Information extraction

5. Taxonomy induction, entity disambiguation, coreference resolution

Simon Razniewski

Winter semester 2019/20

Announcements

- Results assignment 4
  - 0.58 F1
  - Supervised/unsupervised competitive
    - Mapping helps a lot
  - Dataset issues
    - Terminology gap, incomplete sentences
    - Common in distant supervision
    - Upper bound?
    - Solution?
- Extensions not possible
  - (Except medical reasons)
  - Unfair to other participants
  - Studying is about skills as well as meta-skills

Lab 04 Ranking	
2576572	59
2561347	58.66
2558667	55.59
2550421	48.82
2576861	46.11
2568227	36.49
2549786	36.31
2562559	35.62
2548617	35.28
2581370	34.83
2565094	33.49
2579810	31.29
2572706	30.75
2576748	29.76
2576770	25.62
2550309	24.48
2564409	23.32
2571656	21.33
2558462	20.02
2576612	12.22
2571690	10.68
2576610	8.1
2568101	6.03
2576381	0

Outline

1. Taxonomy induction
2. Coreference resolution
3. Entity disambiguation

Taxonomy induction

- Goal: Creating a comprehensive taxonomy from noisy hypernymy relations

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

Taxonomy induction

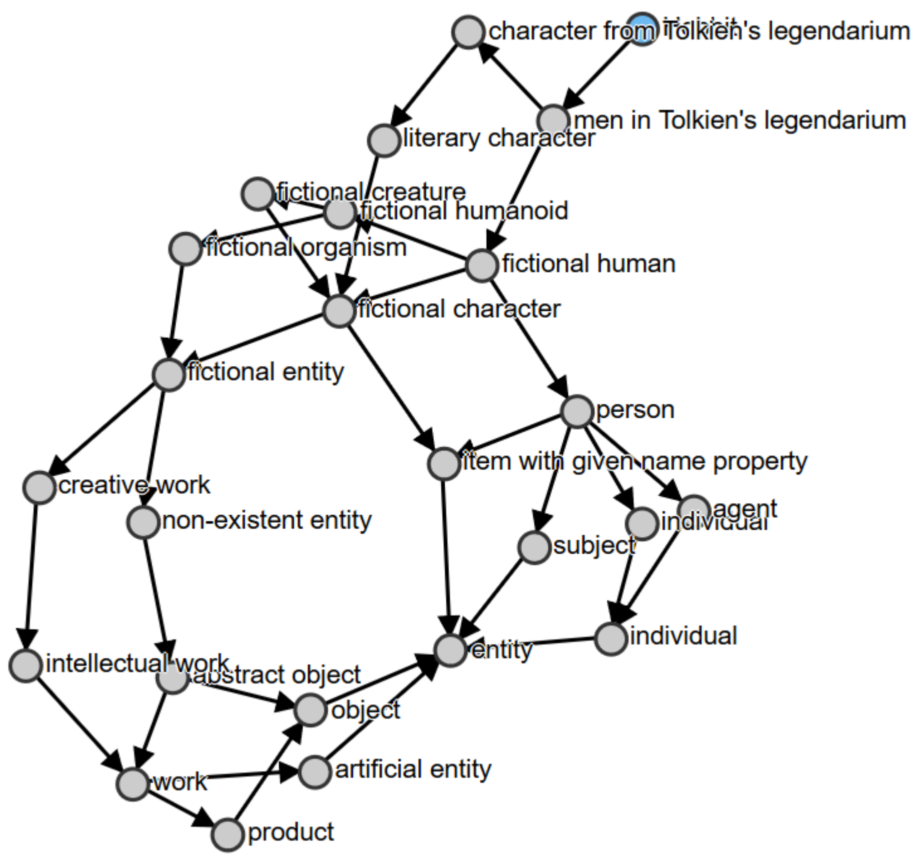
Categories: Featured articles | Characters | Cleanup | Hobbits | Baggins | Ring bearers | Elf friends | Fellowship members | Major characters (The Lord of the Rings) | The Lord of the Rings Characters | Characters that have appeared in the Hobbit and the Lord of the Rings | The Hobbit: An Unexpected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

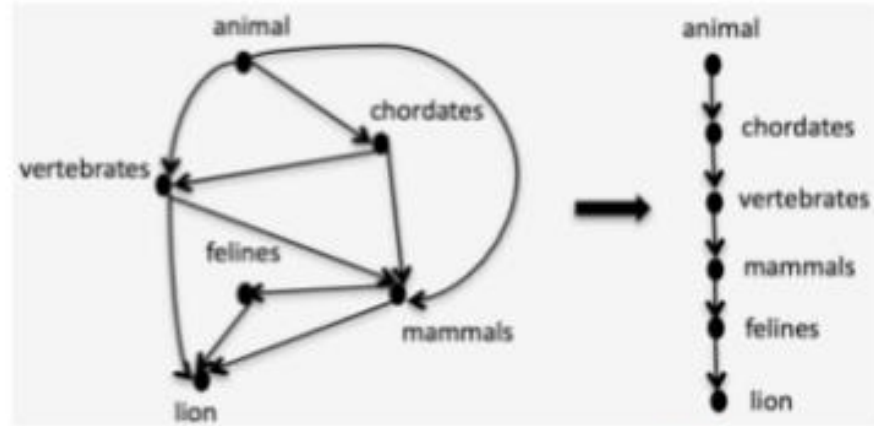
Categories: Swordsmen | Fictional melee weapons practitioners
Hidden categories: Categories requiring diffusion

Desired shape (single leaf)



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

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"?)

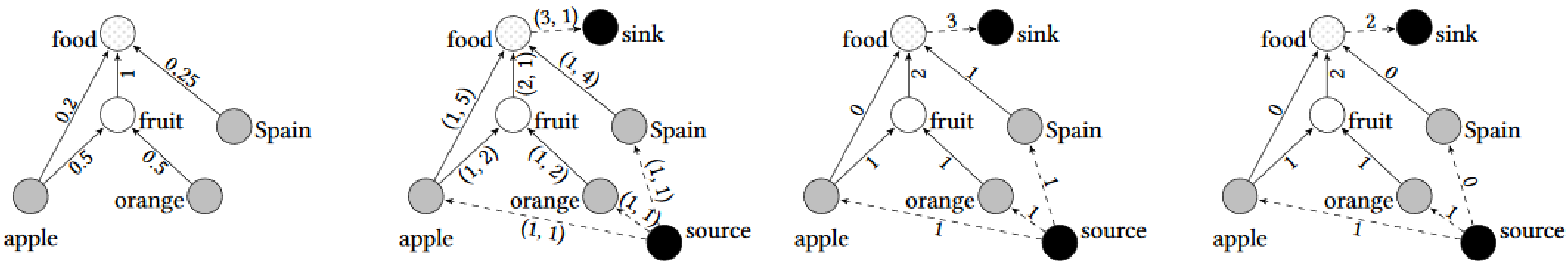
Candidate hypernym	Frequency
company	5536
fruit	3898
apple	2119
vegetable	928
orange	797
tech company	619
brand	463
hardware company	460
technology company	427
food	370

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 $d = 2$.

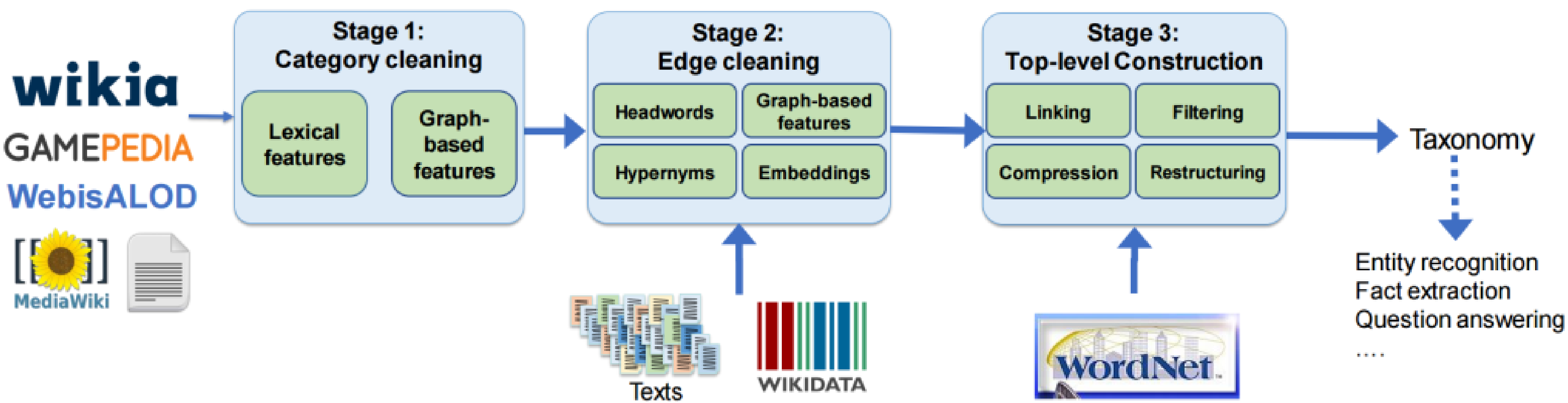
- Method: Find cheapest way to send flow from leaves to root
- Cost inverse proportional to edge weight

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 $subclassOf$ 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.

