Lecture 11.
Knowledge Graph：An Information Retrieval Perspective

Ridho Reinanda, Edgar Meij and Maarten de Rijke (2020). “Knowledge Graphs: An Information Retrieval Perspective”, Foundations and Trends in Information Retrieval: Vol. 14, No. 4, pp 289-444. DOI: 10.1561/1500000063.

Foundations and Trends® in Information Retrieval

Knowledge Graphs: An Information Retrieval Perspective

Suggested Citation: Ridho Reinanda, Edgar Meij and Maarten de Rijke (2020), "Knowledge Graphs: An Information Retrieval Perspective", Foundations and Trends® in Information Retrieval: Vol. 14, No. 4, pp 289–444. DOI: 10.1561/1500000063.

Ridho Reinanda
Bloomberg L.P.
UK
rreinanda@bloomberg.net

Edgar Meij
Bloomberg L.P.
UK
emeij@bloomberg.net

Maarten de Rijke
University of Amsterdam & Ahold Delhaize
The Netherlands
m.derijke@uva.nl

now
the essence of knowledge
Boston — Delft

Two Complementary Angles

- Knowledge Graph (KG)
  - A repository of entities as well as their relationships and attributes that is represented as graph.
- Knowledge graphs for information retrieval
  - Leveraging KGs for information retrieval
- Information retrieval for knowledge graph
  - Enriching KGs using IR techniques
  - How IR can be utilized for the construction and completion of KGs
  - Entity recognition, typing, and relation extraction

Key Concepts

- An *entity* e is an atomic, identifiable object that has a distinct and independent existence.
- A *named entity* is a specific entity for which one or many designators or proper names can be used to refer to it.
- An *entity type* t is a set of classes that is appropriate for an entity based on a pre-defined class hierarchy.
- A *mention* is a text segment that refers to an entity.

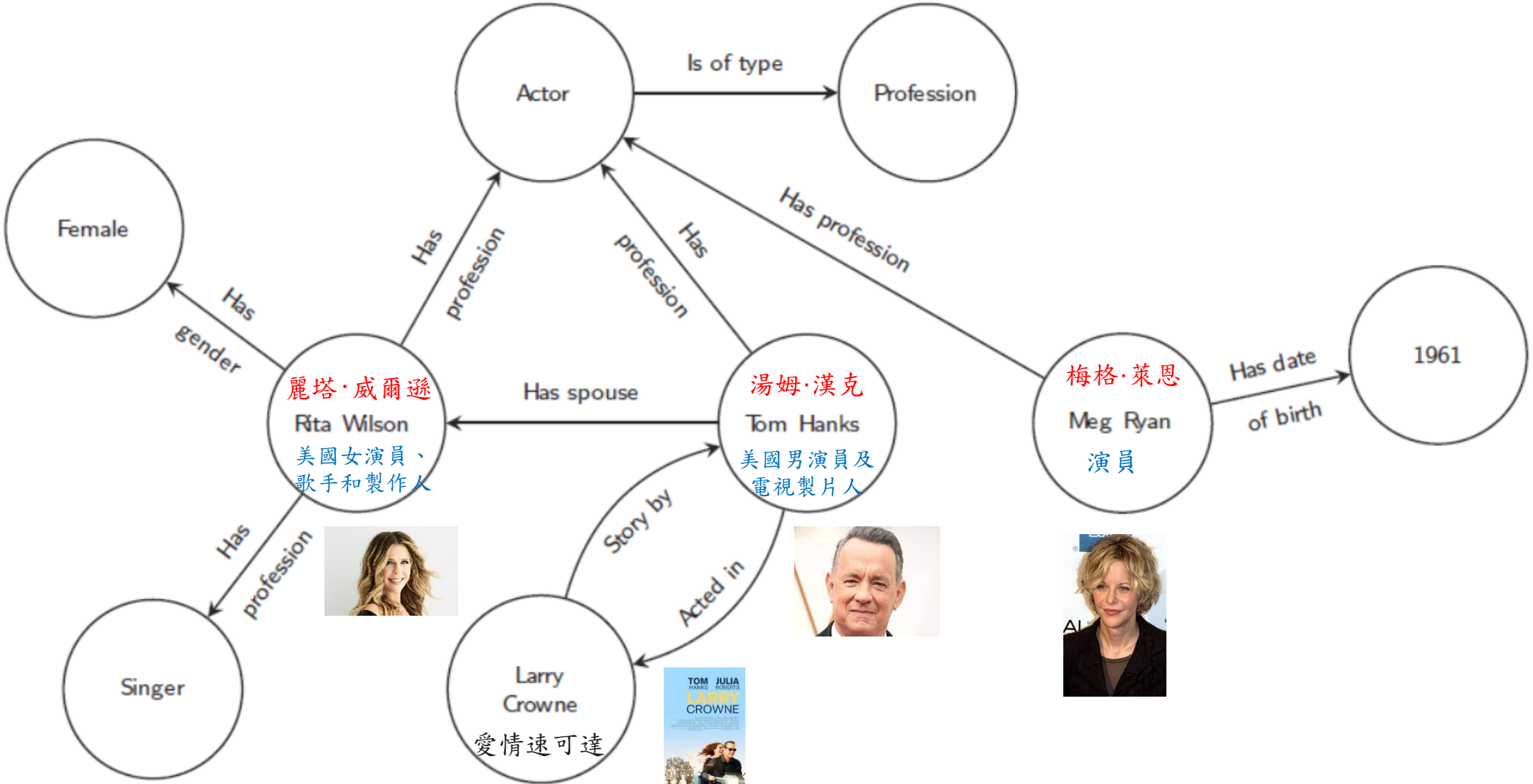
- A *relationship type* is a type of connection between two entities.
- A *relation* r is an instance of a relationship type between two entities, the nature of which can be defined with a label and, can be complemented with attributes and attribute values.
- An *attribute* is a specific characteristic or property of an entity or relationship, with zero or more values.
- A *knowledge base* is a repository of entities with information about their relationships and attributes in a (semi-)structured format.

- A *knowledge graph* is a knowledge base that is specifically represented as a graph. In a knowledge graph, entities, attributes, and relations are represented through the nodes and edges in the graph. Entities are typically represented as nodes, while relationships are represented as edges.
- An *entity profile* is a textual description of an entity.

