COLING 2016
Osaka, Japan

December 2016

The Role of Wikipedia in Text Analysis and Retrieval

Marius Paşca

Google Inc.
mars@google.com

Overview

- Part One: Wikipedia as a Knowledge Resource
- Part Two: Role of Wikipedia in Text Analysis
- Part Three: Role of Wikipedia in Information Retrieval

Part One: Wikipedia as a Knowledge Resource

- Knowledge resources
- Dissecting Wikipedia
- Resources derived from Wikipedia

Knowledge Resources

- Created by experts
  - [Fel98]: C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press 1998.
  - [Len95]: D. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 1995.
- Created collaboratively by non-experts
  - [Rem02]: M. Remy. Wikipedia: The Free Encyclopedia. Journal of Online Information Review 2002.
  - [SLM+02]: P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins and W. Zhu. Open Mind Common Sense: Knowledge Acquisition from the General Public. Lecture Notes In Computer Science 2002.
  - [LS04]: H. Liu and P. Singh. ConceptNet – a Practical Commonsense Reasoning Tool-Kit. BT Technology Journal 2004.

Expert Resources

- WordNet
  - [Fel98]: C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press 1998.
  - lexical database of English created by experts
  - wide-coverage of upper-level conceptual hierarchies
  - replicated or extended to other languages
- Cyc
  - [Len95]: D. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 1995.
  - knowledge repository of common-sense knowledge created by experts over 100+ person-years
  - terms and assertions capturing ground assertions and (inference) rules

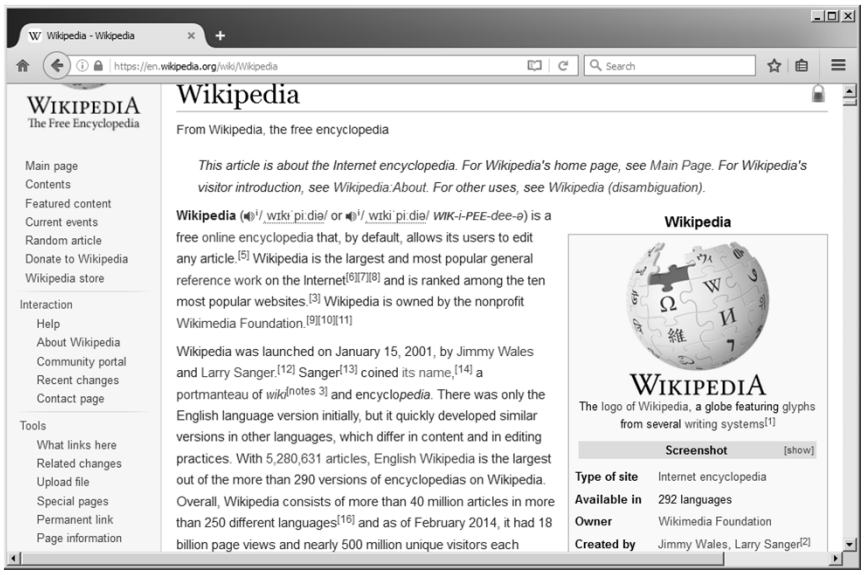
Collaborative, Non-Expert Resources

- Wikipedia
  - [Rem02]: M. Remy. Wikipedia: The Free Encyclopedia. Journal of Online Information Review 2002.
  - free online encyclopedia developed collaboratively by Web volunteers
- Open Mind
  - [SLM+02]: P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins and W. Zhu. Open Mind Common Sense: Knowledge Acquisition from the General Public. Lecture Notes In Computer Science 2002.
  - collect common-sense knowledge from non-expert Web users
  - unlike Cyc, collect and represent knowledge in natural language rather than through formal assertions
- ConceptNet
  - [LS04]: H. Liu and P. Singh. ConceptNet – a Practical Commonsense Reasoning Tool-Kit. BT Technology Journal 2004.
  - introduced as a successor to Open Mind, as a semantic network encoding common-sense knowledge represented through lexical concepts and labeled relations
  - knowledge sources include Open Mind, Wikipedia, WordNet

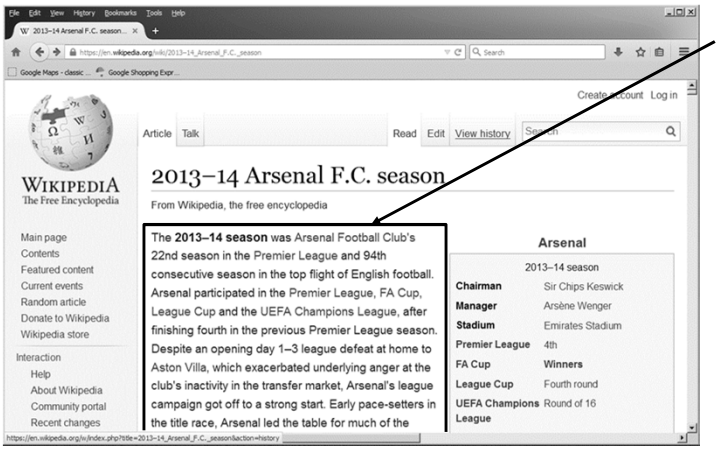
Part One: Wikipedia as a Knowledge Resource

- Knowledge resources
- Dissecting Wikipedia
- Resources derived from Wikipedia

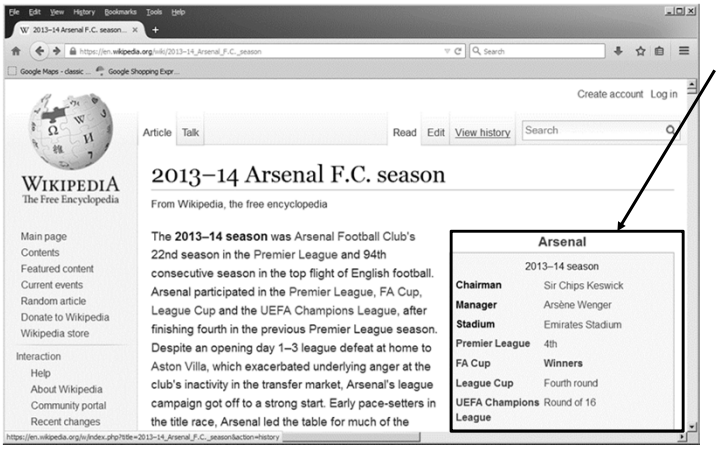
Wikipedia on Wikipedia



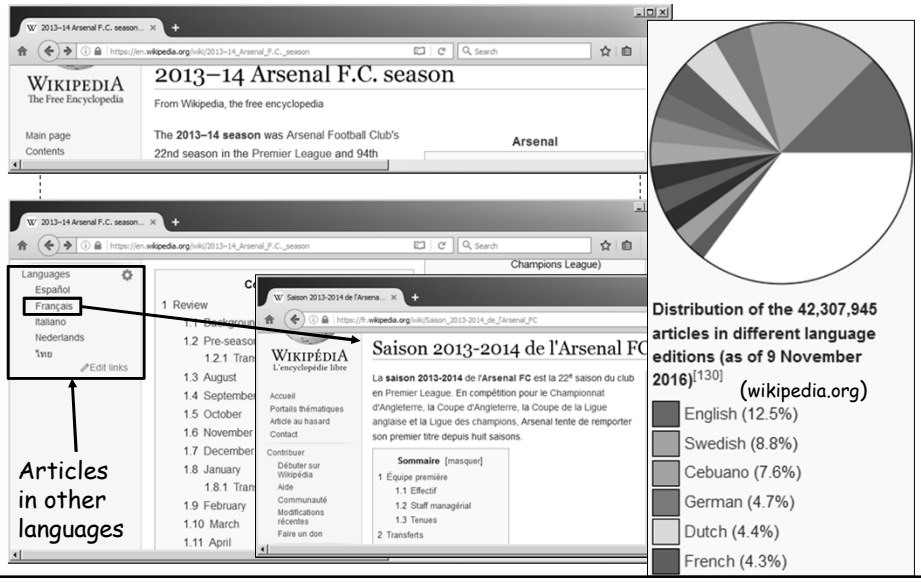
Dissecting Wikipedia Articles



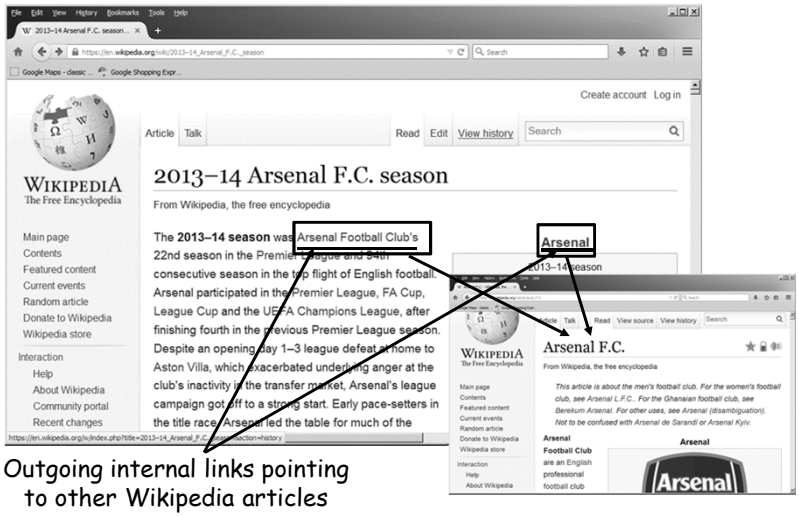
Infoboxes



Language Versions

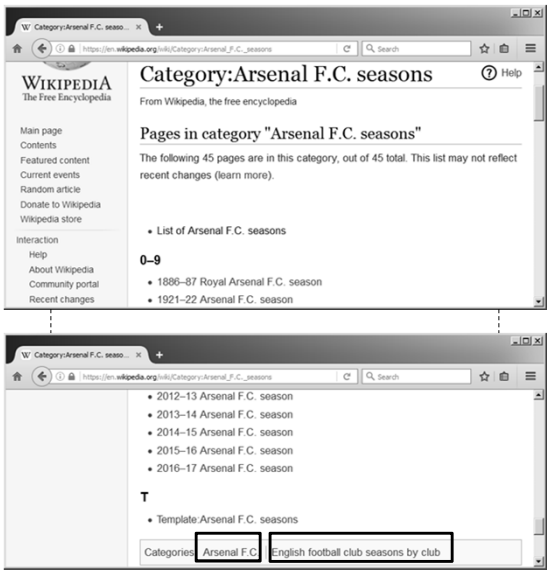


Links to Other Articles



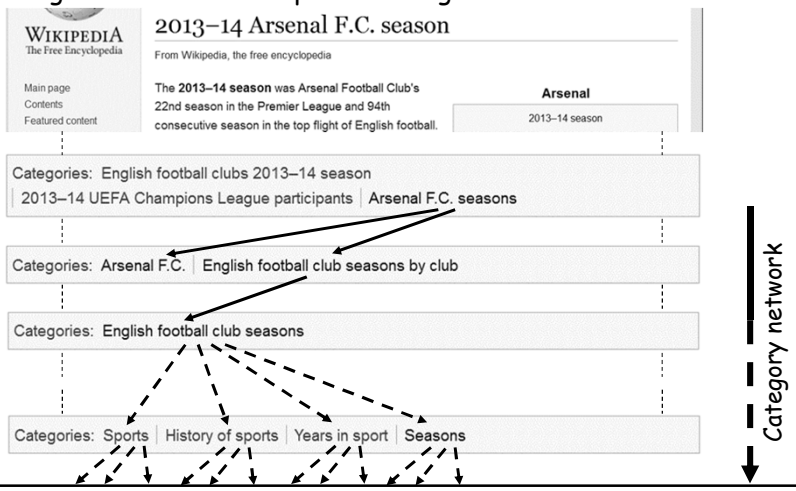
Categories

Category Pages

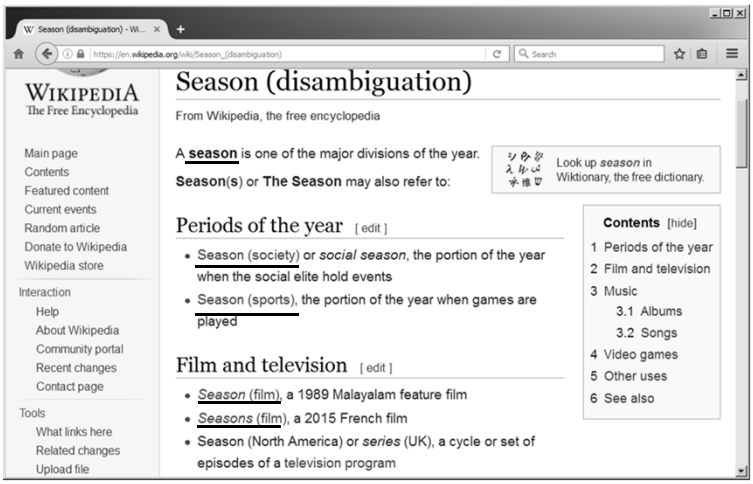


Category Network

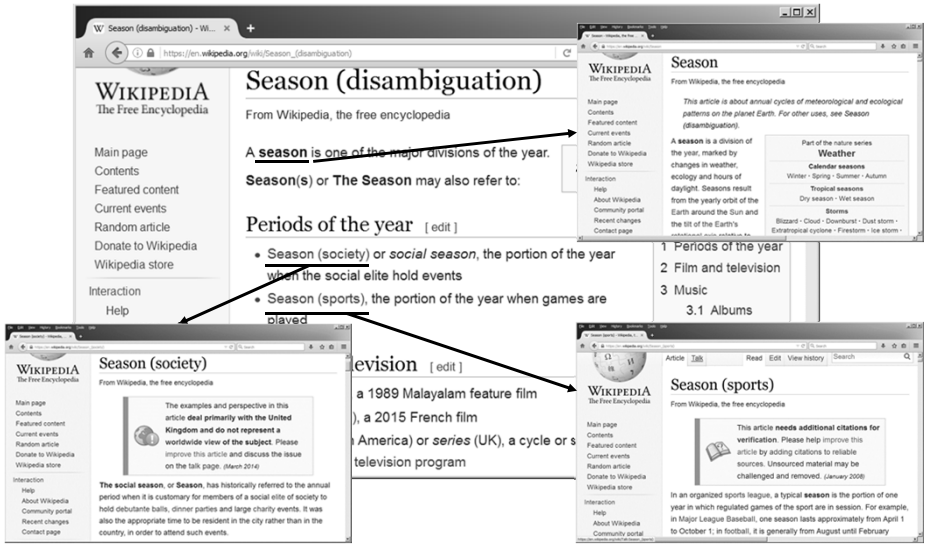
- Edges from articles to their parent categories, and from categories to their own parent categories



