A Comparison on Two Approaches in Machine Comprehension

Qinbo Li

Problem

Machine Comprehension

SQuAD (Stanford Question Answering Dataset)

- Given passages from wikipedia articles
- Answering questions based on the passage
- Answers come from a span of text

SQuAD

The league announced on October 16, 2012, that the two finalists were Sun Life Stadium and Levi's Stadium. The South Florida/Miami area has previously hosted the event 10 times (tied for most with New Orleans), with the most recent one being Super Bowl XLIV in 2010. The San Francisco Bay Area last hosted in 1985 (Super Bowl XIX), held at Stanford Stadium in Stanford, California, won by the home team 49ers. The Miami bid depended on whether the stadium underwent renovations. However, on May 3, 2013, the Florida legislature refused to approve the funding plan to pay for the renovations, dealing a significant blow to Miami's chances.

When were the two finalists for hosting Super Bowl 50 announced?
Ground Truth Answers: October 16, 2012 | October 16, 2012, | October 16, 2012
Prediction: October 16, 2012

How many times has the South Florida/Miami area hosted the Super Bowl?
Ground Truth Answers: 10 | 10 | 10
Prediction: 10

Logistic Regression

1. Preprocess
2. Candidate generation
3. Feature extraction
4. Training

Logistic Regression

1. Preprocess
   - Splitting sentence
   - Word tokenize
   - Save vocabulary

Logistic Regression

2. Candidate generation

All possible candidates: O(L^2)

Constituency parse tree

a. Generate phrase based on the tree

Logistic Regression

```
(ROOT
  (NP
    (NP (NNP Super) (NNP Bowl))
      (SBAR
        (S
          (NP (CD 50))
          (VP (VBD was)
          (NP
            (NP (DT an) (JJ American) (NN football) (NN game))
            (SBAR
              (S
                (VP (TO to)
                  (VP (VB determine)
                    (NP
                      (NP (DT the) (NN champion))
                      (PP (IN of)
                      (NP (DT the) (NNP National) (NNP Football) (NNP League))))))))))))
      (PRN
        (-LRB- -LRB-)
        (NP (NNP NFL))
        (-RRB- -RRB-)))
  (PP
    (IN for)
    (NP (DT the) (CD 2015) (NN season)))
  (. .)
)
```

Logistic Regression

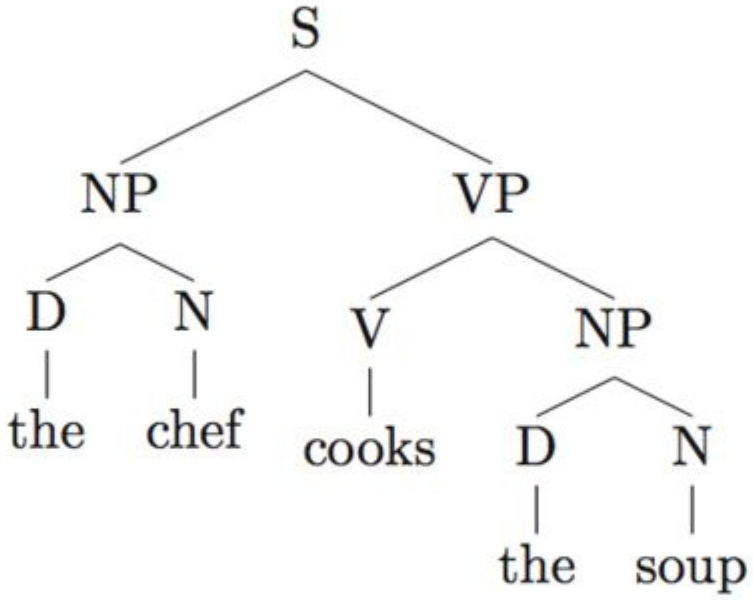
DFS on constituency tree

The chef, the soup, cooks the

Soup, the chef cooks the soup

76%

75% (limitation on length)



Logistic Regression

Add bigram and trigram

Logistic Regression

Approach	Percentage	Avg. candidate size
DFS (no limit on length)	76%	206
DFS	75%	182
Bigram	68%	212
Bigram & Trigram	80%	341
DFS & Bigram	82%	258
DFS & Bigram & Trigram	86%	373

Logistic Regression

3. Feature extraction

a. TF-IDF

b. TF-IDF inside the span

c. word exists in the question

d. length

Logistic Regression

3. Feature extraction

e. Wh-word and constituency label

Combine two components in one feature:

Wh-word & constituency probability

Logistic Regression

Wh-word	Top 1	Top 2	Top 3
which	NP	NNP	JJ
where	NP	NNP	PP
what	NP	NN	NNP
who	NP	NNP	NNS
when	CD	NP	PP
how many	CD	QP	NP
how much	NP	JJ	CD
how old	CD	JJ	PP
how	NP	PP	NNS
none	NP	VP	NN

Logistic Regression

f. Average candidate word similarity
g. Neighbor word similarity vector

Logistic Regression

4. Training and evaluate

91.2% accuracy on training

Feature ablation: wh-word & constituency probability - 85%

Logistic Regression

After each team punted, Panthers quarterback Cam Newton appeared to complete a 24-yard pass Jerricho Cotchery, but the call was ruled an incompletion and upheld after a replay challenge. CBS analyst and retired referee Mike Carey stated he disagreed with the call and felt the review clearly showed the pass was complete. A few plays later, on 3rd-and-10 from the 15-yard line, linebacker Von Miller knocked the ball out of Newton's hands while sacking him, and Malik Jackson recovered it in the end zone for a Broncos touchdown, giving the team a 10–0 lead. This was the first fumble return touchdown in a Super Bowl since Super Bowl XXVIII at the end of the 1993 season.

Which former referee served as an analyst for CBS?
Ground Truth Answers: Mike Carey | Mike Carey | Carey
Prediction: Mike Carey

Logistic Regression

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

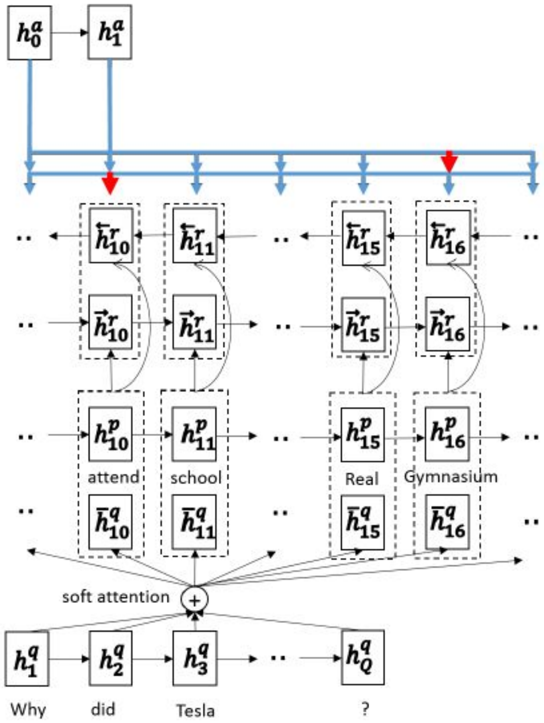
Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos | Denver Broncos | Denver Broncos

Prediction: Levi 's Stadium

Match-LSTM with Answer pointer

1. Preprocess LSTM layer
2. match-LSTM layer
3. Answer pointer layer



Thank you!

Machine Reading Comprehension

CSCE 689 600
Anurag Kapale
927001381

Machine Comprehension

Passage (P) + Question (Q) ⟶ Answer (A)

SQuAD: Stanford Question Answering Dataset

Passage: Selected from Wikipedia
Questions: Crowdsourced
Answer: span in the passage

Machine Comprehension

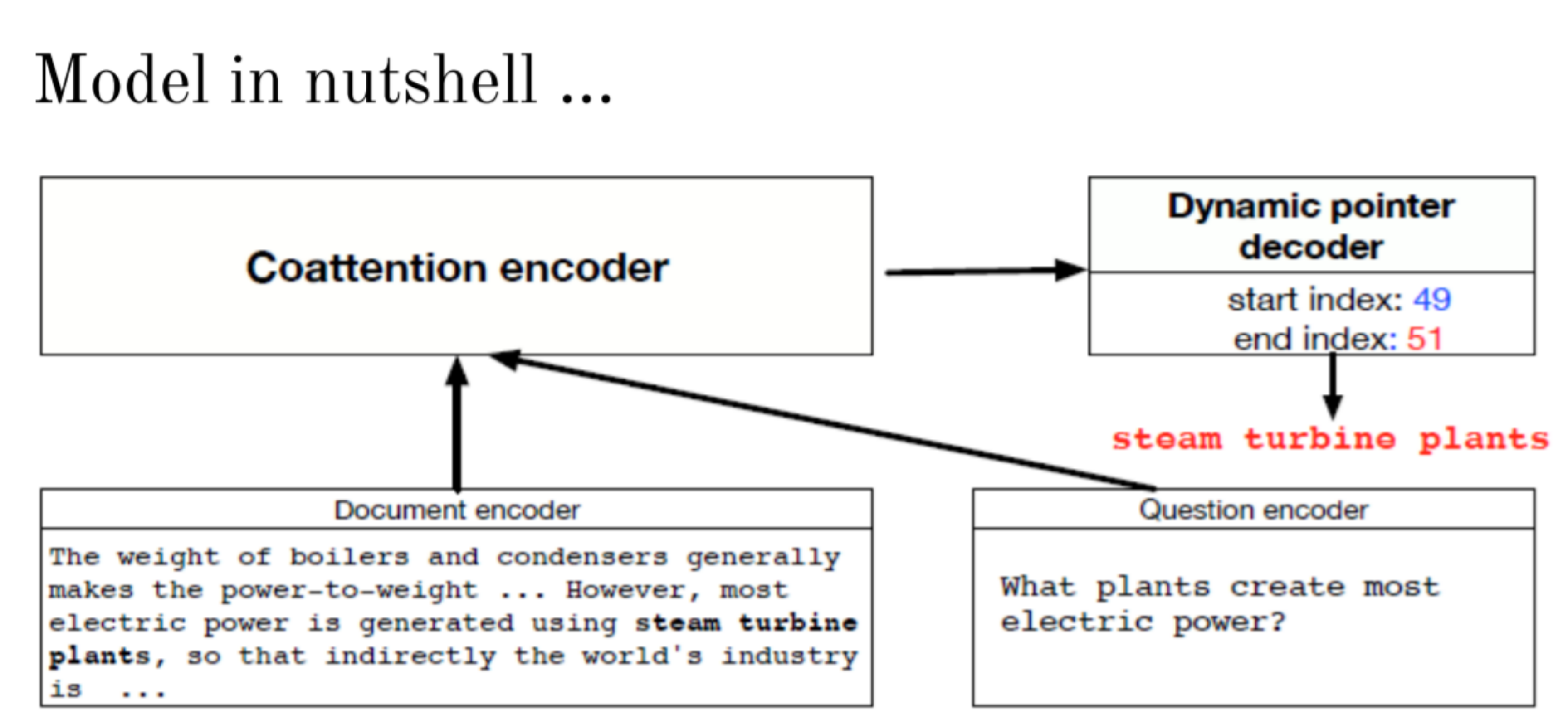
Passage (P) + Question (Q) $\longrightarrow$ Answer (A)

SQuAD: Stanford Question Answering Dataset

Who did **Genghis Khan unite before he** began **conquering** the rest of **Eurasia**?