Examples

Definition	Example
Entity	City
Named entity	London
Entity type	City
Mention	London
Relation type	Capital of
(*Relation*)	(London, capital of, United Kingdom)
Attribute	City: population
Knowledge base	Wikipedia
Knowledge graph	Wikidata

A Fragment of KG in Movie Domain



Availability of KGs

- Created manually from scratch by crowdsourcing or experts
- Public-available, open-domain KGs (see next slide)
- Generated from structures or semi-structured sources or social networks
- Obtained from unstructured textual sources by information extraction
- Constructed by using dialogues and multimedia content

General-Purpose and Domain-Specific KGs

Knowledge Graph	URL
DBpedia (Lehmann *et al.*, 2015)	https://wiki.dbpedia.org
Wikipedia	http://wikipedia.org
YAGO (Suchanek *et al.*, 2007)	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
LittleSis	http://littlesis.org

LittleSis is a free database of who-knows-who at the heights of business and government.

Uses of KGs

- Support users to explore information using entities and relations as navigational aids and for decision support
- Feed conversational search interfaces

Evaluation Metrics

- Recall
- Precision
- F1
- Mean Average Precision (MAP)
- Mean Reciprocal Rank
- NDCG

Corpora Commonly Used in Entity-Related Experiments

Collection	URL
AP	https://catalog.ldc.upenn.edu/LDC93T3B
AQUAINT	https://tac.nist.gov//data/data_desc.html
ClueWeb09	https://www.lemurproject.org/clueweb09.php
ClueWeb12 (Gabrilovich *et al.*, 2013)	https://www.lemurproject.org/clueweb12.php/
GOV2	http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
MSNBC (Cucerzan, 2007)	https://kdd.ics.uci.edu/databases/msnbc/msnbc.data.html
NYT	https://catalog.ldc.upenn.edu/LDC2008T19
TREC Robust	https://trec.nist.gov/data/t13_robust.html
WT10G	http://ir.dcs.gla.ac.uk/test_collections/wt10g.html

Datasets related to Constructing KGs

Task	Datasets
Entity linking	MSNBC (Cucerzan, 2007)
	AQUAINT (Milne and Witten, 2008)
	IITB (Kulkarni *et al.*, 2009)
	AIDA CoNLL-YAGO (Hoffart *et al.*, 2011)
	Twitter-to-concept (Meij *et al.*, 2012)
	FACC (Gabrilovich *et al.*, 2013)
	GERBIL (Usbeck *et al.*, 2015)
	Wikinews/Meantime (Minard *et al.*, 2016)
	GERDAQ4 (Cornolti *et al.*, 2016)
	ERD (Carmel *et al.*, 2014)
	Wikilinks (Singh *et al.*, 2012)
Document retrieval	TREC Robust (Voorhees, 2005)
Entity recommendation	REF (Balog *et al.*, 2009b)
Entity retrieval	INEX-ER (de Vries *et al.*, 2008)
	DBpedia-Entity (Balog and Neumayer, 2013)
	REWQ (Schuhmacher *et al.*, 2015)
Relationship explanation	(Voskarides *et al.*, 2015)

Datasets related to Constructing KGs

Task	Datasets
Entity recognition	CoNLL (Tjong Kim Sang and De Meulder, 2003)
Entity-centric document filtering	TREC-KBA StreamCorpus (Frank *et al.*, 2014)
Entity discovery	TAC-KBP EDL (Ellis *et al.*, 2014)
Relation extraction	SemEval (Girju *et al.*, 2007)
Link prediction	ACE (Doddington *et al.*, 2004) WN18 (Bordes *et al.*, 2013) FB15K (Bordes *et al.*, 2013) FB15K-237 (Toutanova and Chen, 2015)
KG quality estimation	WDVC (Heindorf *et al.*, 2015)

Notation

Notation	Description
e	Entity
f	Fact
m	Mention
d	Document
q	Query
r	Relation
s	Text segment
t	Entity type
T	Entity classification system

Entity Linking and Recognition

Entity Linking

- Semantic grounding of a text using entities in a KG
- Determine which textual spans refer to which specific entities
- Link mentions to Wikipedia
- Main task in Text Analytics Conference (TAC)
  - Evaluate and improve Knowledge Base Population (KBP)

named entity recognition

Definition 3.1 (Entity Linking). Given a text, detect segments of entity mentions m within the text, and link them to an entity e in a knowledge graph KG.

Short Query vs. Long Text

- Entity Linking for Queries
  - Allow for better query understanding
  - Help search engines to retrieve relevant information
- Challenge for entity linking in queries
  - Queries are short, written in a telegraph-style
  - Only very limited context is available

Approaches to Entity Linking

- Two Stages
  - Mention detection: Identify the substrings denoting entities
  - Disambiguation: Link each mention to a specific entity
- Evaluation
  - Document level: list all entities mentioned in the input text
  - Mention level: list all mentions in the input text and identify the most likely entity for each.
  - Identify so-called “Nil” entities, i.e., entities that do not exist in the KG
- Approaches
  - Feature-based Approach
  - Graph-based Approach
  - Neural Approach

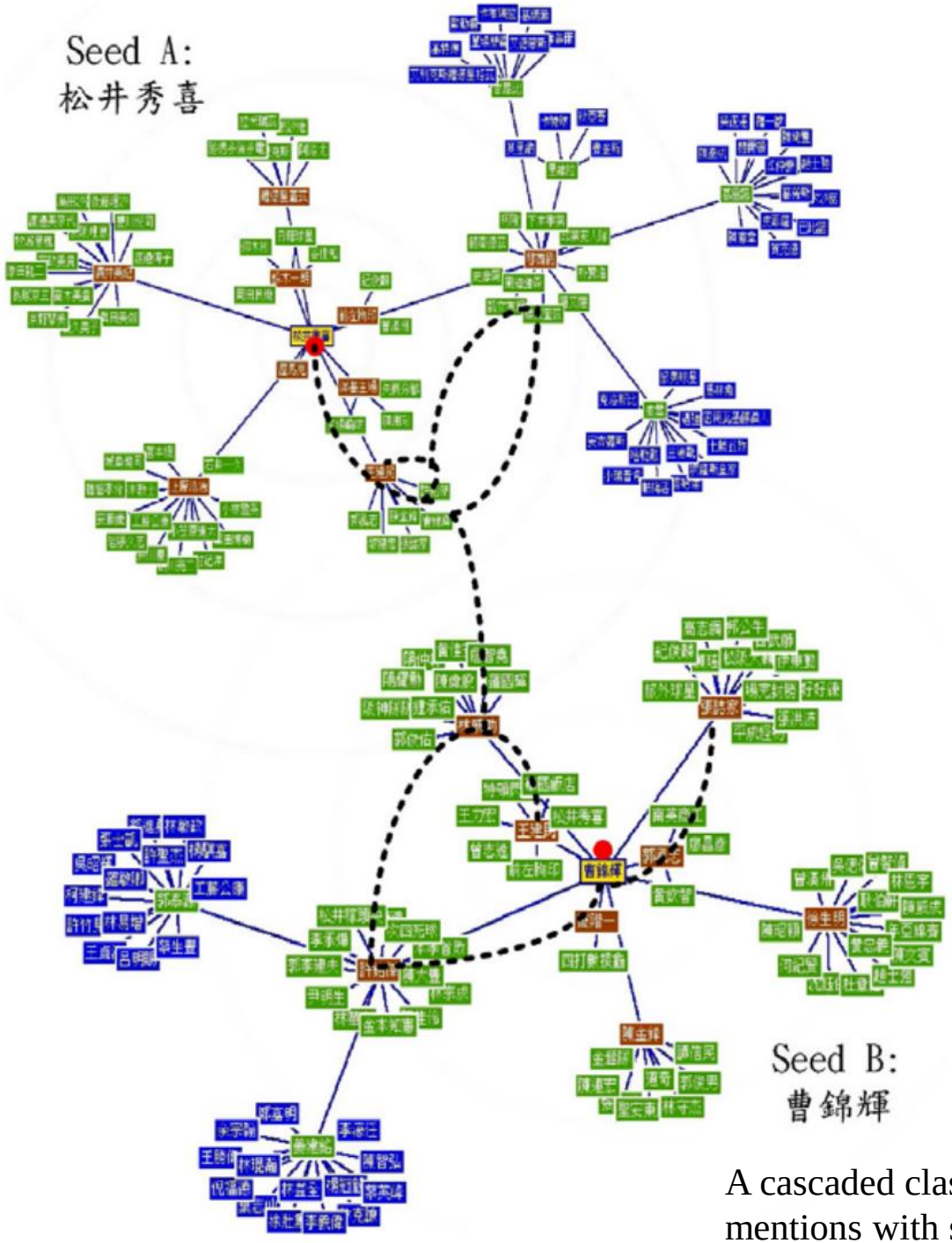
Feature-Based Approaches to Entity Linking