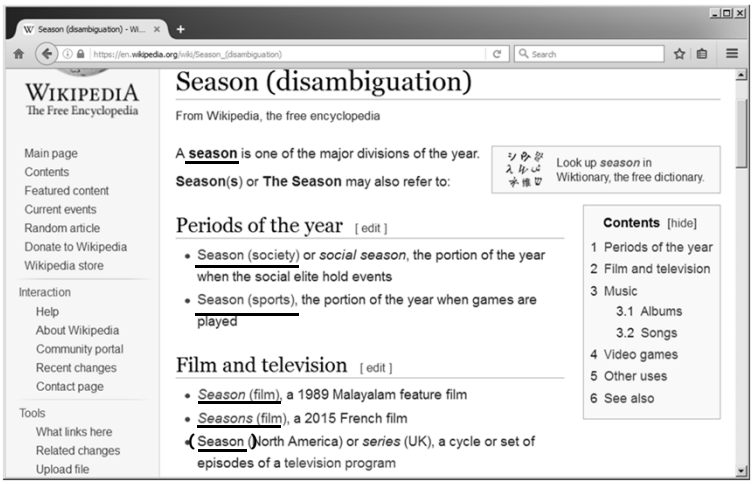
Disambiguation Pages



Connecting Phrases to Articles



(Not) Connecting Phrases to (Missing) Articles



Part One: Wikipedia as a Knowledge Resource

- Knowledge resources
- Dissecting Wikipedia
- Resources derived from Wikipedia

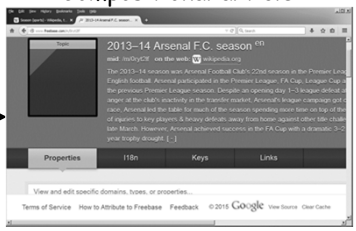
Knowledge Resources

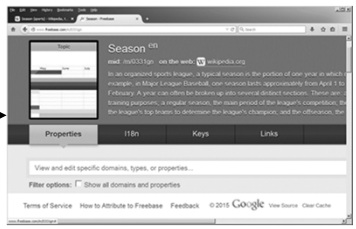
- Derived from Wikipedia
  - [BLK+09] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer et al. DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics 2009.
  - [BEP+08]: K. Bollacker, C. Evans, P. Paritosh et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. SIGMOD-08.
  - [VK14]: D. Vrandecic and M. Krotzsch. Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM 2014.

Deriving from Wikipedia

Compositional topic

Compositional article





Next Part

- Part One: Wikipedia as a Knowledge Resource
- Part Two: Role of Wikipedia in Text Analysis
- Part Three: Role of Wikipedia in Information Retrieval

Role of Wikipedia in Text Analysis

Methods for:

...

- coreference resolution
- word sense disambiguation
- entity disambiguation and linking
- information extraction

...

Role of Wikipedia in Text Analysis

- [PS06]: S. Ponzetto and M. Strube. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. HLT-NAACL-06.
  - determine the relative and combined impact in the task of coreference resolution, for features derived from different resources
- [Mih07]: R. Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. NAACL-HLT-07.
  - over a small evaluation set, validate the hypothesis that manually-created mappings from Wikipedia articles to corresponding WordNet synsets can improve the outcome of word sense disambiguation
- [PS07]: S. Ponzetto and M. Strube. Deriving a Large Scale Taxonomy from Wikipedia. AAAI-07.
  - apply filters to network of Wikipedia categories to extract hierarchy of categories
- [WW07]: F. Wu and D. Weld. Autonomously Semantifying Wikipedia. CIKM-07.
  - extend Wikipedia infoboxes with attributes and values inferred from text
- [SKW07]: F. Suchanek, G. Kasneci and G. Weikum. Yago: a Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW-07.
  - map Wikipedia categories to WordNet to generate hybrid resource of concepts and relations
- [WW08]: F. Wu and D. Weld. Automatically Refining the Wikipedia Infobox Ontology. WWW-08.
  - extend Wikipedia infoboxes with additional attributes and values, by mapping templates of Wikipedia infoboxes to WordNet
- [NS08]: V. Nastase and M. Strube. Decoding Wikipedia Categories for Knowledge Acquisition. AAAI-08.
  - from categories and category network, derive relations among categories or instances, including attributes of categories
- [SSW09]: F. Suchanek, M. Sozio and G. Weikum. Sofie: A Self-Organizing Framework for Information Extraction. WWW-09.
  - extend existing repositories of relations like Wikipedia, with facts acquired from unstructured text

Role of Wikipedia in Text Analysis

- [YTK+09]: I. Yamada, K. Torisawa, J. Kazama, K. Kuroda, M. Murata, S. De Saeger, F. Bond and A. Sumida. Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures. ACL-IJCNLP-09.
  - extract IsA pairs from Web documents, by using lexico-syntactic patterns and distributional similarities, and attach extracted pairs to Wikipedia categories
- [YOM+09]: Y. Yan, N. Okazako, Y. Matsuo et al. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. ACL-IJCNLP-09.
  - identify relevant relations for Wikipedia categories, from parsed Wikipedia articles and from Web documents via search engines
- [PN09]: S. Ponzetto and R. Navigli. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. IJCAI-09.
  - map Wikipedia categories to WordNet synsets, and use mappings to restructure the hierarchy generated in [PS07]
- [BZ10]: R. Blanco and H. Zaragoza. Finding Support Sentences for Entities. SIGIR-10.
  - loosely identify the relation between given a query and a given instance, in the form of explanatory sentences collected from Wikipedia articles
- [WW10]: F. Wu and D. Weld. Open Information Extraction Using Wikipedia. ACL-10.
  - from unstructured text, extract relations whose types are derived from Wikipedia
- [NP10]: R. Navigli and S. Ponzetto. BabelNet: Building a Very Large Multilingual Semantic Network. ACL-10.
  - link Wikipedia articles to WordNet concepts and apply machine translation, to create a multi-lingual repository of relations
- [TP10]: P. Talukdar and F. Pereira. Experiments in Graph-Based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL-10.
  - extract IsA pairs from manually-created or automatically-extracted repositories, via graph propagation, by incorporating structured data derived from Wikipedia

Role of Wikipedia in Text Analysis

- [RRD+11]: L. Ratinov, D. Roth, D. Downey and M. Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. ACL-11.
  - compare the impact of algorithms for disambiguating instances mentioned in a document relative to articles in Wikipedia, using evidence available locally for each mention vs. globally for all mentions
- [RR12]: L. Ratinov and D. Roth. Learning-based Multi-Sieve Co-reference Resolution with Knowledge. EMNLP-CoNLL-12.
  - improve coreference resolution via context-sensitive disambiguation of mentions in text to corresponding Wikipedia articles
- [HSB+13]: J. Hoffart, F. Suchanek, K. Berberich and G. Weikum. YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artificial Intelligence 2013.
  - combine WordNet, Wikipedia and other sources into hierarchically organized instances and their relations, where the data is anchored in time and space
- [FVP+14]: T. Flati, D. Vannella, T. Pasini and R. Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. ACL-14.
  - from network of Wikipedia articles and Wikipedia categories, extract hierarchy of articles and categories
- [DGH+14]: X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao and K. Murphy. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. KDD-14.
  - create a knowledge repository, based on relations extracted from Web documents and knowledge from available repositories
- [VMT+15]: N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke and W. Weerkamp. Learning to Explain Entity Relationships in Knowledge Graphs. ACL-15.
  - from Web documents, extract textual descriptions of relations between entries in pairs of entries from a knowledge repository

Role of Wikipedia in Text Analysis

- [BM16] S. Banerjee and P. Mitra. WikiWrite: Generating Wikipedia Articles Automatically. IJCAI-16.
  - given a new topic not already in Wikipedia, generate a new article whose sections are derived from sections of existing topics similar to the new topic, and filled in with summaries of Web text fragments
- [RBN16] A. Raganato, C. Delli Bovi and R. Navigli. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. IJCAI-16.
  - create links between mentions within Wikipedia articles not already linked, and their corresponding Wikipedia articles

Coreference Resolution

- [PS06]: S. Ponzetto and M. Strube. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. HLT-NAACL-06.

Semantic Features for Coreference

- Input
  - discourse entities and possibly referring expressions in documents
- Data sources
  - collection of documents annotated with coreference chains
  - WordNet
  - collection of articles from Wikipedia
  - category network in Wikipedia, containing edges from articles to their parent categories, and from categories to their own parent categories
- Output
  - coreference chains linking referring expressions and discourse entities to which they refer
- Steps
  - besides lexical, grammatical and distance features, collect features from output of semantic role labeler applied to documents; from WordNet; and from Wikipedia
  - exploit features in classifying pairs of possibly coreferring expressions as being coreferent or not

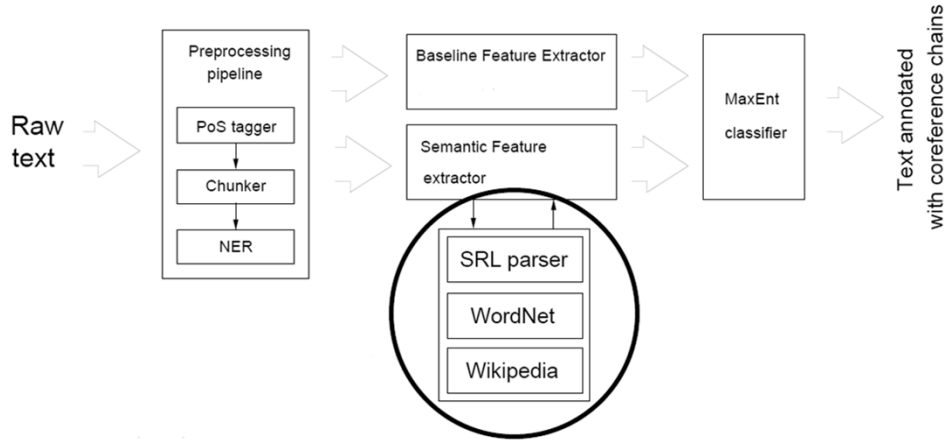
Referring Expressions

But frequent visitors say that given the sheer weight of [the country]'s totalitarian ideology and generations of mass indoctrination, changing [this country]'s course will be something akin to turning a huge ship at sea. Opening [North Korea] up, even modestly, and exposing [people] to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as [his people] begin to get a clearer idea of the deprivation [they] have suffered, especially relative to [their] neighbors. [This] is a [society] that has been focused most of all on stability, [...].

Coreference Chains

But frequent visitors say that given the sheer weight of the country's totalitarian ideology and generations of mass indoctrination, changing this country's course will be something akin to turning a huge ship at sea. Opening North Korea up, even modestly, and exposing people to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as his people begin to get a clearer idea of the deprivation they have suffered, especially relative to their neighbors. This is a society that has been focused most of all on stability, [...].

Finding Coreference Chains



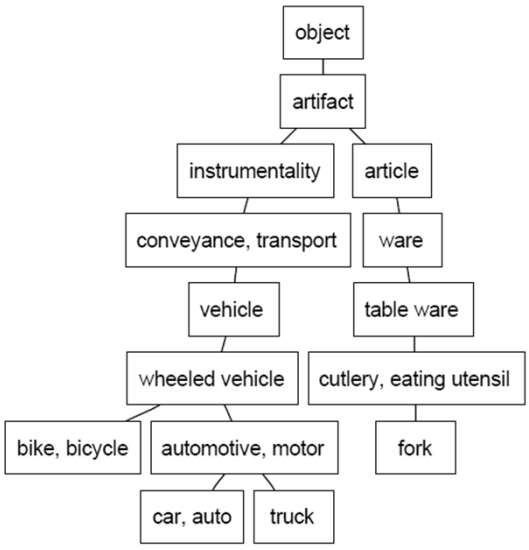
(Courtesy S. Ponzetto)

Evidence from Semantic Role Labeling

- Features: semantic role argument-predicate pairs of the referring expressions

A state commission of inquiry into the sinking of the Kursk will convene in *Moscow* on Wednesday,
the Interfax news agency A0/report reported. It A0/say said that the diving operation will be completed by the end of next week.