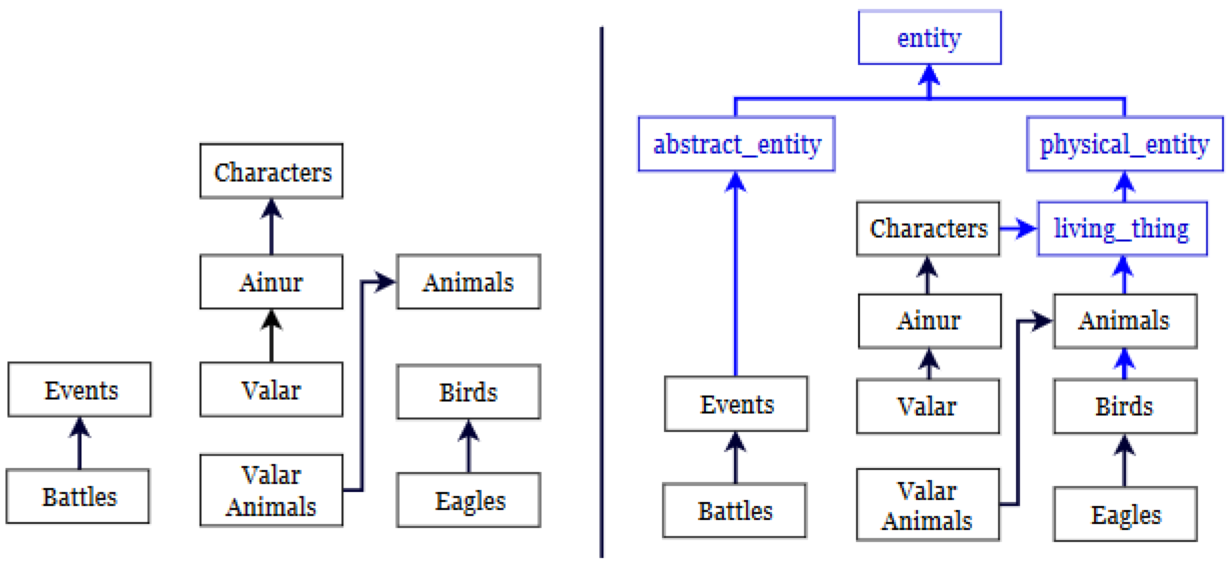
← Embedding only

← 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.

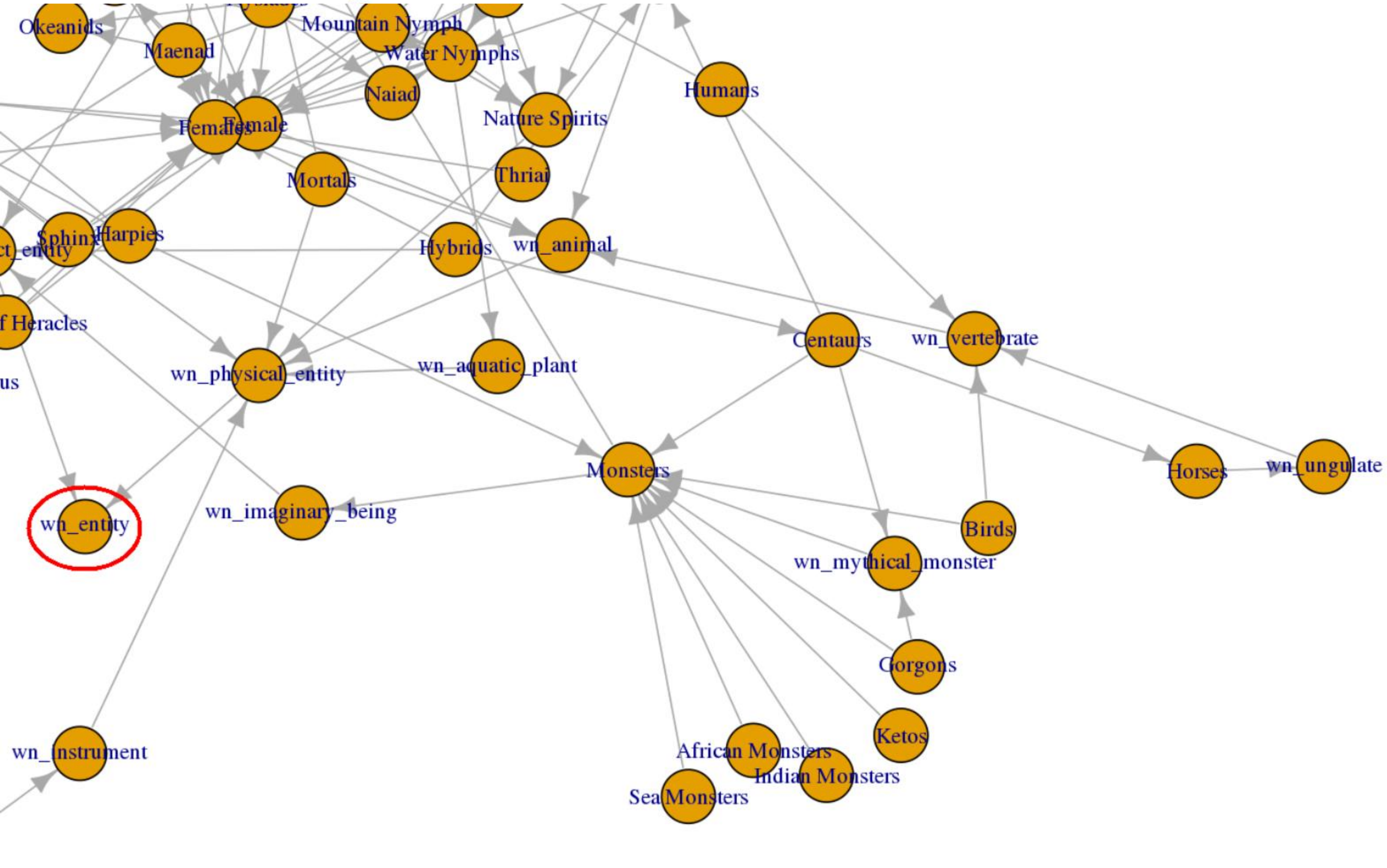
TiFi – Relevance for entity search

	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

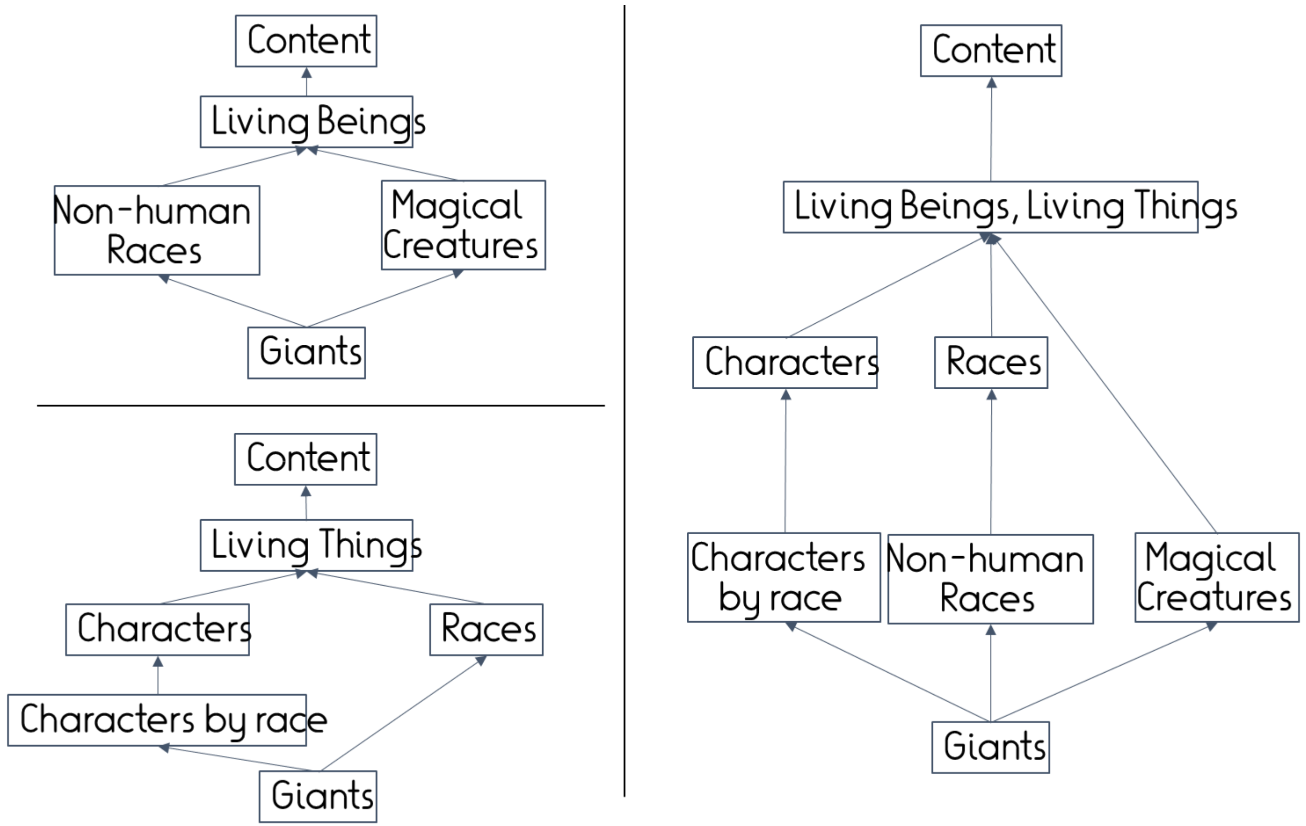
Table 12. Example queries and results for the entity search evaluation.