- Introduce features that have become commonplace in both mention detection and disambiguation
- Keyphraseness
  - The probability of a phrase to be selected as a keyword, i.e., a mention, in a document
- Commonness 王建民 (職業棒球選手 vs. 中研院研究員)
  - The relative popularity of each candidate entity as a target given the same mention
- Relatedness
  - The semantic similarity between two entities
  - Maximize the relatedness of relevant entities will minimize disambiguation errors
  - Learn entity relatedness functions to improve entity linking

Graph-Based Approaches to Entity Linking

- The networks (KG) formed not only by relationships between entities, but also by relationships between mentions in a document carry signal for disambiguation
- A related set of entities will provide context for disambiguation, as they tend to appear together in the text
- Perform entity disambiguation collectively, optimizing the coherence between candidate entities

  - 物以聚類



A cascaded classification approach to disambiguating polysemous mentions with social chains (Wei, Lin, Chen, 2010)

Neural Approach

- Approaches entity linking end-to-end in a deep learning framework
- Propagating information between the mention detection and entity disambiguation subtasks

Kolitsas, N., O.-E. Ganea, and T. Hofmann (2018). “End-to-end neural entity linking”. In: *CoNLL ’18*.

Entity Recognition and Classification

- Named entity recognition

Definition 3.2. Given a text segment s, the *entity recognition* task is to detect segments of entity mentions m within s. Given a type classification system T and an entity mention m within a text segment s, the *entity mention classification* task is to decide whether m belongs to a type $t \in T$ and, if so, which type.

Approaches to Entity Recognition and Mention Classification

- Rule-Based Approaches to Entity Recognition and Classification
  - Rely on dictionaries and handcrafted rules
- Feature-Based Approaches to Entity Recognition and Classification
  - Learn to classify entities from data using contextual clues around the entity mentions
  - Formulated as a sequential classification problem: tagging words in a sentence sequentially to indicate whether they are a part of a named entity or not
- Embedding-Based Approaches to Entity Recognition and Classification

KGs for IR

How KB Enhances User's Search Experience

- Better Understanding
  - Improve the understanding of intent of queries and documents
- Direct Answer
  - Allow to answer information needs that might be more amendable to be answered directly
- Enhanced Exploration Facilities
  - Enable the exploration of related entities mentioned in a document collection or a search engine
  - Help to provide explanations of entities and relationships in context

How entities detected in queries and documents can be used to improve document retrieval

Document retrieval (Subsection 4.1)	Rank documents given a query.
Expansion-based	Expand queries and/or documents with entity-based information.
Latent factor modeling	Model and leverage a latent space between query and documents.
Language modeling	Incorporate term sequences marked as entities when building language models of a query and a document.
Deep learning	Incorporate KG-based embeddings to improve query/document representations and steer the retrieval process.
Entity retrieval (Subsection 4.2)	Rank entities in text or KG given a query.
Language modeling	Retrieve entities by matching a query with entity descriptions or mentioning documents.
Neural language modeling	Learn latent representations of query and entities, compare for retrieval.
Multi-fielded representation	Represent an entity as a multi-fielded document and use document retrieval techniques.

Entity recommendation (Subsection 4.3)	Recommend related entities given an entity and/or context.
Heuristic	Estimate statistical associations between entities from text.
Behavioral	Recommend entities based on similar users' interest.
Graph-based	Recommend entities based on the structural connections in a graph.
Relationship explanation (Subsection 4.4)	Explain the relationship between a pair of entities.
Instance-based	Explain the relationship by selecting a set of key related entities.
Description ranking	Generate and rank candidate explanations from external corpora.

Document Retrieval

Definition 4.1 (Document Retrieval). Given a query q and a collection of documents D, score and rank each document $d \in D$ based on its relevance to q.

- Approaches
  - Expansion-based approach
  - Latent factor modeling approach
  - Language modeling approach
  - Deep learning approach

Expansion-Based Approach

- Explicitly incorporate entity-oriented information as features in the retrieval process
- Enrich the query with features from entities and their links to KGs
- Entity query feature expansion (EQFE)
  - Preprocessing
    - Documents are preprocessed with entity linking and additional information obtained from knowledge graphs is indexed as different fields of the document.
  - Query annotation
    - At query time, the query is also preprocessed with entity linking, providing annotations for all entity mentions in the query.
  - Expansion from feedback
    - Two types of relevance feedback are considered.