Evidence from WordNet



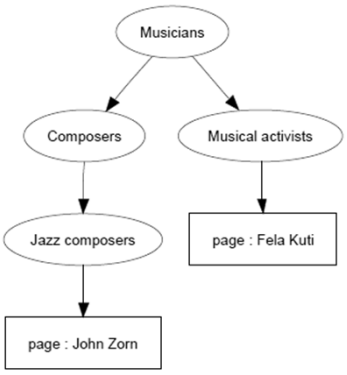
- Features: largest, average similarity score among all pairs of WordNet synsets to which pair of referring expressions belong
- Based on path length:
  sim(car, auto) = 1
  sim(car, bike) = 0.25

Evidence from Wikipedia

- Features from pairs of Wikipedia articles, if any, retrieved for the two referring expressions
  - feature: whether first paragraph of the article of one expression contains the other expression
  - feature: whether an outgoing internal link of the article of one expression is an article that contains the other referring expression
  - feature: whether categories of the article of one expression contain the other expression as an article
  - feature: ngram overlap score between first paragraphs of the articles of the two expressions

Evidence from Wikipedia

- Relatedness among Wikipedia articles
  - path length, via ancestor categories, between Wikipedia articles



- Features from Wikipedia category network
  - features: largest, average relatedness score

Relative and Combined Contribution

- Evaluation set of ACE news documents

Method	R	P	F_1	A_p	A_{cn}	A_{pn}
Baseline	54.5	88.0	67.3	34.7	20.4	53.1
+SRL	56.3	88.4	68.8	35.6	28.5	49.6
+WordNet	56.7	87.1	68.6	34.8	26.0	50.5
+Wikipedia	55.8	87.5	68.1	38.9	21.6	51.7
(All)	61.0	84.2	70.7	38.9	29.9	51.2

(A=accuracy rate; p=pronouns; cn=common nouns; pn=proper nouns)

Word Sense Disambiguation

- [Mih07]: R. Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. NAACL-HLT-07.

Wikipedia in Word Sense Disambiguation

- Input
  - ambiguous words in documents
- Data sources
  - collection of documents annotated with word senses
  - WordNet
  - collection of articles from Wikipedia
- Output
  - disambiguation of words into WordNet synsets
- Steps
  - for a small set of ambiguous words, manually create mappings from Wikipedia articles that likely capture possible senses of the words, and WordNet synsets
  - based on the mappings, collect additional features to disambiguate the ambiguous words as they occur in documents

Mappings from Wikipedia to WordNet

- For a given word of interest:

Titles of Wikipedia articles pointed to from anywhere within Wikipedia, where the anchor text of the pointing link is the word

Definition from Wikipedia disambiguation page of the word

Manually selected WordNet synset corresponding to Wikipedia article

Word Sense	Labels in Wikipedia	Wikipedia definition	WordNet definition
bar (counter)	bar (counter)	the counter from which drinks are dispensed	a counter where you can obtain food or drink
bar (music)	bar (music), measure music, musical notation	a period of music	musical notation for a repeating pattern of musical beats
bar (landform)	bar (landform)	a type of beach behind which lies a lagoon	a submerged (or partly submerged) ridge in a river or along a shore

(Courtesy R. Mihalcea)

Word Sense Disambiguation via Wikipedia

- Set of 30 nouns with at least two WordNet senses for Wikipedia labels

	#s	#ex	MFS	LeskC	WSD
argument	2	114	70.17%	73.63%	**89.47%**
arm	2	291	61.85%	69.31%	**84.87%**
bank	3	1074	**97.20%**	**97.20%**	**97.20%**
bar	10	1108	47.38%	68.09%	**83.12%**
circuit	4	327	85.32%	85.62%	**87.15%**
degree	7	849	58.77%	73.05%	**85.98%**
stress	3	565	53.27%	54.28%	**86.37%**
Average	3.31	316	72.58%	78.02%	**84.65%**

Number of senses

Number of examples

Most frequent sense

Lesk corpus

Wikipedia vs. WordNet

- Comparison of Senseval word senses vs. Wikipedia word senses

	#s	#ex	MFS	LeskC	WSD
Senseval	4.6	226	51.53%	58.33%	**68.13%**
Wikipedia	3.31	316	72.58%	78.02%	**84.65%**

Documents annotated with WordNet synsets

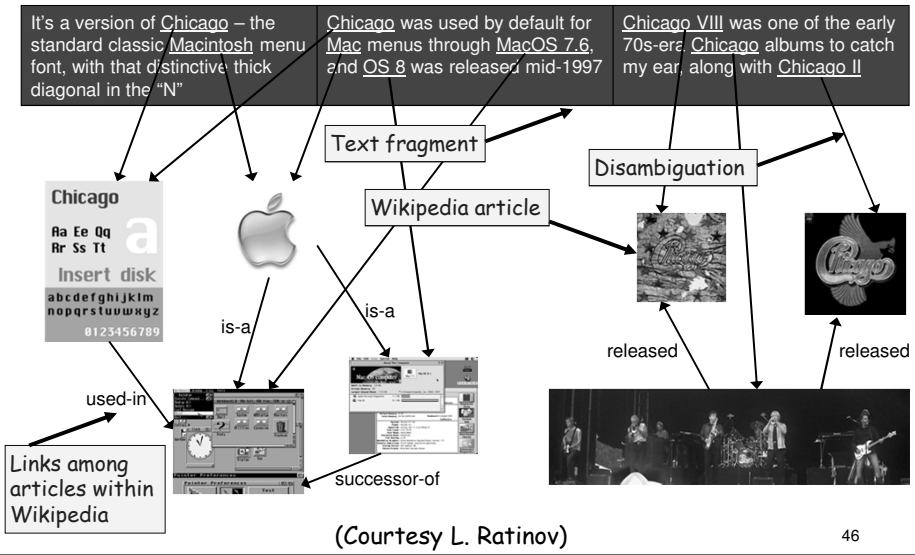
Entity Disambiguation and Linking

- [RRD+11]: L. Ratinov, D. Roth, D. Downey and M. Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. ACL-11.

Disambiguation to Wikipedia

- Task
  - given a text fragment containing mentions (substrings) to be disambiguated, "wikifi" the mentions by identifying the Wikipedia article, if any, corresponding to each mention
  - mapping from mentions to Wikipedia articles relies on evidence available in the text fragment
- Scope of available evidence
  - local: separately available for each mention in the text fragment
  - global: collectively available for all mentions in the text fragment
- Goal
  - investigate impact of local vs. global evidence on accuracy of disambiguation

Mapping Mentions to Wikipedia



(Courtesy L. Ratinov)

Disambiguation Strategy

Algorithm: Disambiguate to Wikipedia
Input: document d, Mentions $M = \{m_1, \ldots, m_N\}$
Output: a disambiguation $\Gamma = (t_1, \ldots, t_N)$.
1) Let $M' = M \cup$ { *Other potential mentions in* d}
2) For each mention $m'_i \in M'$, *construct a set of disambiguation candidates* $T_i = \{t^i_1, \ldots, t^i_{k_i}\}, t^i_j \neq$ null
3) **Ranker**: *Find a solution* $\Gamma = (t'_1, \ldots, t'_{|M'|})$, *where* $t'_i \in T_i$ *is the best non-null disambiguation of* m'_i.
4) **Linker**: *For each* m'_i, *map* t'_i *to* null *in* Γ iff *doing so improves the objective function*
5) Return Γ *entries for the original mentions* M.

- Two stages
  - ranker: compute best Wikipedia article that potentially disambiguates the mention
  - linker: determine whether the mention should be mapped to the Wikipedia article or should not be mapped to any article

Ranker: Local vs. Global Disambiguation

Accuracy: fraction of mentions for which ranker identifies correct disambiguation

Dataset	Baseline	Baseline+ Lexical	Baseline+ Global Unambiguous	Baseline+ Global NER	Baseline+ Global, All Mentions
ACE	94.05		94.56	96.21	96.75
MSNBC News	81.91		84.46	84.04	88.51
AQUAINT	93.19		95.40	94.04	95.91
Wikipedia Test	85.88		89.67	89.59	89.79

Previous methods

Ranker: Local vs. Global Disambiguation

Accuracy: fraction of mentions for which ranker identifies correct disambiguation

Dataset	Baseline	Baseline+ Lexical	Baseline+ Global Unambiguous	Baseline+ Global NER	Baseline+ Global, All Mentions
ACE	94.05	96.21			96.75
MSNBC News	81.91	85.10			88.51
AQUAINT	93.19	95.57			95.91
Wikipedia Test	85.88	93.59			89.79

Over test set of Wikipedia documents, local performs better than global

Local disambiguation

Global disambiguation

Overall: Local vs. Global Evidence

Combined precision and recall (F1 score)

Dataset	Baseline	Baseline+ Lexical	Baseline+ Lexical+ Global
ACE	94.05	96.21	97.83
MSNBC News	81.91	85.10	87.02
AQUAINT	93.19	95.57	94.38
Wikipedia Test	85.88	93.59	94.18

(Comparing set of Wikipedia articles output by algorithm for a document, with gold set of Wikipedia articles for the document)

Role of Wikipedia in Text Analysis

Methods for:

...

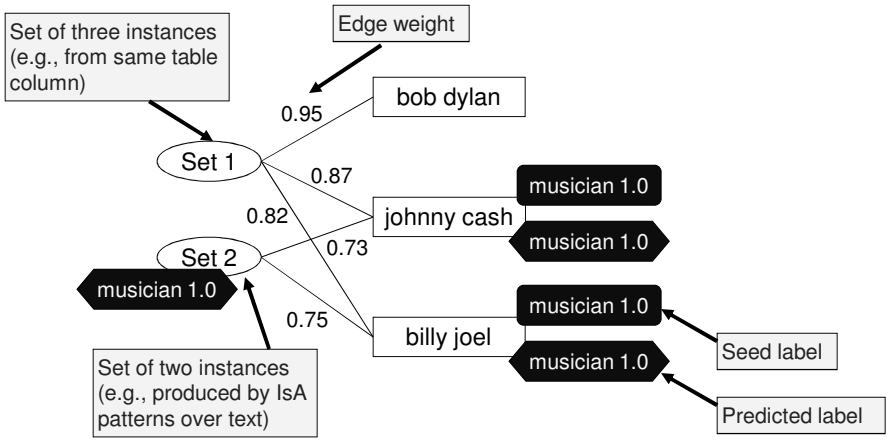
- coreference resolution
- word sense disambiguation
- entity disambiguation and linking
- information extraction

...

Information Extraction

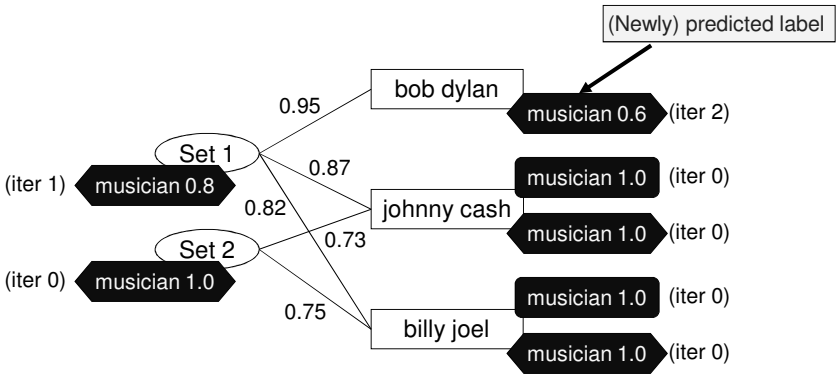
- [TP10]: P. Talukdar and F. Pereira. Experiments in Graph-Based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL-10.

Graph-Based Assignment of Class Labels

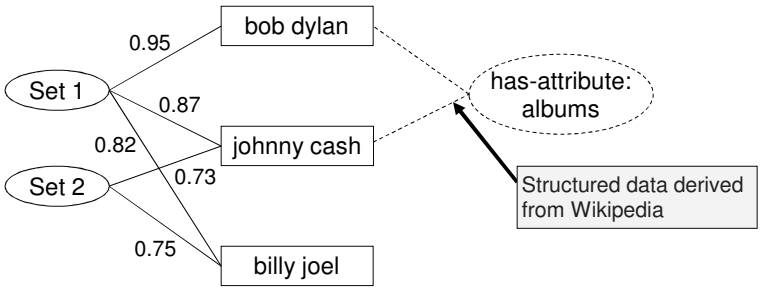


(Courtesy P. Talukdar)

Graph-Based Assignment of Class Labels



Incorporating Semantic Constraints

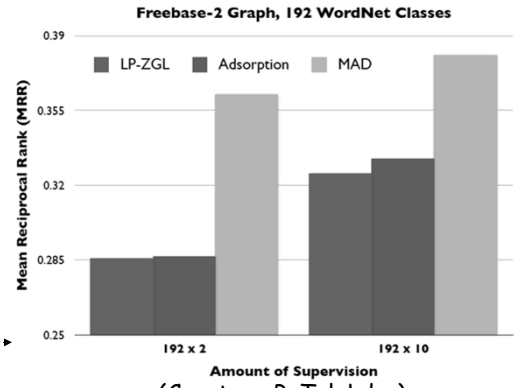


Extracted Class Labels

- Input data = graphs derived from existing sources
- Evaluate lists of class labels assigned to instances: mean reciprocal rank of gold class labels

Source	Derived Graph	
	Vertices	Edges
Freebase-1	32K	957K
Freebase-2	301K	2310K
TextRunner	175K	529K
Yago	142K	777K
TextRunner+Yago	237K	1307K

(LP-ZGL, Adsorption = earlier methods)



(Courtesy P. Talukdar)

Information Extraction

- [PS07]: S. Ponzetto and M. Strube. Deriving a Large Scale Taxonomy from Wikipedia. AAAI-07.