He came to power by **uniting** many of the nomadic tribes of Northeast Asia. **After** founding the Mongol Empire and being proclaimed "**Genghis Khan**", he started the Mongol invasions that resulted in the **conquest** of most of **Eurasia**. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.

Dynamic Co-attention Networks



MCTest Dataset

Passage: Children's stories

Questions: Crowdsourced

Answer: Multiple Choice Questions

Passage: ...John asked Tim if he could play on the slide. Tim said no. John was very upset and started crying. A girl named Susan saw him crying. Susan told the teacher Ms. Tammy. ...

Question: Who saw John crying and told Ms. Tammy?

A) Tim

B) Susan

C) John

D) Ms. Tammy

Approaches

A) Neural Networks Based

- Generalizable soft features
- LSTMs with attention based mechanisms
- Requires huge training data and time

B) Feature Based

- Explainability
- Related to the techniques learned in the class
- Works with relatively less data

Implementation: Features

Goal: Maximize Pr(P, Q, A_i)

1. Sliding Window:

- Within a sliding window in P, number of word matches to the Q+A.
- To prevent boosting by trivial words, weight using inverse frequency
- Window size k = 2 to 30. (weighted sum)

Implementation: Features

Goal: Maximize Pr(P, Q, A_i)

2. Distance Features:

- Minimize the distance between question and the answer.

$$d_i = \min_{q \in S_Q, a \in S_{A,i}} d(q, a),$$

with

$$S_Q = (Q \cap PW) \backslash U,$$

and

$$S_{A,i} = (A_i \cap PW) \backslash (U \cup Q).$$

Implementation: Features

Goal: Maximize Pr(P, Q, A_i)

3. Word Embeddings:

- Find similarities between Q-A pairs and sentence *s* in the passage.
- Cosine between:
  - sum of word embeddings of Q-A
  - sum of embeddings of sentence in passage

$$F'(P, Q, A_i) = \max_{s \in P} g(P, Q, A_i, s).$$

Implementation: Features

Goal: Maximize Pr(P, Q, A_i)

4. Coreference Resolution:

- Should not differentiate between ‘Mr. Trump’ and ‘the president’
- Enhance previous word matching by pre-processing with Coreference resolution.
- Library from StanfordCoreNLP

Implementation: Features

Goal: Maximize Pr(P, Q, A_i)

5. Word Dependencies:

- Transform Q-A pair into statement using grammar rules.
- Compare dependency tree parsing between sentences.

Q: What did he do on Tuesday?
A: He went to school.
Generated: He went to school on Tuesday.

Implementation: Classifier

- Classifier: Shallow Neural Network
- Use above 5 features and 1 hidden layer.

Results

Feature	Accuracy
Random Guess	0.231
Sliding Window (SW)	0.456
SW + Distance (D)	0.468
SW + D + Word Embeddings (WE)	0.514
SW + D + WE+ Coreference (C)	0. 513
SW + D + WE + C + Word Dependencies	0.521
Human Performance	0.92

Analysis

Q: What animals dropped on his ice-cream cone?
*A) A spider and a fly
B) Spider and a pig
C) Fly and a bee
D) Spider and a bee

P: ...she sometimes let him get a treat if he was helpful
Q: What can James get at the store if he is well behaved?

Type	Accuracy
Who	0.48
How	0.43
When	0.66
Which	0.35
Why	0.31
Count	0.38
What	0.76
Where	0.63
Other	0.52

Thank You!!!

TEXAS A&M UNIVERSITY

Project Presentation

Using NLP to Automatically Map the Networks within the Plans of Houston

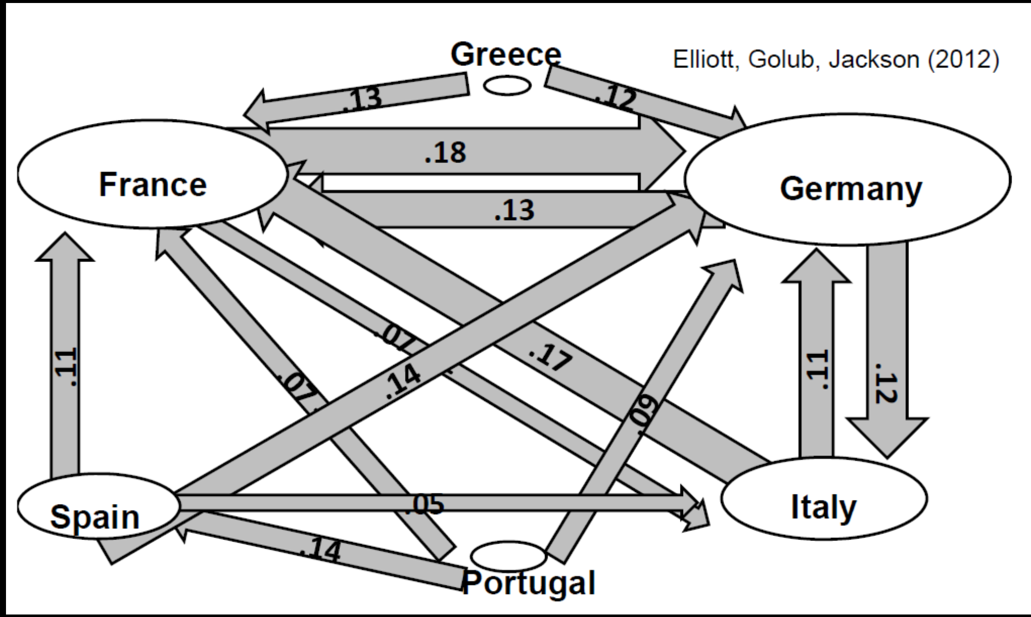
Qingchun Li 4.19.2018

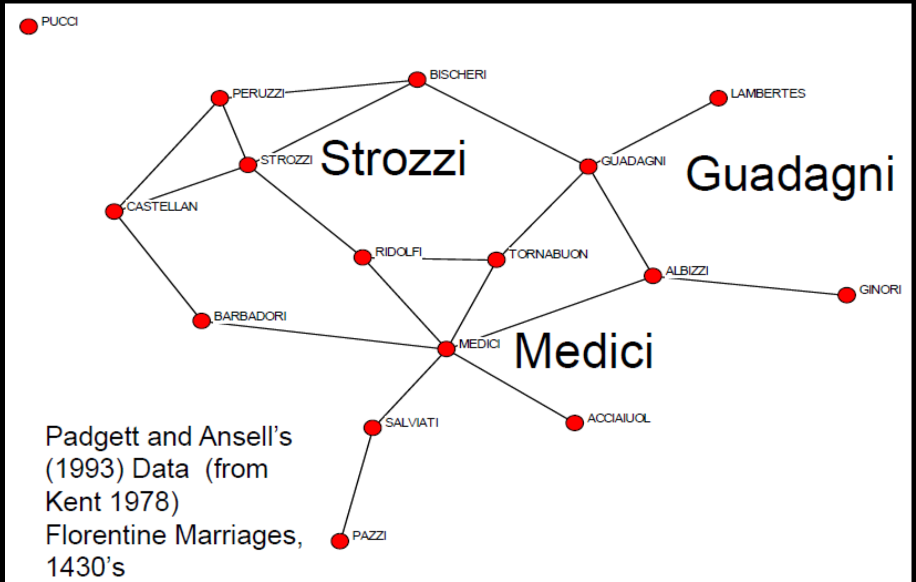
Outline

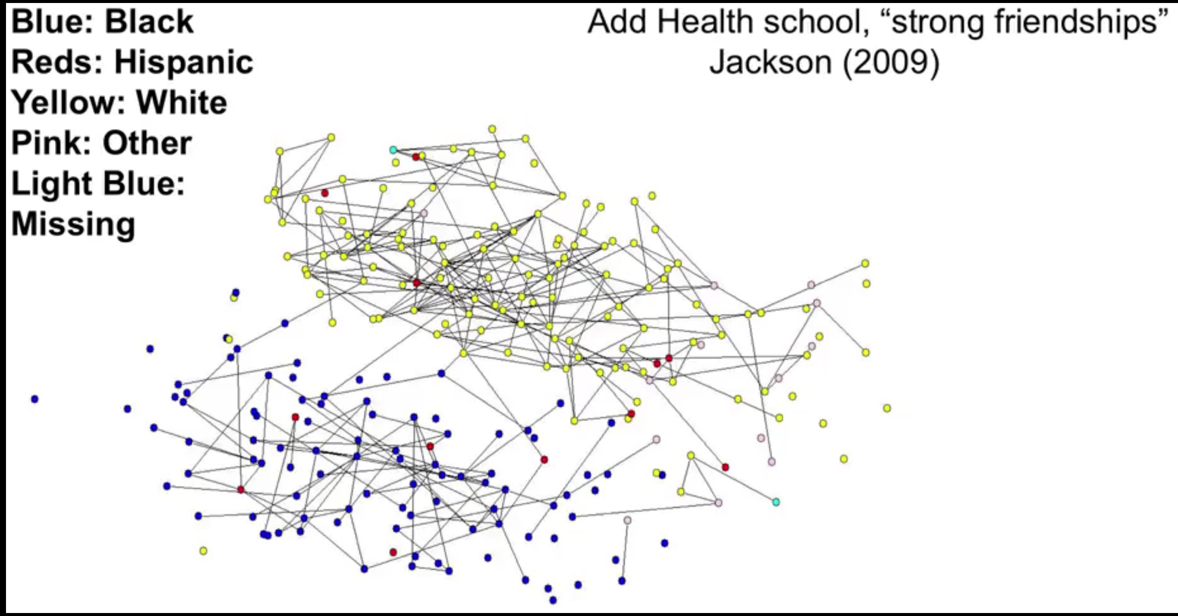
- Motivation
- Data and Methodology
- Detail Steps to Implement the Algorithm
- Results and Evaluation
- Discussion and Future Work

Motivation:

- Mapping the network automatically:
  - Why to map network
  - Why automatically

