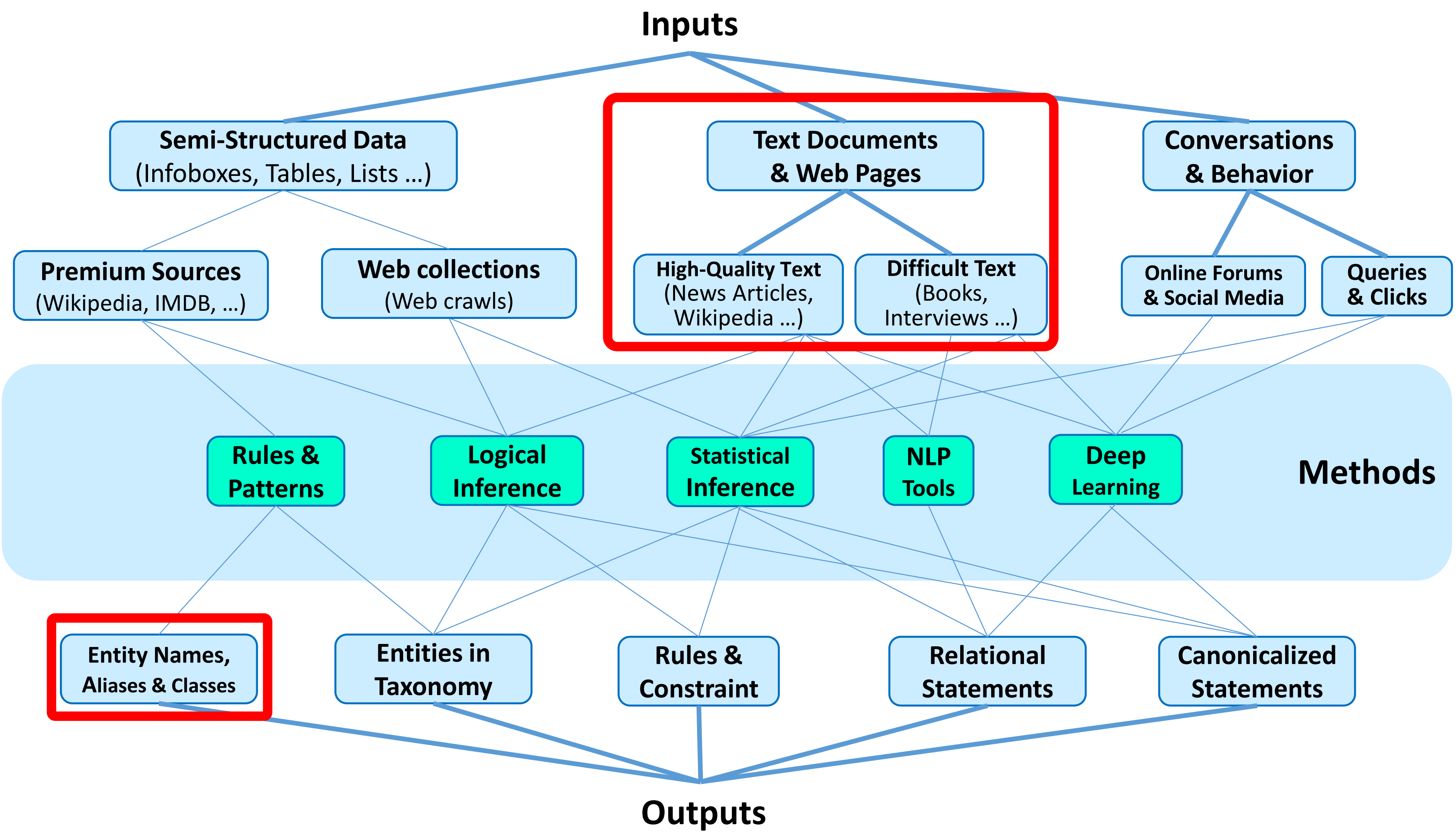
Ready for fact extraction?

Homer is the main character of the TV series "Simpsons".

Homer is the author of the Odyssey.