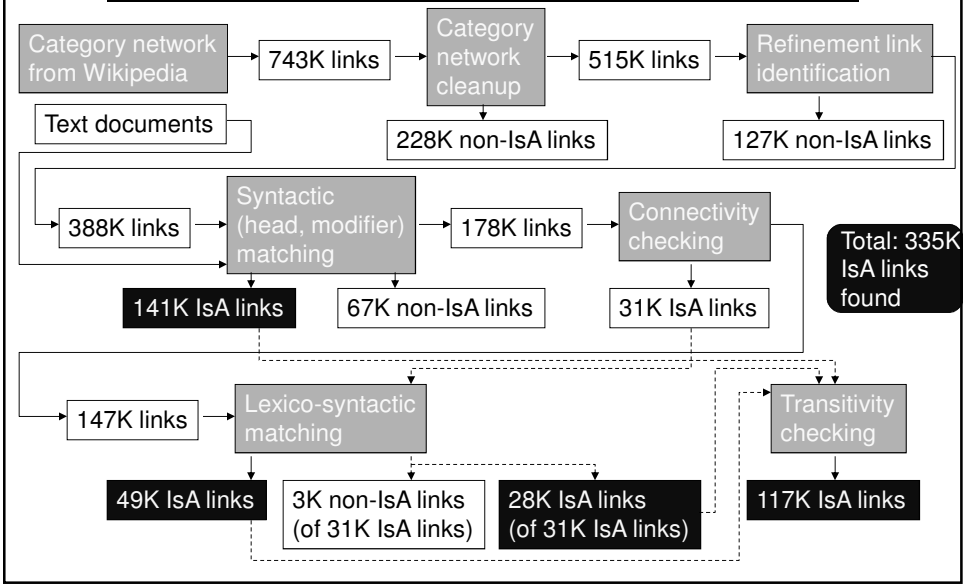
Hierarchy of Wikipedia Categories

- Input
  - category network in Wikipedia, containing edges from articles to their parent categories, and from categories to their own parent categories
    - e.g., Semantic relatedness belongs to the categories Statistical distance measures and Computational linguistics; Statistical distance measures belongs to the category Statistics
- Data sources
  - collection of text documents
  - collection of Wikipedia articles
- Output
  - hierarchy of Wikipedia categories
    - e.g., [Computer scientists IsA Scientists, Caffeine IsA Stimulant,…] organized hierarchically
- Steps
  - apply various types of filters to network of categories, to identify IsA (vs. non-IsA) relations among the pairwise category relations

Filters

- Category network cleanup
  - mark meta-categories (e.g., Disambiguation pages, Weather templates) as non-IsA links
- Identification of refinement links
  - identify links from category Y X (e.g., French cuisine) to category X by Z (e.g., Cuisine by Nationality) as refinement links, and mark them as non-IsA links
- Syntactic matching (head of category, modifier of category)
  - mark links from category Y X (e.g., British computer scientists) to category Z X (e.g., Computer scientists) as IsA links
  - mark links from category X Y (e.g., Islamic dress) to category Z X, where Z is optional (e.g., Islam) as non-IsA links
- Connectivity checking
  - mark links from any leaf (non-category) article to a category as IsA links
- Lexico-syntactic matching
  - identify links from category X to category Y, where Y and X occur in text documents in one of a few PartOf patterns (e.g., Y's X, Y with X)
  - identify links from category X to category Y, where Y and X occur in text documents in one of a few IsA patterns (e.g., Y such as X)
  - if PartOf patterns occur less frequently than IsA patterns, then mark links from category X to category Y as IsA links, otherwise as non-IsA links
- Transitivity checking
  - if link from category X to category Y was marked as IsA link, and link from category Y to category Z was marked as IsA link, then mark link from category X to category Z as IsA link

Filters



Information Extraction

- [PN09]: S. Ponzetto and R. Navigli. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. IJCAI-09.

Hierarchy Refinement and Restructuring

- Input
  - hierarchy of Wikipedia categories generated in [PS07]
    - e.g., [Computer scientists IsA Scientists, Caffeine IsA Stimulant,...] organized hierarchically
  - hierarchy from WordNet
- Output
  - restructured hierarchy of Wikipedia categories, with categories mapped to WordNet synsets
- Steps
  - map Wikipedia categories to WordNet synsets
  - use mappings to restructure hierarchy of Wikipedia categories

Hierarchy Refinement and Restructuring

- Map categories from Wikipedia hierarchy to WordNet synsets
  - identify sets of candidate WordNet synsets
    - full match, e.g., category Plants to synsets available for plant
    - head match, e.g., category Edible plants to synsets available for plant
  - select best candidate WordNet synsets
    - e.g., for category Plants, select second synset available for plant (a living organism lacking the power of locomotion)
    - equivalent to finding most appropriate sense for Wikipedia category, from senses available in WordNet
    - as an alternative to choosing the most frequent of the available WordNet senses
- Use mappings to restructure hierarchy of Wikipedia categories
  - identify links in Wikipedia hierarchy whose corresponding links in WordNet hierarchy are the most inconsistent
    - e.g., [Fruits-->Crops-->Edible plants-->Plants] in Wikipedia hierarchy, but no [fruit-->crop] or [fruit-->plant] in WordNet
  - identify replacements of most inconsistent links, by finding alternative parent categories
    - e.g., check for possible replacements of [Fruits-->Crops]

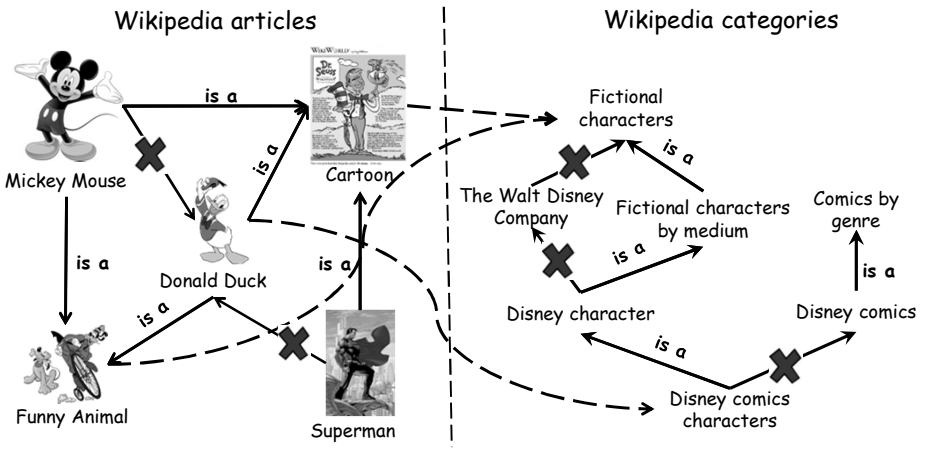
Information Extraction

- [FVP+14]: T. Flati, D. Vannella, T. Pasini and R. Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. ACL-14.

Constructing Hierarchies from Text

- Data source
  - category network in Wikipedia, containing edges from articles to their parent categories, and from categories to their own parent categories; also edges from articles to other articles
- Output
  - an hierarchy containing IsA edges between Wikipedia articles, and another hierarchy containing IsA edges between Wikipedia categories
- Steps
  - analyze the text of each article to select a candidate hypernym phrase for the article (e.g., character for Mickey Mouse)
  - disambiguate the hypernym phrase into another article corresponding to the desired sense of the hypernym phrase (e.g., character into Character (arts))
  - organize pairs of an article and its disambiguated hypernym article into hierarchy of articles
  - starting from the hierarchy of articles (e.g., Real Madrid IsA Football club), exploit Wikipedia edges between articles and categories to iteratively infer IsA edges between categories of articles (e.g., Football clubs in Madrid IsA Football clubs), and then IsA edges between articles of categories (e.g., Atletico Madrid IsA Football club)
  - expand the hierarchy of categories to increase its coverage

Edges in Category Network



(Courtesy T. Flati)

Constructing Hierarchy of Articles

- Select main (first) sentence in article (Courtesy T. Flati)

Scrooge McDuck

From Wikipedia, the free encyclopedia

Scrooge McDuck is a cartoon character created in 1947 by Carl Barks and licensed by The Walt Disney Company. Scrooge is an elderly Scottish anthropomorphic white duck with a yellow-orange bill, legs, and feet.

- From the dependency parse of main sentence, select candidate hypernym phrases

Scrooge McDuck is a character […] ⟶ Scrooge McDuck is a character[…] ⟶ Hypernym phrase: character

(nn, nsubj, cop)

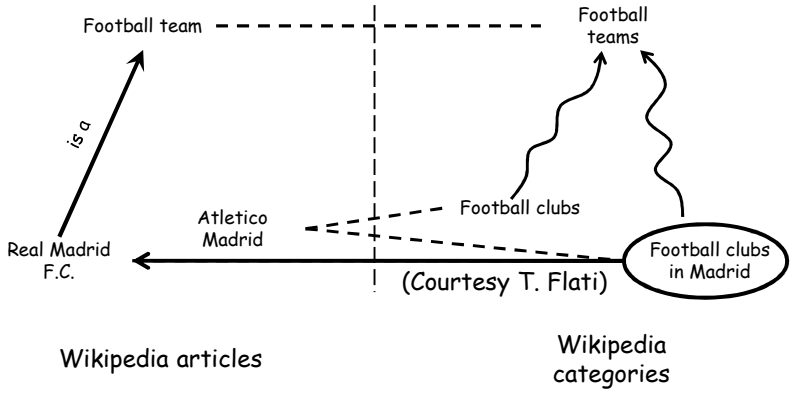
- Using heuristics, disambiguate the hypernym phrase into another article corresponding to the desired sense
  - heuristics rely on links among articles, common categories, context around links

Hypernym phrase: character ⟶ Character (arts)

- Retain pairs of an article and its hypernym article, as IsA edges

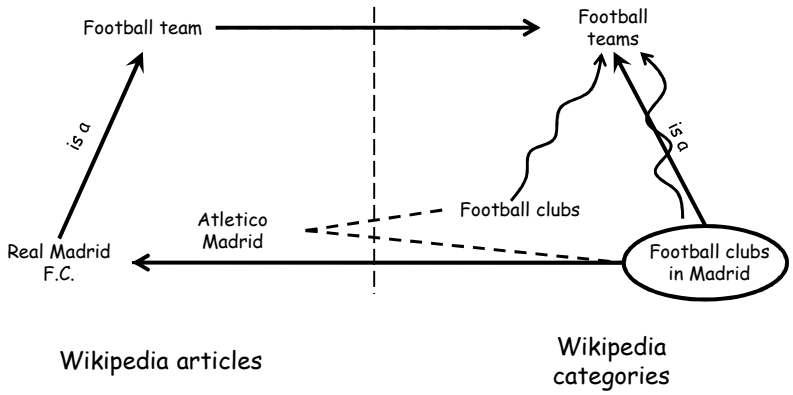
Scrooge McDuck ⟶ Character (arts)

Constructing Hierarchy of Categories



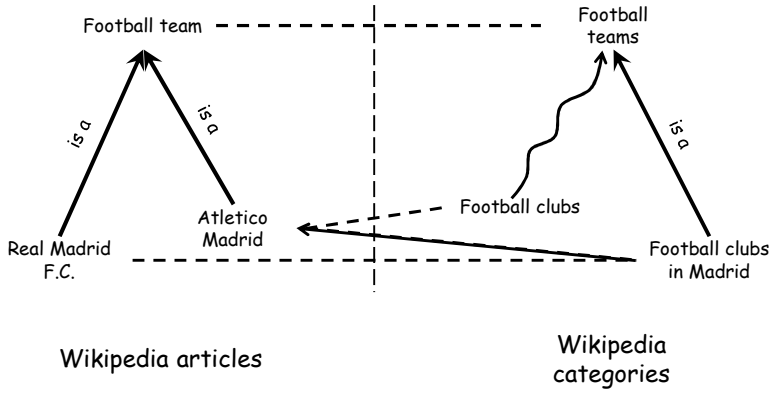
Start from the hierarchy of articles

Constructing Hierarchy of Categories



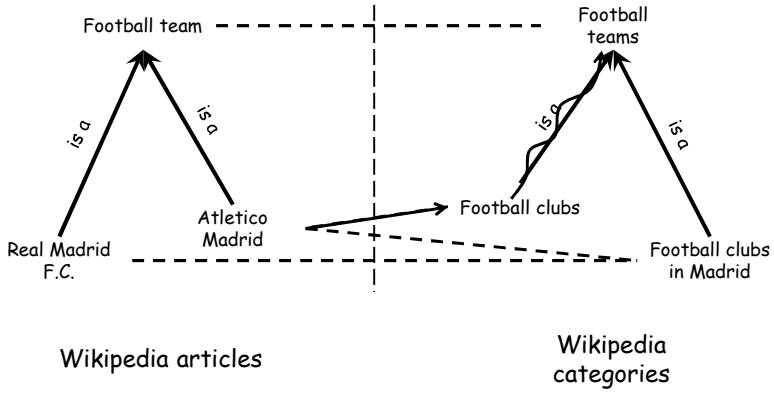
Exploit the article to category edges to infer hypernym relations in the category hierarchy

