AutoSUM: Automating Feature Extraction and Multi-user Preference Simulation for Entity Summarization

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

Dongjun Wei [1,2] , Yaxin Liu[1,2] , Fuqing Zhu[1], Liangjun Zang[1] , Jizhong Han[1], Wei Zhou[1], Yijun Lu[3], and Songlin Hu[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences

[2] School of Cyber Security, University of Chinese Academy of Sciences

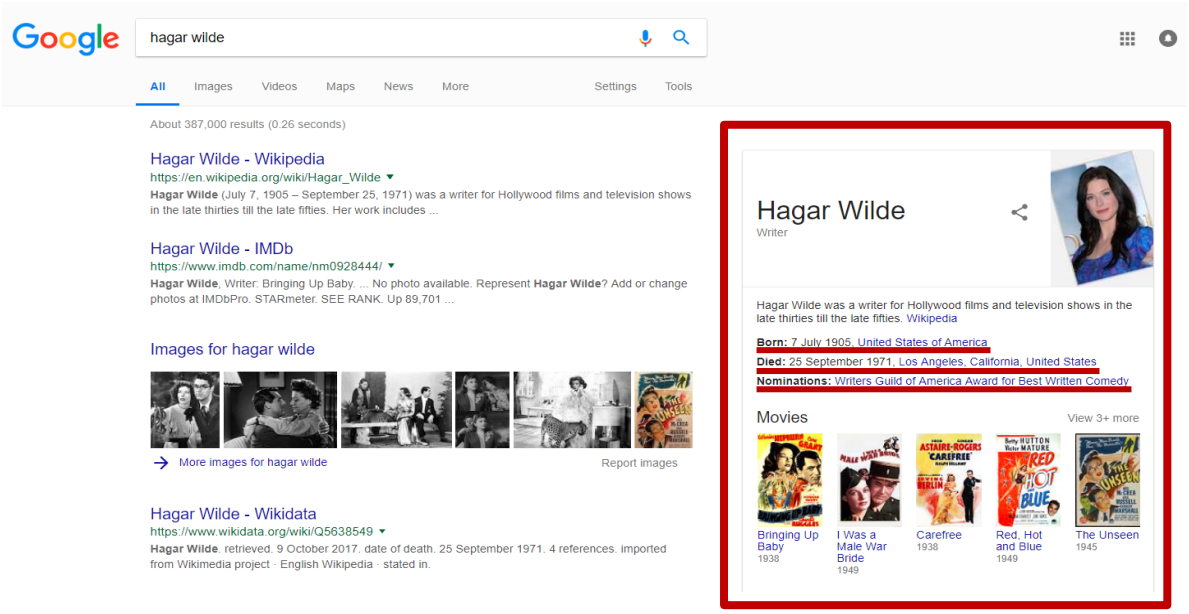
[3] Alibaba Cloud Computing Co. Ltd.

Introduction

Task Definition

Entity summarization is the problem of identifying a limited number of ordered RDF triples that summarize an entity in the best way—the result is typically presented in knowledge panels.

Application



- *A description card based on entity summarization*

Introduction

- Example

  Query a subject named ***Hagar Wilde*** in DBpedia with SPARQL：

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://dbpedia.org

Query Text

```
Select distinct ?Predicate ?Object where
{
<http://dbpedia.org/resource/Hagar_Wilde>
?Predicate
?Object
}
```

Predicate	Object
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Thing
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://xmlns.com/foaf/0.1/Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#Agent
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#NaturalPerson
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.wikidata.org/entity/Q215627
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.wikidata.org/entity/Q24229398

- ***RDF triples***: *<Subject, Predicate, Object>*
- *Search Hager Wilde in DBpedia: <**Hagar Wilde**, Predicate, Object>*
- *Total **65** RDF triples*