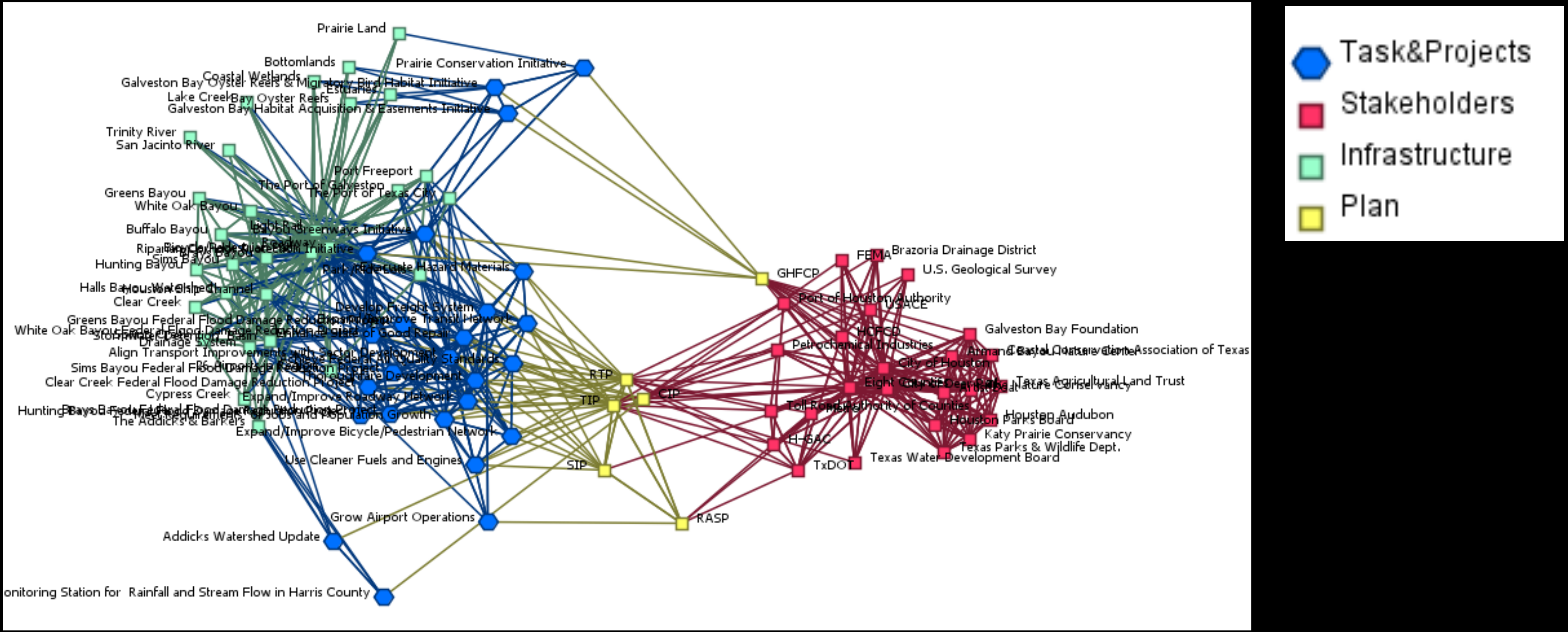






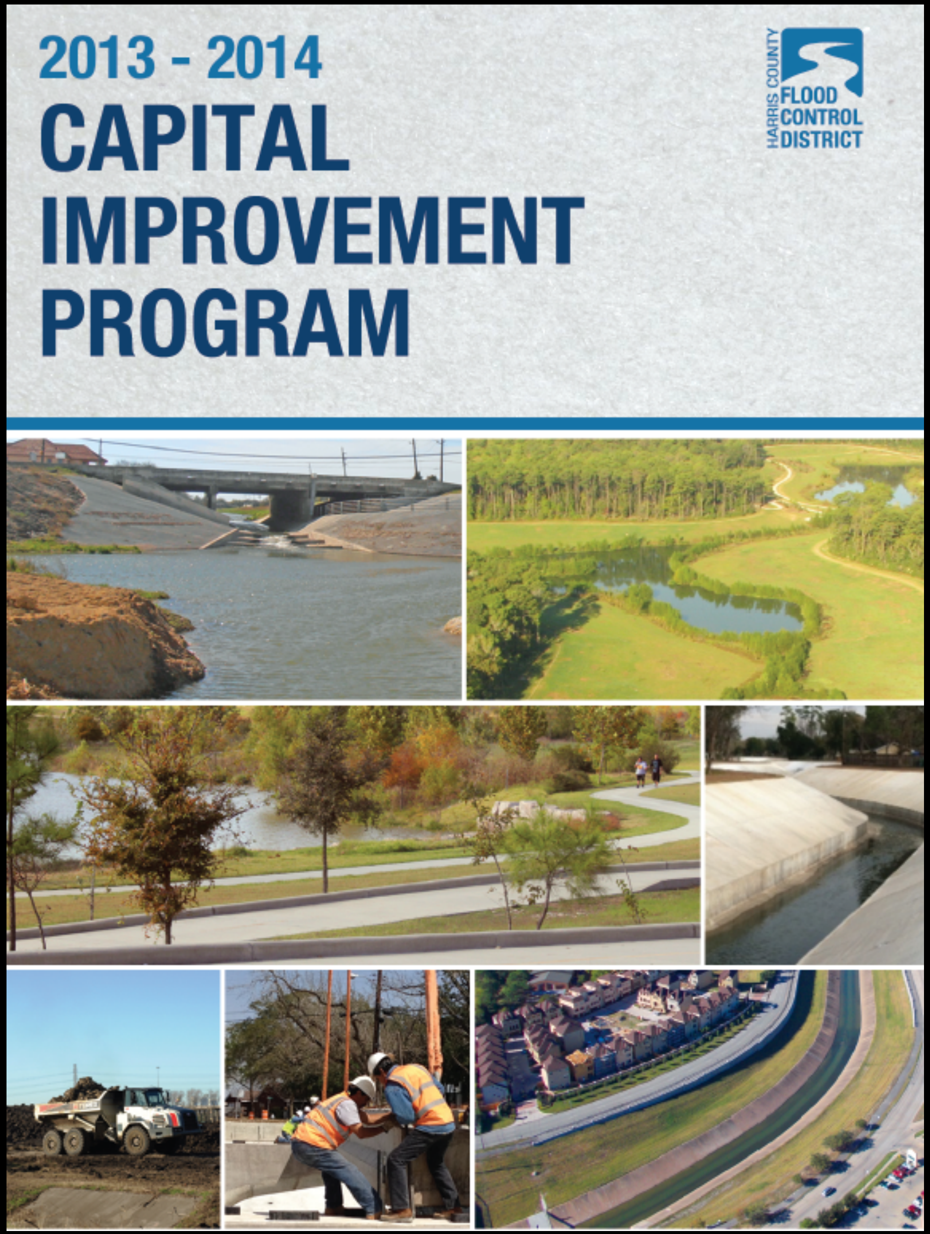
Motivation-ctd:

- Mapping the network automatically:
  - Why to map network
  - Why automatically



Data and Methodology:

- 2040 Regional Transportation Plan
- Capital Improvement Program: Harris County Flood Control District
- Gulf-Houston Regional Conservation Plan



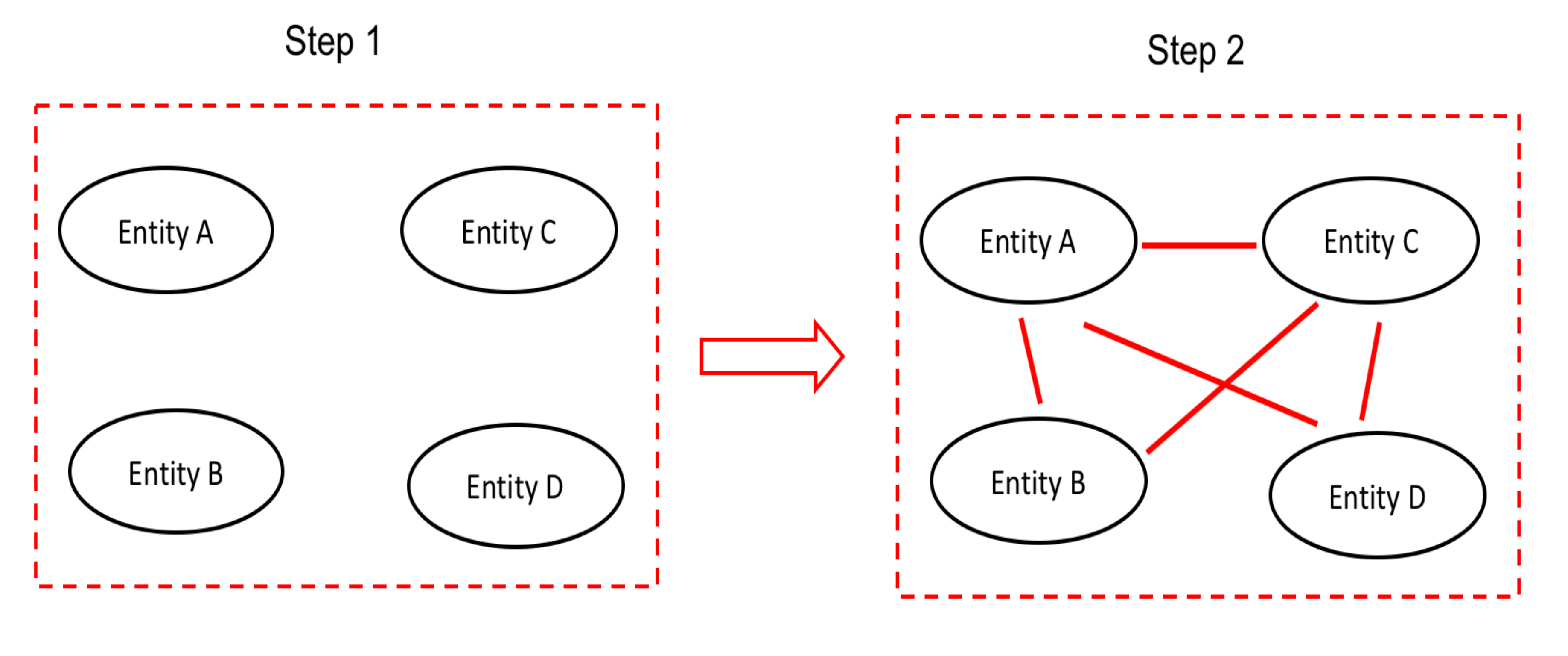
Data and Methodology-ctd:

- Entities and relationships of the urban system network

	Agents & Organizations	Plans & Policies	Tasks & Projects	Infrastructure
Agent & Organizations	Social Network	Plan-Development Network	*Left blank*	Infrastructure Support Network
Plans & Policies		Institutional Network	Task-Assignment Network	*Left blank*
Tasks & Projects			Task-Flow Network	Infrastructure Renewal & Retrofit Network
Infrastructure				Infrastructure System Network

Data and Methodology-ctd:

- Framework of Methodology to automapping the network



Data and Methodology-ctd:

- Obstacles to extract the relationships:
  - Distance
  - Not implied in the context
  - Relationships in a new domain
  - Etc……
- Focus on the first step:
  - No annotated data
  - A weakly supervised bootstrapping algorithm

Batista, D. S., Martins, B., & Silva, M. J. (2015). Semi-supervised bootstrapping of relationship extractors with distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 499-504).

Choubey, P. K., & Huang, R. (2017). Event Coreference Resolution by Iteratively Unfolding Inter-dependencies among Events. *arXiv preprint arXiv:1707.07344*.

De Marneffe, M. C., & Manning, C. D. (2008). *Stanford typed dependencies manual* (pp. 338-345). Technical report, Stanford University.

Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. *arXiv preprint arXiv:1710.07394*.

Huang, R., & Riloff, E. (2013). Multi-faceted event recognition with bootstrapped dictionaries. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 41-51).

Jain, A., Kasiviswanathan, G., & Huang, R. (2016). Towards Accurate Event Detection in Social Media: A Weakly Supervised Approach for Learning Implicit Event Indicators. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)* (pp. 70-77).

Riloff, E. (1996, August). Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence* (pp. 1044-1049).

Riloff, E., & Jones, R. (1999, July). Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI* (pp. 474-479).

Riloff, E., & Phillips, W. (2004). *An introduction to the sundance and autoslog systems*. Technical Report UUCS-04-015, School of Computing, University of Utah.