	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%)	$\boxed{\text{6 (79\%)}}$

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


Okeanids
Mountain Nymph
Maenad
Water Nymphs
Naiad
Humans
Females
Female
Nature Spirits
Mortals
Thriai
Sphinx
Harpies
Hybrids
wn_animal
f Heracles
Centaurs
wn_vertebrate
wn_physical_entity
wn_aquatic_plant
Monsters
Horses
wn_ungulate
wn_entity
wn_imaginary_being
Birds
wn_mythical_monster
Gorgons
Ketos
African Monsters
Indian Monsters
Sea Monsters
wn_instrument

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
- Top-level most challenging
- Relevance for IE:
  - Types can power search
  - Types can guide relation extraction
  - Taxonomies allow to detect compatibility/conflicts
    - givesPresent(person, item)

      → givesPresent(politician, suitcase) ✓

      → givesPresent(cat, deadMouse) ?

      → givesPresent(song, location) x

Outline

1. Taxonomy induction
2. Coreference resolution
3. Entity disambiguation

Ready for fact extraction?

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Nominated(Barack Obama, Hillary R. Clinton)

Chose(He, her)?

Coreference Resolution

Task: Identify all noun phrases (mentions) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

A couple of years later, Vanaja met Akhila at the local park. Akhila's son Prajwal was just two months younger than her son Akash, and they went to the same school. For the preschool play, Prajwal was chosen for the lead role of the naughty child Lord Krishna. Akash was to be a tree. She resigned herself to make Akash the best tree that anybody had ever seen. She bought him a brown T--shirt and brown trousers to represent the tree trunk. Then she made a large cardboard cutout of a tree's foliage, with a circular opening in the middle for Akash's face. She attached red balls to it to represent fruits. It truly was the nicest tree.

From The Star by Shruthi Rao, with some shortening.

Coreference Resolution

- Noun phrases refer to entities in the world, many pairs of noun phrases co--refer, some nested inside others

John Smith, CFO of Prime Corp since 1986,

saw his pay jump 20% to $1.3 million as

the 57--year--old also became

the financial services co.'s president.

Kinds of Reference

- Names and noun phrases
  - John Smith
  - President Smith
  - the president
  - the company's new executive

More common in news, generally harder in practice, more world knowledge needed

- Pronouns
  - She/he/it
- Demonstratives
  - This, that

More interesting grammatical constraints, more linguistic theory, easier in practice "anaphora resolution"

Information Status

- Some expressions (e.g. indef NPs) introduce new info
- Others refer to old referents (e.g. pronouns)

- Theories link form of refexp to given/new status

The givenness hierarchy:

in focus >	activated >	familiar >	uniquely identifiable >	referential >	type identifiable
{it}	{ *that*, *this*, *this* N }	{that N}	{the N}	{indef. *this* N}	{*a* N}

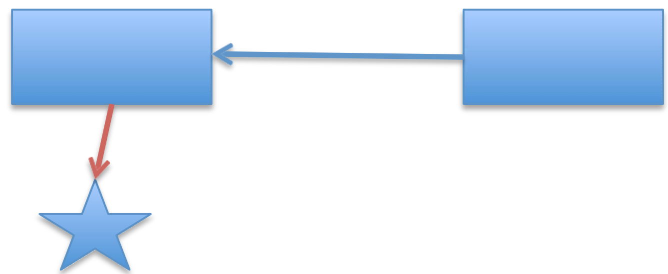
- Accessibility:
    - More salient elements easier to call up

Anaphora vs. coreference

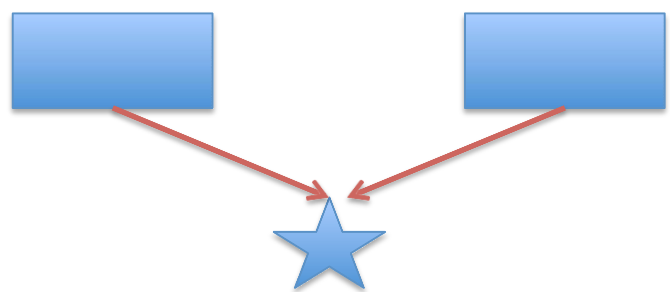
- Coreference is when two mentions refer to the same entity in the world
- Anaphora is when a term refers to another term and the interpretation of the second is in some way determined by the interpretation first
  - Anaphora, no coreference:
    We went to see a concert last night. The tickets were really expensive.
  - Conversely, multiple identical full NP references are typically coreferential but not anaphoric.
    Smith was looking forward to the concert. Smith therefore couldn't wait until …

Two different things…

- Anaphora
  - Text
  - World



- (Co)Reference
  - Text
  - World



How to approach (pronoun) coreference resolution?

- Baselines
  - Pick closest previous entity?
  - Pick closest previous entity that agrees in gender and cardinality?

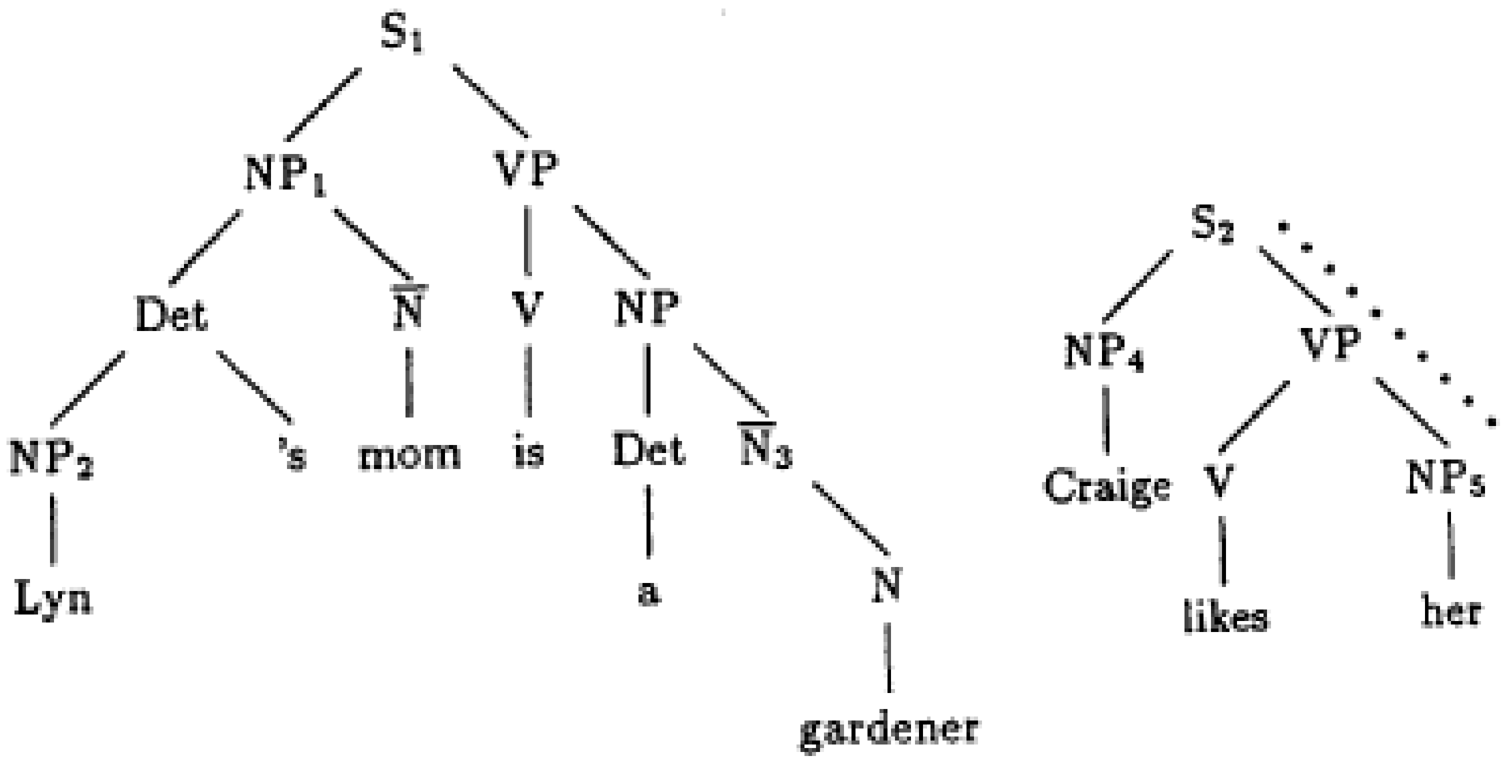
Hobbs' Resolution Algorithm

- Requires:
  - Syntactic parser
  - Gender and number checker
- Input:
  - Pronoun
  - Syntactic parse of current and previous sentences
- Captures:
  - Preferences: Recency, grammatical role
  - Constraints: binding theory, gender, person, number

Hobbs' Algorithm

- Intuition:
  - Start with target pronoun
  - Climb parse tree to sentence (S) root
  - For each NP or S
    - Do breadth-first, left-to-right search of children
      - Restricted to left of target
    - For each NP, if another NP or S appears before root check agreement with target
  - Repeat on earlier sentences without in-between condition, until matching NP found

