Disclaimer

- The material provided in this document is not my original work and is a summary of some one else's work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
- Downloaded from najeebkhan.github.io

Identifying Implicit Relationships

Presented By
나집
20135501
3rd June 2013

Outline

- Introduction
- Spreading Activation for Concept Expansion
- Knowledge Resources Used
- Application to Common Bond Questions
- Application to Missing Link Questions
- Experimental Results
- Conclusion

Introduction

Introduction

- Answering natural-language questions may often involve identifying hidden associations and implicit relationships

Introduction

- Answering natural-language questions may often involve identifying hidden associations and implicit relationships
- What Teddy Roosevelt and Barack Obama have in common?

Introduction

- Answering natural-language questions may often involve identifying hidden associations and implicit relationships
- What Teddy Roosevelt and Barack Obama have in common?
- How old was the youngest U.S. president when he took office?

Introduction Contd…

Introduction Contd…

- To answer these questions we need to identify concepts that are closely related to those given in the question

Introduction Contd…

- To answer these questions we need to identify concepts that are closely related to those given in the question
- IBM Watson uses recursive spreading-activation algorithm, which identifies related concepts based on a collection of heterogeneous underlying data resources

Spreading Activation for Concept Expansion

Spreading Activation for Concept Expansion

- Spreading activation refers to the idea that concepts in a semantic network may be activated through their connections with already active concepts based on a certain spreading strategy

Spreading Activation for Concept Expansion

- Spreading activation refers to the idea that concepts in a semantic network may be activated through their connections with already active concepts based on a certain spreading strategy
- This process allows us to identify concepts closely related to a given concept and to score the relatedness between two concepts

Spreading Activation for Concept Expansion

Spreading Activation for Concept Expansion

- The spreading-activation interface allows for the specification of fan size f and depth d, which causes the process to identify the f most related concepts to the current active concept and to recursively invoke the activation process on these f new concepts another d-1 times

Spreading Activation for Concept Expansion

- The spreading-activation interface allows for the specification of fan size f and depth d, which causes the process to identify the f most related concepts to the current active concept and to recursively invoke the activation process on these f new concepts another d-1 times
- Watson uses three different underlying resources to measure relatedness: an n-gram corpus, the PRISMATIC knowledge base, and Wikipedia

Using an N-gram Corpus

Using an N-gram Corpus

- Lexical collocation in naturally occurring text is a simple and natural measure for concept relatedness

Using an N-gram Corpus

- Lexical collocation in naturally occurring text is a simple and natural measure for concept relatedness
- JFK is semantically strongly related to ‘airport’ and ‘assassination’, and these relationships are represented in the n-gram corpus by the high collocation frequency between the terms ‘JFK’ and ‘airport’

Using an N-gram Corpus

- Watson uses a 5-gram corpus with frequency counts from Watson's primary unstructured sources
- The n-gram based spreading activation implementation uses Lucene
- Given term t the f most frequent 5-grams that include t are retrieved from the corpus
- Given two terms, the NGD semantic similarity metric was used to compute the semantic distance between two given terms based on the underlying n-gram corpus

Using the PRISMATIC Knowledge Base

Using the PRISMATIC Knowledge Base

- A knowledge base of extracted frames and slots based on syntactic and semantic relationships

Using the PRISMATIC Knowledge Base

- A knowledge base of extracted frames and slots based on syntactic and semantic relationships

- SVO query (Ford, ?v, ?o) → a count of all SVO tuples for which Ford is the subject

Using the PRISMATIC Knowledge Base

- A knowledge base of extracted frames and slots based on syntactic and semantic relationships

- SVO query (Ford, ?v, ?o) → a count of all SVO tuples for which Ford is the subject
- We use PRISMATIC for estimating the degree of relatedness between two concepts with the frequency of how often they co-occur

Using the PRISMATIC Knowledge Base

- A knowledge base of extracted frames and slots based on syntactic and semantic relationships

- SVO query (Ford, ?v, ?o) → a count of all SVO tuples for which Ford is the subject
- We use PRISMATIC for estimating the degree of relatedness between two concepts with the frequency of how often they co-occur
- "Ford did not act hastily but did finally pardon Nixon in September" → (Ford, pardon, Nixon)

Using the PRISMATIC Knowledge Base

- A knowledge base of extracted frames and slots based on syntactic and semantic relationships
- Provides quick access to statistics over tuples
- SVO query (Ford, ?v, ?o) → a count of all SVO tuples for which Ford is the subject
- We use PRISMATIC for estimating the degree of relatedness between two concepts with the frequency of how often they co-occur
- "Ford did not act hastily but did finally pardon Nixon in September" → (Ford, pardon, Nixon)

Using Wikipedia Links

Using Wikipedia Links

- The metadata encoded in Web documents is used, rather than the texts of the documents themselves

Using Wikipedia Links

International Business Machines (IBM) (NYSE: IBM) is an ⟨American | United States⟩ multinational ⟨technology⟩ and ⟨consulting⟩ firm headquartered in ⟨Armonk, New York⟩. IBM manufactures and sells computer ⟨hardware | Personal computer hardware⟩ and ⟨software | Computer software⟩ and it offers ⟨infrastructure⟩, ⟨hosting | Internet hosting service⟩ and ⟨consulting services | Consultant⟩ in areas ranging from ⟨mainframe computers | Mainframe computer⟩ to ⟨nanotechnology⟩.