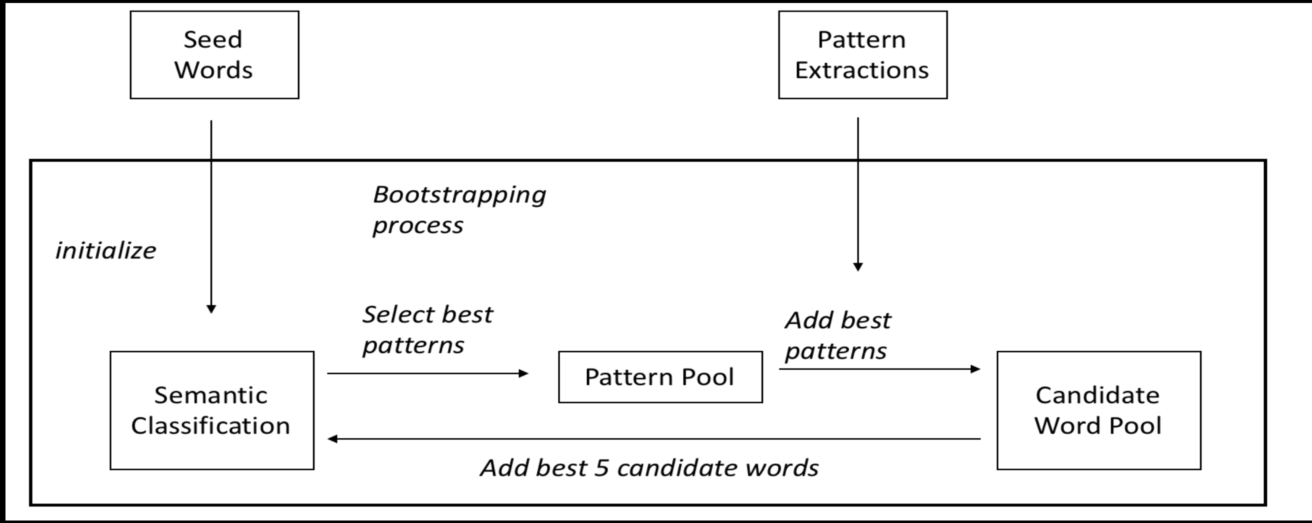
Thelen, M., & Riloff, E. (2002, July). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*(pp. 214-221). Association for Computational Linguistics.

Data and Methodology-ctd:

- A weakly supervised bootstrapping algorithm:
  - Pattern Based
  - Multiple Categories
  - Based on high accurate seeds

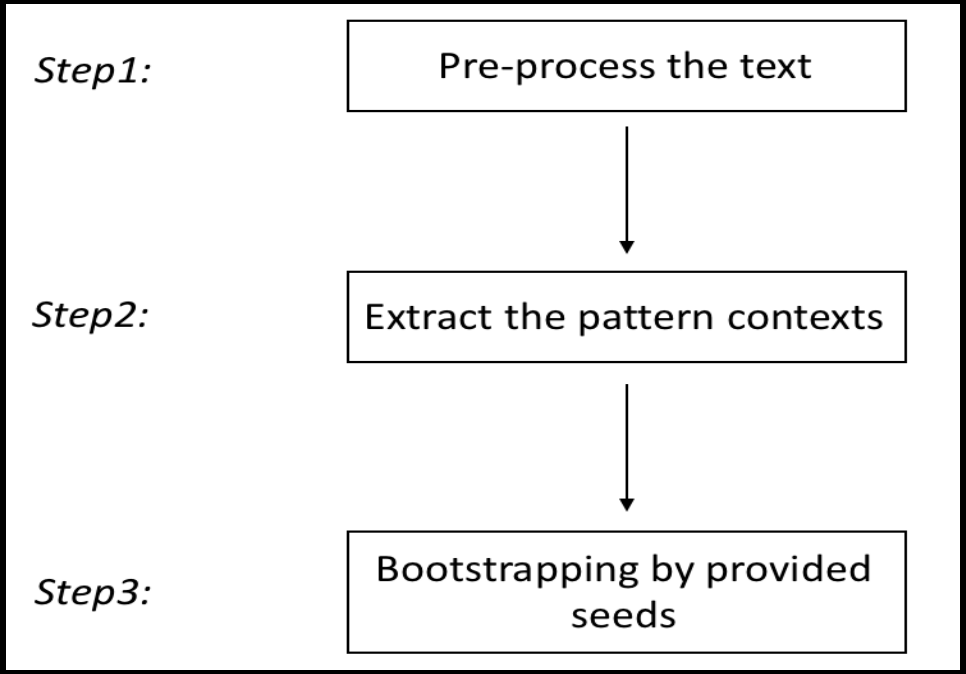
$$RlogF(pattern_i) = \frac{F_i}{N_i} * log_2(F_i)$$

$$Score(word_i) = \frac{\sum_{j=1}^{Pi} log_2(F_j+1)}{P_i}$$



Detail Steps to Implement the Algorithm:

- Pre-process the plans
- Extract the pattern contexts
- Bootstrapping by provided seeds



Autoslog's pattern: a noun phrase in one of three syntactic roles: *subject, direct object, or prepositional phrase object*

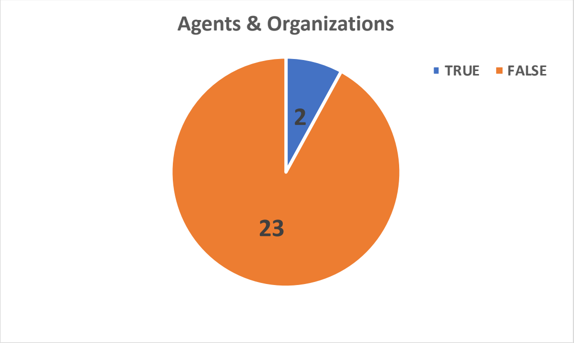
<subject> was murdered, murdered <direct_subject>, collaborated with <pp_object>

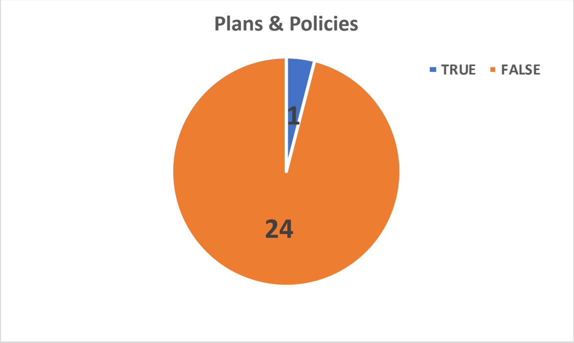
Extract pattern from Stanford Dependent Parser: *the Universal dependency relation*, *the part of speech tag and the head or dependent word.*

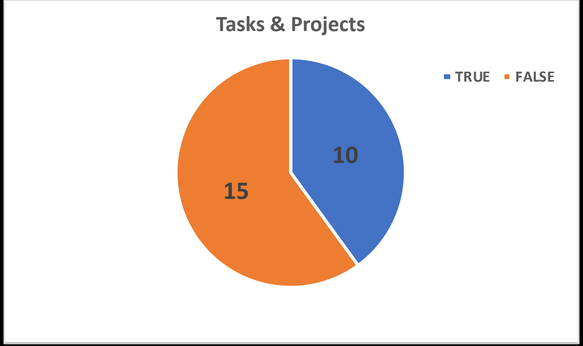
<conj_and>:<head>:StrategiesMeasures

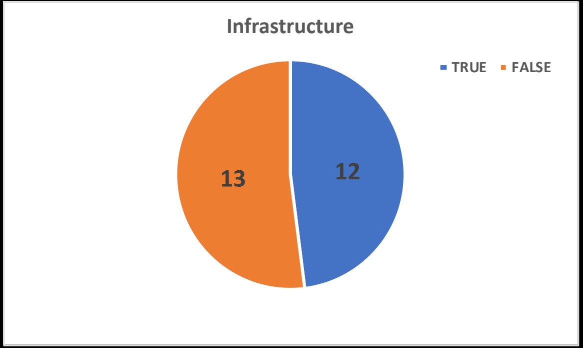
Results and Evaluation:

- Five iterations and ten iterations:









	True	False	Total	Precision
Agents & Organizations	2	23	25	8%
Plans & Policies	1	24	25	4%
Tasks & Projects	10	15	25	40%
Infrastructure	12	13	25	48%

	True	False	Total	Precision
Agents & Organizations	2	48	50	4%
Plans & Policies	1	49	50	2%
Tasks & Projects	18	34	50	36%
Infrastructure	22	28	50	44%

Discussion and Future Work:

- Different precision for categories
- Conflicts between categories
- Pre-process of Plans
- The way to extract pattern in contexts
- Compare results of different pattern
- Relationships extraction

Thank you!!

- Qingchun Li 04.19.2018

Extracting and Classifying Keyphrases from Scientific Publications

Qingqing Li
Luxing Shen

Motivation

- Provide keyphrases of a document to help reader understand the material
    - **Automatically classify label for each keyphrase**
        - PROCESS (P), e.g ‘nuclear reaction’
        - TASK (T), e.g ‘predict the gas exchange processes’
        - MATERIAL (M), e.g ‘water’
    - **Identify and highlight keyphrases in the document**
- Provide a multi-domain system for scientific area
    - Domain-independent
    - Scientific areas are involved: Physics, Computer Science, Material Science

Dataset

- Includes 500 journal articles evenly distributed from above domains
  - Plain text documents
  - Standoff annotation files for paragraphs
  - Xml documents with the original full article text
- Dataset composition
  - Train (350 documents)
  - Dev (50 documents)
  - Test (100 documents)

Data Preparation

- Left-and-Right context method:
  - Provide the context of given keypharse in a document
  - Fixed input_size for each keypharse
  - Input_size = *left_keyphrase_token* + *right_keyphrase_token* + *keyphrase_token*
- Embedding Matrix
  - 100 dimension GloVe embedding
  - Input length of 20

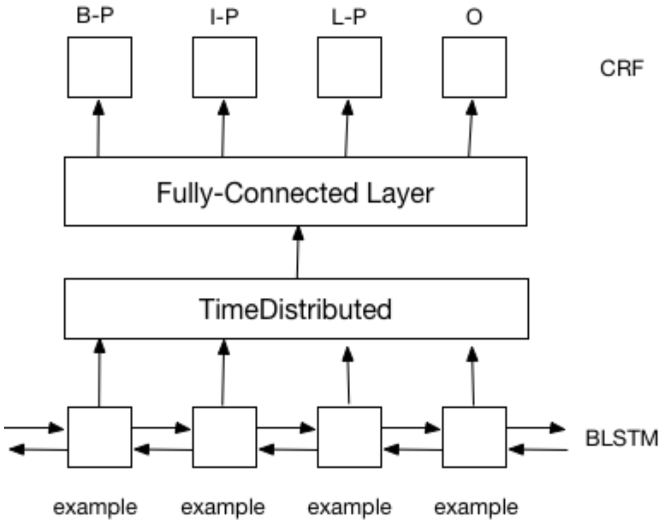
Data Preparation

- Sequence Tagging
  - Find the actual keyphrase within the given document
  - Index token within a sentence with BILOU method

In	chiral	soliton	or	Skyrme	models	the	parity	is	positive	.
O	B-Process	L-Process	O	B-Process	L-Process	O	O	O	O	O