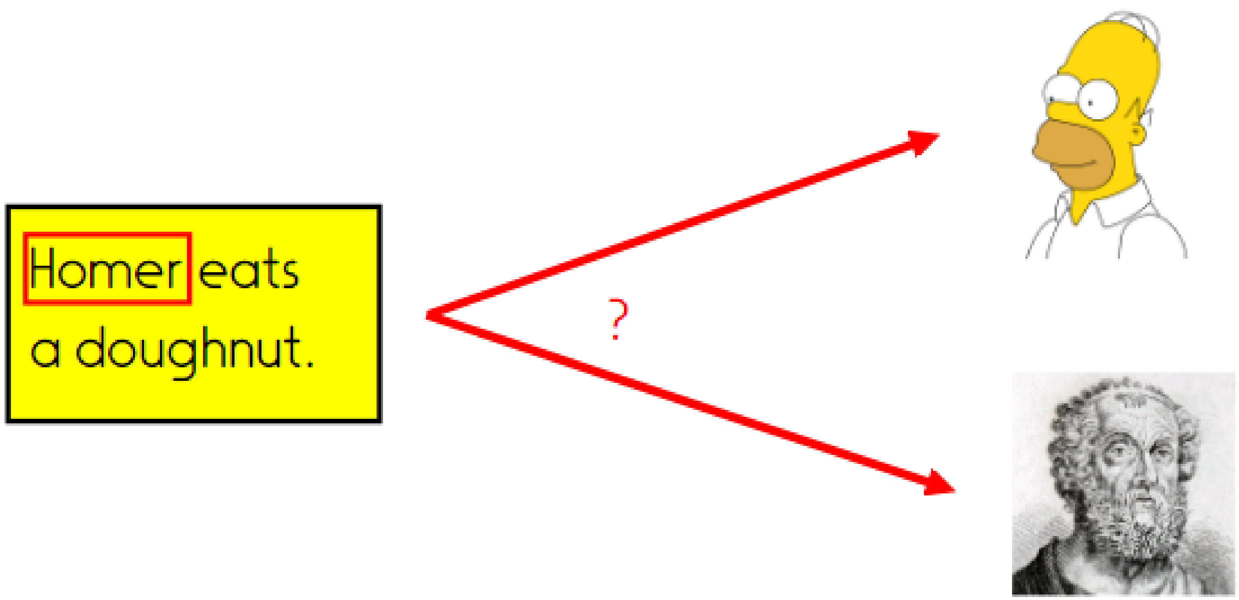
appearsIn(Homer, Simpsons)

wrote(Homer, Odyssey)?


Inputs
Semi-Structured Data
(Infoboxes, Tables, Lists …)
Text Documents
& Web Pages
Conversations
& Behavior
Premium Sources
(Wikipedia, IMDB, …)
Web collections
(Web crawls)
High-Quality Text
(News Articles,
Wikipedia …)
Difficult Text
(Books,
Interviews …)
Online Forums
& Social Media
Queries
& Clicks
Rules &
Patterns
Logical
Inference
Statistical
Inference
NLP
Tools
Deep
Learning
Methods
Entity Names,
Aliases & Classes
Entities in
Taxonomy
Rules &
Constraint
Relational
Statements
Canonicalized
Statements
Outputs

Def: Disambiguation

Given an ambiguous name in a corpus and its meanings, disambiguation is the task of determining the intended meaning.



Disambiguati

Usually Named Entity Recogn

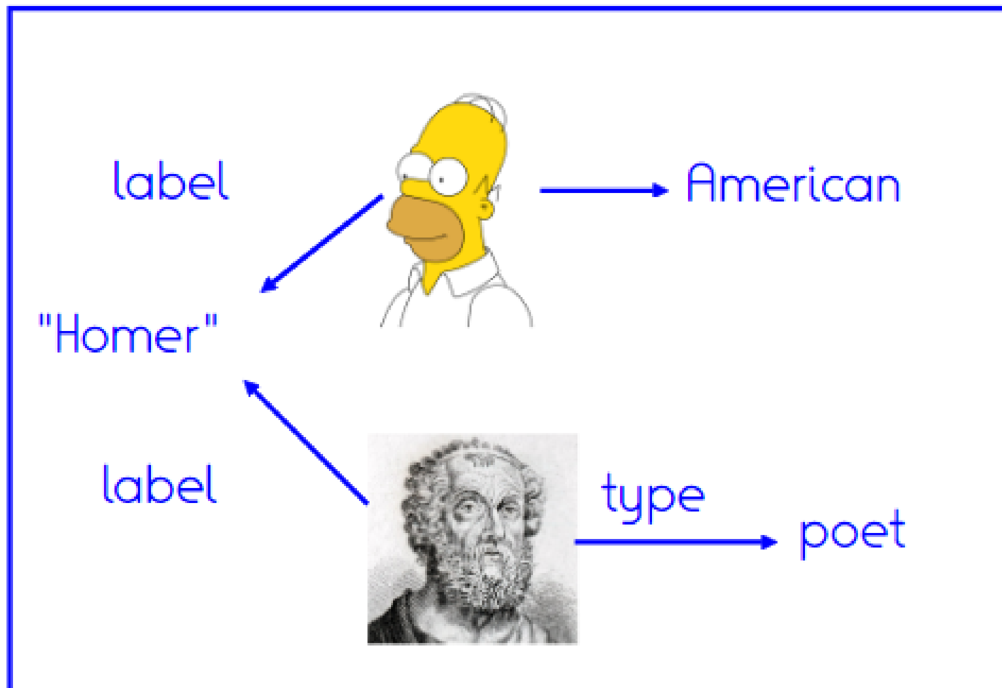
is to map the names to entities i

Also called "Wikification", because everyone links to Wiki[pedia | data]