Expansion from Feedback

- KG feedback (KB 中鄰近的entities)
  - The query is issued against an index of a KG in order to retrieve related entities.
- Corpus-based feedback (documents中含有的entities)
  - Related entities are obtained from retrieved documents.
- Expansion strategies
  - Related words, entities, mentions, types, categories, and neighbors
- Feature weights
  - Learned for each of these different expansions with a log-linear learning to rank approach

Entity-Oriented Query Expansion by Using KG

- Two main steps
  - Object linking
    - Generate ranked lists of related KG entities
    - Issue a query to the Google Search API and select entities from FACC1 annotations in the top-ranked documents
    - FACC1 annotations are automatic annotations of English web pages from ClueWeb09 and ClueWeb12 to Freebase entities
  - Term Selection
    - Rank the related terms from the linked objects' descripts for expansion

Latent Factor Approaches to Document Retrieval

- Extract concepts inherent in queries and documents
  - Extract latent factors for queries and documents vs. enrich query or document representations from a KG
- From a graph representation of a query G, which contains the query terms, an expanded graph H can be derived by adding single and multiple terms concepts. (就是傳統query expansion)
  - A probability distribution over latent concepts is inferred from a small number of relevant or pseudo-relevant documents for query *q*. (就像local documents)
  - To perform query expansion, *k* latent concepts with the highest likelihood are selected.
  - A new graph G' is constructed by expanding the original graph G with the selected concepts.

Latent Factor Approach Using KB Information

- Expansion using external data in knowledge graph
  - Consider entity relationships as a latent space
  - Treat vocabularies, terms, and entities from external data as a means to connect a query and documents
  - Rerank an initial set of documents with the help of related entities and feature vectors
  - The feature vectors are derived from the relationships between entities and documents, and another feature vector, which represents the relationship between the entity and the query (entity as bridge, see next page)
- Three strategies to find entities given a query and document
  - Query annotation: entity linking on queries
  - Entity search
  - Document annotation (標記entity在document的位置)

Relationships between Query, Documents, and Entities

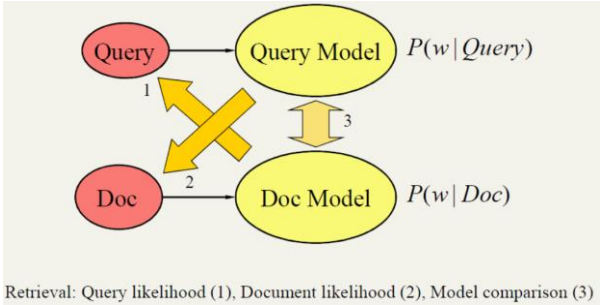
- A set of features describing the relationship between query and entities
  - entity selection score, textual similarity score, ontology overlap, entity frequency, etc.
- A set of features describing the relationship between entities and documents
  - textual similarity, ontology overlap, graph connection, and document quality
- The best combination of query representation and document ranking is learned from these features.

Entity Selection

- Finding relevant entities for query and documents is an important step in using external data for ranking.
- Query annotation (entity linking) on queries
  - Select related objects to improve document retrieval
  - Utilize the TAGME entity linking method

Latent Entity Space (LES)

- Map queries and documents to a high-dimensional latent entity space (很像query和document以entities來表示)
- Each dimension in the latent space corresponds to one entity
- Information around an entity in each dimension is captured by a profile of the entity (entity以profile來表示)
- Two approaches to build the entity profile
  - Combining information around the entity across multiple documents in the corpus (entity出現在多個document，以entity周遭的information來當作entity的profile)
  - Using the entity profile from an external KGs (entity在KG周遭的資訊)



Retrieval: Query likelihood (1), Document likelihood (2), Model comparison (3)

- Estimate query and document models and compare
- Suitable measure is KL divergence $D(M_Q \| M_d)$

$$R(d;Q) = KL(M_Q \| M_d) = \sum_{t \in V} P(t \mid M_Q) \log \frac{P(t \mid M_Q)}{P(t \mid M_d)}$$

Use of Latent Space

- At query time, only few entities that are highly related to the query are used in the construction of the latent space.
- Query projection
  - The weighted sum of the similarity between each entity in the query and each entity profile is computed with a query likelihood model.
- Document projection
  - After the entities in the latent space have been selected, document projections into the latent entity space are computed.
- Similarity between the entity model and document model
  - KL-divergence
  - Compare the language model of the document and the entity profiles.

Language Modeling Approaches to Document Retrieval

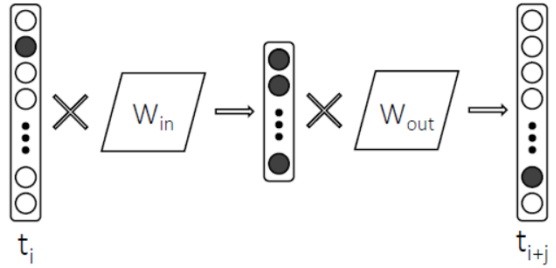
- An entity-based language model that uses entity linking methods
  - The uncertainty inherent in the entity linking process
  - Incorporate a balance between using entity based and term-based information
- Apply entity linking to obtain entities along with the linking confidence score (就是前面的uncertainty)
- Define an entity-based language model over a token space
  - The token space includes the set of all terms in the document collection and the set of entities that were linked at least once within a document
  - Entity-Document Matrix * (Term-Document Matrix)^t
  - Entity in terms of terms (tokens)

Pseudo Count

- Hard confidence thresholding
  - A threshold is placed on the confidence score of each annotation
  - Those mentions linked with a certain confidence score are considered for pseudo-counts
- Soft confidence thresholding
  - The confidence score of linking a particular mention is taken as pseudo-count during the estimation, interpolated using an importance parameter.
- Experiments
  - The entity-based language model with hard and soft thresholding improves over the standard term-based language model.

Deep Learning Approaches to Document Retrieval

- Explicit Semantic Ranking (ESR)
  - A ranking technique that leverages knowledge graph embeddings
  - Represents queries and documents by embeddings of entities in the knowledge graph. (以前是以word embedding來建立query和document embedding，現在延伸以entity embedding來建立query和document embedding)
  - Semantic relatedness between the representations of query and documents is then computed in the embedding space.
- Embeddings of entities
  - Trained from edges in a knowledge graph using a skip-gram based approach



Detail Procedure

- Represent query and documents as bags of entities using an entity linking method
- Construct a translation matrix that captures the relatedness between entities in the query and documents (query和document entity不一定直接match，例如alias，藉由translation matrix對應)
- Compute histogram features of entity relatedness by grouping them by strength into several bins
- Use the histogram-based features to rank documents with a learning to rank-based approach

Follow-up Approach

- Enrich queries and documents with entity information from knowledge graphs (前面的做法僅看表面形式的對應，頂多僅用 translation 來延伸，這裡把KG考慮進來)
- Model query and documents as word-based representations and entity-based representations simultaneously
- Consider four types of interaction
  - Build bag-of-words and bag-of-entities representations of query and documents
  - Extract matching features between (query words, documents words), (query entities, document words), (query words, document entities), and (query entities, document entities)

Entity-Duet Neural Ranking Model

- Incorporate semantic information from knowledge graphs in neural ranking systems
- Inspired by the improvements brought about by entity-based models to feature-based ranking systems
- Follow the same search framework as the word-entity duet
  - Building bag-of words and bag-of-entities representations of queries and documents
  - Use handcrafted features
- EDRM captures the matching between queries and documents through a translation layer in a neural architecture

Relation of Document Retrieval to Entity Linking/Retrieval/Recommendation Tasks

- Entity linking
  - Use entity information primarily depend on performing entity linking on the queries and the candidate documents
- Entity retrieval/entity recommendation
  - Relevant documents are first retrieved, and then entities found in these documents are ranked.