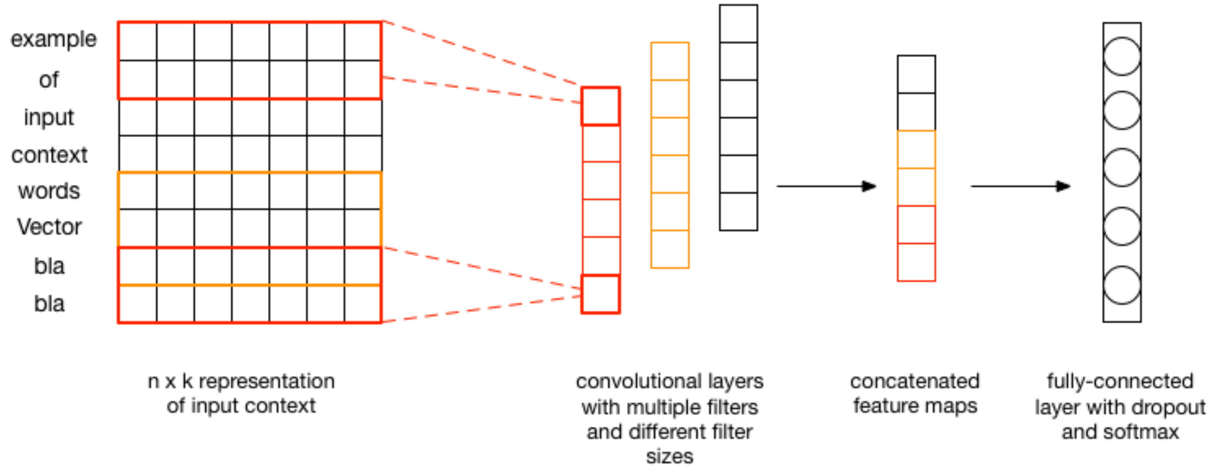
Approach - Task A

- BLSTM-CRF-based sequence tagging
  - Dense word representation - GloVe
  - Contextual word representation - BLSTM
  - Decoding tagging score - FC
  - Considering neighboring tagging decisions - linear-chain CRF



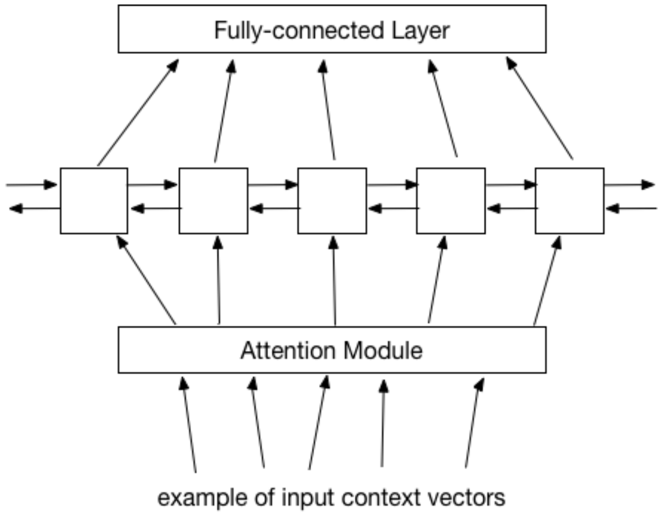
Approach - Task B

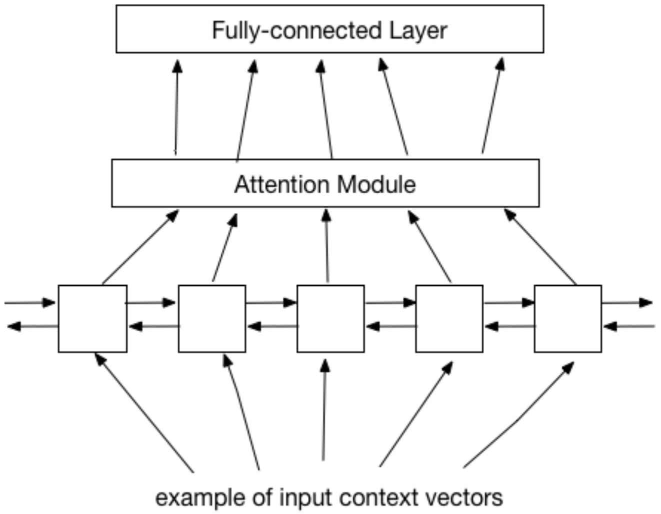
- CNN-based keyphrases classifier
    - Capture feature maps in the input context



Approach - Task B

- attention-BLSTM-based keyphrases classifier





Results - Task A

- Dev set f1 score: 35.166 with std deviation 4.1
- Test set f1 score: 33.705 with std deviation 0.607
- Keyphrases are much more challenging to identify than named entity recognition, since they vary significantly between different domains, lack clear contexts and can consist of many tokens.

Results - Task B

- CNN-based classifier outperforms all BLSTM-based classifiers.

set	f1-metrics	CNN	BLSTM	attention-before-BSLTM	attention-after-BLSTM
dev	macro	**0.6**	0.443	0.465	0.473
	micro	**0.68**	0.627	0.641	0.638
	weighted	**0.68**	0.588	0.607	0.605
test	macro	**0.534**	0.406	0.437	0.446
	micro	**0.632**	0.594	0.614	0.611
	weighted	**0.628**	0.552	0.582	0.581

Table 1: average f1 score on development set and test set for different architectures

Future Work

- BLSTM+CRF might need to be additionally augmented in order to highlight keyphrases.
    - Additional feature sets (n-grams, lexical features, etc) could be included to augment the system.

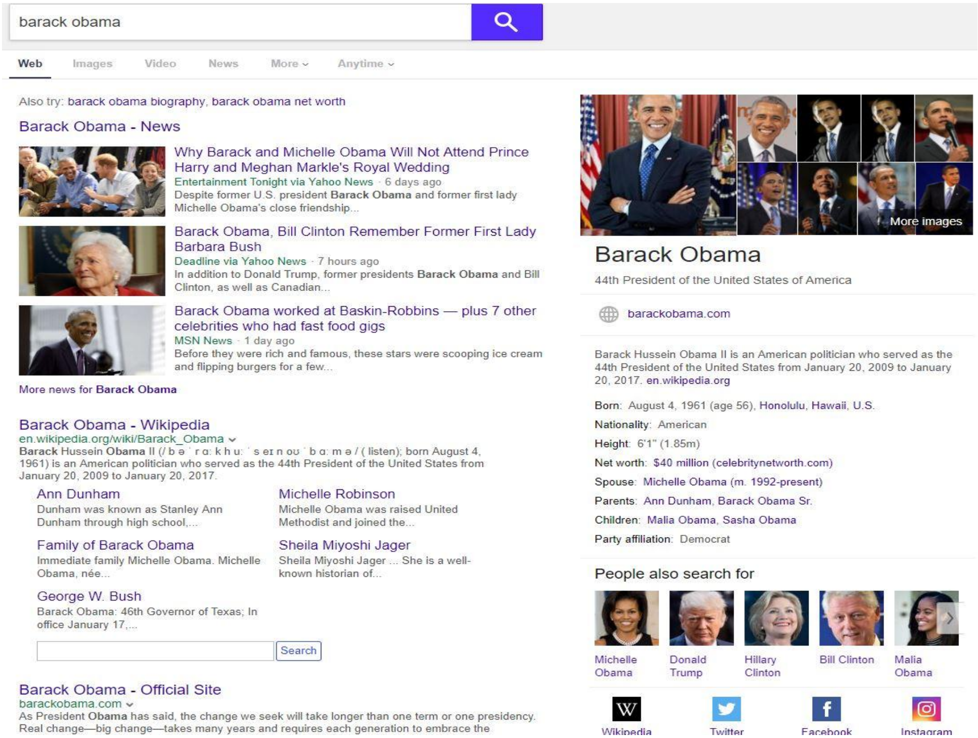
Questions

Automatic Trivia Fact Extraction from Wikipedia

Qiancheng Li, Aniket Bonde
- Under Supervision of Prof. Ruihong Huang

Motivation

- Web Search is now an exploratory activity
- Improving engagement is a key goal
- Most queries are entity related



The Contribution

Did you know? Grammy Award winner

Wikipedia Category

- Natural Language - obstacle of detecting trivia
- Wikipedia also has structured information
- **Categories:** set of articles with a shared topic

Categories: Barack Obama | 1961 births | 20th-century American writers | 20th-century scholars | 21st-century American politicians | 21st-century American writers | 21st-century scholars | Activists from Illinois | African-American feminists | African-American non-fiction writers | American Christians | American Protestants | American non-fiction writers | African-American people in Illinois politics | African-American United States presidential candidates | African-American United States Senators | American civil rights lawyers | American community activists | American feminist writers | American feminists | American legal scholars | American Nobel laureates | American people of English descent | American people of French descent | American people of German descent | American people of Irish descent | American people of Luo descent | American people of Scottish descent | American people of Swiss descent | American people of Welsh descent | American political writers | American politicians of Luo descent | Columbia University alumni | Democratic Party (United States) presidential nominees | Democratic Party Presidents of the United States | Democratic Party United States Senators | Grammy Award winners | Harvard Law School alumni | Illinois Democrats | Illinois lawyers | Illinois State Senators | Living people | Male feminists | Nobel Peace Prize laureates | Obama family | Occidental College alumni | Politicians from Chicago | Politicians from Honolulu | Presidents of the United States | Punahou School alumni | United States presidential candidates, 2008 | United States presidential candidates, 2012 | United States Senators from Illinois | University of Chicago Law School faculty | Writers from Chicago

Wikipedia Category

Categories: Barack Obama | 1961 births | 20th-century American writers | 20th-century scholars | 21st-century American politicians | 21st-century American writers | 21st-century scholars | Activists from Illinois | African-American feminists | African-American non-fiction writers | American Christians | American Protestants | American non-fiction writers | African-American people in Illinois politics | African-American United States presidential candidates | African-American United States Senators | American civil rights lawyers | American community activists | American feminist writers | American feminists | American legal scholars | American Nobel laureates | American people of English descent | American people of French descent | American people of German descent | American people of Irish descent | American people of Luo descent | American people of Scottish descent | American people of Swiss descent | American people of Welsh descent | American political writers | American politicians of Luo descent | Columbia University alumni | Democratic Party (United States) presidential nominees | Democratic Party Presidents of the United States | Democratic Party United States Senators | Grammy Award winners | Harvard Law School alumni | Illinois Democrats | Illinois lawyers | Illinois State Senators | Living people | Male feminists | Nobel Peace Prize laureates | Obama family | Occidental College alumni | Politicians from Chicago | Politicians from Honolulu | Presidents of the United States | Punahou School alumni | United States presidential candidates, 2008 | United States presidential candidates, 2012 | United States Senators from Illinois | University of Chicago Law School faculty | Writers from Chicago

- Rank these categories by how trivia-worthy they are
- **Challenge:** Formalize notion of trivia-worthy

How a 23-year-old makes $500,000 a year tweeting random facts

Madeline Stone

Feb. 12, 2015, 2:10 PM 183,522

FACEBOOK LINKEDIN TWITTER EMAIL COPY LINK

When Kris Sanchez joined Twitter in 2009, he didn't expect much to come of it.