Iteratively Refining the Hierarchies



Traverse article to category edges to infer back is-a relations in the hierarchy of articles

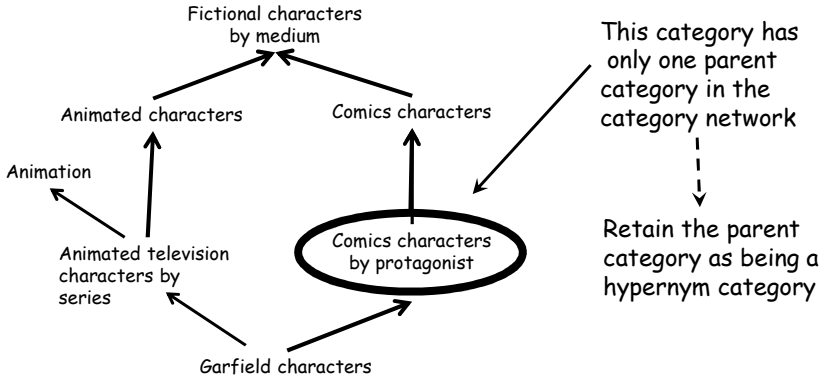
Iteratively Refining the Hierarchies



Iteratively use the relations found in previous step to infer new hypernym edges

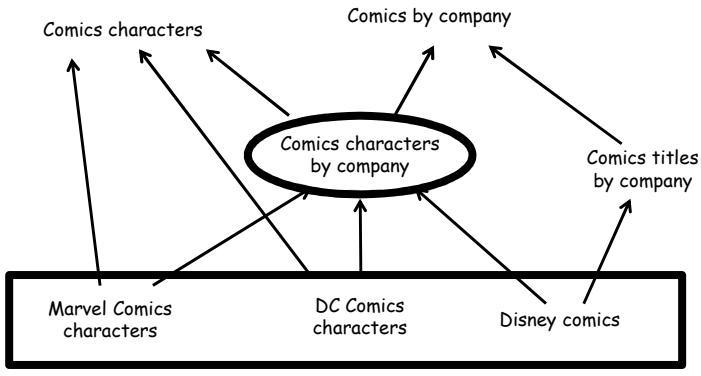
Expanding the Hierarchy of Categories

- Expand for categories with a single parent category in the category network



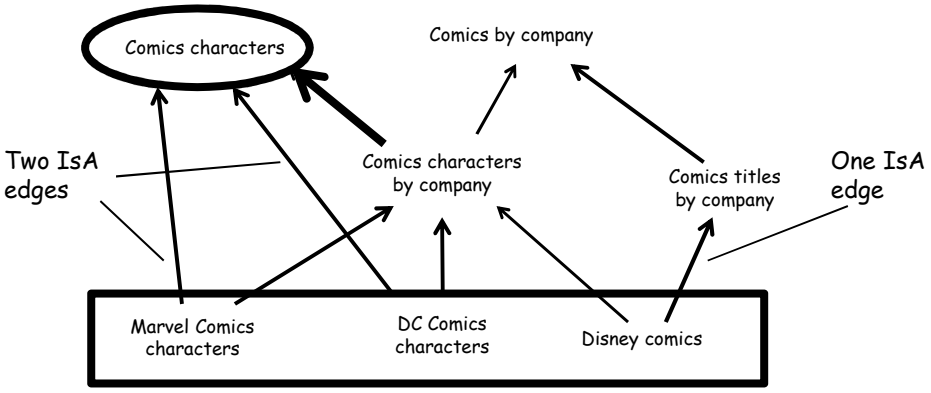
Expanding the Hierarchy of Categories

- Infer hypernym categories of a category, from hypernym categories already extracted for child categories in the category network



Expanding the Hierarchy of Categories

- Infer hypernym categories of a category, from hypernym categories already extracted for child categories in the category network



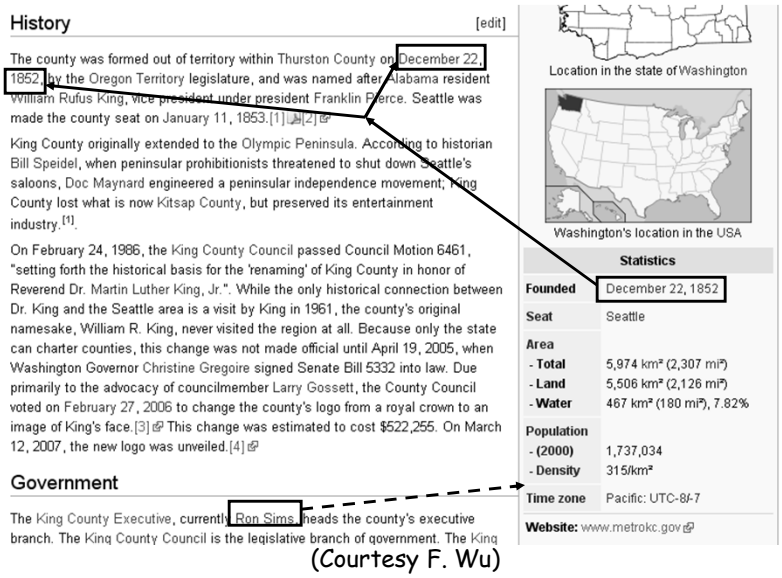
Information Extraction

- [WW07]: F. Wu and D. Weld. Autonomously Semantifying Wikipedia. CIKM-07.

Creation and Augmentation of Infoboxes

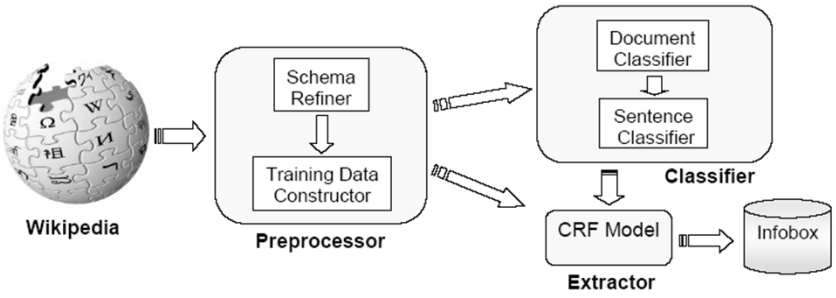
- Motivation
  - Wikipedia articles may lack or have incomplete infoboxes
  - content from an infobox may contradict content from article
  - similar articles may incorrectly use different infobox templates
- Data source
  - collection of articles from Wikipedia
- Output
  - Wikipedia infoboxes extended with inferred attributes and values

Processing Infoboxes



(Courtesy F. Wu)

Extraction from Wikipedia Articles



(Courtesy F. Wu)

Extraction Components

- Preprocessor
  - identify relevant attributes from articles using the same infobox template
  - generate training data for classification and extraction
- Document classifier
  - determine whether an article belongs to a certain class
  - one classifier per class of articles
- Sentence classifier
  - determine whether a sentence contains the value of an attribute
  - one classifier per attribute per infobox template
- Extractors
  - given a sentence deemed to contain a value, extract the value
  - aggregate across sentences, return first/multiple values for single/multiple-valued attributes

Information Extraction

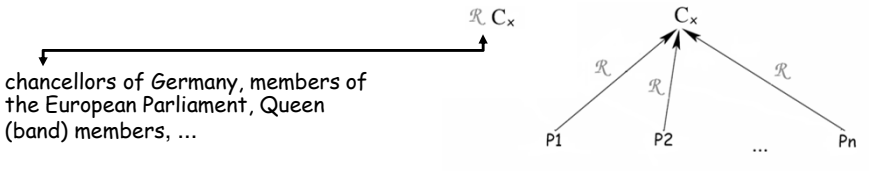
- [NS08]: V. Nastase and M. Strube. Decoding Wikipedia Categories for Knowledge Acquisition. AAAI-08.

Extraction from Wikipedia Categories

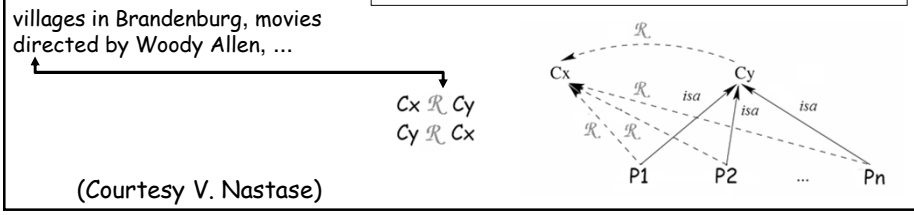
- Data source
  - Wikipedia category network
- Output
  - relations among categories or instances
    - e.g., <kind of blue-artist-miles davis>, <deconstructing harry-directed-by-woody allen>
  - ... including attributes of categories
    - e.g., {band, community, ethnicity, genre, instrument, language, region, ...} for Musician
  - filter the category network and analyze the categories
  - extract relations from categories, by matching category names to pre-defined patterns
  - extract attributes from categories, by matching category names to pre-defined patterns

Extraction from Wikipedia Categories

- Match category name with patterns identifying either a relation and a category, or a relation and two categories
- Given matches and the category network, derive relations

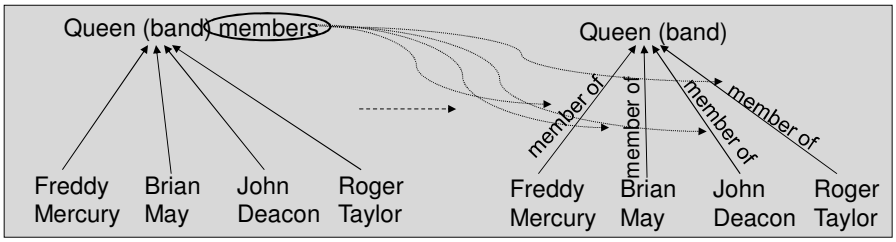


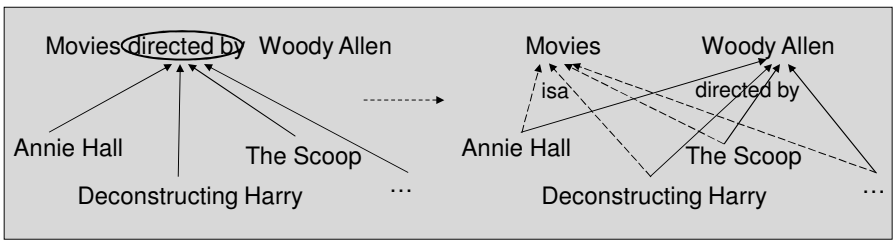
Cx, Cy = categories; P1, P2, ..., Pn = pages (articles)



(Courtesy V. Nastase)

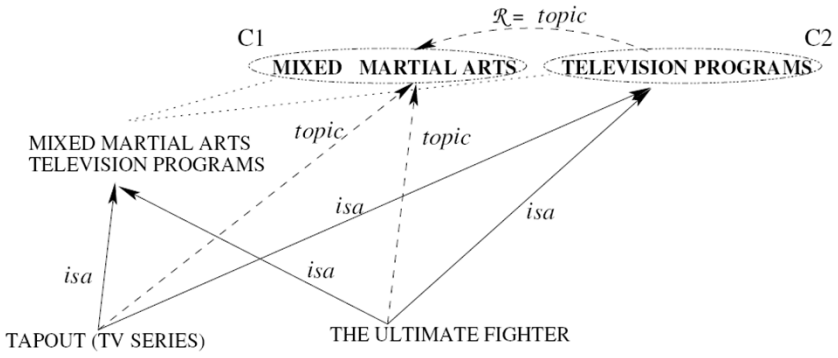
Extraction of Labeled Relations





Extraction of Unlabeled Relations

- Category name: mixed martial arts television programs



Information Extraction

- [VMT+15]: N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke and W. Weerkamp. Learning to Explain Entity Relationships in Knowledge Graphs. ACL-15.

