Generating Templates of Entity Summaries with an Entity-Aspect Model and Pattern Mining

Peng Li, Jing Jiang, Yinglin Wang

Shanghai Jiao Tong University

Singapore Management University

An Example Entity Summary

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article

Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help

Toolbox

Print/export

Languages
Bân-lâm-gú
Afrikaans
Alemannisch
العربية
Aragonés

Albert Einstein

From Wikipedia, the free encyclopedia

"Einstein" redirects here. For other uses, see Einstein (disambiguation).

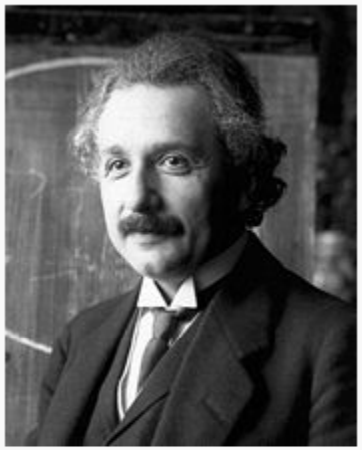
Albert Einstein (pronounced /ˈælbərt ˈaɪnstaɪn/; German: [ˈalbɐt ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a theoretical physicist, philosopher and author who is widely regarded as one of the most influential and best known scientists and intellectuals of all time. He is often regarded as the father of modern physics.[3] He received the 1921 Nobel Prize in Physics "for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect."[4]

His many contributions to physics include the special and general theories of relativity, the founding of relativistic cosmology, the first post-Newtonian expansion, the explanation the perihelion precession of Mercury, the prediction of the deflection of light by gravity (gravitational lensing), the first fluctuation dissipation theorem which explained the Brownian motion of molecules, the photon theory and the wave-particle duality, the quantum theory of atomic motion in solids, the zero-point energy concept, the semiclassical version of the Schrödinger equation, and the quantum theory of a monatomic gas which predicted Bose–Einstein condensation.

Einstein published more than 300 scientific and over 150 non-scientific works; he additionally wrote and commentated prolifically on various philosophical and political subjects.[5] His great intelligence and originality has made the word "Einstein" synonymous with genius.[6]

Contents [hide]

Albert Einstein



Albert Einstein, 1921

Born	14 March 1879 Ulm, Kingdom of Württemberg, German Empire
Died	18 April 1955 (aged 76) Princeton, New Jersey, USA
Resting place	Grounds of the Institute for Advanced Study, Princeton, New Jersey.

上海交通大學
Shanghai Jiao Tong University

SMU
SINGAPORE MANAGEMENT UNIVERSITY

An Example Entity Summary

Albert Einstein (pronounced /ˈælbərt ˈaɪnstaɪn/; German ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was physicist, philosopher and author who is widely regarded as most influential and best known scientists and intellectuals of time. He is often [3] He received the 1921 Nobel tical Physics, and especially for his discov he law of the photoelectric effect."[4]

His many contributions to physics include the special and general theories of relativity, the founding of relativistic cosmology, the first post-Newtonian expansion, the explanation the perihelion precession of Mercury, the prediction of the deflection of light by gravity (gravitational lensing), the first fluctuation dissipation theorem which explained the Brownian motion of molecules, the photon the theory of atomic motion in semiclassical version of the nger equation, and the quantum theory of a monatomic gas wh predicted Bose–Einstein condensation.