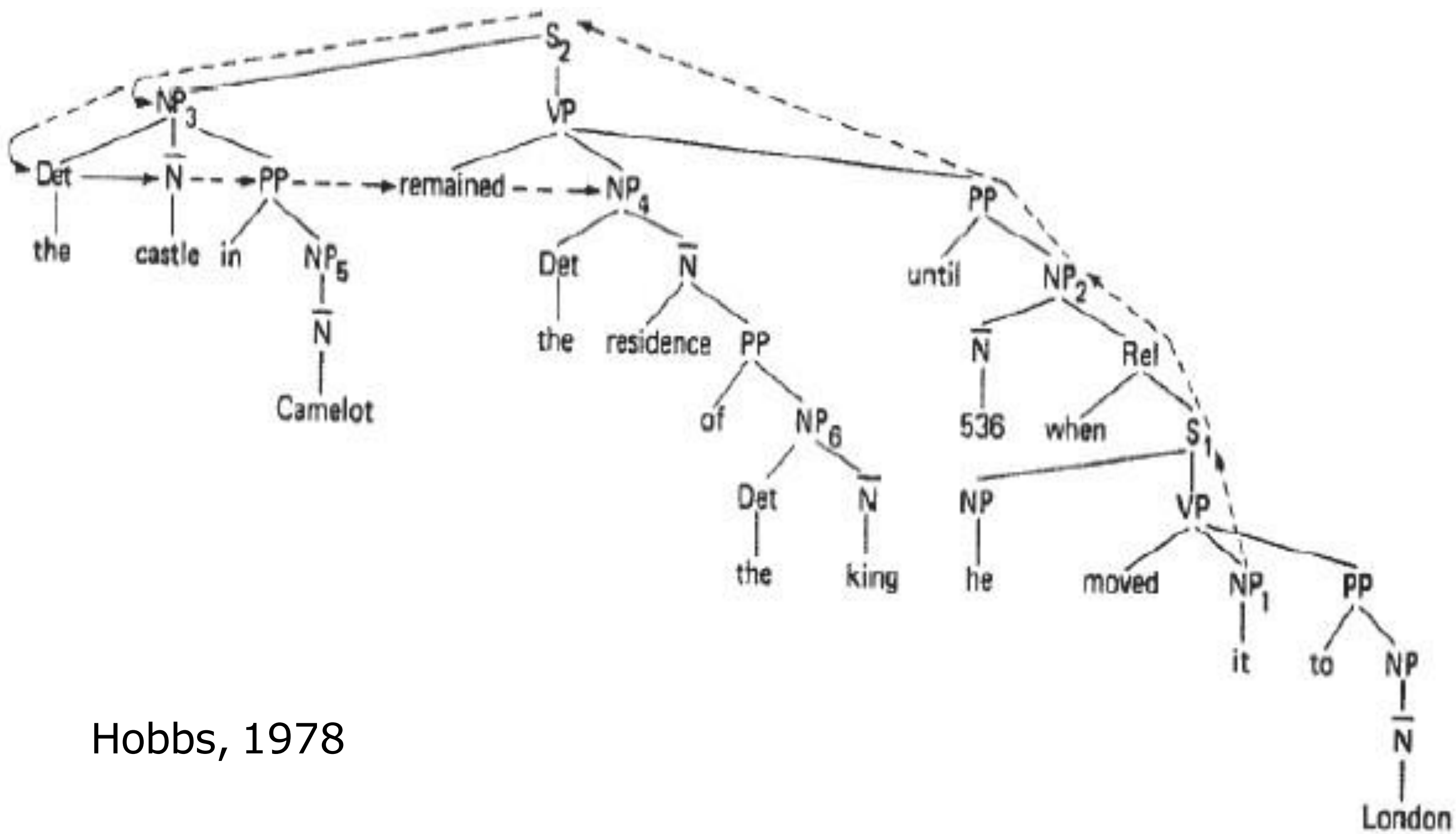
Hobbs' Example



Lyn’s mom is a gardener. Craige likes her.

Another Hobbs' Example

The castle in Camelot remained the residence of the King until 536 when he moved it to London.



Hobbs, 1978

Hobbs' Algorithm

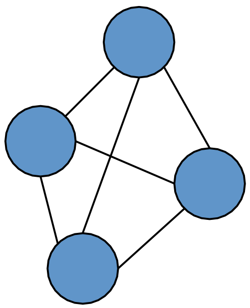
- Results: 88% accuracy; 90+% intrasentential
  - On perfect, manually parsed sentences
- Useful baseline for evaluating pronominal anaphora
- Issues:
  - Parsing:
    - Informal language
    - Parsers are not always accurate

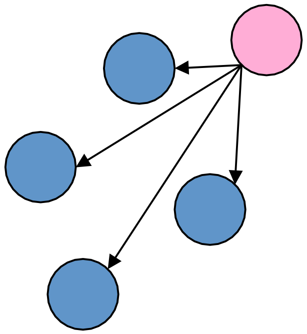
But it's complicated …

- Common nouns can differ in number but be coreferent:
  - a patrol … the soldiers
- Common nouns can refer to proper nouns
  - George Bush … the leader of the free world
- Pleonastic expressions
  - It is raining.
- Split antecedence
  - John waited for Sasha. Then they went out.

Data-driven Coreference Resolution

- Data-driven machine learning approach
  - Coreference as classification, clustering, ranking problem
    - Mention-pair model:
      - For each pair NP_i,NP_j, do they corefer?
      - Cluster/split to form equivalence classes
    - Entity-mention model
      - For each pair NP_k and cluster C_j, should the NP be in the cluster?
    - Ranking models
      - For each NP_k, and all candidate antecedents, which highest?





Mention-pair model

- Given a mention and an entity mentioned earlier, classify whether the pronoun refers to that entity or not given the surrounding context (yes/no)

? ? ?

Obama visited the city. The president talked about Milwaukee's economy. He mentioned new jobs.

- Obtain positive examples from training data, generate negative examples by pairing each positive example with other (incorrect) entities
- This is naturally thought of as a binary classification (or ranking) task

Features in the mention-pair model

- Constraints:
  - Number agreement
    - Singular pronouns (it/he/she/his/her/him) refer to singular entities and plural pronouns (we/they/us/them) refer to plural entities
  - Person agreement
    - He/she/they etc. must refer to a third person entity
  - Gender agreement
    - He → John; she → Mary; it → car
    - Jack gave Mary a gift. She was excited.
  - Certain syntactic constraints
    - John bought himself a new car. [himself → John]
    - John bought him a new car. [him can not be John]

Features in the mention-pair model

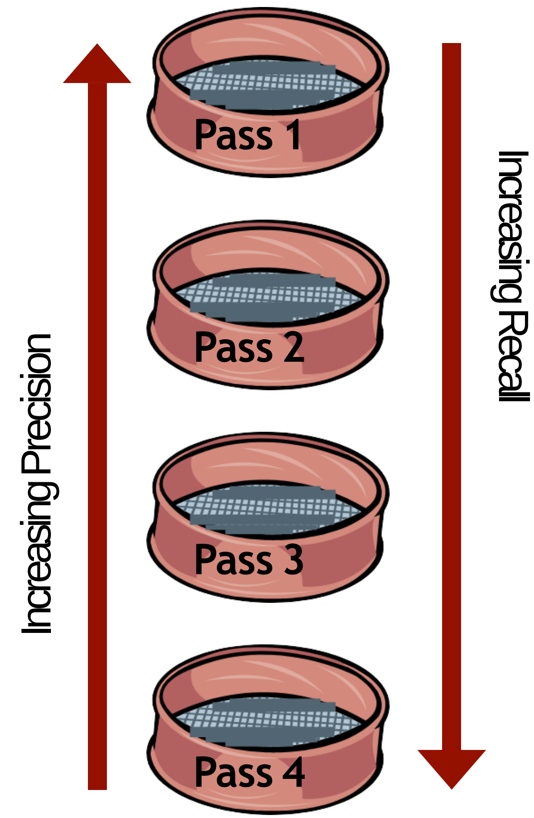
- Preferences:
  - Recency: More recently mentioned entities are more likely to be referred to
    - John went to a movie. Jack went as well. He was not busy.
  - Grammatical Role: Entities in the subject position is more likely to be referred to than entities in the object position
    - John went to a movie with Jack. He was not busy.
  - Parallelism:
    - John went with Jack to a movie. Joe went with him to a bar.