Using Wikipedia Links

- The metadata encoded in Web documents is used, rather than the texts of the documents themselves

Using Wikipedia Links

- The metadata encoded in Web documents is used, rather than the texts of the documents themselves
- Anchor texts are oftentimes the same as the target document titles in Wikipedia

Using Wikipedia Links

- The metadata encoded in Web documents is used, rather than the texts of the documents themselves
- Anchor texts are oftentimes the same as the target document titles in Wikipedia
- In cases where they differ, we attempt to capture semantic relatedness using the target document titles for two reasons

Using Wikipedia Links

- The metadata encoded in Web documents is used, rather than the texts of the documents themselves
- Anchor texts are oftentimes the same as the target document titles in Wikipedia
- In cases where they differ, we attempt to capture semantic relatedness using the target document titles for two reasons
  - Anchor texts frequently co-occur with the source document title

Using Wikipedia Links

- The metadata encoded in Web documents is used, rather than the texts of the documents themselves
- Anchor texts are oftentimes the same as the target document titles in Wikipedia
- In cases where they differ, we attempt to capture semantic relatedness using the target document titles for two reasons
  - Anchor texts frequently co-occur with the source document title
  - the target document title represents the canonical form

Using Wikipedia Links

Using Wikipedia Links

- Given term t, we identify the Wikipedia document whose title best matches t and return all target document titles from links in that document

Application to Common-Bond Questions

Application to Common-Bond Questions

- Common-Bond questions generally refer to questions that seek the hidden relationship among multiple entities

Application to Common-Bond Questions

(1) COMMON BONDS: Bobby, bowling, rolling. (Answer: “pins”)

(2) COMMON BONDS: Your legs, your T’s, the Rubicon. (Answer: “things you cross”)

(3) CULINARY COMMON BONDS: Grinder, hero, submarine. (Answer: “sandwiches”)

(4) COMMON BONDS: Shirts, TV remote controls, telephones (Answer: “things with buttons”)

Application to Common-Bond Questions

Application to Common-Bond Questions

- Focusing on the commonality among these examples, the answers are all semantically closely related to the given entities.

Application to Common-Bond Questions

- Focusing on the commonality among these examples, the answers are all semantically closely related to the given entities.
- This observation of semantic relatedness enables us to adopt the spreading-activation mechanism previously outlined as a principal method for answering common-bond questions

Application to Common-Bond Questions

Application to Common-Bond Questions

- Spreading activation is used to answer common-bond questions in two ways

Application to Common-Bond Questions

- Spreading activation is used to answer common-bond questions in two ways
  - Identify concepts that are closely related to each given entity

Application to Common-Bond Questions

- Spreading activation is used to answer common-bond questions in two ways
  - Identify concepts that are closely related to each given entity
  - Score each concept on the basis of their degrees of relatedness to all given entities

Common-Bond Candidate Generation

Common-Bond Candidate Generation

- Analysis of common-bond questions in Jeopardy! indicates that the answer is typically directly related to the given entities

Common-Bond Candidate Generation

- Analysis of common-bond questions in Jeopardy! indicates that the answer is typically directly related to the given entities
- So depth d=1 and f is empirically evaluated to be 50

Common-Bond Candidate Generation

- Analysis of common-bond questions in Jeopardy! indicates that the answer is typically directly related to the given entities
- So depth d=1 and f is empirically evaluated to be 50
- The spreading-activation process is invoked on each entity given in the question

Common-Bond Candidate Generation

- Analysis of common-bond questions in Jeopardy! indicates that the answer is typically directly related to the given entities
- So depth d=1 and f is empirically evaluated to be 50
- The spreading-activation process is invoked on each entity given in the question
- For most questions, the common bond can be found in lexical proximity to the given entities so only n-gram corpus was used for common bond questions

Common-Bond Answer Scorer

Common-Bond Answer Scorer

- NGD similarity score is computed using the n-gram corpus

Common-Bond Answer Scorer

- NGD similarity score is computed using the n-gram corpus
- The scores representing the candidate's semantic relatedness to the given entities are multiplied together to represent the overall goodness of the candidate as a common-bond answer

Application to Missing-Link Questions

Application to Missing-Link Questions

- Questions in which a missing entity is either explicitly or implicitly referred to and the identification of this missing entity facilitates answering the question