Knowledge Base

NER'ed corpus

Homer eats a doughnut.



Def: Context of a word

The context of a word in a corpus is the multi-set of the words in its vicinity without the stopwords.

(The definition may vary depending on the application)

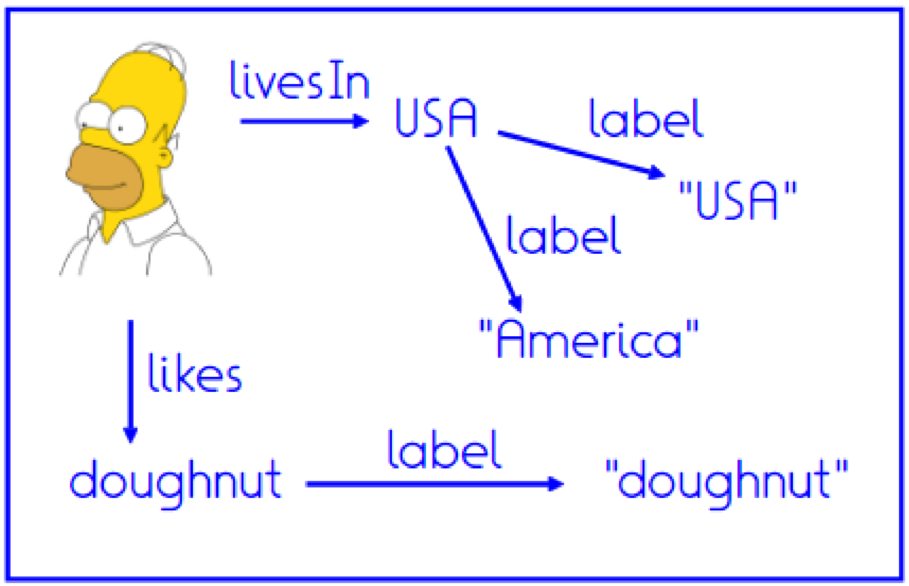
$\boxed{\text{Homer}}$ eats a doughnut.

Context of "Homer":
{eats, doughnut}

Def: Context of an entity

The context of an entity in a KB is the set of all labels of all entities in its vicinity.

(The definition may vary depending on the application)



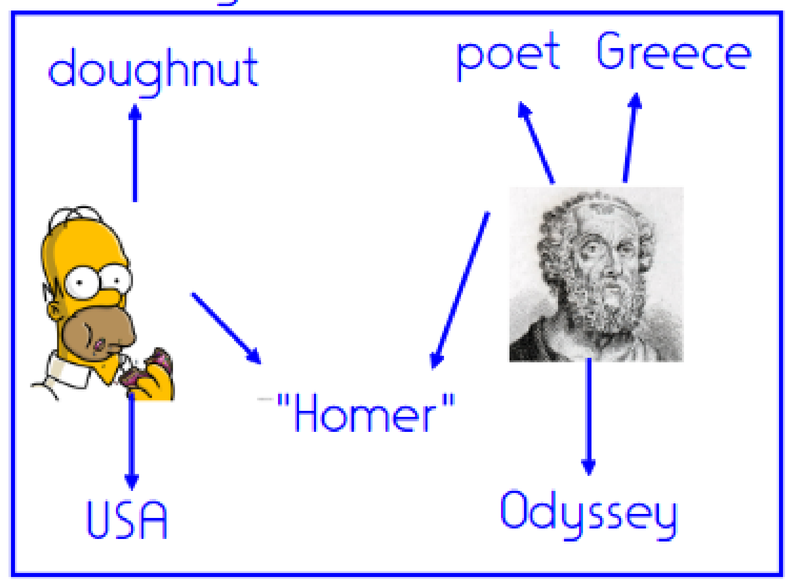
Context of Homer:
{doughnut, USA,
America}

Def: Context-based disambiguation

Context-based disambiguation (also: bag of words disambiguation) maps a name in a corpus to the entity in the KB whose context has the highest overlap to the context of the name.