Entity Retrieval

- Expert finding track at TREC

Definition 4.2 (Entity Retrieval). Given a query q and a document collection D, retrieve and rank entities mentioned in or associated with each document $d \in D$ according to their relevance to q.

Definition 4.3 (Entity Retrieval from KG). Given a query q and a knowledge graph KG, retrieve and rank entities in the KGs according to their relevance to q.

- Forms of Entity Retrieval
  - Term based entity retrieval, ad-hoc object retrieval, and list retrieval

Approaches to Entity Retrieval

- Term-based Entity Retrieval (王建民)
  - Language modeling
    - The task of expert finding
  - Neural approaches
    - Learn distributed word representation of entities optimized for retrieval
- ad-hoc object retrieval (紐約洋基隊王建民)
  - The entities are considered as objects with attributes and relationships
  - Multi-fielded representation approaches, which represent entities and documents as a set of fields

Language Modeling Approaches to Entity Retrieval

- Candidate-centric model
  - Represent an entity as a virtual document
  - Rank the documents given the query mentioning certain entities
- Main Idea
  - Represent a candidate expert as a multinomial distribution over a vocabulary of terms
  - Predict how likely a candidate would generate the query
- Document-centric model
  - (1) Find documents that are relevant to the query
  - (2) Identify the experts associated with these documents

Neural Approaches to Entity Retrieval

- Expert finding
  - Learn distributed word representations in an unsupervised way from textual evidence
  - Learn a log-linear model of probabilities of a candidate entity given the word
- Latent Semantic Entities (LSE) approach
  - Learn term and entity representations in a different space, by adjusting the representations so that they are close in the entity space
  - Used in combination with query-independent features and the query likelihood model
  - This neural approach and term-based retrieval make very different errors

Multi-Fielded Representations

- ad-hoc object retrieval task
  - Retrieve a list of resource objects (i.e., entities) with respect to a user query
  - Focus on entities, their attributes, and their relationships in KG
- Multi-fielded representations
  - An entity is represented as a set of fields with bag-of-words values
- A probabilistic approach for ranking multi-fielded documents
  - Compute the posterior probability that a query will be mapped to a certain field in the document
  - Estimated by considering how often a certain term appears in a certain field
  - Used to weight the score computed for each field in the entity representation

Ad-hoc Object Retrieval in An Entity-based Context

- Baseline
  - Consider TF.IDF over the entity properties in the graph
  - Compute term frequency (TF) and inverse document frequency (IDF) statistics for every property of the entity in the graph
  - Several methods aim to learn appropriate weights for each field
- Use of structured search on top of standard IR approaches
  - Combine IR and structured search techniques
  - Represent each object in the graph with entity names in URIs, entity names in labels, and attribute values of the entity
  - Index information as a structured, multifielded index on top of which multi-fielded retrieval algorithms such as BM25F

Leverage relationship information

- Leverage relationship information to improve entity retrieval
  - Represent an entity as common fields such as names, attributes, and outgoing links
  - Representing the entity relationship graph as a tensor
  - Factorize the tensor into a number of latent factors
  - Enrich the fielded representations of the entity with top-related entities obtained through latent factor modeling
- Adapt term dependency models
  - Consider the dependencies between the query terms and fields
  - Parameterized as a set of features based on the contribution of query concepts matched in a field towards the retrieval score

Dynamic Representations of Entities

- Static and dynamic description sources
  - **Knowledge base**: anchor text, redirects, category titles, and titles of linked entities in a KG;
  - **Web anchors**: anchor text of links to Wikipedia pages;
  - **Twitter**: tweets with links to Wikipedia pages;
  - **Delicious**: references to entities through social tags; and
  - **Queries**: queries that can be linked to Wikipedia pages.

Enhancing the Representation from Various External Sources

- Entities are modeled as fielded documents where each field is a term vector that represents the entity's content from a description source
- Learn feature weights for query-field similarity, field importance, and entity importance score based on each field and description

Relation of Entity Retrieval to Other Tasks

- Entity retrieval depends on having reliable entity recognition and/or entity linking systems
- Entity retrieval can be used as a query understanding strategy analogously to entity linking to support document retrieval
- Entity retrieval vs entity recommendation
  - Do not have an explicit query to take into account in entity retrieval

Outlook on Entity Retrieval

- Term-based entity retrieval from documents
  - (1) Retrieve documents first, then extracting entities from the retrieved documents
  - (2) Represent candidate entities as documents themselves, by concatenating all documents for an entity
- Neural approaches
  - Considered as complements to the language-modeling approaches
- Entity retrieval from KGs
  - Multi-fielded representation, in which each entity is represented as a collection of named and nested fields
  - Representation strategy and field weighting strategy

Entity Recommendation

- Related entity finding
- Recommending related entities in response to a textual query and a set of query entities
- In all the major web search engines, entities related to an entity relevant to the query are shown.

Definition 4.4 (Entity Recommendation). Given a query q (where q can be in the form of an entity or an entity plus some additional context keywords), rank each entity $e \in KG$ based on their relatedness to the query.

Approaches to Entity Recommendation

- Heuristic approach
  - Statistical associations of the entities estimated from a data source
  - Co-occurrence statistics
- Behavioral approach
  - Utilize signals derived from user interactions or feedback to generate the recommendations
  - e.g., clicks on related entities, documents, or entity panes
- Graph-based approach
  - Rely on semantic relationships without any explicit user feedback

Entity Relationship Explanation

- Explanations are required to describe entity pairs, paths between entities, as well as other relationships observed in, e.g., query logs.

Definition 4.5 (Entity Relationship Explanation). Given a pair of entities e and e', provide an explanation, i.e., a textual description, supported by a KG, of how the pair of entities is related.

Approaches to Entity Relationship Explanation

- Instance-Based Entity Relationship Explanations
  - Provide an explanation of a relationship by returning a set of related entities
- Entity Relationship Explanations Based on Ranking Descriptions
  - Come up with candidate textual descriptions of a relationship and provide a ranking of possible explanations

Instance-Based Entity Relationship Explanations

- Explaining connections between entities by utilizing KGs.
- (1) Explanation enumeration phase
  - Generate all path instances of a specified length found in a knowledge graph.
- (2) Explanation ranking
  - Structure-based measures
  - Aggregate measures

Entity Relationship Explanations Based on Ranking Descriptions

- Explaining relationships between pairs of knowledge graph entities with human readable descriptions.
  - Extract and enrich sentences that refer to an entity pair.
  - Rank the sentences according to how well they describe the relationship between the entities.