Application to Missing-Link Questions

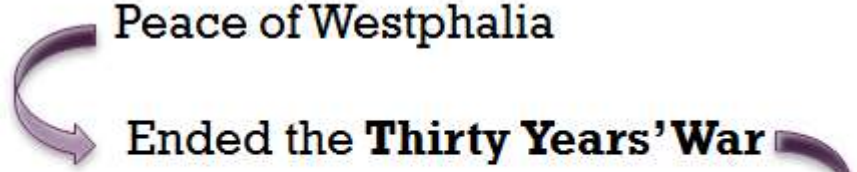
(5) THE 17th CENTURY: The 1648 Peace of Westphalia ended a war that began on May 23 of this year. (Answer: “1618”)

(6) EXPLORERS: On hearing of the discovery of George Mallory’s body, this explorer told reporters he still thinks he was first. (Answer: “Edmund Hillary”)

Application to Missing-Link Questions

(5) THE 17th CENTURY: The 1648 Peace of Westphalia ended a war that began on May 23 of this year. (Answer: "1618")

(6) EXPLOR discovery of George xplorer told reporters rst. (Answer: "Edmund Hillary")



Application to Missing-Link Questions

(5) THE 17th CENTURY: The 1648 Peace of Westphalia ended a war that began on May 23 of this year. (Answer: “1618”)

(6) EXPLORERS: On hearing of the discovery of George Mallory’s body, this explorer told reporters he still thinks he was first. (Answer: “Edmund Hillary”)

Application to Missing-Link Questions

(5) THE 17th Cl 1648 Peace of Westphalia e began on May 23 of this year. ")

George Mallory

Mount Everest

Edmund Hillary

(6) EXPLORERS: On hearing of the discovery of George Mallory's body, this explorer told reporters he still thinks he was first. (Answer: "Edmund Hillary")

Application to Missing-Link Questions

(5) THE 17th CENTURY: The 1648 Peace of Westphalia ended a war that began on May 23 of this year. (Answer: “1618”)

(6) EXPLORERS: On hearing of the discovery of George Mallory’s body, this explorer told reporters he still thinks he was first. (Answer: “Edmund Hillary”)

Application to Missing-Link Questions

Application to Missing-Link Questions

- Our approach to answering these missing-link questions is to

Application to Missing-Link Questions

- Our approach to answering these missing-link questions is to
  - First hypothesize the missing links

Application to Missing-Link Questions

- Our approach to answering these missing-link questions is to
  - First hypothesize the missing links
  - Invoke the system again by including these missing links in the search process, with the hope that the new search results will include some correct answers that we previously failed to generate as candidate answers

Missing-Link Identification

Missing-Link Identification

- For an entity to be a good missing link

Missing-Link Identification

- For an entity to be a good missing link
  - It must be highly related to concepts in the question

Missing-Link Identification

- For an entity to be a good missing link
  - It must be highly related to concepts in the question
  - It must be ruled out as a possible correct answer

Missing-Link Identification

- For an entity to be a good missing link
  - It must be highly related to concepts in the question
  - It must be ruled out as a possible correct answer
- The existing WATSON's components are used to find highly associated answers

Missing-Link Identification

- For an entity to be a good missing link
  - It must be highly related to concepts in the question
  - It must be ruled out as a possible correct answer
- The existing WATSON's components are used to find highly associated answers
- Then the missing links are the ones which are of the wrong type

Missing-Link Identification

- For an entity to be a good missing link
  - It must be highly related to concepts in the question
  - It must be ruled out as a possible correct answer
- The existing WATSON's components are used to find highly associated answers
- Then the missing links are the ones which are of the wrong type
- 'thirty year war', highly associated with the question but its type is not year

Candidate Generation using Missing Links

Candidate Generation using Missing Links

- If one or more missing links exist for the question, we invoke part of the question-answering process again, taking the missing link(s) into consideration

Candidate Generation using Missing Links

- If one or more missing links exist for the question, we invoke part of the question-answering process again, taking the missing link(s) into consideration
- The scoring of the candidate answers is done based on its relatedness to the missing link

Experimental Evaluation

Experimental Evaluation

- Common Bond Questions

	Binary recall	*Accuracy*	*Precision@70*
Baseline	69%	48%	62%
+Common bond	73%	58%	73%
Percentage change	4%	10%	11%

Experimental Evaluation

Experimental Evaluation

- Missing Link Questions

	All questions (1,112)		*Missing-link subset (259)*	
	Binary recall	*Accuracy*	*Binary recall*	*Accuracy*
Baseline	74.82%	51.08%	74.1%	45.6%
+Missing link	75.63%	51.53%	76.5%	47.1%
Percentage change	0.81%	0.45%	2.4%	1.5%

Conclusion

Conclusion

- We have described a spreading-activation approach for concept expansion and for measuring semantic relatedness

Conclusion

- We have described a spreading-activation approach for concept expansion and for measuring semantic relatedness
- We have shown how this technique can be adopted in an end-to-end question-answering system to more effectively address two types of Jeopardy! questions, i.e., common-bond and missing link questions