For USA Today, Homer is among the top 25 most influential people of the past 25 years.

Knowledge Base

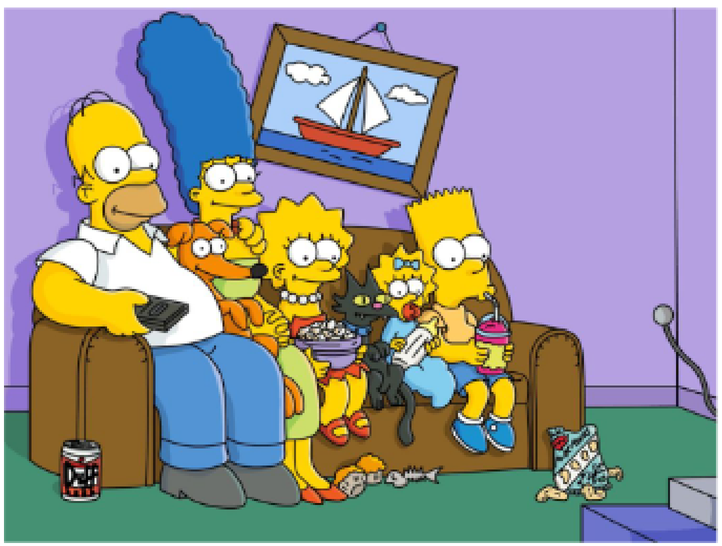


Who wins?

What if there is little context?

This is very important for the Simpsons.

?

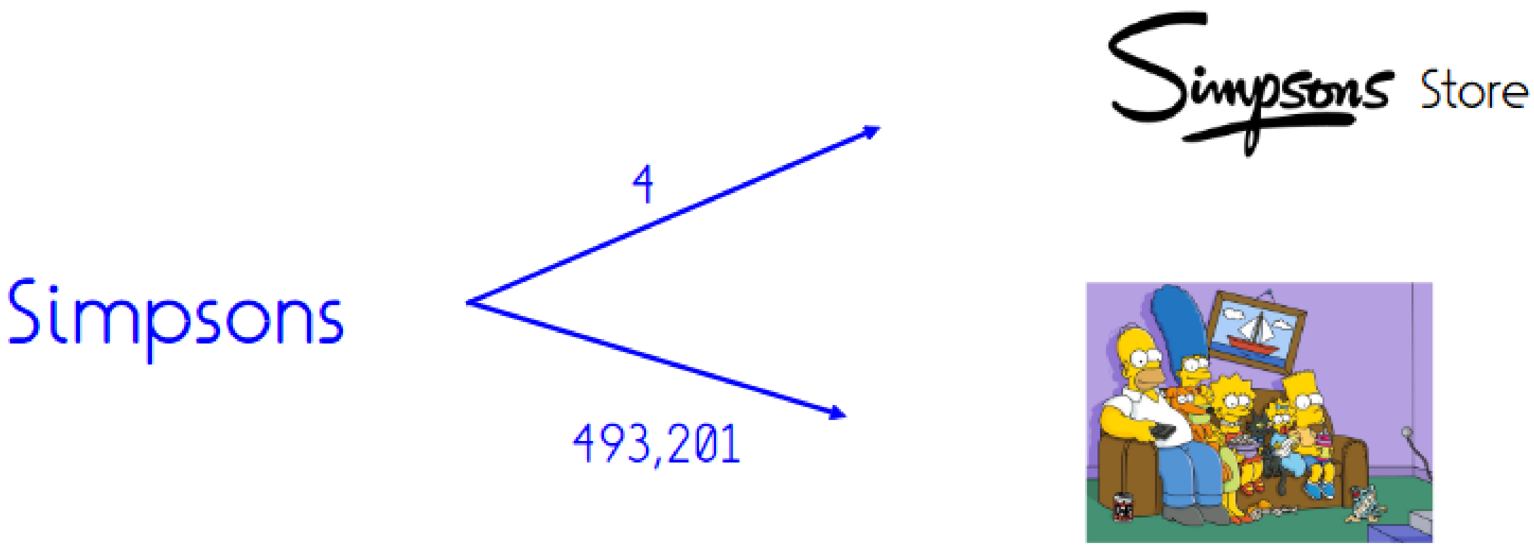


Simpsons

The Robert Simpson
Department Store.
Defunct since 1990.

Def: Disambiguation Prior

A disambiguation prior is a mapping from names to their meanings, weighted by the number of times that the name refers to the meaning in a reference corpus.