Introduction

- Example

Select *Top-5* triples from total 65 triples mentioned above：

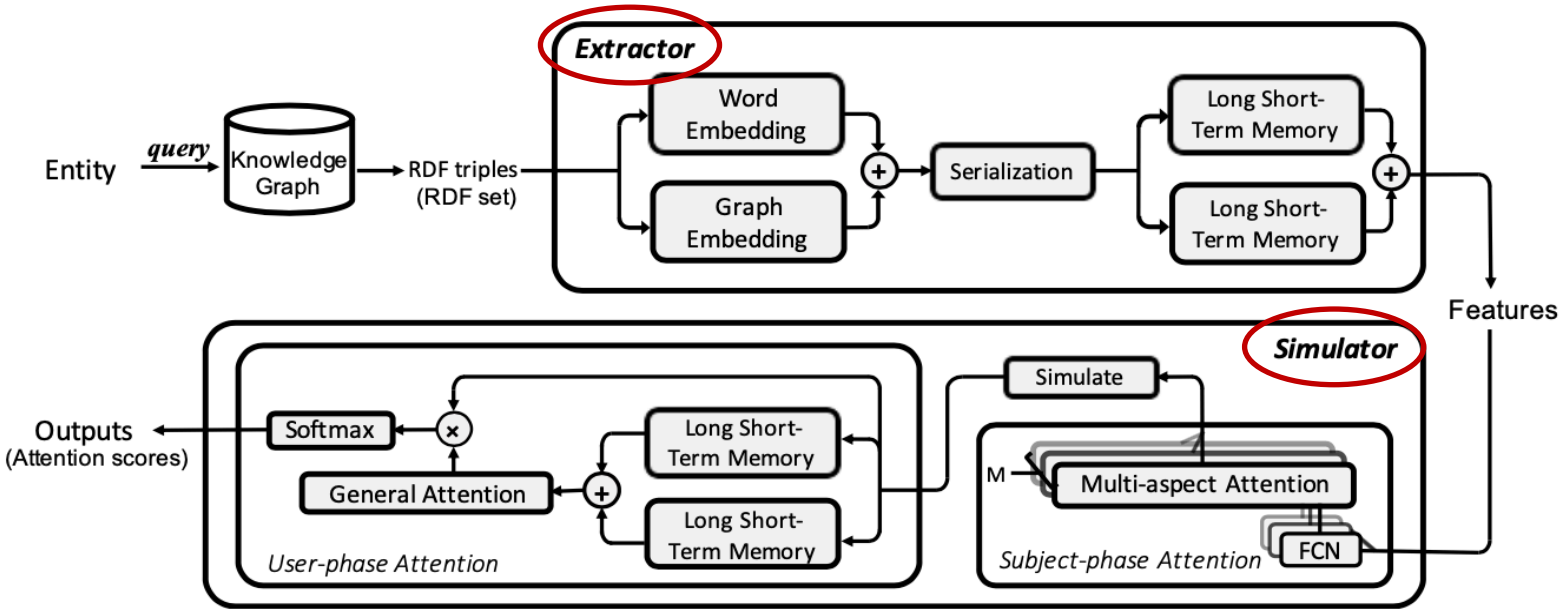
```
<http://dbpedia.org/ontology/basedOn> <http://dbpedia.org/resource/Hagar_Wilde> .
<http://dbpedia.org/ontology/birthDate> "1905-07-07"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/ontology/deathDate> "1971-09-25"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://xmlns.com/foaf/0.1/name> "Hagar Wilde"@en .
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/class/yago/WomenTelevisionWriters> .
```

Top-5 predicate-object pairs of subject ***Hagar Wilde***

Predicate	Object
Name	Hagar Wilde
Type	WomenTelevisonWriters
Birthdate	1905-07-07
Deathdate	1971-09-25

The Architecture of AutoSUM

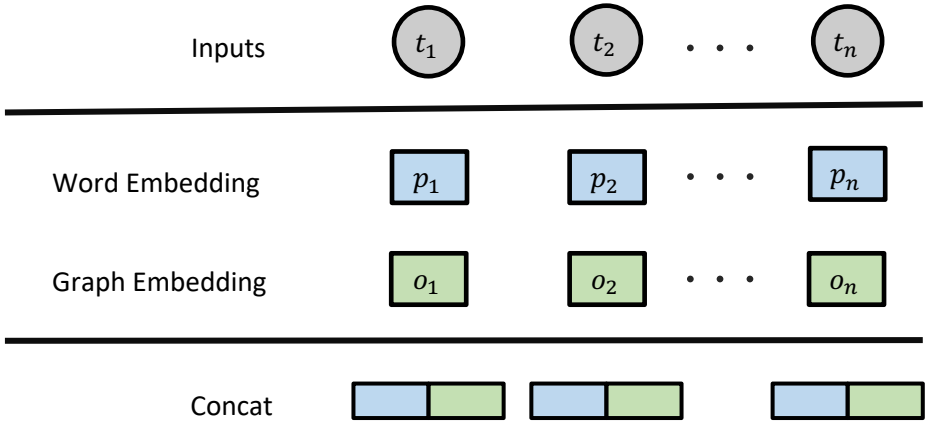
- Overview



- ***Extractor:*** Feature Extraction
- ***Simulator:*** Multi-user Preference Simulation