Information Retrieval for Knowledge Graphs

KG-Related Tasks

- Use information retrieval (IR) techniques to help construct or interact with KGs.
- What are typical KG-related tasks?
  - Text classification and trend detection techniques from IR are useful for discovering novel entities.
  - Select items relevant to particular entities in a stream of documents for KG construction and completion. The filtered documents can then be fed to a relation extraction system to extract novel facts and triples for the entities.
  - Estimate the quality of individual KG statements based on authority and classification models.
- How IR approaches can be used for creating, improving, and updating KGs?

KG-related Tasks and Approaches

判斷是否有新的entity。

Task and Approaches	Description
Entity discovery (Subsection 5.1)	Decide whether an entity should be added as a new entry to a KG.
Linking-based	Utilize the confidence score from an entity linking system to detect unlinkable entities.
Feature-based	Train a classifier based on features from timestamp and text features in a supervised on semi-supervised fashion.
Expansion-based	Discover new entities similar to a number of seed entities.

判斷entity的型態。

Entity typing (Subsection 5.2)	Decide which type should be assigned to an entity.
Constraint-based	Define a set of class constraints and optimize through, e.g., integer linear programming.
Embedding-based	Learn the association between an entity and a type embedding.
Graph-based	Represent entities' associations with other entities, type context descriptions, and entity descriptions as a graph.
Generative	Build a co-occurrence dictionary of entities and context nouns using translation and generation probabilities.

判斷文件是否包含指定entity的資訊。

Document filtering (Subsection 5.3)	Decide whether a document contains important information about an entity.
Entity-dependent	Learn a model for every entity based on lexical and distributional features.
Entity-independent	Learn a single model for all entities based on distributional features.

給定兩個entities，判斷他們間的關係。

Relation extraction (Subsection 5.4)	Extract entity relationships from text.
Supervised, feature-based	Extract features based on a context within sentences and use supervised machine learning.
Distantly-supervised, feature-based	Extract features based on relation context aggregated from multiple sentences, learn in distantly-supervised fashion.
Semi-supervised pattern extraction	Learn and apply relation patterns in a semi-supervised fashion.

給予一新的entity，判斷可能有的relations。

Link prediction
(Subsection 5.4)

Predict new entity relations given existing relations.

Latent feature models

Learn latent features of entities that explain observable facts and apply to new entities.

由entity下手，應具有類似型態entity擁有的relations

Graph feature models

Predict new edges by learning features from the observed edges in the graph.

由edge下手，可以接在那些edges上

KG correctness estimation (Subsection 5.5) *Sampling-based*	Estimate the quality of set of facts in the KG.
	Predict overall KG accuracy by sampling the facts efficiently.
Triple correctness prediction (Subsection 5.5) *Fusion-based*	Estimate the likelihood of a predicted KG statement.
	Predict triple correctness by aggregating predictions of individual extractors.

評估KG

評估一個KG statement

評估編輯者的信賴度

Contribution quality estimation (Subsection 5.5)	Predict the quality of (parts of an) KG item.
Feature-based	Predict contribution quality based on user contribution history, relation difficulty, and user contribution expertise.
Graph-based	Leverage the graph connecting profiles and editors to estimate the quality of contributions.

故意破壞

Vandalism detection (Subsection 5.5)	Predict whether an edit in a KG is malicious.
Feature-based	Predict vandalism based on content and context features.

評估是否為惡意編輯

Entity Discovery

- The set of entities in a knowledge graph tends to evolve as new entities emerge over time.
- Continuously discover emerging entities in news and other streams.
- TAC-KBP, an evaluation campaign on entity linking and relation extraction.

Definition 5.1 (Entity Discovery). Given a set of documents D and a knowledge graph KG, detect E, i.e., a set of new entities that should be added to KG.

Approaches to Entity Discovery

- Linking-based (運用entity linking，不能連到KG，就間接隱含有新的entity)
  - Discover new entities based on the confidence during linking.
  - Do not aim to solve the discovery of novel entities specially.
- Feature-based (根據candidate entities周遭的特徵，判斷是不是)
  - Treat entity discovery as a prediction problem based on features extracted around the candidate entities.
- Expansion-based (針對有興趣的類型，給予一組entity當作seed，挖掘更多類似的entities)
  - Start with a seed of entities for each type that needs to be populated, and try to extract similar entities.

Linking-Based Approaches to Entity Discovery

- Utilize a global threshold to recognize entities not found in the knowledge base by an entity linking method.
- Entity linking systems, when attempting to link entity mentions in a text segment to KG entities, often generate confidence scores in the linking process.
  - Extract candidates for emerging entities from out-of-KG entities, i.e., ones with low scores with respect to a disambiguation score.
  - Such out-of-KG entities typically occur in similar contexts as known entities of a certain type.
  - Finding a reliable threshold for novel entities might be impractical.

Feature-Based Approaches to Entity Discovery

- Detecting new entities based on their usage characteristics in a corpus over time.
  - Task: Perform classification of unlinkable text segments as *entity*, *non-entity*, or *unclear*.
  - Define an unlinkable text segment to be a noun phrase that cannot be linked to Wikipedia.
  - Entities have different usage characteristics over time than non-entities.
  - Train a classifier with features primarily derived from a time-stamped document collection.
  - Leverage various entity usage statistics from this longitudinal corpus, and augment the feature set by word features of the noun phrases such as capitalization and numeric modifiers.
  - Two annotators labeled 250 unlinked bigrams extracted from OpenIE assertions as *entity*, *non-entity*, or *unclear*.

假設名詞片語是可能的entity candidate。

相對來說，entity會隨著時間有不同的使用情境，而non-entity的使用不會因為時間有很大的變動。

New Entities Are Ambiguous

同一mention字串可能指涉KG中的entity，也可能是新的entity。

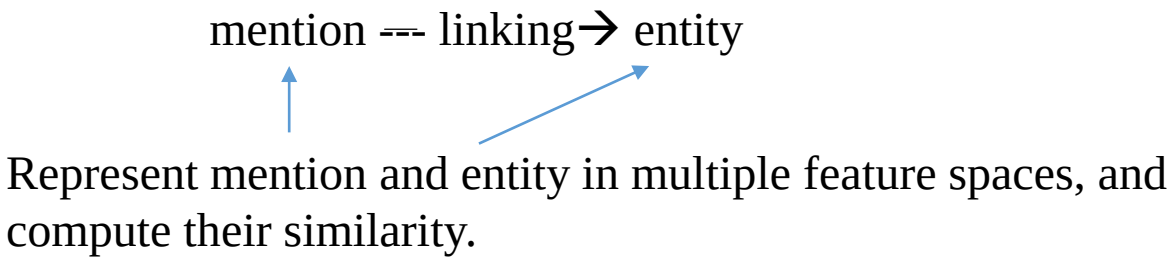
- A mention can refer to not only new but also known entities.
  - Measure the confidence of mapping an ambiguous mention to an existing entity.