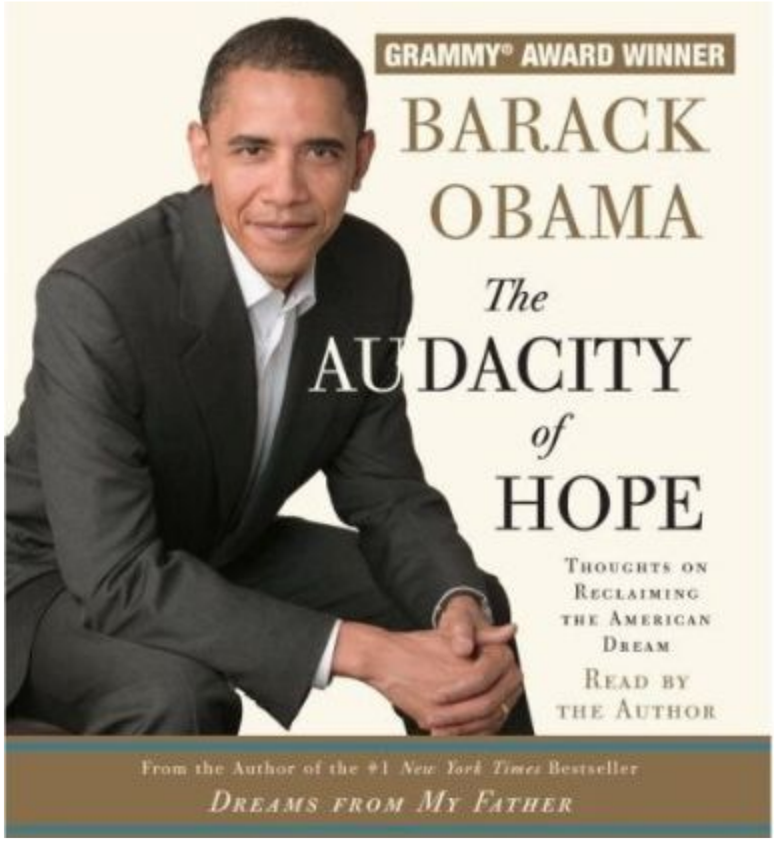
"I really started my Twitter account because I wanted to follow Britney Spears," Sanchez told Business Insider. "I'm a huge fan."

He found he didn't have much to tweet about in his daily life, so he started sharing random facts he found while procrastinating on the internet.

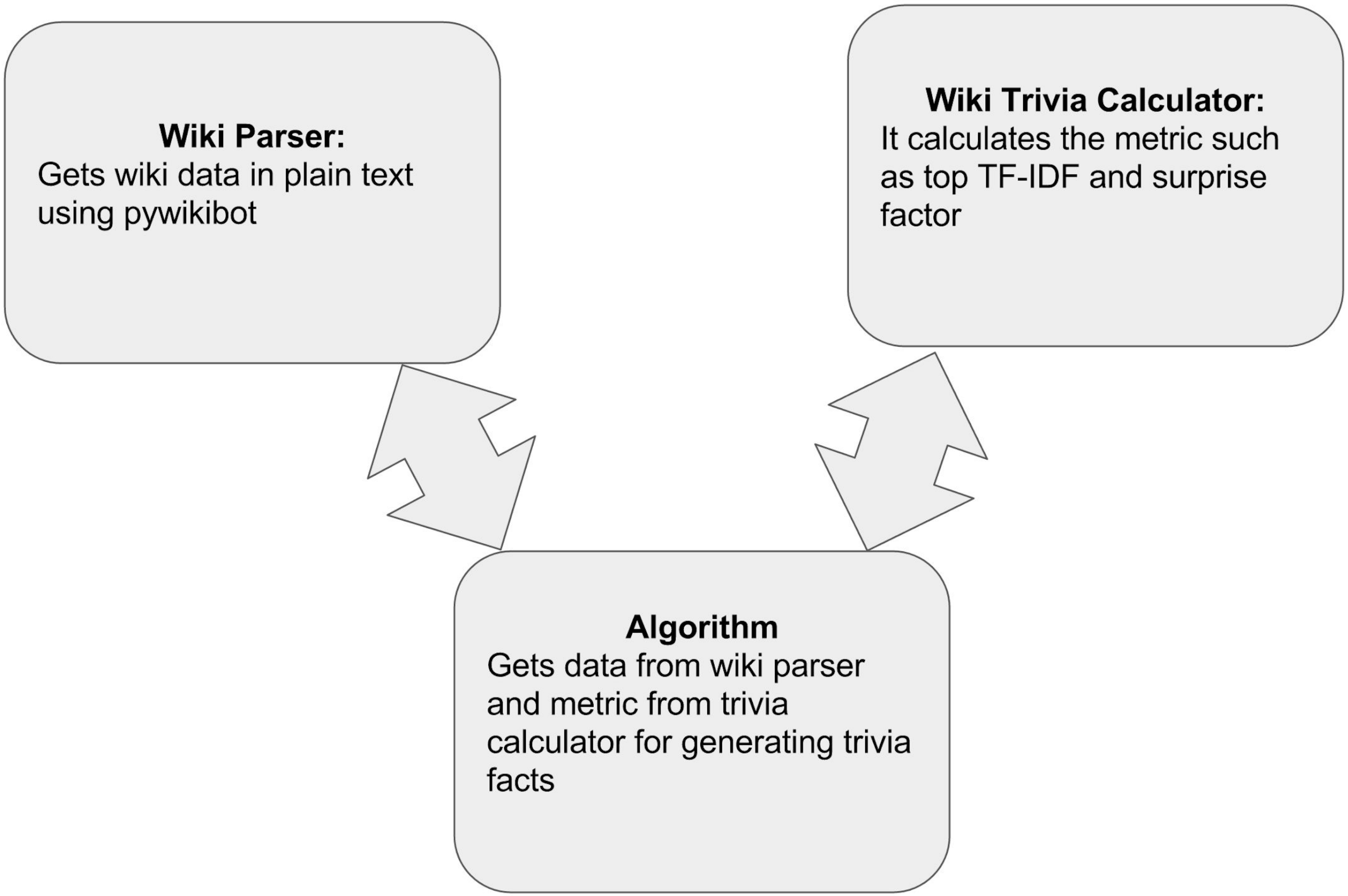
Kris Sanchez and his UberFacts team. Eric Charbonneau

Not All Facts Are Trivia

- Obama was a US president
- Obama was born in 1961
- Obama won a Grammy award



Architecture



Trivia Worthy

- Surprise: get people's attention

$$\sigma(a, C) = \frac{1}{|C| - 1} \sum_{a \neq a' \in C} \sigma(a, a') \qquad surp(a, C) = \frac{1}{\sigma(a, C)}$$

- Cohesiveness

$$cohesive(C) = \frac{1}{\binom{|C|}{2}} \sum_{a \neq a'} \sigma(a, a')$$

- Trivia Worthy

$$trivia(a, C) = cohesive(C) \cdot surp(a, C)$$

Top Surprise:

20th-century Austrian people

Women in technology

Radio pioneers

American anti-fascists

American people of Hungarian-Jewish descent

Top Cohesiveness:

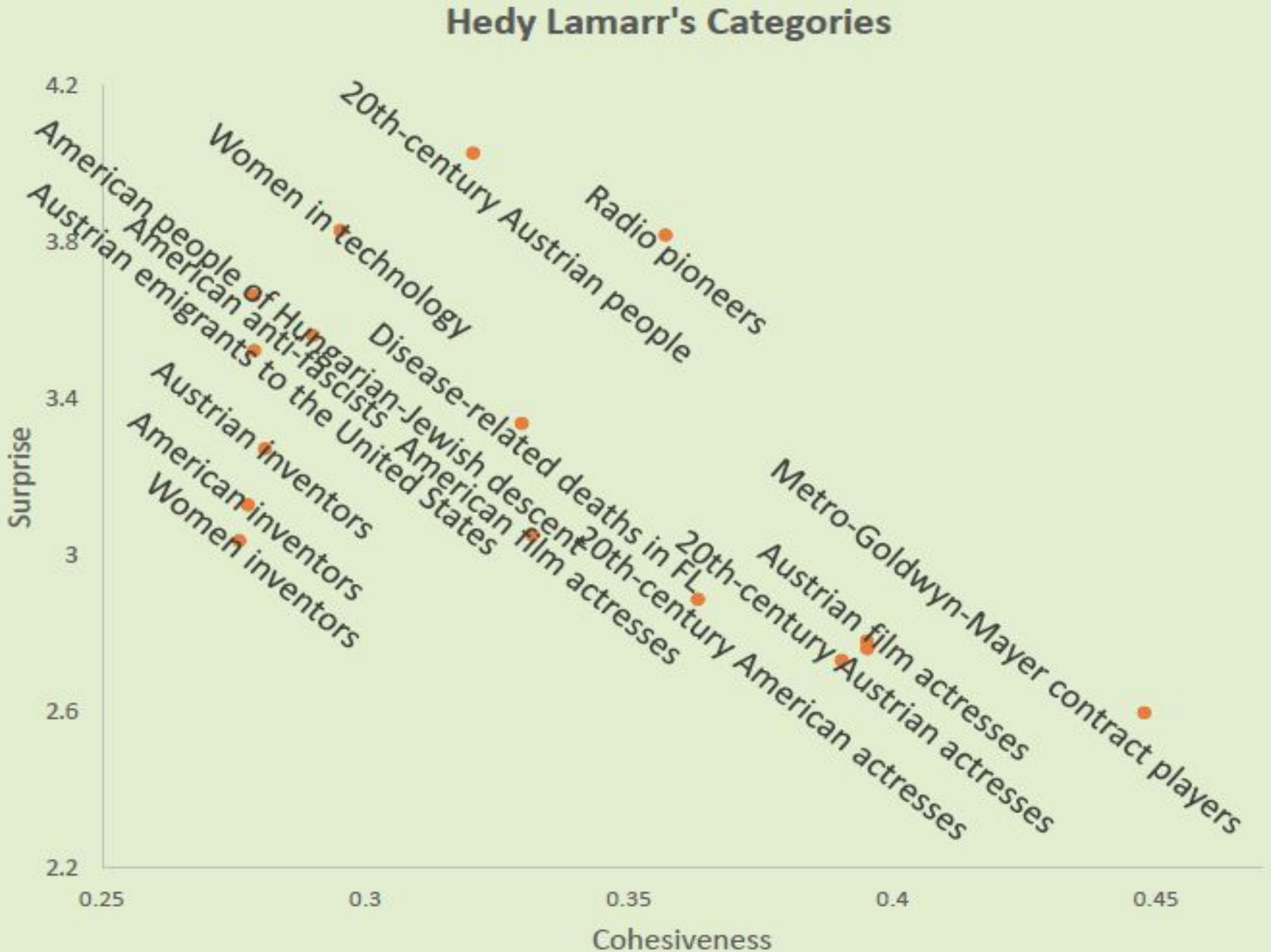
Metro-Goldwyn-Mayer contract players

Actresses from Vienna

Austrian film actresses

20th-century Austrian actresses

American film actresses


Hedy Lamarr's Categories
Surprise
Cohesiveness
4.2
3.8
3.4
3
2.6
2.2
0.25
0.3
0.35
0.4
0.45
20th-century Austrian people
Radio pioneers
Women in technology
American people of Hungarian-Jewish descent
American anti-fascists
Austrian emigrants to the United States
Disease-related deaths in FL
Austrian inventors
American inventors
Women inventors
American film actresses
20th-century American actresses
20th-century Austrian actresses
Austrian film actresses
Metro-Goldwyn-Mayer contract players

Algorithm

Algorithm 4 Article Similarity

function ARTICLESIM($article1, article2$)