From Relations to Explanatory Sentences

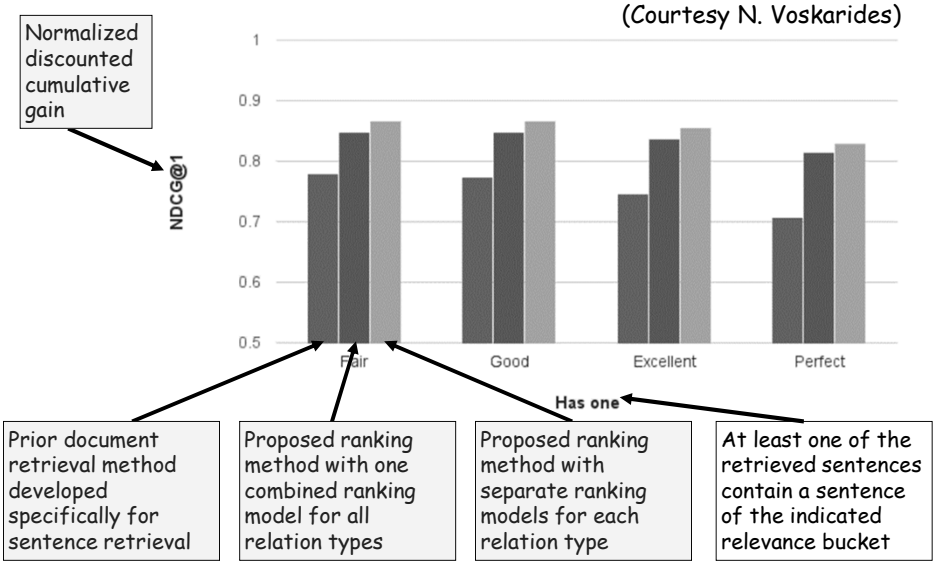
- Input
  - pairs of instances connected by a relation, as available in a knowledge repository (e.g., the pair Brad Pitt, Seven (1995 movie))
- Data source
  - Wikipedia articles corresponding to instances from the knowledge repository
- Output
  - ranked list of sentences extracted from documents, which explain the relations within the pairs of instances (e.g., "Brad Pitt gave critically acclaimed performances in the crime thriller Seven")
- Steps
  - extract surface forms for each instance in a pair of instances (e.g., Brad Pitt, Brad, Pitt)
  - extract candidate sentences from Wikipedia articles, using both surface forms and instances to which surface forms are disambiguated
  - rank candidate sentences using a variety of features

Ranking Candidate Sentences

- Textual features
  - length of candidate sentence
  - fractions of sentence tokens that are verbs vs. nouns vs. adjectives
  - ...
- Instance features
  - count of instances in candidate sentence
  - distance in tokens between last match of the two input instances in the candidate sentence
  - ...
- Relation features
  - whether candidate sentence contains tokens of the input relation
  - ...
- Document features
  - position of candidate sentence in source document
  - whether the source document of the candidate sentence is the article of one of the two input instances
  - ...

Accuracy of Explanatory Sentences

(Courtesy N. Voskarides)



Next Part

- Part One: Wikipedia as a Knowledge Resource
- Part Two: Role of Wikipedia in Text Analysis
- Part Three: Role of Wikipedia in Information Retrieval

Role of Wikipedia in Information Retrieval

Methods for:

...

- Document analysis and understanding
- Query analysis and understanding
- Onebox search results

...

Role of Wikipedia in Information Retrieval

- [Cuc07]: S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP-07.
  - disambiguate instances mentioned in a document relative to articles in Wikipedia, using evidence available in the form of local context for each mention
- [HWL+09]: J. Hu, G. Wang, F. Lochovsky, J. Sun and Z. Chen. Understanding User's Query Intent with Wikipedia. WWW-09.
  - model query intent domains as areas in the Wikipedia category network situated around manually-provided seed articles in Wikipedia, and map queries into those domains
- [PF11]: P. Pantel and A. Fuxman. Jigs and Lures: Associating Web Queries with Structured Entities. ACL-11.
  - compute mappings from queries into instances from a structured database, for the purpose of identifying relevant products from a product catalog and recommending them in response to queries
- [SMF+12]: U. Scaiella, A. Marino, P. Ferragina and M. Ciaramita. Topical Clustering of Search Results. WSDM-12.
  - take advantage of mappings from instances mentioned in documents to Wikipedia articles, in order to cluster search results and their result snippets into sets associated with descriptive labels
- [FS12]: P. Ferragina and U. Scaiella. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. IEEE Software, vol. 29, 2012.
  - disambiguate mentions in short text fragments to their Wikipedia articles, by scoring candidates of a given mention higher, if they are more strongly related to candidates of other mentions
- [HMB13]: L. Hollink, P. Mika and R. Blanco. Web Usage Mining with Semantic Analysis. WWW-13.
  - compute mappings from query fragments into instances from an existing knowledge repository, to better identify patterns of Web usage

Role of Wikipedia in Information Retrieval

- [YV14]: X. Yao and B. Van Durme. Information Extraction over Structured Data: Question Answering with Freebase. ACL-14.
  - in response to fact-seeking questions, extract answers from unstructured text from Web documents and from relations available in a knowledge repository
- [DAD14]: J. Dalton, J. Allan and L. Dietz. Entity Query Feature Expansion using Knowledge Base Links. SIGIR-14.
  - compute mappings from query fragments into instances from an existing knowledge repository, to expand queries for better search results
- [BMH+15]: B. Bi, H. Ma, B. Hsu, W. Chu, K. Wang and J. Cho. Learning to Recommend Related Entities to Search Users. WSDM-15.
  - given a query, compute and recommend related entries from a knowledge repository
- [BOM15]: R. Blanco, G. Ottaviano and E. Meij. Fast and Space-Efficient Entity Linking in Queries. WSDM-15.
  - compute mappings from query fragments into instances from an existing knowledge repository, under strong latency constraints
- [CFC+16] M. Cornolti, P. Ferragina, M. Ciaramita, S. Rud and H. Schutze. A Piggyback System for Joint Entity Mention Detection and Linking in Web Queries. WWW-16.
  - link mentions within queries to corresponding Wikipedia articles by jointly exploiting evidence available for multiple, co-occurring mentions being disambiguated
- [GGL+16] O. Ganea, and M. Ganea, A. Lucchi, C. Eickhoff and T. Hofmann. Probabilistic Bag-Of-Hyperlinks Model for Entity Linking. WWW-16.
  - link mentions within documents to corresponding Wikipedia articles by jointly exploiting evidence available for multiple, co-occurring mentions being disambiguated

Role of Wikipedia in Information Retrieval

- [RXC+16] M. Rizoiu, L. Xie, T. Caetano and M. Cebrian. Evolution of Privacy Loss in Wikipedia. WSDM-16.
  - recover privacy-sensitive traits of human contributors to Wikipedia, from their editing contributions over time
- [SF16] A. Sil and R. Florian. One for All: Towards Language Independent Named Entity Linking. ACL-16.
  - based on evidence from one language edition of Wikipedia, link mentions within documents to corresponding articles in a variety of language editions of Wikipedia
- [XRF16] K. Xu, S. Reddy, Y. Feng, S. Huang and D. Zhao. Question Answering on Freebase via Relation Extraction and Textual Evidence. ACL-16.
  - in response to fact-seeking questions, extract candidate answers from relations available in a knowledge repository and filter the candidates using evidence from unstructured text within Wikipedia

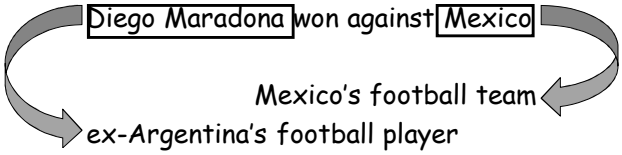
Document Understanding

- [FS12]: P. Ferragina and U. Scaiella. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. IEEE Software 2012.

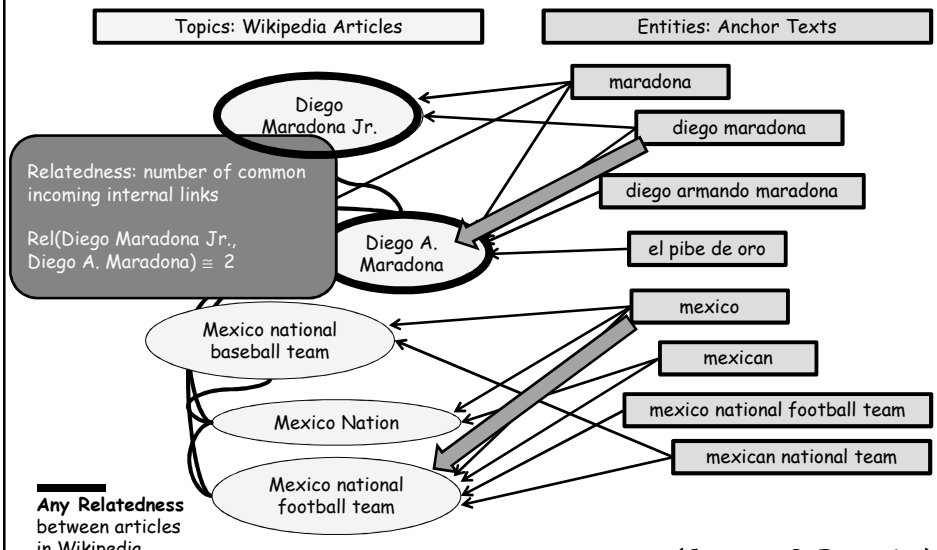
Linking in Short Text Fragments

- Input
  - short fragment of text
- Data source
  - Wikipedia articles and categories, connected via the category network
- Output
  - disambiguation of relevant mentions in text to their corresponding Wikipedia articles
- Steps
  - identify candidate mentions in text fragment
  - score candidate Wikipedia articles of a given mention higher, if they are more strongly related to candidate articles of other mentions
  - select highest-scoring Wikipedia article of a given mention, as candidate annotation of the mention
  - possibly discard the candidate annotation, if the anchor text is usually not linked within Wikipedia, or the candidate annotation is not coherent with candidate annotations selected for other mentions

Disambiguating Mentions in Text



Linking Mentions in Text to Wikipedia



(Courtesy P. Ferragina)

(Result) Document Understanding

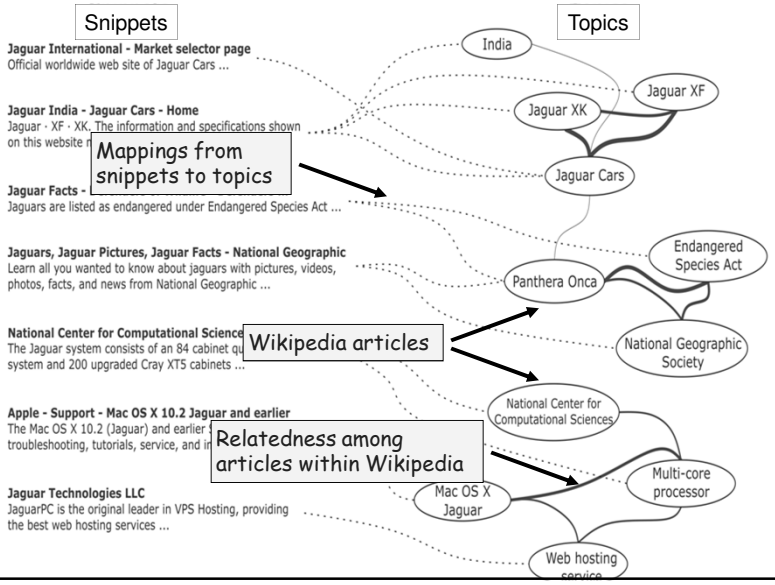
- [SMF+12]: U. Scaiella, A. Marino, P. Ferragina and M. Ciaramita. Topical Clustering of Search Results. WSDM-12.

Clustering of Search Results

- Input
  - search results and their snippets, returned in response to queries
- Data source
  - Wikipedia articles and categories, connected via the category network
- Output
  - decomposition of search results into topically coherent subsets associated with labels derived from Wikipedia
  - on the fly, without analysis of full content of search results
- Steps
  - annotate snippets with corresponding Wikipedia articles ("topics")
  - analyze graph of snippets and topics, to determine most significant topics
  - partition graph around most significant topics, and cut into ~10 clusters
  - for each cluster, select centroid topic as label for the entire cluster

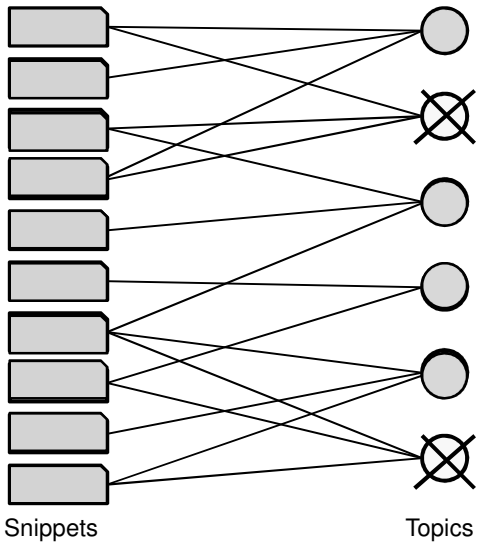
Annotation of Snippets with Topics

(Courtesy P. Ferragina)