Features in the mention-pair model

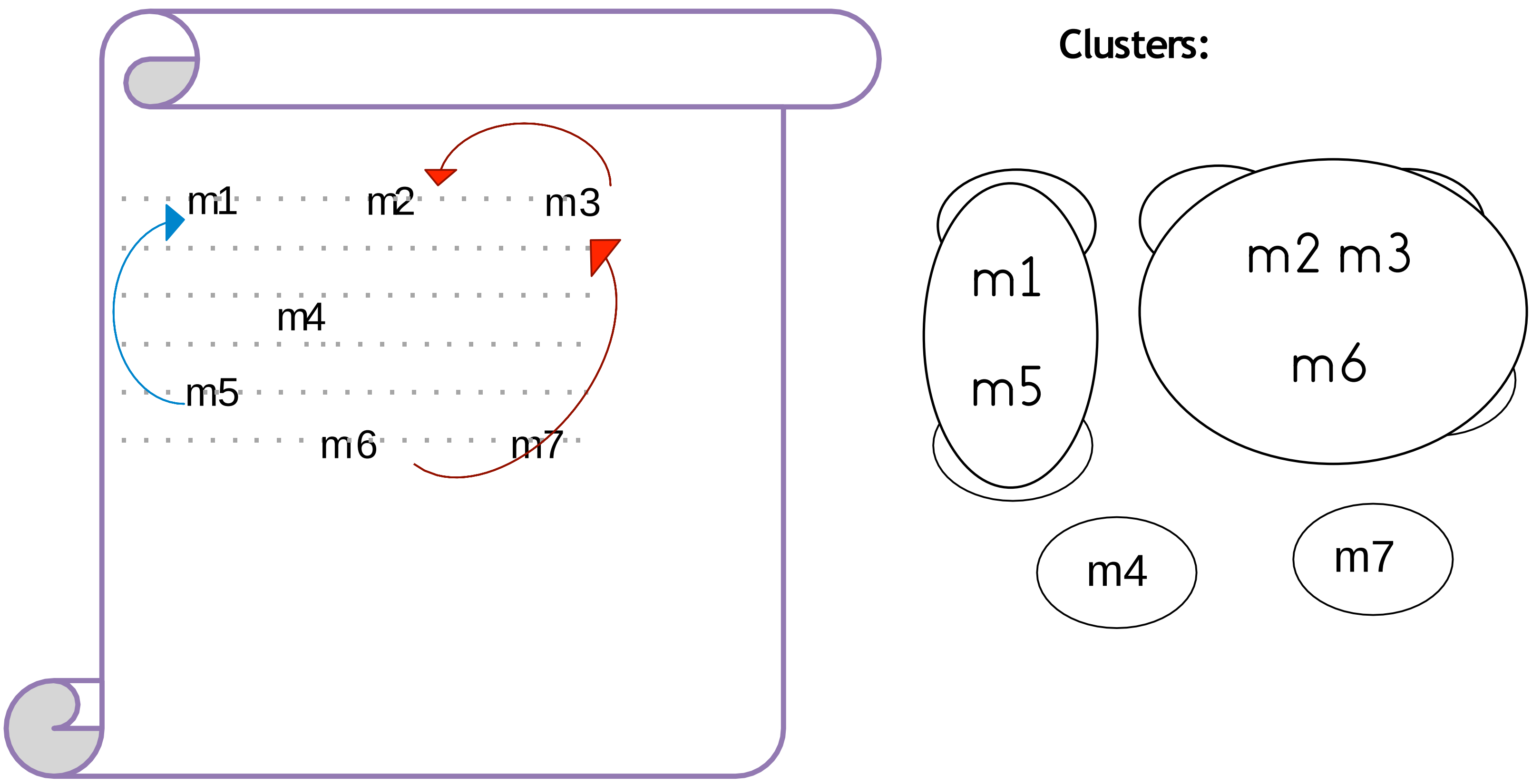
- Preferences:
  - Verb Semantics: Certain verbs seem to bias whether the subsequent pronouns should be referring to their subjects or objects
    - John telephoned Bill. He lost the laptop.
    - John criticized Bill. He lost the laptop.
  - Selectional Restrictions: Restrictions because of semantics
    - John parked his car in the garage after driving it around for hours.
- Encode all these and may be more as features

Lee et al. (2010): Stanford deterministic coreference

- Cautious and incremental approach
- Multiple passes over text
- Precision of each pass is lesser than preceding ones
- Recall keeps increasing with each pass
- Decisions once made cannot be modified by later passes
- Rule-based ("unsupervised")



Entity-mention model: Clusters instead of mentions



Detailed Architecture

The system consists of seven passes (or sieves):

- Exact Match
- Precise Constructs (appositives, predicate nominatives, …)
- Strict Head Matching
- Strict Head Matching – Variant 1
- Strict Head Matching – Variant 2
- Relaxed Head Matching
- Pronouns

Subsequent sieves extend earlier found coreferences

Approach: start with high precision clumpings

E.g.

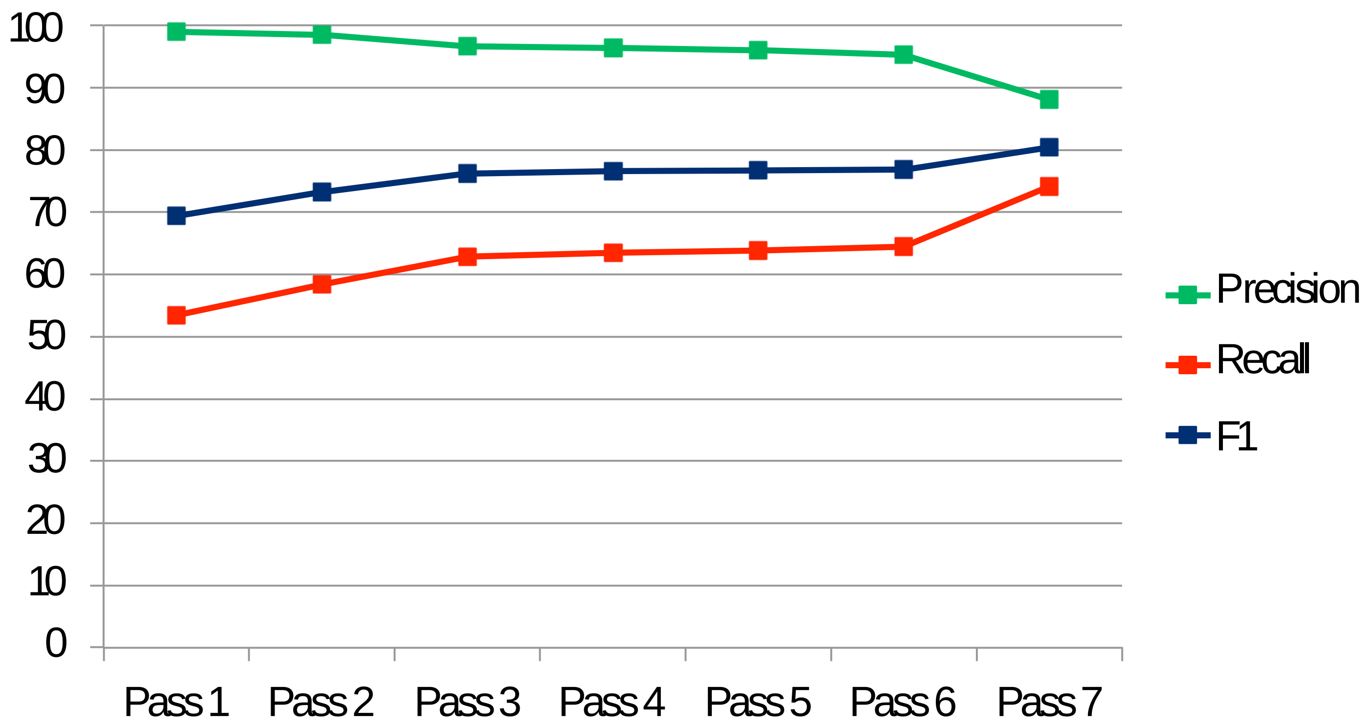
Pepsi hopes to take Quaker oats to a new level ….. Pepis says it expects to double Quaker's snack food growth rate. … the deal will give Pepsi access to Quaker oats Gatorade drink as well as …

…
Angela Merkel, the leader of the free world, …
…

Multi-pass Sieve Modules

- Pass 3: Strict head matching
  - Matches cluster head noun AND all non-stop cluster wds AND modifiers AND non i/I
- Pass 4 & 5: Variants of 3: drop one of above
- Pass 6: Relaxed head match
  - Head matches any word in cluster AND all non-stop cluster wds AND non i/I
- Pass 7: Pronouns
  - Enforce constraints on gender, number, person, animacy, and NER labels