  $K = 10$

  $T1 = TopTFIDF(article1, K)$

  $T2 = TopTFIDF(article2, K)$

  $sim = (article1, article2)$

return sim

Algorithm

Algorithm 3 Cohesiveness

function COHESIVENESS($category$)
    $sum, count = 0$
    for $every\ article\ pair\ a1 \neq a2\ in\ category\ C$ **do**
        $sim = ArticleSim(a1, a2)$
        $sum = sum + sim$
        $count = count + 1$
    $cohesiveness = sum/count$
return cohesiveness

Algorithm

Algorithm 2 Surprise

function SURPRISE($inputArticle, C$)
    $sum, count = 0$
    for $every\ article\ a \neq inputArticle\ in\ category\ C$ **do**
        $sim = ArticleSim(inputArticle, a)$
        $sum = sum + sim$
        $count = count + 1$
    $similarityToCategory = sum/count$
    $surprise = 1/similarityToCategory$
return surprise

Algorithm

Algorithm 1 Trivia Extract

function TRIVIAEXTRACT($inputArticle$)

  for every category C of $inputArticle$ **do**

    $surp = Surprise(inputArticle, C)$

    $cohes = Cohesiveness(C)$

    $C.trivia = cohes\ .\ surp$

return category C with maximum trivia score

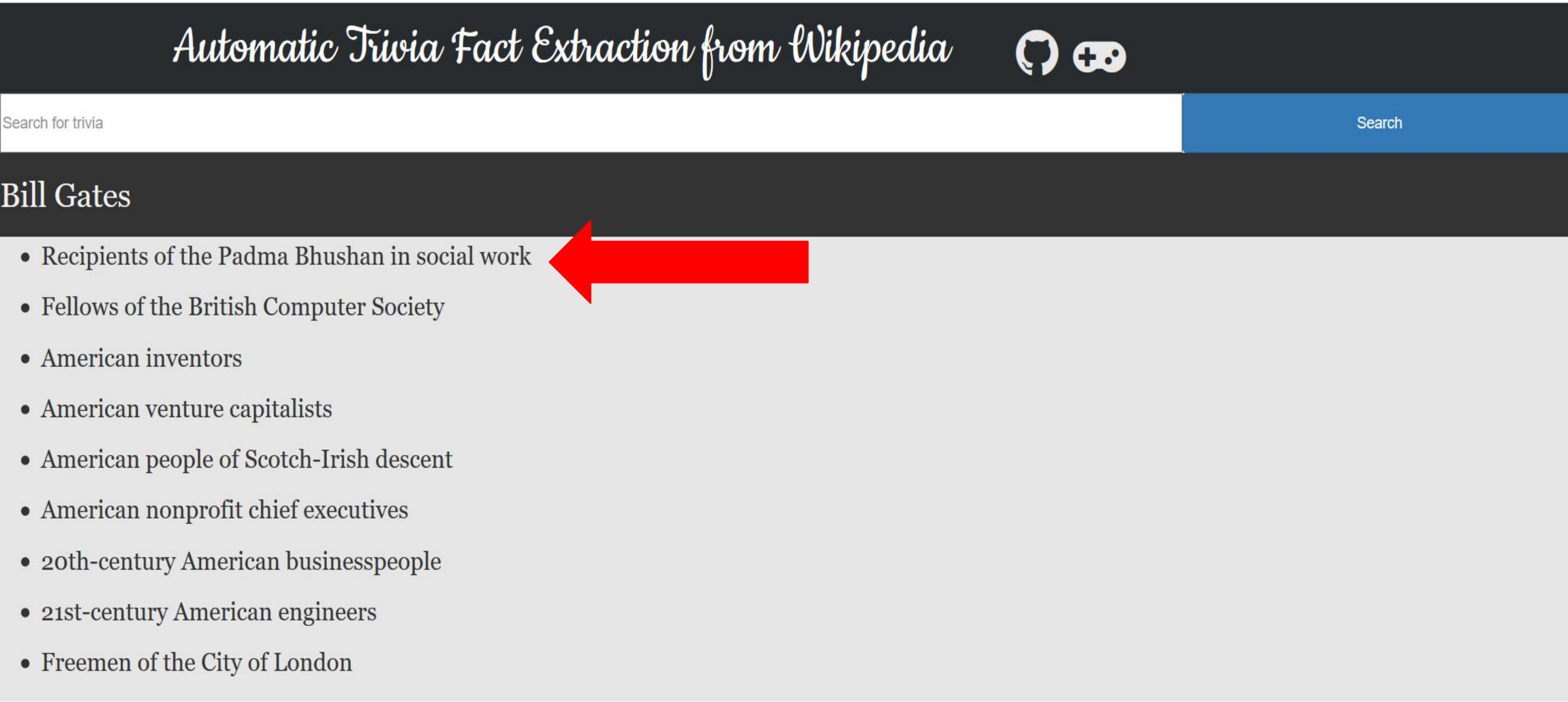
Result: Barack Obama for example

Punahou School alumni	1.3387910646917047
Grammy Award winners	1.2859668279303453
American Nobel laureates	1.2565331843053882
Nobel Peace Prize laureates	1.1786827577086749
American feminist writers	1.1551973277907599
African-American feminists	1.1167987835685782
American feminists	1.0896562059932569
21st-century American politicians	1.0057797842693141
Democratic Party United States Senators	0.99517280067274383
Harvard Law School alumni	0.994574666360582

Evaluation

- Wikipedia Trivia Miner (WTM, IJCAI ‘15)
- Compared top 5 (Didn’t get 100% overlap)
- Due to randomly sampling ‘k’ entities for a category
- Best Way - Run user studies (Not done due to lack of time/resources)

App!


Automatic Trivia Fact Extraction from Wikipedia
Search for trivia
Search
Bill Gates
Recipients of the Padma Bhushan in social work
Fellows of the British Computer Society
American inventors
American venture capitalists
American people of Scotch-Irish descent
American nonprofit chief executives
20th-century American businesspeople
21st-century American engineers
Freemen of the City of London

App!

Automatic Trivia Fact Extraction from Wikipedia

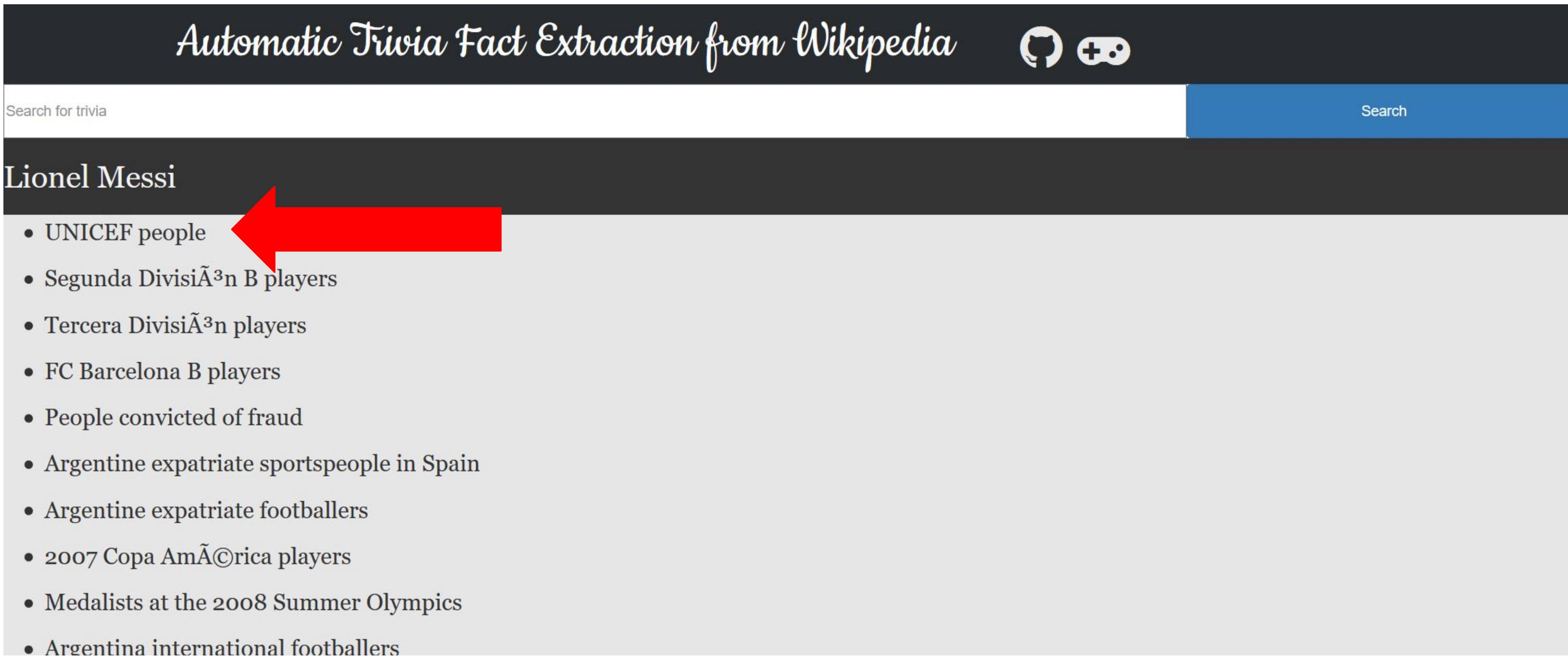
Search for trivia

Search

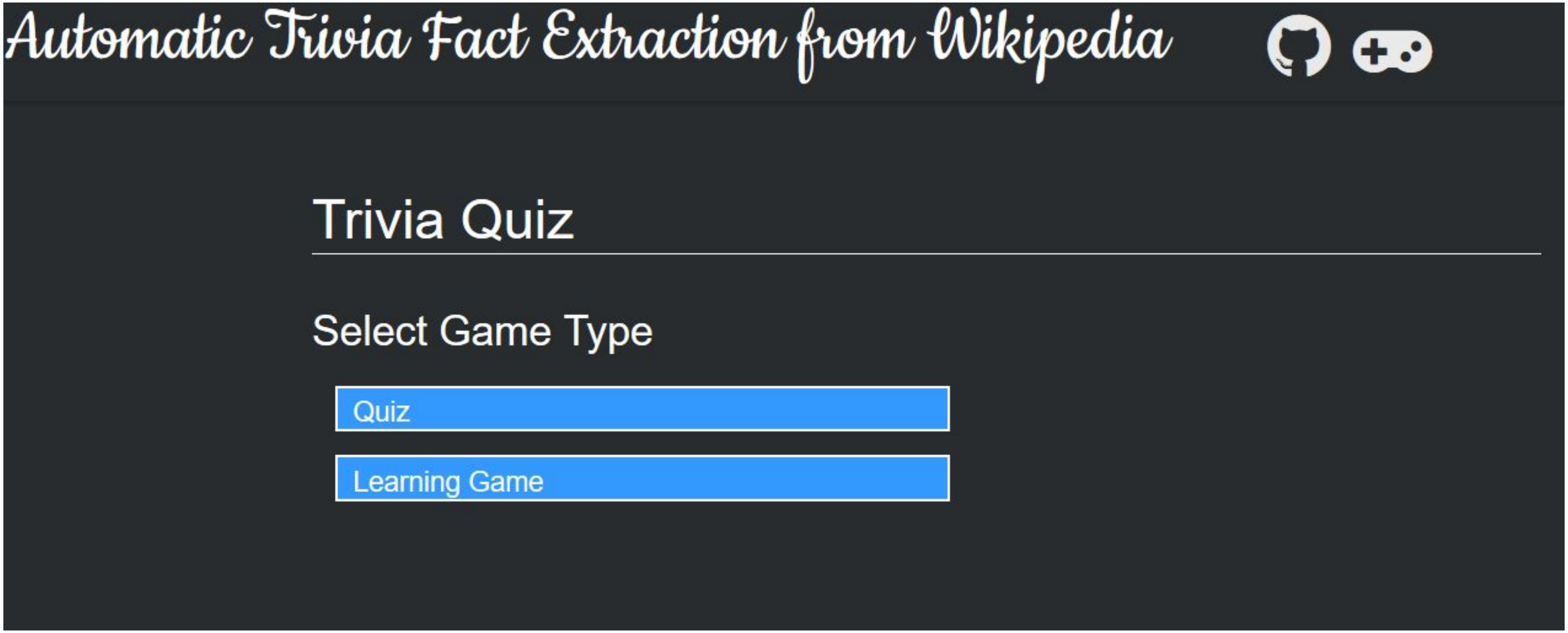
Donald Trump

- WWE Hall of Fame
- 21st-century American politicians
- American Presbyterians
- American billionaires
- New York Republicans
- American political writers
- 20th-century American businesspeople
- People named in the Panama Papers
- American business writers
- American hoteliers