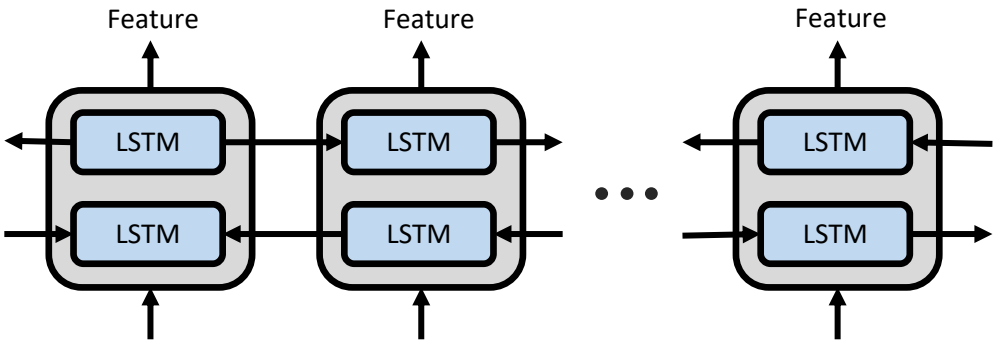
Extractor: Feature Extraction

- Example
  - Input representation



- *Word embedding + Graph embedding = Input representation*

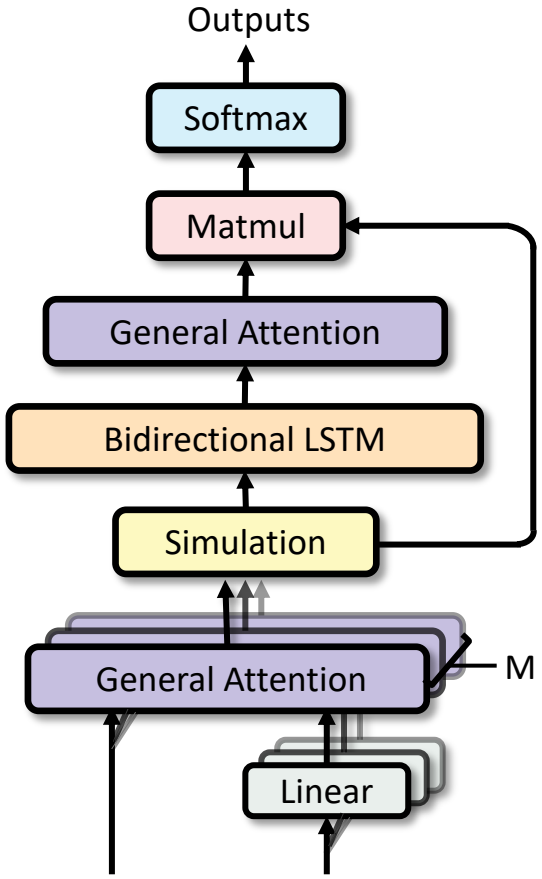
  - Automatic feature extraction



- *Bidirectional Long-short Term Memory Network -> extraction*

Simulator: Multi-user Preference

- Simulator
  - Entity & User Phase Attention



User Phase Attention

- *Attention scores = User preference*
- *BiLSTM + Attention -> allocate different scores to different users*

Entity Phase Attention

- *General Attention Score Funtion -> allocate different score to different entities*

Experiment

- Settings
  - ESBM Datasets
    - DBpedia
    - LMDB
  - Metrics
    - F-measure
    - MAP
  - Baselines
    - RELIN [3]
    - DIVERSUM [4]
    - CD [5]
    - FACES [6]
    - FACES-E [7]
    - LINKSUM [8]
    - ESA [2]