Can be computed e.g. from Wiki[pedia | a]
by link disambiguation or page views

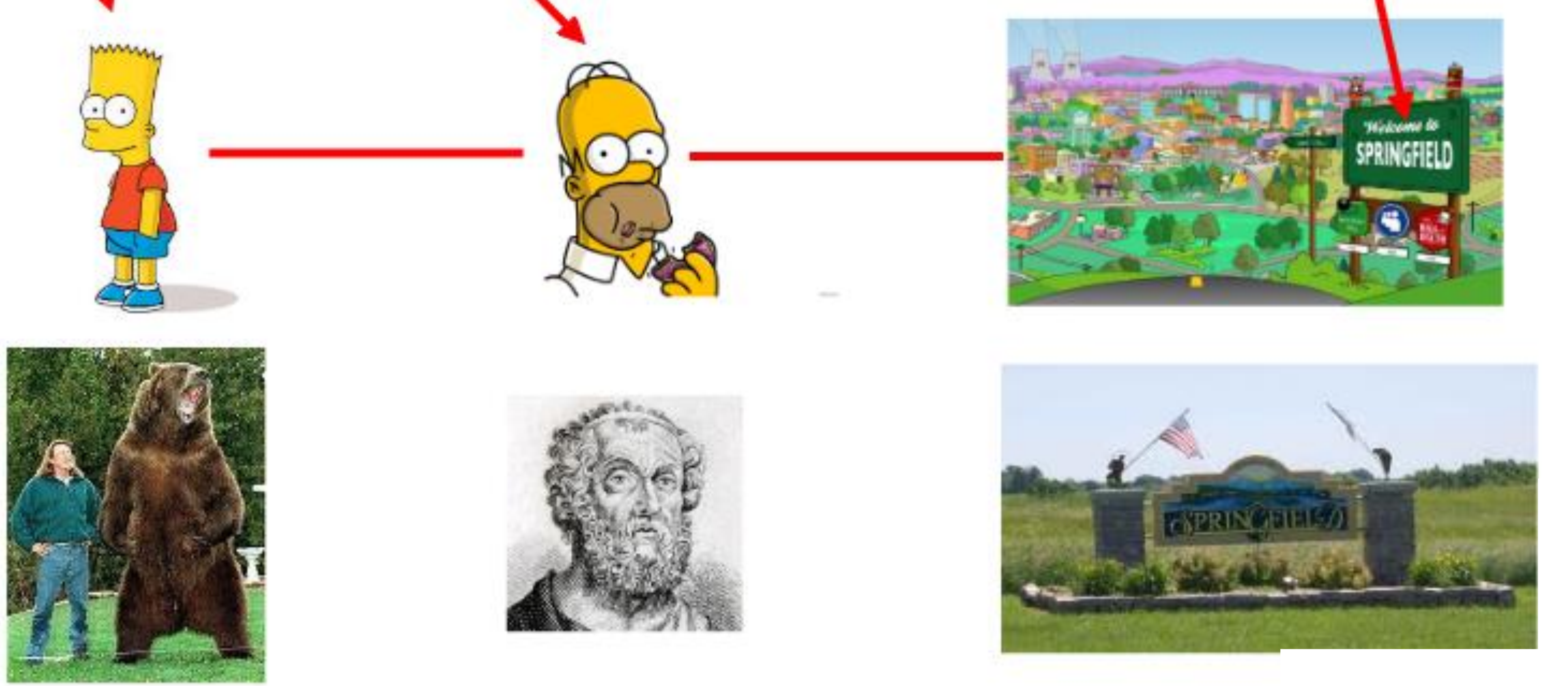
Local or global solution?

- Features so far local (one entity mention at a time)
  - Context-similarity
  - Disambiguation prior
- Do disambiguations influence each other?

Def: Coherence Criterion

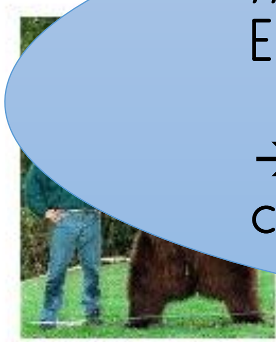
The Coherence Criterion postulates that entities that are mentioned in one document should be related in the KB.

Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.



Possible implementation (2)

Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.