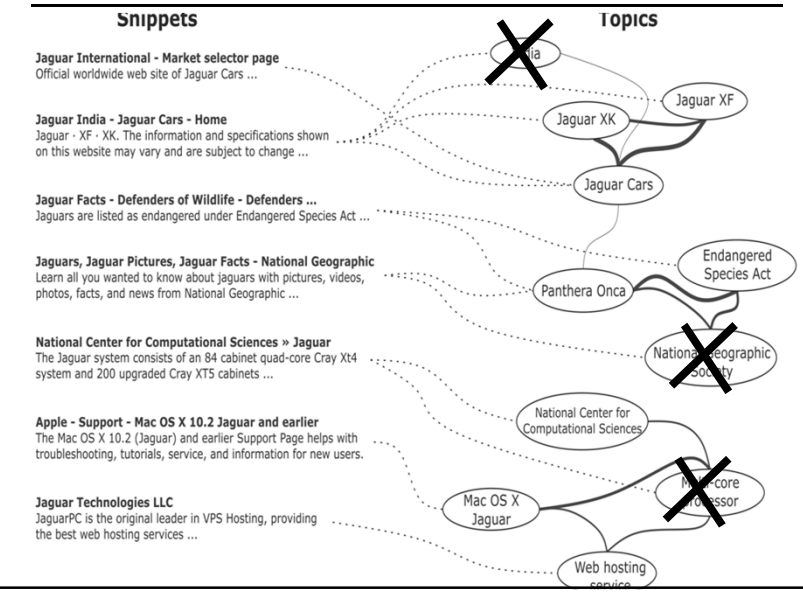


Selection of Significant Topics

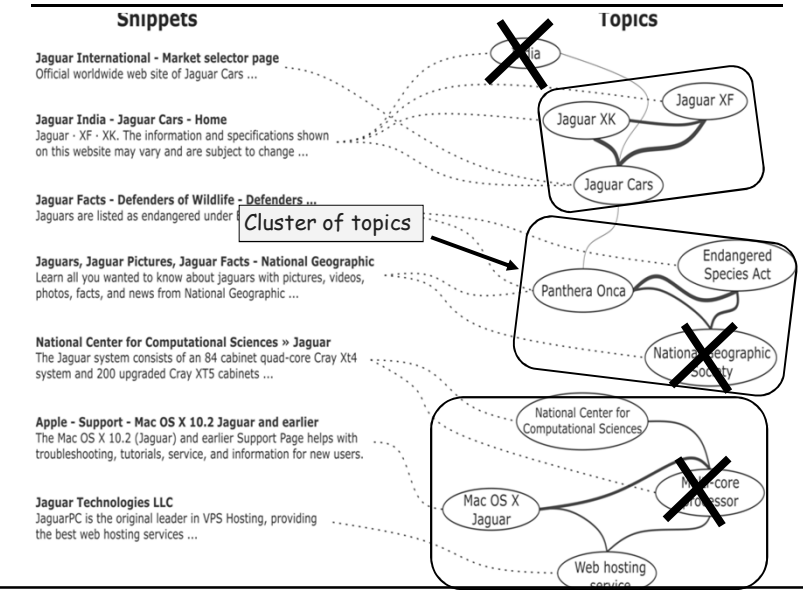
- Exploit weights assigned during annotation to mappings from snippets to topics
- Process topic nodes iteratively, in order of the sum of weights of connecting edges
- As topic nodes are marked as significant, ignore corresponding snippet nodes and their outgoing mappings in subsequent iterations



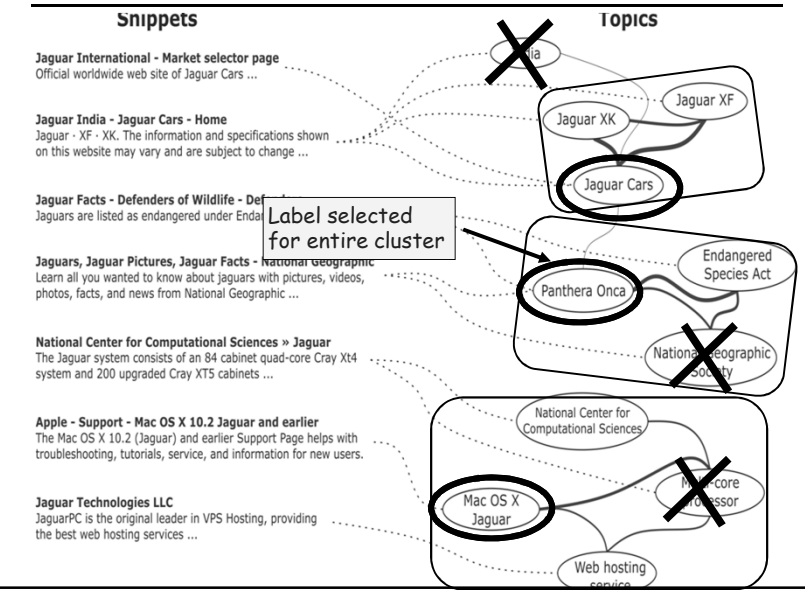
Selection of Significant Topics



Partition into Clusters



Selection of Cluster Labels



Role of Wikipedia in Information Retrieval

Methods for:

...

- Document analysis and understanding
- Query analysis and understanding
- Onebox search results

...

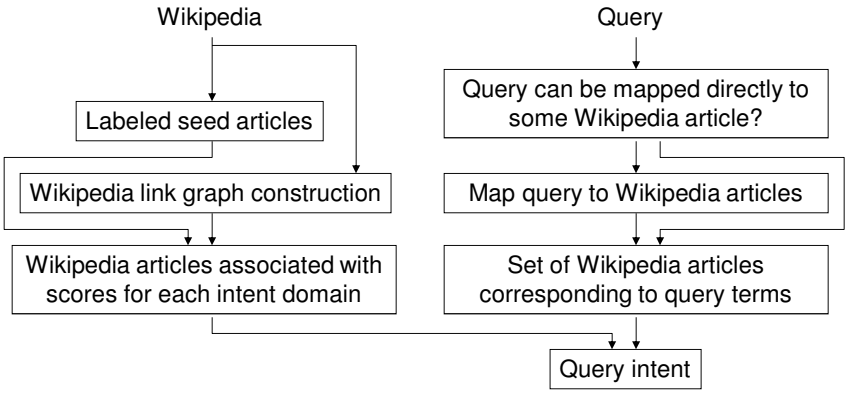
Query Understanding

- [HWL+09]: J. Hu, G. Wang, F. Lochovsky, J. Sun and Z. Chen. Understanding User's Query Intent with Wikipedia. WWW-09.

Modeling Query Intent with Wikipedia

- Input
  - queries
- Data source
  - Wikipedia articles and categories, connected via the category network
- Output
  - intent domains identified for queries, modeled as areas in the Wikipedia category network situated around manually-provided seed articles in Wikipedia
- Steps
  - independently from input queries, manually identify a small set of seed queries for each domain of interest
  - given set of seed queries, manually identify seed Wikipedia articles that correspond to the domain of interest
  - for each domain, expand seed Wikipedia articles into more Wikipedia articles, using connections between articles (article links, category network)
  - map queries into intent domains, taking into consideration manually-provided mappings from sets of seed queries

System Architecture



Modeling of Intent Domains

- Construct link graph for Wikipedia articles
  - nodes: Wikipedia articles, Wikipedia categories
  - edges: links between articles, links in Wikipedia category network between articles and categories; edges added between two nodes only when bi-directional links exist between the two nodes
  - edge weights: counts of links between the two nodes
- Associate Wikipedia articles with score for each intent domain
  - manually select seed Wikipedia articles deemed to belong to intent domain

Intent Type	Examples of Seed Queries	# Seed Queries
Travel	travel, hotel, tourism, airline tickets, expedia	2389
Person Name	britney spears, david beckham, george w. bush	10000
Employment	employment, monster, career	2543

  - iteratively propagate intent from seed articles to their neighbors articles in the link graph, assigning gradually lower intent scores

Determining Query Intent

- Case 1: query can be mapped directly to a Wikipedia article
  - retrieve intent domain whose intent score associated with the Wikipedia article is highest
- Case 2: query cannot be mapped directly to a Wikipedia article
  - map query into its more related Wikipedia articles, by disambiguating ("wikifying") mentions (substrings) from query to corresponding Wikipedia articles
  - retrieve intent domain for which the combination of intent scores, associated with the related Wikipedia articles, is highest

Query	Top Articles to Which Query is Mapped	Query Intent
employment guide	employment website, job search engine, careerlink, job hunting, eluta.ca, types of unemployment, airline tickets, expedia	Employment
job builder	job search engine, jobserve, falcon's eye, careerbuilder, eluta.ca, monster (website)	Employment

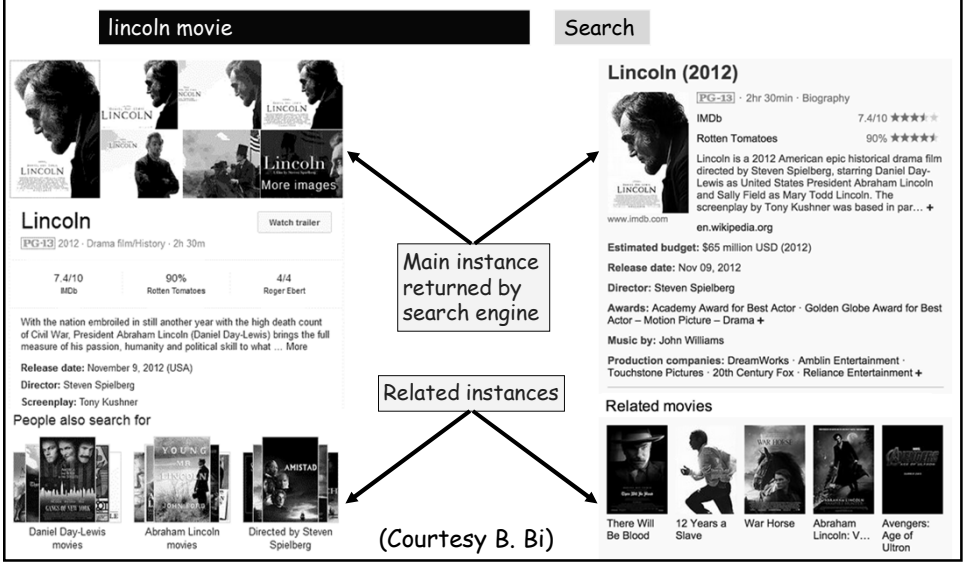
Query Understanding

- [BMH+15]: B. Bi, H. Ma, B. Hsu, W. Chu, K. Wang and J. Cho. Learning to Recommend Related Entities to Search Users. WSDM-15.

Recommending Related Instances

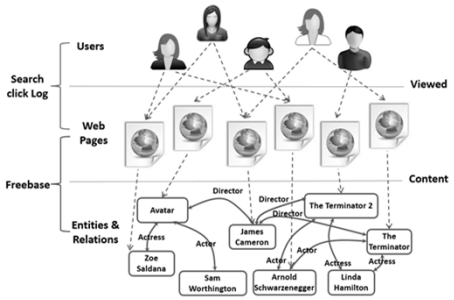
- Input
  - query submitted by a user
  - main structured instance, if any, returned by search engine as a side result for the query
- Data sources
  - click data for search results returned in response to queries
  - click data for main instances, if any, returned as side results in response to queries
  - collection of instances available within a structured knowledge repository (e.g., Freebase)
- Output
  - list of instances from knowledge repository deemed relevant to the query and the main instance, based on click data available for the user
  - similar to query suggestion, but suggestions are instances not strings, and suggestions are for instances rather than queries
- Steps
  - exploit three types of evidence, namely edges between instances within knowledge repository; edges between main instances in side results and clicked documents; and edges between main instances in side results and clicked related instances in side results
  - given a main instance, recommend a list of related entities based on the user's interests

Main Instance and Related Instances



(Courtesy B. Bi)

Evidence Towards Related Instances



Edges between instances and clicked documents

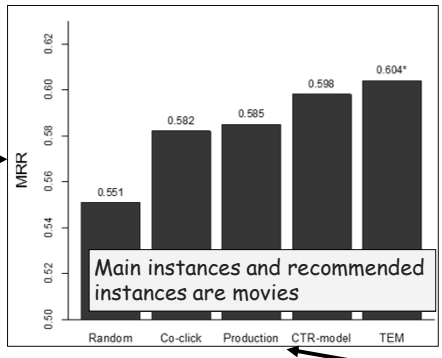
Edges between instances within knowledge repository

Edges between instances and clicked related instances in side results

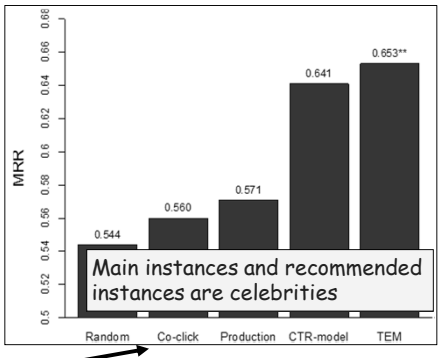
User ID	Time	Main entity	Related entity	Rank	Click
32	7/9/2013 10:32:26	Pacific Rim	Man of Steel	1	0
32	7/9/2013 10:32:26	Pacific Rim	The Wolverine	2	0
32	7/9/2013 10:32:26	Pacific Rim	The Lone Ranger	3	1
498	6/16/2013 15:16:41	Leonardo DiCaprio	Kate Winslet	1	0
498	6/16/2013 15:16:41	Leonardo DiCaprio	Johnny Depp	2	1

Recommended Instances

Mean reciprocal rank of relevant instance in computed ranked list of related instances



Main instances and recommended instances are movies



Main instances and recommended instances are celebrities

Co-click: evidence from user clicks on both main instance and related instance

CTR-model: evidence from click-through rate for related instances being returned

TEM: all sources of evidence

Query Understanding

- [HMB13]: L. Hollink, P. Mika and R. Blanco. Web Usage Mining with Semantic Analysis. WWW-13.