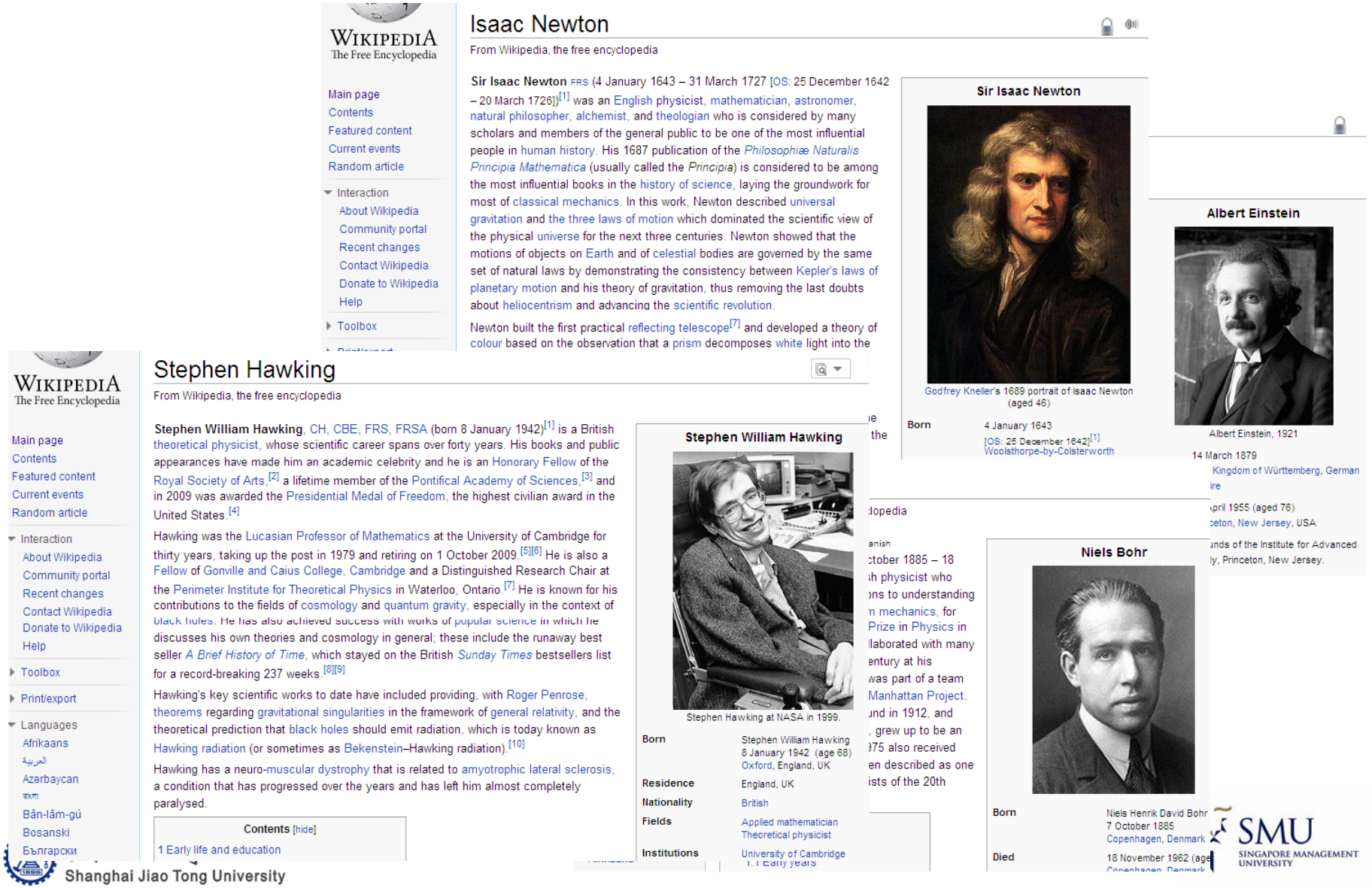
Einstein published more than 300 scientific and over 150 non-scientific works; he additionally wrote and commentated prolifically on various philosophical and political subjects.[5] His great intelligence and originality has made the word "Einstein" synonymous with genius.[6]

… received … Nobel Prize …

His … contributions … include …

… published … works …

Entity Summaries in the Same Category



Entity Summaries in the Same Category

Aspect	Representative Sentence Patterns
Education	X received his PhD from ___ University X studied ___ under ___ X earned his ___ in physics from University of ___
Awards	X was awarded the medal in ___ X won the ___ award X received the Nobel Prize in physics in ___
Career	X was ___ director X was the head of ___ X worked for ___
Contributions	X made contributions to ___ X is best known for work on ___ X is noted for ___

Our task is to automatically generate such entity summary templates.

Why Is It Useful?

- Better organizes information units
  - Wikipedia infoboxes
- Provides a structured template for humans to create new entity summaries
- Facilitates automatic entity summary generation

Outline

- Task and motivation
- **Related work**
- Our approach
  - Sentence clustering
  - Pattern mining
- Evaluation
- Conclusions