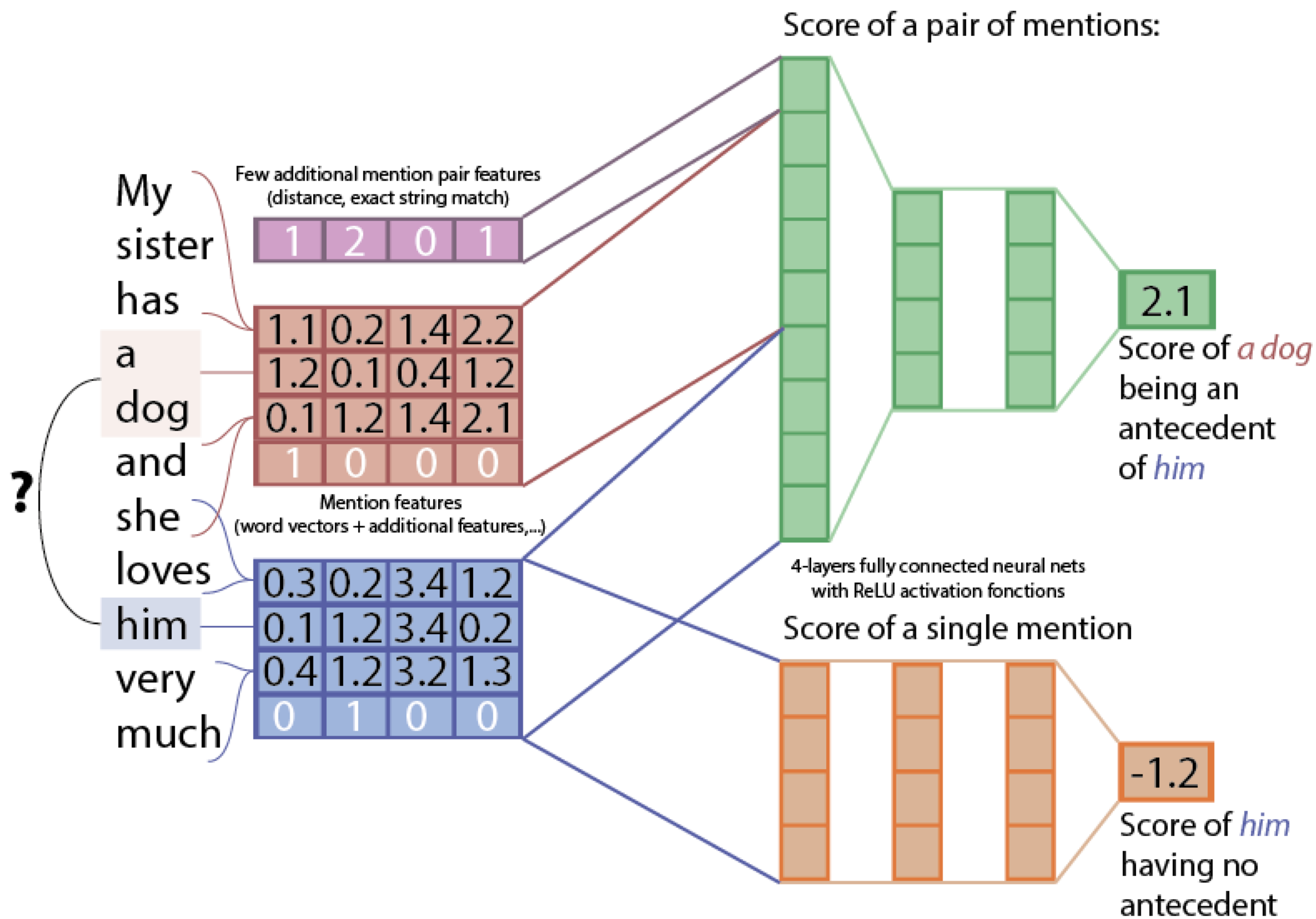
Cumulative performance of passes



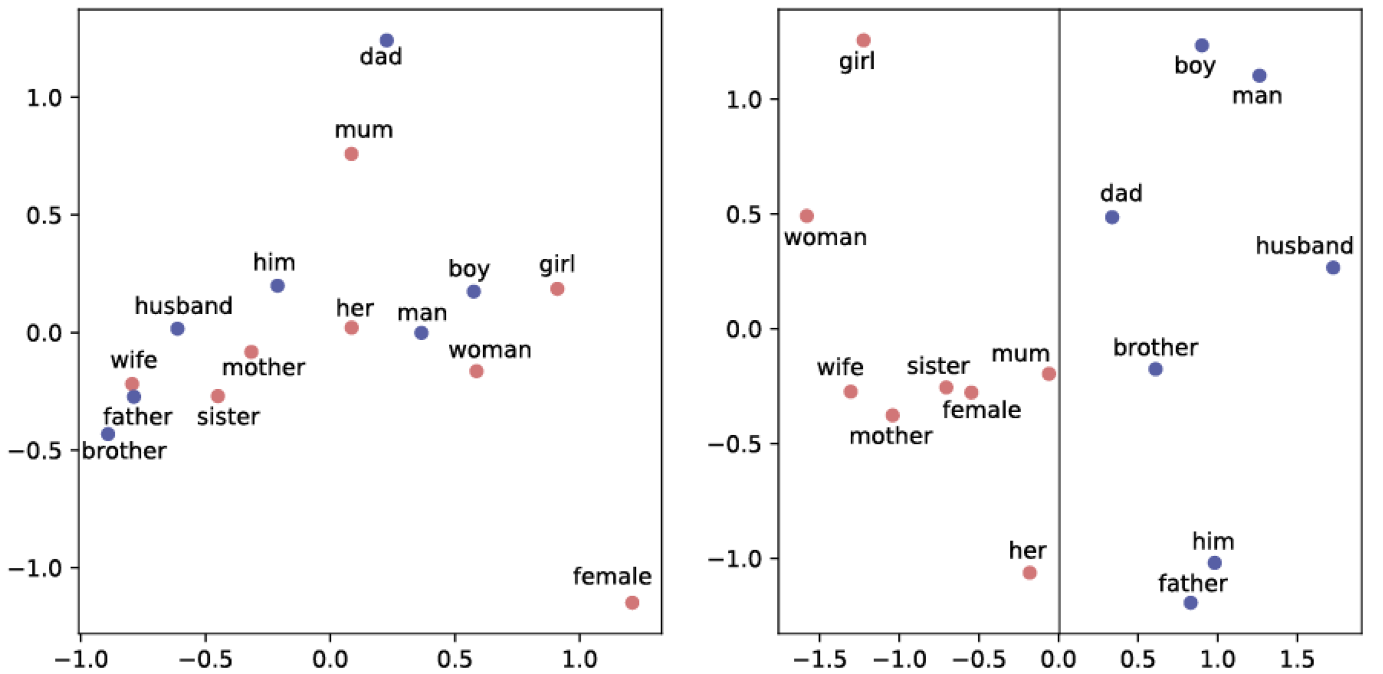
Graph showing the system's B3 Precision, Recall and F1 on ACE2004-DEV after each additional pass

State of the art: Neural coref

Mention-ranking model



Neural coref: Embedding retraining



Word embeddings before and after retraining on coref task

Neural coref

- Python extension of spaCy available

- Demo online
  - https://huggingface.co/coref/

- (let's try)

Outline

1. Taxonomy induction
2. Coreference resolution
3. Entity disambiguation

Ready for fact extraction?

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

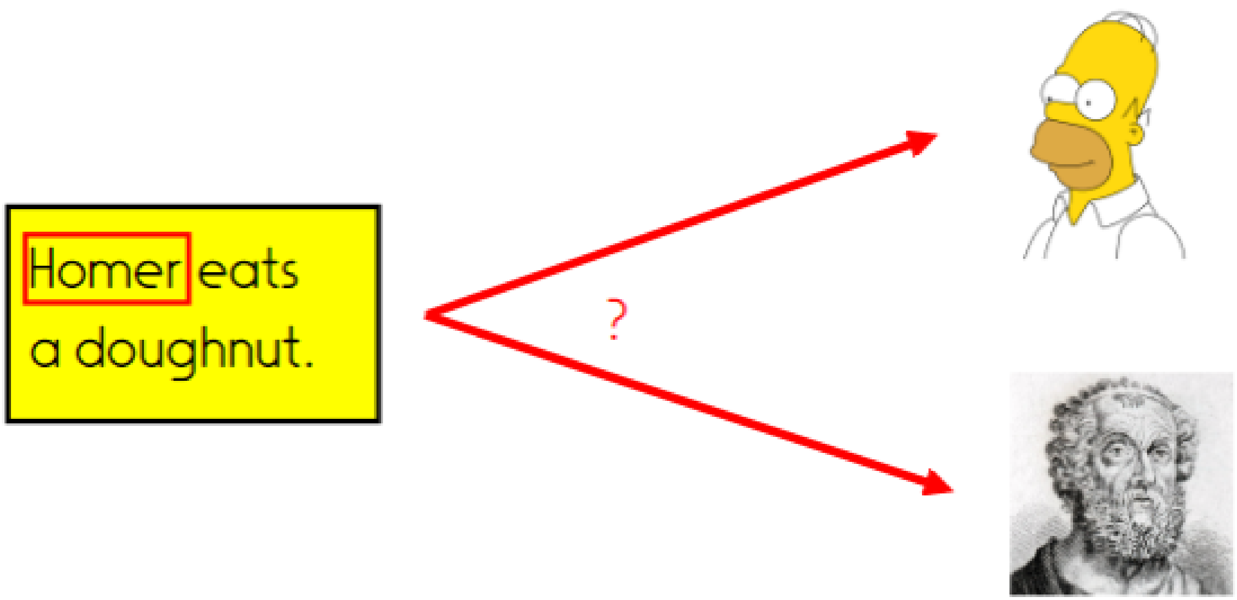
Homer is the author of the Odyssey.

appearsIn(Homer, Simpsons)

wrote(Homer, Odyssey)?

Def: Disambiguation

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



Disambiguati

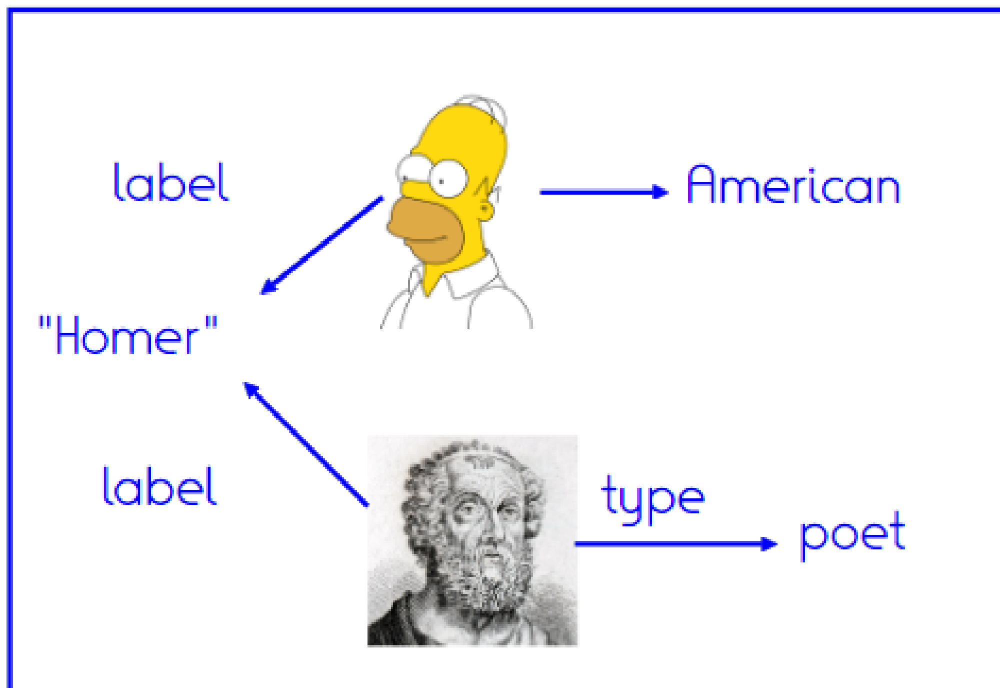
Usually Named Entity Recogn
is to map the names to entities th

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

NER'ed
corpus

Homer eats
a doughnut.