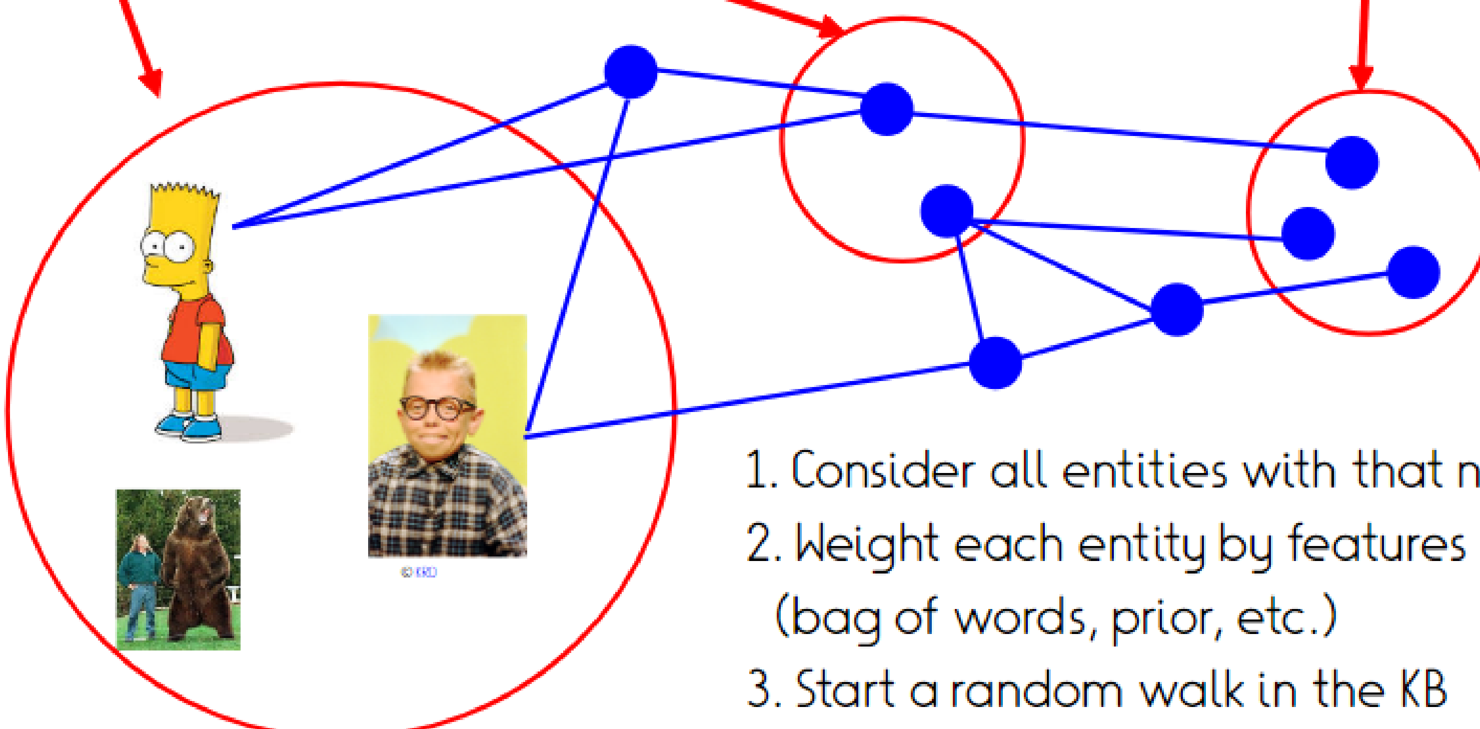


n entity mentions
Each with *m* candidate KB entities

→ Compute coherence scores for m^n combinations

Possible implementation (2)

Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.



1. Consider all entities with that name
2. Weight each entity by features (bag of words, prior, etc.)
3. Start a random walk in the KB