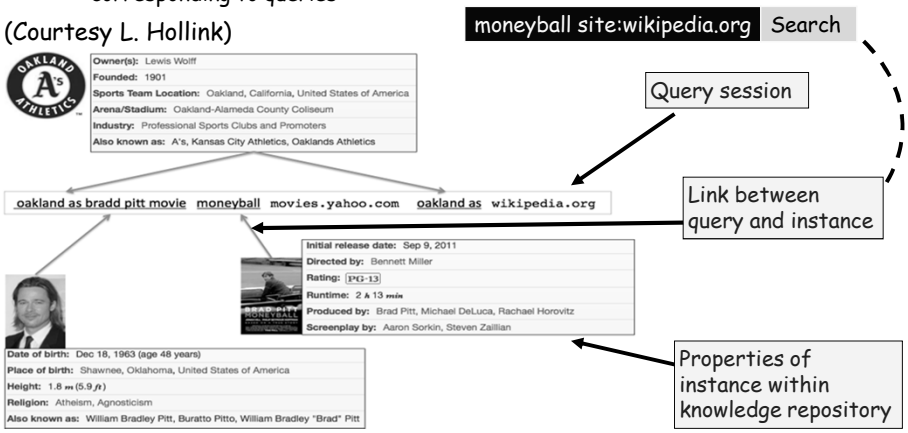
Web Usage Understanding

- Data sources
  - query sessions including queries and clicked documents
  - collection of instances available within a structured knowledge repository (Freebase and DBpedia)
- Output
  - semantic rather than lexical patterns of Web usage mining
- Steps
  - annotate queries, by linking query fragments to corresponding instances from knowledge repository
  - use properties available for instances within knowledge repository to generalize and categorize queries into patterns of usage

Linking Queries to Instances

- Link queries to instances
  - from search results returned for site-restricted queries, identify instances corresponding to queries

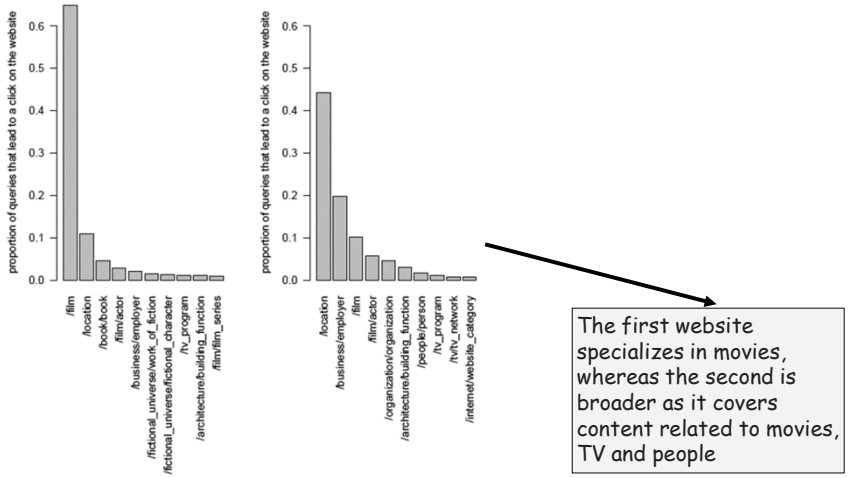
(Courtesy L. Hollink)



- Generalize from linked queries to query types
  - using instance types from knowledge repository, generalize queries into query types that lead to websites

Patterns of Web Usage

- Estimate the differences in the content of two websites, by comparing the top query types that lead to clicks on the sites



Patterns of Web Usage

- Compare query patterns obtained by generalizing from queries linked to instances that are relatively recent movies; vs. instances that are relatively older movies

Level	Pattern	Support
L1	*movie*	0.396
	2011	0.15
	trailer	0.103
	movies	0.095
	moviedict : 2011	0.063
	cast	0.053
	new	0.049
	dvd	0.035
	release	0.032
	reviews	0.028
L2	*movie* ⟶ *movie*	0.165
	2011 ⟶ 2011	0.042
	movie ⟶ 2011	0.04
	2011 ⟶ *movie*	0.038
	movies ⟶ *movies*	0.038
	trailer ⟶ *trailer*	0.027
	movie ⟶ *movie* 2011	0.026
	movies ⟶ *movie*	0.025
	movie ⟶ *trailer*	0.024
	movie 2011 ⟶ *movie*	0.023

vs.

Level	Pattern	Support
L1	*movie*	0.391
	cast	0.096
	movies	0.091
	quotes	0.038
	trailer	0.034
	new	0.027
	2011	0.027
	free	0.023
	soundtrack	0.021
	watch	0.021
L2	*movie* ⟶ *movie*	0.169
	movies ⟶ *movies*	0.038
	cast ⟶ *cast*	0.028
	movies ⟶ *movie*	0.025
	movie ⟶ *movies*	0.023
	quotes ⟶ *quotes*	0.019
	movie ⟶ *cast*	0.018
	movie ⟶ *trailer*	0.012
	movie ⟶ *moviecast*	0.011
	new ⟶ *new*	0.01

Role of Wikipedia in Information Retrieval

Methods for:

...

- Document analysis and understanding
- Query analysis and understanding
- Onebox search results

...

Retrieval of OneBox Results

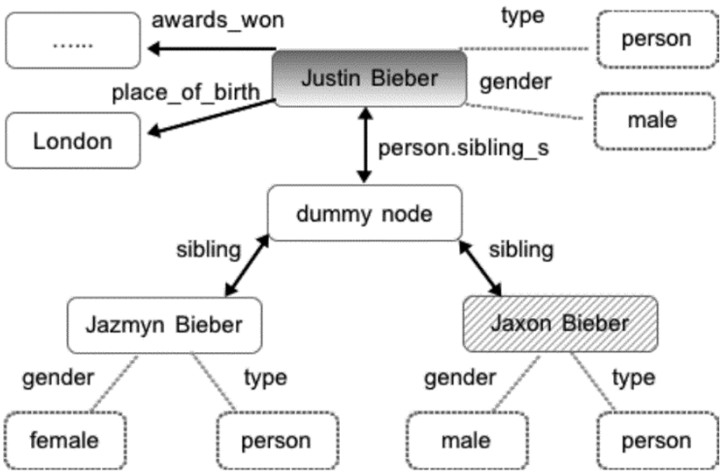
- [YV14]: X. Yao and B. Van Durme. Information Extraction over Structured Data: Question Answering with Freebase. ACL-14.

Question Answering as Binary Classification

- Input
  - natural-language questions (What is the name of the Justin Bieber's brother?)
- Data sources
  - knowledge repository of inter-connected topics (Freebase)
  - collection of Web documents
- Output
  - topics that answer the questions (Jaxon Bieber)
- Steps
  - convert the question into a question graph
  - based on the node from the question graph corresponding to the question topic (Justin Bieber), assemble a topic graph of inter-connected topics up to a few hops away from the question topic
  - using individual and combination features from question graph and topic graph, determine whether each node from the topic graph is or is not an answer to the question

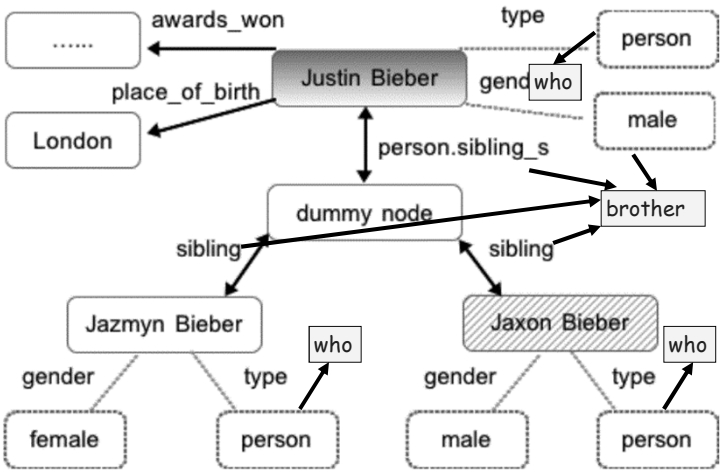
Answers from Knowledge Repositories

(Courtesy X. Yao)



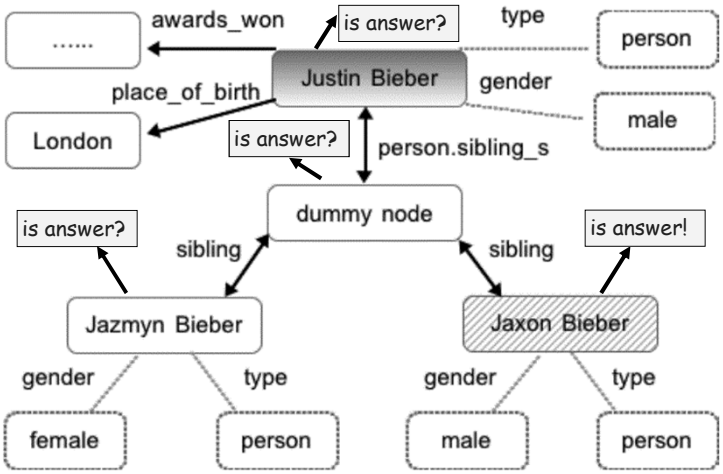
Question: What is the name of Justin Bieber's brother?

Answers from Knowledge Repositories



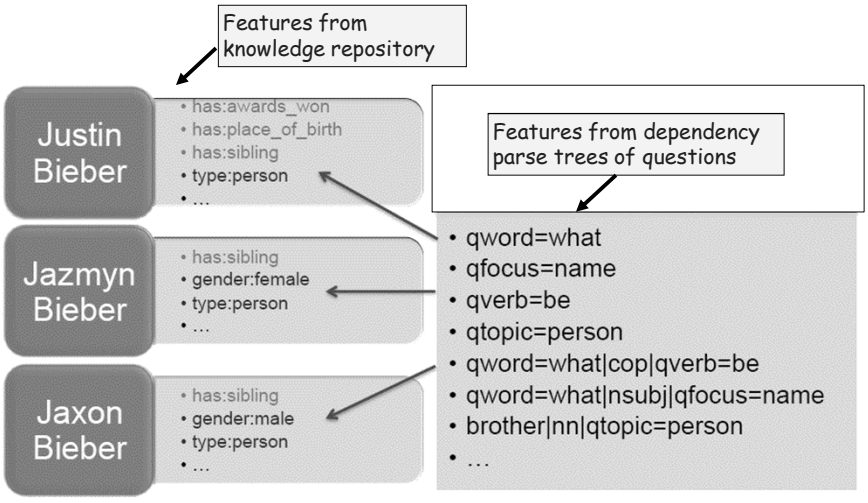
Question: What is the name of Justin Bieber's brother?

Question Answering as Classification

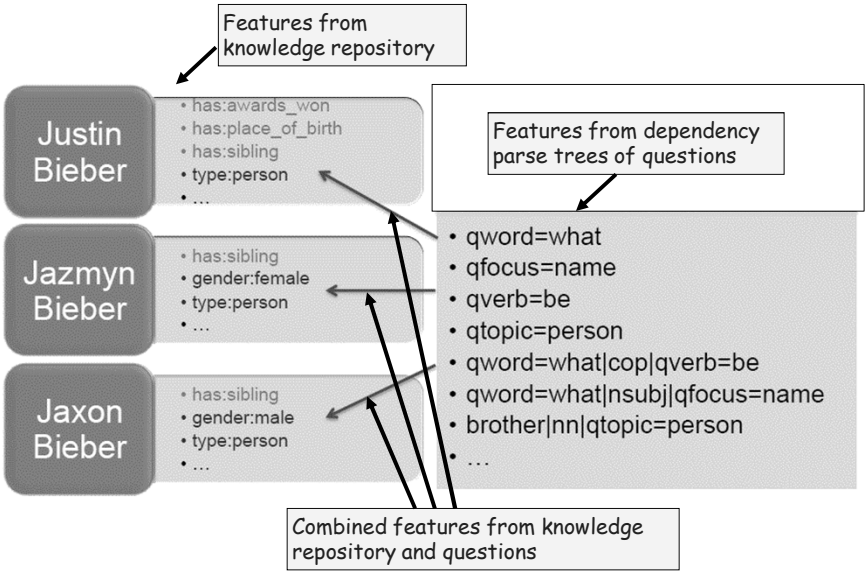


Question: What is the name of Justin Bieber's brother? Answer: Jaxon Bieber

Features for Classification



Features for Classification



Estimated Utility of Features

Justin Bieber	expected weights
• has:awards_won | qword=what	• medium
• has:place_of_birth | qword=what	• low
• has:sibling | qfocus=name	• low
• type:person | qfocus=name	• medium
• …	• …
Jazmyn Bieber	**expected weights**
• has:sibling | brother | nn | qtopic=person	• high
• gender:female | brother | nn | qtopic=person	• low
• type:person | brother | nn | qtopic=person	• high
• …	• …
Jaxon Bieber (is answer)	**expected weights**
• has:sibling | brother | nn | qtopic=person	• high
• gender:male | brother | nn | qtopic=person	• high
• type:person | qword=what | nsubj | qfocus=name	• high
• …	• …

Mapping Relations to Phrases

- Align relations from knowledge repository to phrases that may express the relations in document sentences
  - film/starring (Gravity, Sandra Bullock)
    vs.
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity
- Use alignments to predict relevant relations when answering questions

feature	weight	feature	weight
qfocus=religion|type=Religion	8.60	qword=when|type=datetime	5.11
qfocus=money|type=Currency	5.56	qverb=border|rel=location.adjoins	4.56
qverb=die|type=CauseOfDeath	5.35	qverb=go|qtopic=location|type=Tourist attraction	2.94