Experiment

- Results

F-measure

Model	DBpedia		LinkedMDB		ALL		↑ %		
	k = 5	k = 10	k = 5	k = 10	k = 5	k = 10	min	max	avg
RELIN	0.242	0.455	0.203	0.258	0.231	0.399	25	118	72
DIVERSUM	0.249	0.507	0.207	0.358	0.237	0.464	12	114	54
CD	0.287	0.517	0.211	0.328	0.252	0.455	10	110	52
FACES	0.270	0.428	0.169	0.263	0.241	0.381	23	162	73
FACES-E	0.280	0.488	0.313	0.393	0.289	0.461	17	48	38
LINKSUM	0.274	0.479	0.140	0.279	0.236	0.421	18	216	80
ESA	0.310	0.525	0.320	0.403	0.312	0.491	8	38	26
AutoSUM	**0.387**$^{+}$	**0.569**$^{+}$	**0.443**$^{+}$	**0.556**$^{+}$	**0.403**$^{+}$	**0.565**$^{+}$	-	-	-
AutoSUM^{1}	0.303^{-}	0.425^{-}	0.316	0.442^{-}	0.290^{-}	0.462^{-}	22	40	31
AutoSUM^{2}	0.316^{+}	0.538	0.375^{+}	0.463^{-}	0.333^{-}	0.517^{+}	6	22	16
AutoSUM^{3}	0.221^{-}	0.390^{-}	0.330^{+}	0.406^{-}	0.252^{-}	0.394^{-}	34	75	49
AutoSUM^{4}	0.254^{-}	0.417^{-}	0.309	0.394^{-}	0.270^{-}	0.411^{-}	36	52	43
AutoSUM^{5}	0.325^{+}	0.532^{+}	0.343^{-}	0.413^{+}	0.323	0.502^{+}	7	35	21

- *Traditional Methods*
- *Neural Network*
- *AutoSUMs*

MAP

Model	DBpedia		LinkedMDB		ALL		↑ %		
	k = 5	k = 10	k = 5	k = 10	k = 5	k = 10	min	max	avg
RELIN	0.342	0.519	0.241	0.335	0.313	0.466	25	115	55
DIVERSUM	0.310	0.499	0.266	0.390	0.298	0.468	30	94	53
CD	-	-	-	-	-	-	-	-	-
FACES	0.255	0.382	0.155	0.273	0.227	0.351	69	234	114
FACES-E	0.388	0.564	0.341	0.435	0.375	0.527	15	64	36
LinkSUM	0.242	0.271	0.141	0.279	0.213	0.345	68	267	132
ESA	0.392	0.582	0.367	0.465	0.386	0.549	11	41	23
AutoSUM	**0.459**$^{+}$	**0.647**$^{+}$	**0.517**$^{+}$	**0.600**$^{+}$	**0.476**$^{+}$	**0.633**$^{+}$	-	-	-
AutoSUM^{1}	0.419^{-}	0.508^{-}	0.420^{+}	0.522^{+}	0.389^{-}	0.563	10	27	18
AutoSUM^{2}	0.404	0.598^{-}	0.431^{+}	0.525^{+}	0.412^{-}	0.578^{+}	8	20	14
AutoSUM^{3}	0.291^{-}	0.456^{-}	0.383^{+}	0.488^{+}	0.317^{-}	0.465^{-}	23	58	41
AutoSUM^{4}	0.333^{-}	0.486^{-}	0.376^{-}	0.467	0.346^{-}	0.480^{-}	28	38	34
AutoSUM^{5}	0.405^{+}	0.582	0.368	0.473	0.412^{+}	0.550	11	40	21

- *Traditional Methods*
- *Neural Network*
- *AutoSUMs*

Conclusions

- Conclusions
  - a novel integration model
  - the performance is significantly better than other methods in both F-measure and MAP.
  - sufficient ablation studies are provided to demonstrate the effectiveness of each module in AutoSUM
- Future Works
  - expand the ESBM dataset
  - introduce the notion of AutoSUM into other applications such as recommender systems

References

[1] Sneha Chaudhari et al. An Attentive Survey of Attention Models. In arXiv:1904.02874, 2019.

[2] Dongjun Wei et al. Entity Summarization with Attention. In EYRE@CIKM, 2019.

[3] Gong Cheng et al. RELIN: Relatedness and Informativeness-Based Centrality for Entity Summarization. In ESWC, 2011.

[4] Marcin Sydow et al. DIVERSUM: Towards diversified summarisation of entities in knowledge graphs. In ICDEW, 2010.

[5] Danyun Xu et al. Generating Characteristic and Diverse Entity Summaries. In ENSEC, 2016.

[6] Kalpa Gunaratna et al. Gleaning Types for Literals in RDF Triples with Application to Entity Summarization. In ESWC, 2016.

[7] Kalpa Gunaratna et al. FACES: Diversity-Aware Entity Summarization using Incremental Hierarchical Conceptual Clustering. In AAAI, 2015.

[8] Andreas Thalhammer et al. LinkSUM: Using Link Analysis to Summarize Entity Data. In ICWE, 2016.

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Thanks!
Q&A.