State of the art

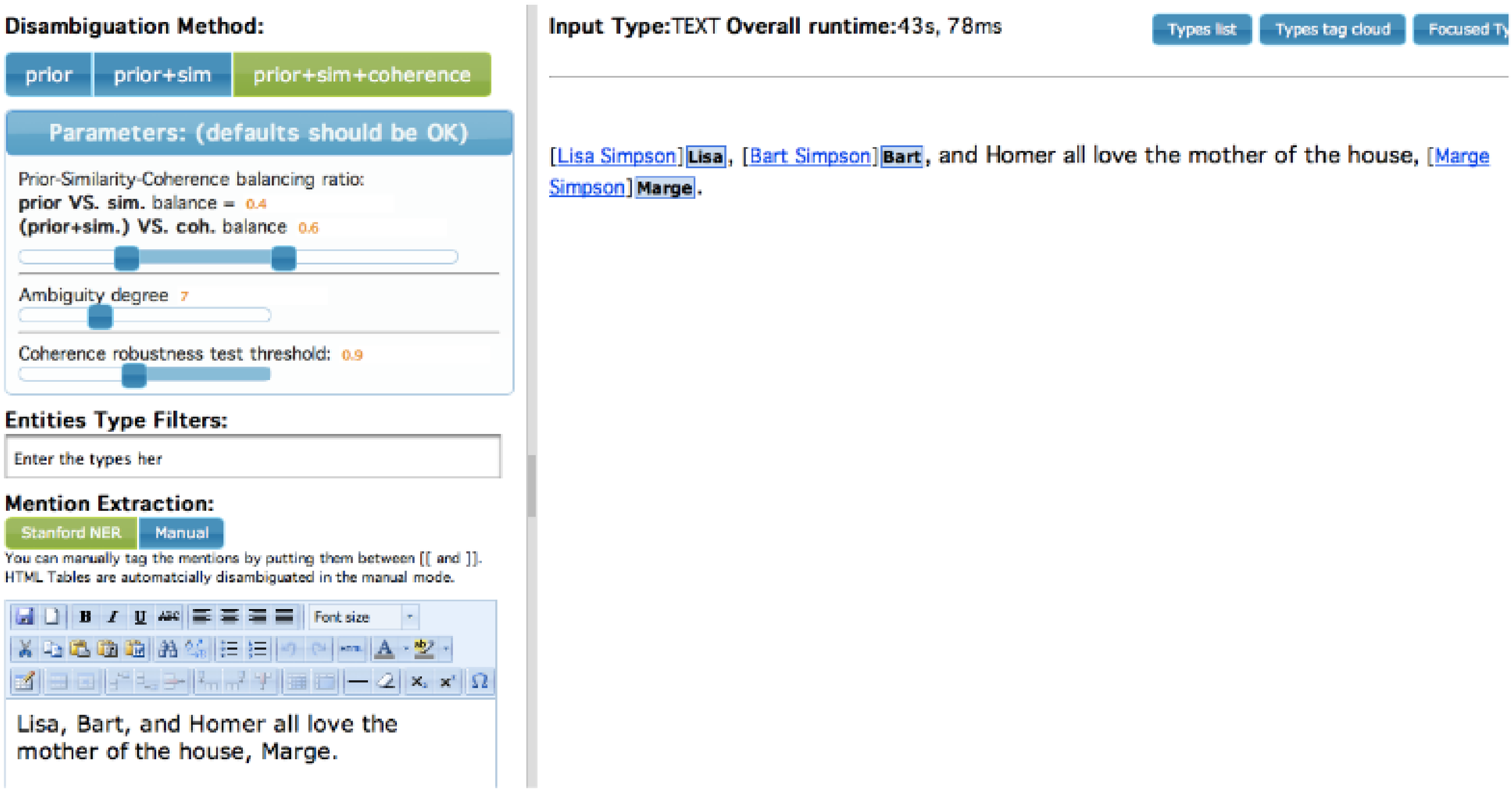
- Pre-trained neural models again
  - Encode KB context
  - Encode text context
  - Predict match likelihood
  - …or, predict KB identifier directly (GENRE, de Cao, ICLR 2021)

- Automated training data: Wikidata text links

Example systems (1): Opentapioca

https://opentapioca.org/

Example systems (2): AIDA



[Explicit parameter tuning – no more functioning ☹](https://gate.d5.mpi-inf.mpg.de/webaida/)
https://gate.d5.mpi-inf.mpg.de/webaida/

Further solutions

- spaCy can do this
  - https://spacy.io/usage/linguistic-features#entity-linking
  - Though more complex setup, KB
- Commercial APIs
  - https://try.rosette.com/
  - https://cloud.google.com/natural-language/docs/analyzing-entities
  - https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

Summary: Disambiguation

We saw 3 indicators for disambiguation:

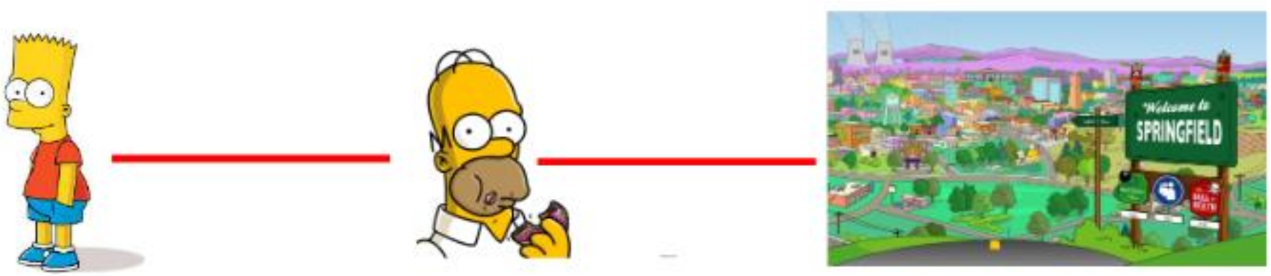
1. Context

Homer eats a doughnut.