這裡的keyphrase可以想像成contextual clue。

  - Represent an ambiguous new entity as a set of weighted keyphrases.
  - Extract descriptive keyphrases of a candidate emerging entity and compute the set difference of those keyphrases and entities already covered in the KG.
  - Cluster different mentions with similar keyphrases as a new emerging entity.

Modeling Mention and Entity Representations into Multiple Feature Spaces

- Incorporate features based on contextual, topical, lexical, neural embedding, and query spaces
  - Semantic relatedness between an entity and a mention: the relatedness of the entity in each embedding space
  - Contextual features: supportive entities, alien entities, and dependent words
  - Query space: include context words found in users' search history surrounding the entities.



Expansion-Based Approaches to Entity Discovery

- Discover more entities within a specific category.
- Leverage existing type and attribute information found in KGs and discover new entities with similar attributes.
  - Take existing entities from a particular Wikipedia category as a seed set and explore their attribute in infoboxes to obtain clues for the discovery of more entities belonging to the same category.
  - Leverages IR techniques by using the clues from Wikidata infoboxes for constructing a query to retrieve web pages that might contain new entities belonging to the same category as the seed entities.
  - The retrieved documents are then considered as candidates for entity extraction.

Relation of Entity Discovery to Other Tasks

- Entity linking: entity discovery can be performed alongside.
- Entity typing: how a specific type of such attributes, i.e., the entity type, can be extracted using entity discovery clues in a corpus.

  entity discovery 和 entity typing可能運用相同的線索
- Document filtering: automatically build an initial profile for a new entity identified through entity discovery.

  Document filtering的目標是選取 (濾出) 含有指定entity的文件，所以可以從中建立profile，作為entity discovery使用

Entity Typing

- Related to the problem of discovering new entities is deciding to which entity type(s) they belong

Definition 5.2 (Entity Typing). Given a KG and a set of documents D_e mentioning an entity e, decide whether type $t \in T$ should be assigned to annotate entity e, where T is a type system in the knowledge graph. The entity type assignment could be a hard, binary assignment or a soft assignment with a relevance score.

Approaches to Entity Typing

- Graph-based approaches
  - Model the relationship between mention, entities, and types as relationships in a graph.
  - Perform inference based on the constructed graphs.
- Constraint-based approaches
  - Define a set of constraints that need to be satisfied.
  - Solve the optimization problem to decide the entity type assignments.
- Embedding-based methods
  - Use contextual features to learn embeddings of entities and mentions.
  - Use the representations to decide on the typing.
- Generative approaches
  - Model the assignment of entities, mentions, and types as a generative process.
  - Learn the parameters of the generative model.

mention鏈結到entity，entity屬於某個type，因此這個mention屬於這個type。

屬於某個type，應具備哪些條件
當mention具備某些條件時，猜測其形態

Graph-Based Approaches to Entity Typing

related context → related entities → related types

- Propagates class labels from labeled to unlabeled instances.
  - Find similar entities that share the same textual relations with a target entity.
  - Predict the types of the target entity from the types of the related entities.
- Joint bootstrapping approach for entity linking and typing
  - A bipartite graphical model for joint type-mention inference, relying on three signals:
  - Entity neighborhood
    - Leverage the direct or indirect information of type information from known parts of the knowledge graph.
    - Infer an entity's type based on types of related entities.
  - Language model
    - Utilize mention contexts from Wikipedia annotated text.
  - Neighborhood match with snippet
    - Utilize the linked related entities in context.

目標：決定mention的type
方法：mention context中出現一些entities (稱為related entities)，運用這些related entities鏈結到KG的type資訊，來決定這個mention的type。

Constraint-Based Approaches to Entity Typing

- Consider the task of both discovering and semantically typing newly emerging out-of-KB entities.
- Use type signatures of relational phrases and type correlation or disjointness constraints
  - Relation patterns that are organized in a type signature taxonomy.
  - The candidate types to be assigned to an entity are determined based on the entity's co-occurrence with a type relational pattern.
  - Start by generating a number of confidence-weighted candidate types for entity *e*.
  - The compatible subsets of candidate entities for an entity *e* are decided with an integer linear programming (ILP) formulation.

不同的relational phrase會接不同的type，例如beBornIn會接人和地
Type和type之間會有關聯性或互斥性

Embedding-Based Approaches to Entity Typing

- Learn classifiers over sparse high-dimensional feature spaces that result from the conjunctive features of the entity mention and its context of occurrence. 給定一個entity，猜測corpus每一次提到時的type
- Aggregate contextual information about an entity from the corpus and then perform classification for each possible candidate type.
- Utilize an entity embedding $\vec{v}(e)$ and make class predictions based on a neural approach.

Entity-Centric Document Filtering

Document filtering 的目的

- Knowledge Base Acceleration (KBA)
  - The maintenance of KGs is performed by periodically recommending a number of relevant documents to KG editors.
  - KG editors will decide whether a document actually contains new facts and formulate the specific inclusion of facts in the KG.
  - Document filtering　提供合適的documents給KG editors
- Knowledge Base Population (KBP)
  - Automatically populate a KG without the involvement of human editors.
  - The selection or filtering of documents from which specific relations are to be extracted.　提供合適的documents給knowledge extractor
  - Applying relation extraction techniques to the selected document collection.

Entity-Centric Document Filtering

- The task of analyzing an ordered stream of documents and selecting those that are relevant to a specific set of entities.

Definition 5.3. Given a stream of documents $\hat{D}$ and an entity e, the *entity-centric document filtering* task is to decide whether a document $d \in \hat{D}$ contains important information about e.

Approaches to Entity-Centric Document Filtering

- Entity-Dependent Approaches
  - Learn a single model for each entity. 判斷相關或不相關 (含有或不含)
  - Detect particular features of specific entities in a stream.
- Entity-Independent Approaches
  - Do not learn the specifics of each entity directly.
  - Compare the distributional features of the entity against incoming documents to decide the relevance.