Knowledge Base



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)

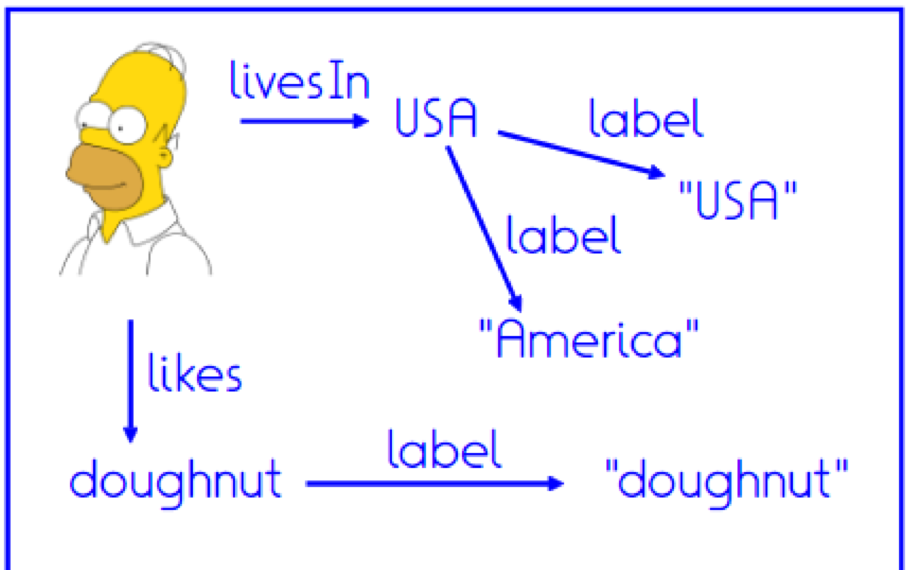
Homer eats
a doughnut.

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)

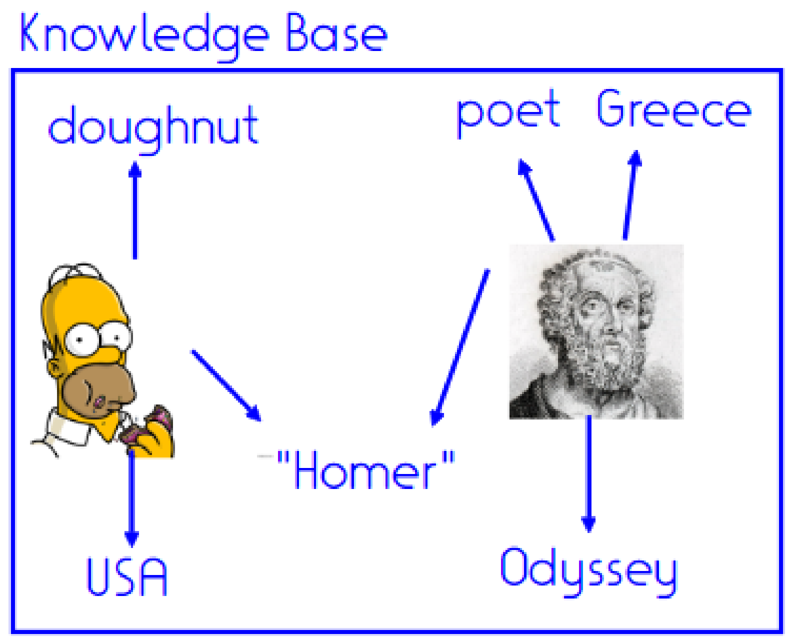


Context of Homer:
{doughnut, USA, America}

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.

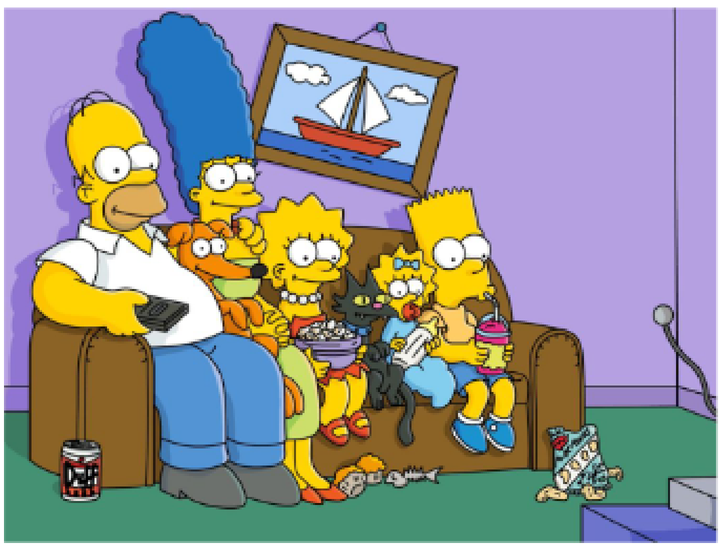
For USA Today, Homer is among the top 25 most influential people of the past 25 years.



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.

Simpsons Store

4

Simpsons

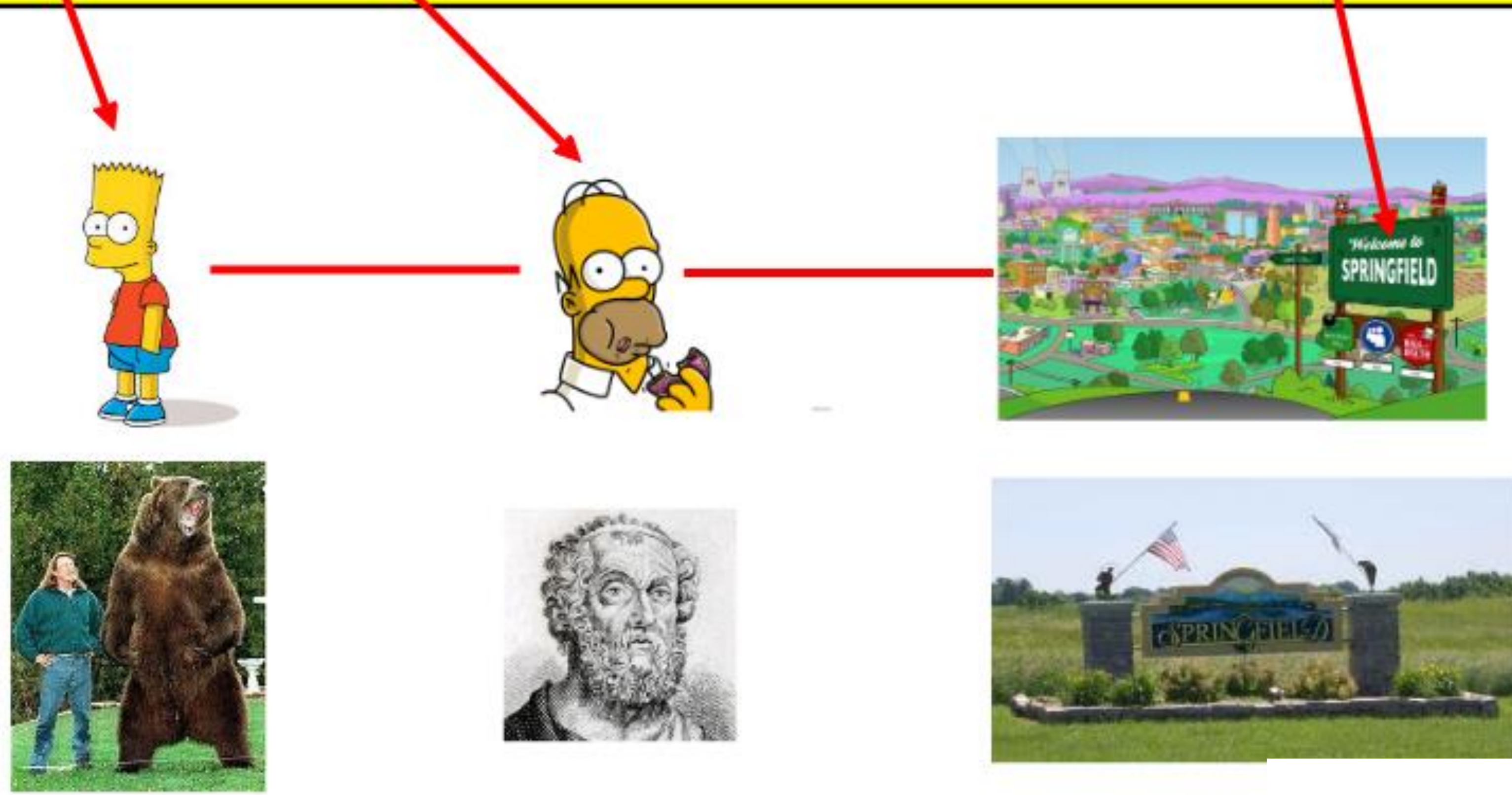
493,201

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

Def: Coherence Criterion

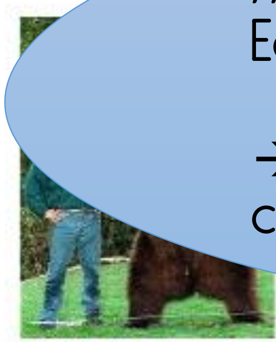
The Coherence Criterion postulates that entities that are mentioned in one document should be related in the KB.