2. Disambiguation prior

> Simpsons

3. Coherence



Disambiguation vs. mention typing

- Like for typing, context is decisive

- Unlike typing, no chance for supervised approach
  - Can train classifiers that predict "Politician-ness" of a mention
  - Cannot train classifier to predict "Einstein-ness"
- Disambiguation is ranking problem (single solution), not multiclass classification

- Type predictions can be used as intermediate features for context-based disambiguation
- Type prediction can augment disambiguation, if KB has sparse content

References

- Panchenko, Alexander, et al. Taxi at SEMEVAL-2016 Task 13: A taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. SemEval 2016.
- Gupta, Amit, et al. "Taxonomy induction using hypernym subsequences." CIKM 2017.
- Chu, Cuong Xuan, et al. "TiFi: Taxonomy Induction for Fictional Domains." WWW 2019.
- Yosef, Mohamed Amir, et al. Aida: An online tool for accurate disambiguation of named entities in text and tables. VLDB 2011.

- Slides adapted from Fabian Suchanek, Gina-Anne Levow and Chris Manning

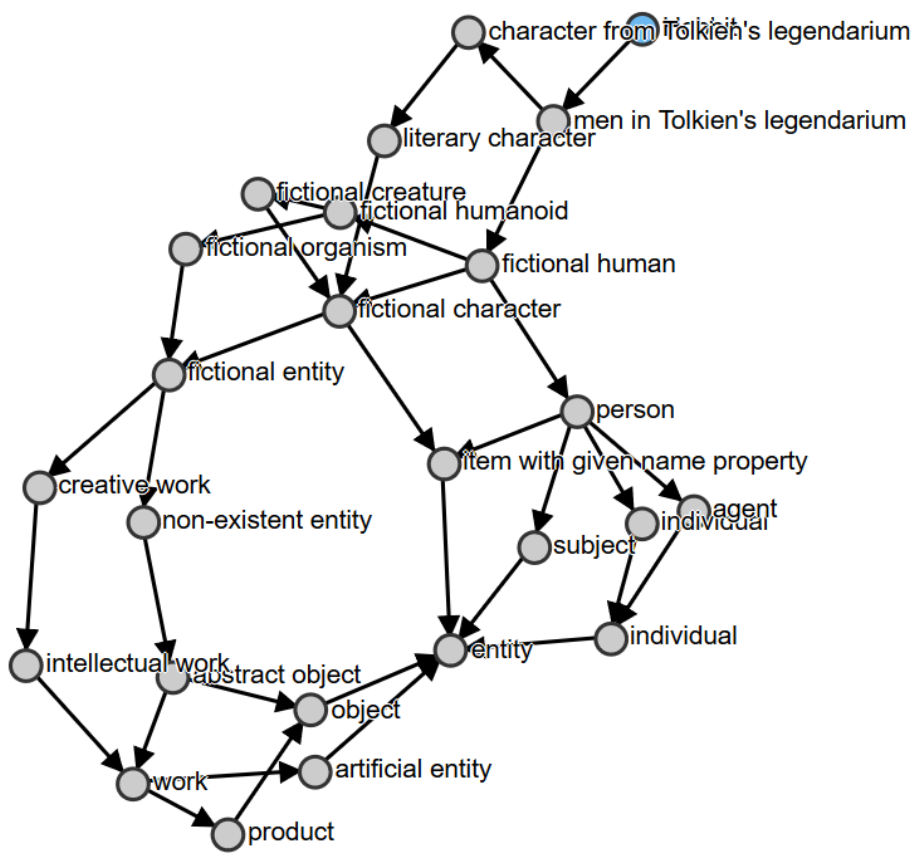
Assignment 5 – Taxonomy induction

- Given: Set of terms
- Task: Build a small taxonomy that organizes them
  - Can become both leafs or inner nodes
- Noisy input provided from WebIsALOD
  - Cleaning, filtering, etc. highly recommended
  - Other inputs allowed too
- Evaluation:
  - Two known term sets
  - One unseen set (robustness)

Take home

- Taxonomy induction:
  - Structure matters
  - Important features: Lexical/semantic matches, structural properties
- Entity disambiguation
  - Context seen already in typing
  - Coherence as additional feature
- Meta-observation:
  - Both problems are better approached globally than locally
  - Both problems are complementary

Playing with the Wikidata taxonomy



https://angryloki.github.io/wikidata-graph-builder/?property=P279&item=Q74359