Entity-Dependent Approaches to Entity-Centric Document Filtering

- Highly-supervised approaches utilizing related entities and bag-of-word features.
- The training data is used to identify keywords and related entities, and classify the documents in the test data.
- A typical entity-dependent approach
  - Pool related entities from the profile page of a target entity.
  - Estimate the weight of each related entity with respect to the query entity. 就是target entity
  - Apply the weighted related entities to estimate confidence scores of streaming documents.
- A query expansion-based approach on relevant entities from the KG
  - Augment the original query terms (i.e., entity name) with other terms that are likely to indicate relevant documents.

就是target entity

Relation Extraction and Link Prediction

Definition 5.4 (Relation Extraction). Given a sentence s containing a pair of entities e_1 and e_2, decide whether e_1 and e_2 are connected through a relation of type r.

Definition 5.5 (Link Prediction). Given a set of facts F where each fact is a triple of entity relations in KG, predict the existence of other relations between two entities e_i and e_j within relation type r, where e_i, e_j is in KG.

Approaches to Relation Extraction and Link Prediction

- Supervised approaches
  - Having training labels for each relation instance and contextual text expression of the relations.
  - Apply either features or learns kernel functions to classify the training data correctly.
- Distant supervision approaches
  - Utilize known relations, but without having the context in which the relations are expressed in a piece of text.
  - Any extracted context will be considered and will contribute in the prediction.
- Semi-supervised approaches
  - Extract textual patterns around known relations.
  - Use them to discover more relations.

Supervised, Feature-Based Approaches to Relation Extraction

- Feature-based methods
  - Syntactic and semantic features are extracted from the text.
  - Syntactic features: the entities, the types of the entities, word sequences between the entities, and the number of words between the entities.
  - Semantic features: derived from the path in the parse tree containing the two entities.

Distantly-Supervised, Feature-Based Approaches to Relation Extraction

- Generate training data for distant supervision.
- Pseudo-training data.
- Take into account the uncertainty of instance labels during training.
- Feature engineering vs. CNN with piecewise max pooling.

Semi-Supervised Pattern Extraction

- Bootstrapping/semi-supervised approaches
  - Start with a small number of seed relation instances.
  - Learn a general textual pattern that will apply to these relations.
  - Apply the newly discovered patterns to discover more relations.
- Open relation extraction (Open IE)
- Two challenges
  - Incoherent extraction
    - Cases where an extracted relation phrase has no meaningful interpretation.
  - Uninformative extraction
    - Extractions omit critical information.

Approaches to Link Prediction

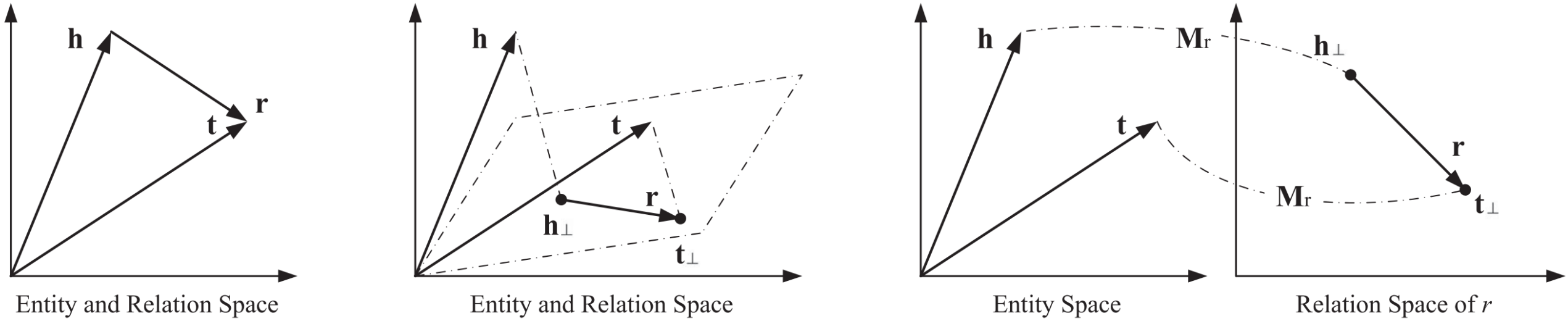
- Latent-feature approaches
  - Model the attributes of entities (including relationships to other entities) to learn latent representations of entities used to predict links between two entities.
- Graph-based approaches
  - Apply graph algorithms (e.g., random walks) to discover potential connections between entities.
  - Compute the likelihood of the relation.

Latent Feature Models to Link Prediction

- Factorization models
  - Learn a distributed representation for each entity and each relation.
  - Make predictions by taking the inner products of the representations.
- TransE
  - Model relationships by interpreting them as translations operating on the low-dimensional embeddings of entities.
  - For two entities *e* and *e'*, the embedding of entity *e* should be close to the embedding of entity *e'* plus some vector that depends on the relationship between the two entities.
  - Learn only one low-dimensional vector for each entity and each relationship.
- TransH (an improvement of TransE)
  - Consider certain mapping properties of relations including reflexive, one-to-many, many-to-one, and many-to-many relations.

Latent Feature Models to Link Prediction

- TransR
  - Build entity and relation embeddings in separate spaces
  - Build translations between projected entities



(a) TransE. (b) TransH. (c) TransR.

Graph-Based Models for Link Prediction

- Path Ranking Algorithm (PRA)
  - Perform random walks of bounded lengths to predict relations.
  - Learn the likelihood of each relation path, combining a bounded number of adjacent relations.
- Apply PRA to perform inference over the augmented graph
  - Enriching KGs with additional edges
    - Label additional edges with latent features mined from a large dependency-parsed corpus of 500 million web documents.
  - Bridging entities
    - Augment a KG not only with edges but also with so-called bridging entities mined from web text corpus.

KG Quality Estimation

- Automatic quality estimation of KGs is a relatively new area.
- Ensure the quality of facts contained in knowledge graphs.
- Approaches to KG Quality Estimation
  - Quality estimation at the set level
    - Evaluate the overall quality of a set of facts in a KG
  - Quality estimation at the unit level
    - Evaluate the quality of each individual unit in the KG, i.e., at the triples or edit level.

Sampling-Based Approaches to KG Correctness Estimation

- Sampling a number of facts and evaluate them manually.
- Discover the right sampling strategies that would minimize the number of annotations required while obtaining a reliable estimate.
- KGEval
  - Rely on coupling constraints
  - The idea that some facts in the KG are connected and, for each group of connected facts, evaluating a representative subset of the group facts would be sufficient.
- Coupling constraints can be derived from the ontology of the KG and also link prediction algorithms.

References

- Ridho Reinanda, Edgar Meij and Maarten de Rijke (2020). “Knowledge Graphs: An Information Retrieval Perspective”, Foundations and Trends in Information Retrieval: Vol. 14, No. 4, pp 289-444. DOI: 10.1561/1500000063.