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



Possible implementation (2)

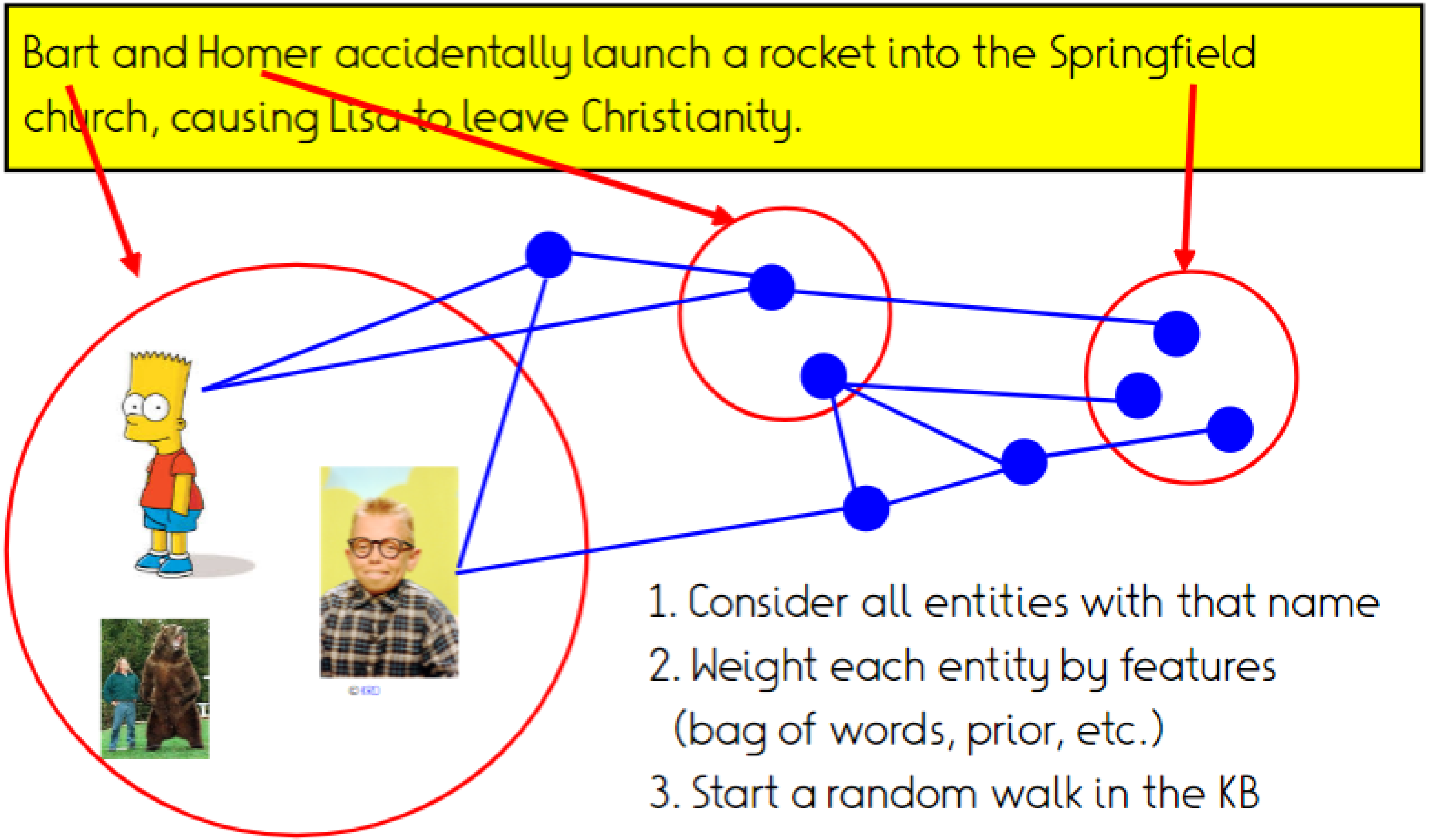
Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.



n entity mentions
Each with m candidate KB entities

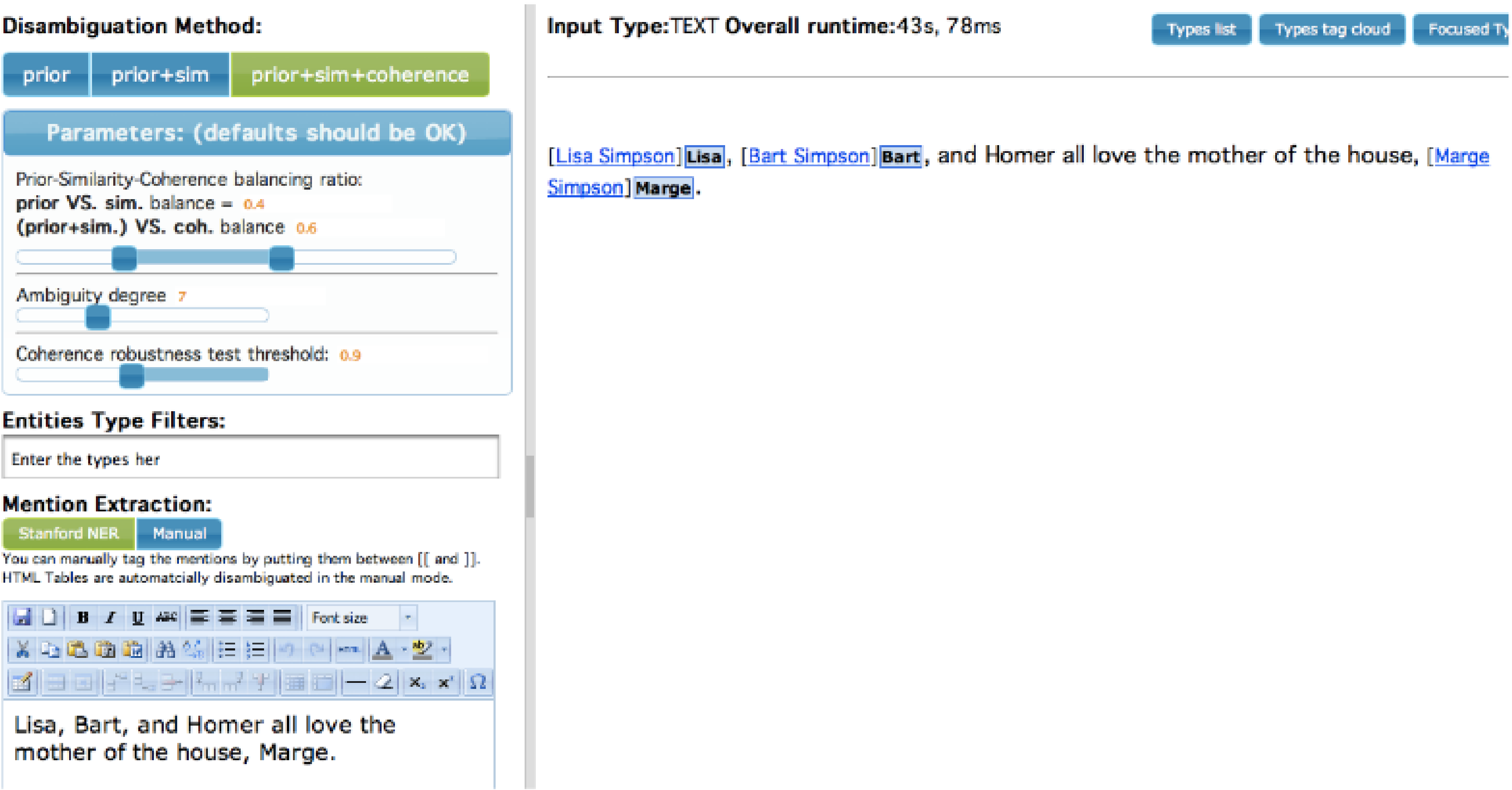
→ Compute coherence scores for m^n combinations

Possible implementation (2)



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

Example: Disambiguation by AIDA



Explicit parameter tuning
https://gate.d5.mpi-inf.mpg.de/webaida/

Further solutions

- spaCy can
  - 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:

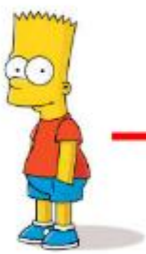
1. Context

Homer eats a doughnut.

2. Disambiguation prior

> Simpsons

3. Coherence



Disambiguation vs. coreference

- Closely related problems
- KBs can provide world knowledge for coreference
  - Gender, animatedness, etc.
- Coreference clusters give larger context for disambiguation
  - *He is the leader of a country. He also has orange hair. He is ridiculed frequently in social media.*

→ Ideally approached jointly

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

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 be both leafs or classes already
- 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
- Coreference resolution
    - Mention-pair classification/ranking
    - Recency and grammatical roles strong features
- Entity disambiguation
    - Context seen already in typing
    - Coherence as additional feature
- Meta-observation:
    - Each problem is better approached globally than locally
    - All three problems interact