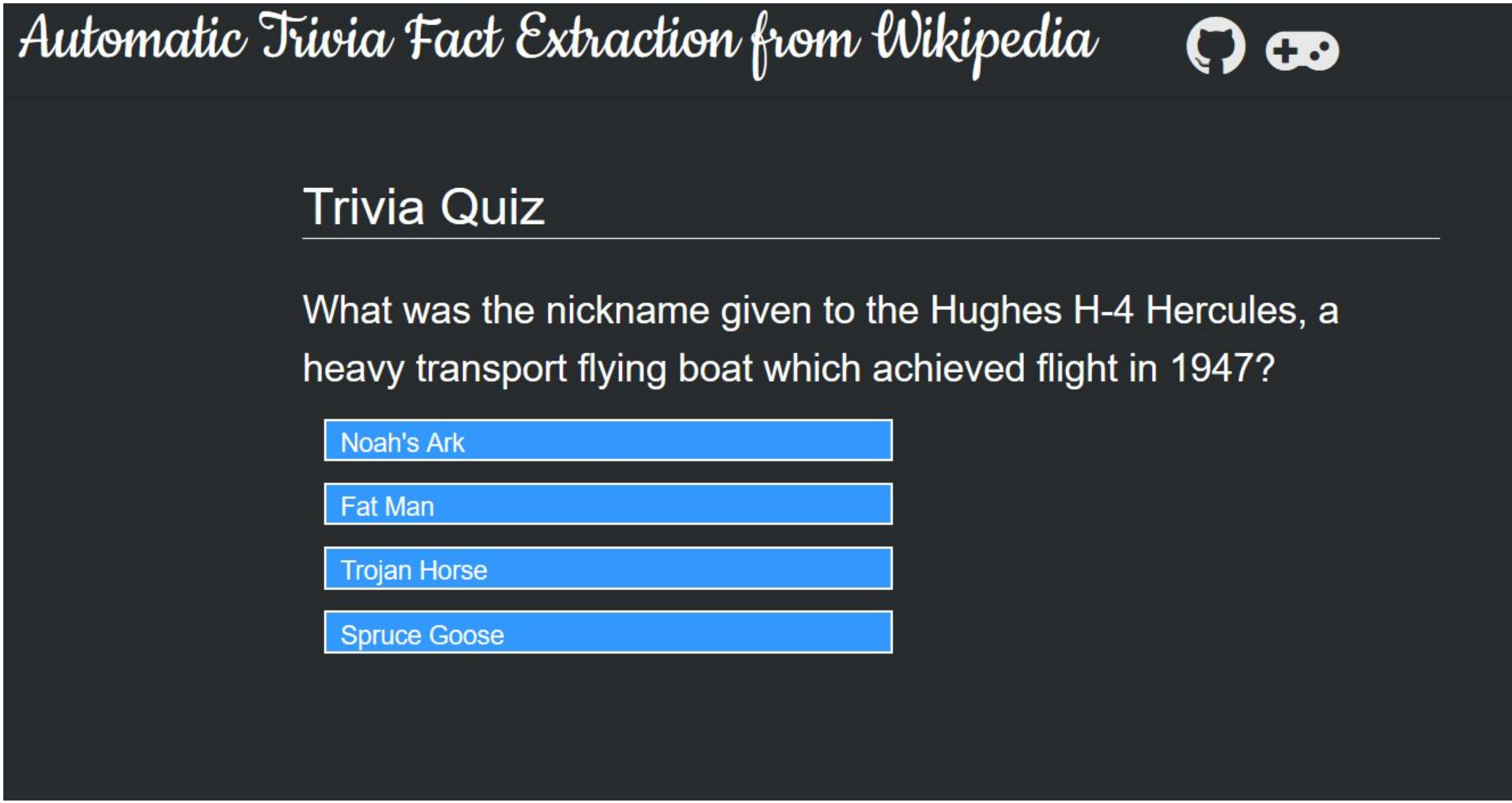
App!


Automatic Trivia Fact Extraction from Wikipedia
Search for trivia
Search
Lionel Messi
UNICEF people
Segunda DivisiÃ³n B players
Tercera DivisiÃ³n players
FC Barcelona B players
People convicted of fraud
Argentine expatriate sportspeople in Spain
Argentine expatriate footballers
2007 Copa AmÃ©rica players
Medalists at the 2008 Summer Olympics
Argentina international footballers

App!

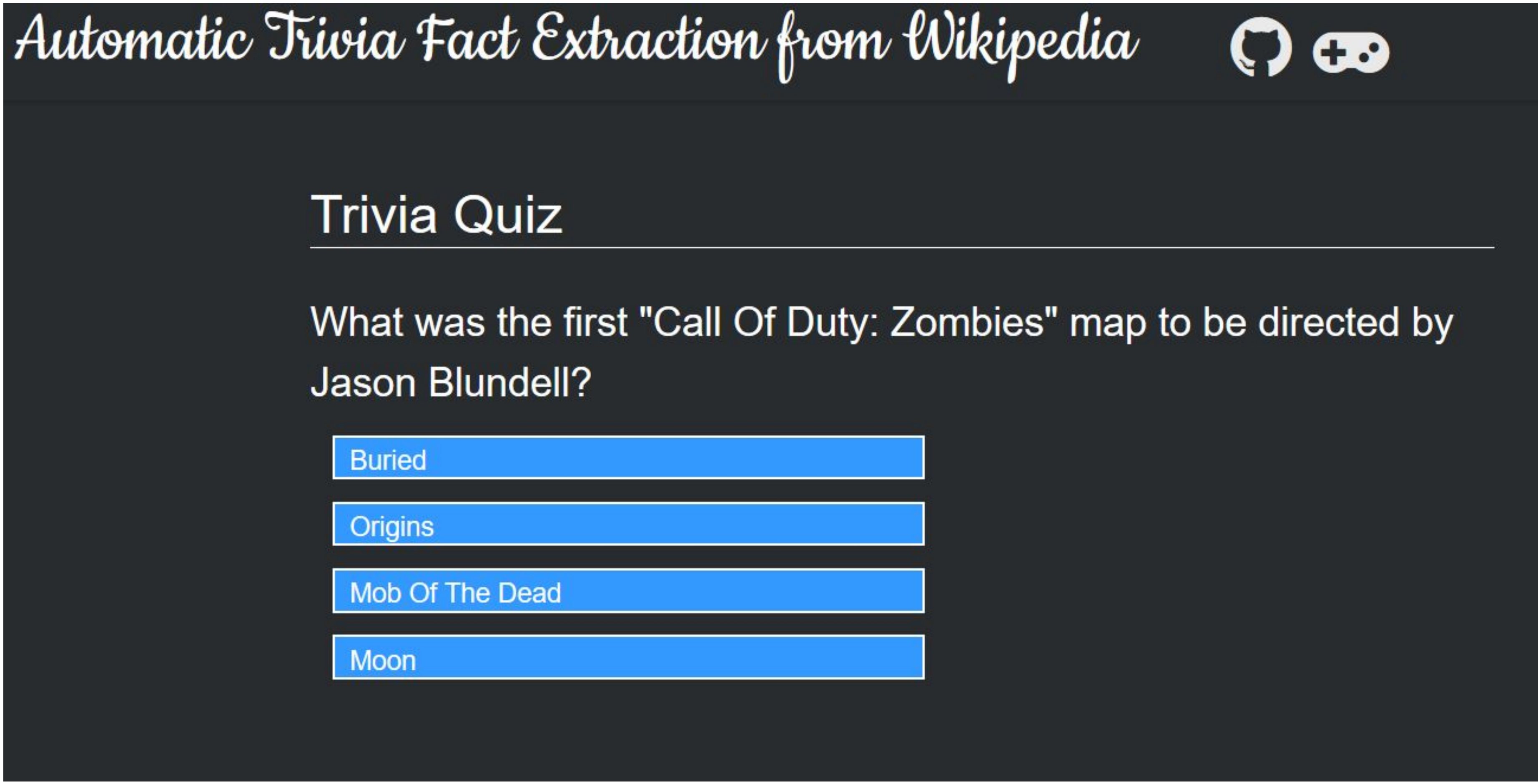

Automatic Trivia Fact Extraction from Wikipedia
Trivia Quiz
Select Game Type
Quiz
Learning Game

App!



Automatic Trivia Fact Extraction from Wikipedia
Trivia Quiz
What was the nickname given to the Hughes H-4 Hercules, a heavy transport flying boat which achieved flight in 1947?
Noah's Ark
Fat Man
Trojan Horse
Spruce Goose

App!


Automatic Trivia Fact Extraction from Wikipedia
Trivia Quiz
What was the first "Call Of Duty: Zombies" map to be directed by Jason Blundell?
Buried
Origins
Mob Of The Dead
Moon

Conclusion

- Detect good trivia
- Introduced formulation: Surprise, cohesiveness
- Increase user engagement

References

Mika, Peter. "Entity search on the web." Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013.

Tsurel, D., Pelleg, D., Guy, I., Shahaf, D.: Fun facts: automatic trivia fact extraction from wikipedia (2016). arXiv preprint arXiv:1612.03896

Yin, Xiaoxin, and Sarthak Shah. "Building taxonomy of web search intents for name entity queries." Proceedings of the 19th international conference on World wide web. ACM, 2010.

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. Extracting semantics relationships between Wikipedia categories. SemWiki, 206, 2006.

Abhay Prakash, Manoj K. Chinnakotla, Dhaval Patel, and Puneet Garg. Did you know?: Mining interesting trivia for entities from wikipedia. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, pages 3164--3170. AAAI Press, 2015.

References

Matthew Merzbacher. Automatic generation of trivia questions. In International Symposium on Methodologies for Intelligent Systems, pages 123-130.Springer, 2002.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of Association for Computational Linguistics (ACL), volume 1, 2014.

Tom Kenter and Maarten de Rijke. Short text similarity with word embeddings. In Proceedings of the 4th ACM International on Conference on 353 Information and Knowledge Management, CIKM '15, pages 1411–1420, New York, NY, USA, 2015. ACM.

Tom Kenter and Maarten de Rijke. Short text similarity with word embeddings. In Proceedings of the 4th ACM International on Conference on 353 Information and Knowledge Management, CIKM '15, pages 1411–1420, New York, NY, USA, 2015. ACM.