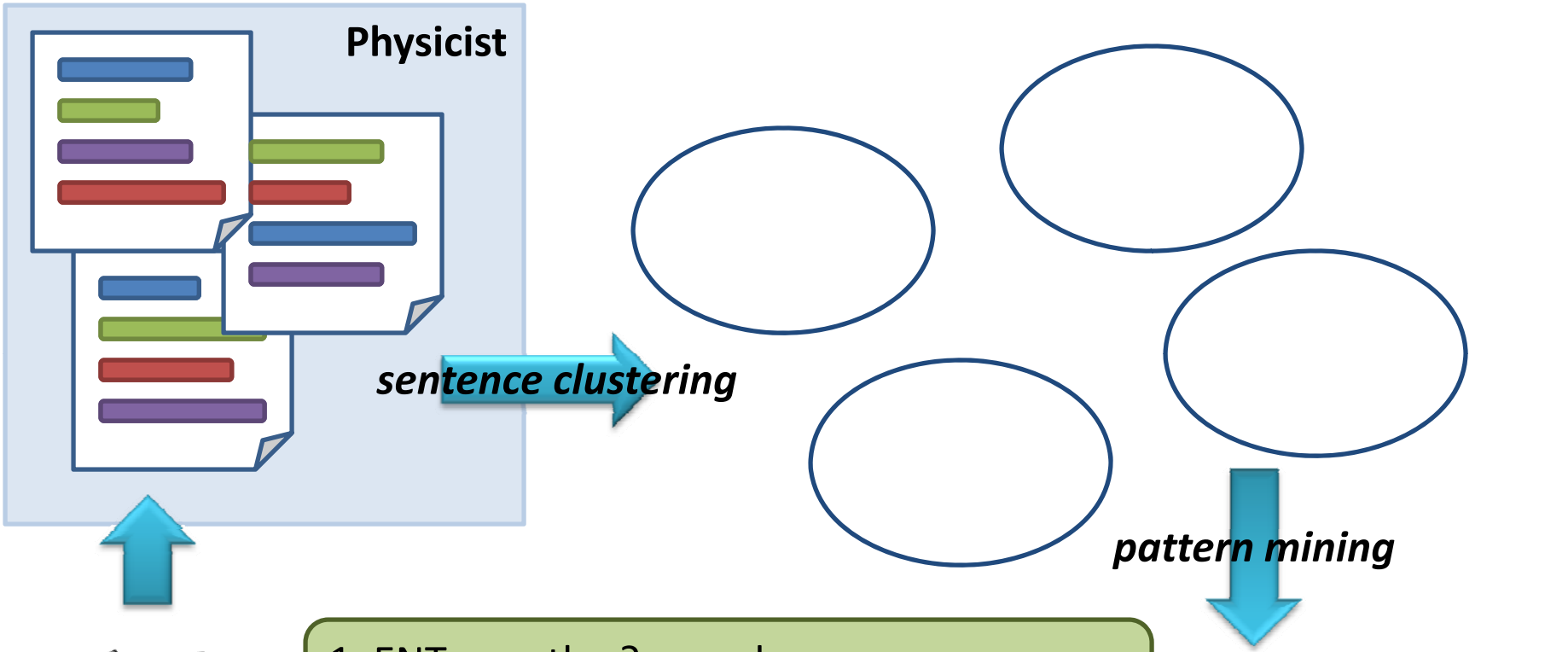
Related Work

- Filatova et al. (2006) “Automatic creation of domain templates”
  - Patterns must contain a non-auxiliary verb
  - Patterns are not clustered into aspects
  - Pattern slots are identified through heuristics
- Sauper & Barzilay (2009) “Automatically generating Wikipedia articles”
  - Long, comprehensive articles
- LDA extensions
  - Chemudugunta et al. (2007)
  - Titov & McDonald (2008)
  - Daume III & Marcu (2006), Haghi & Vanderwende (2009)

Outline

- Task and motivation
- Related work
- **Our approach**
  - Sentence clustering
  - Pattern mining
- Evaluation
- Conclusions

Overview of Our Approach



WIKIPEDIA
The Free Encyclopedia

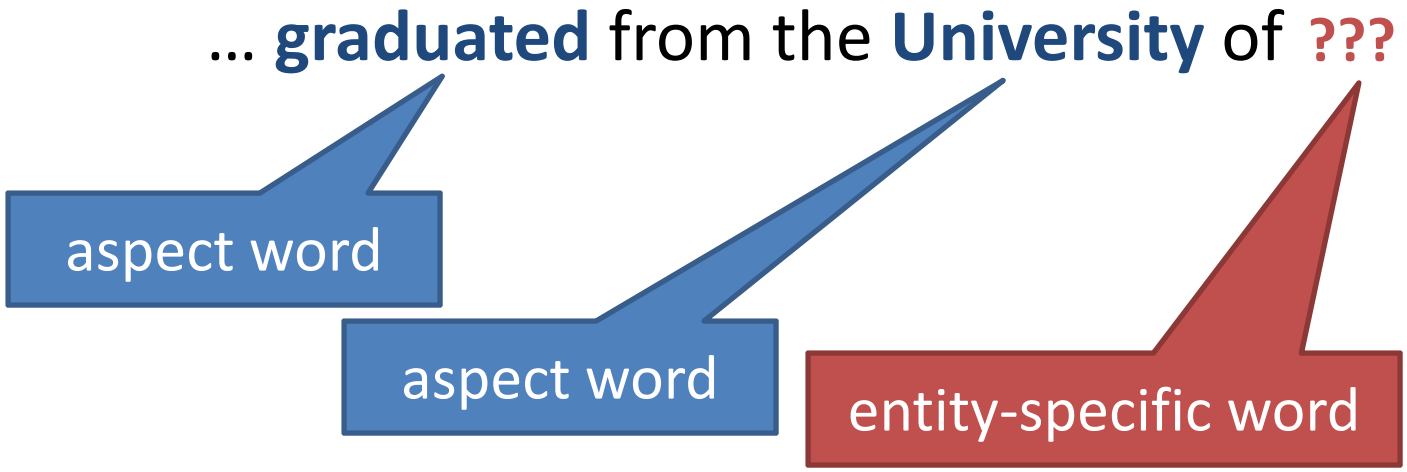
上海交通大學
Shanghai Jiao Tong University

Outline

- Task and motivation
- Related work
- Our approach
  - **Sentence clustering**
  - Pattern mining
- Evaluation
- Conclusions

Sentence Clustering

- To group sentences related to the same aspect together
- To distinguish between aspect words and entity-specific words



Motivating Example

… … Venturi was a professor of physics at the University of Modena … …

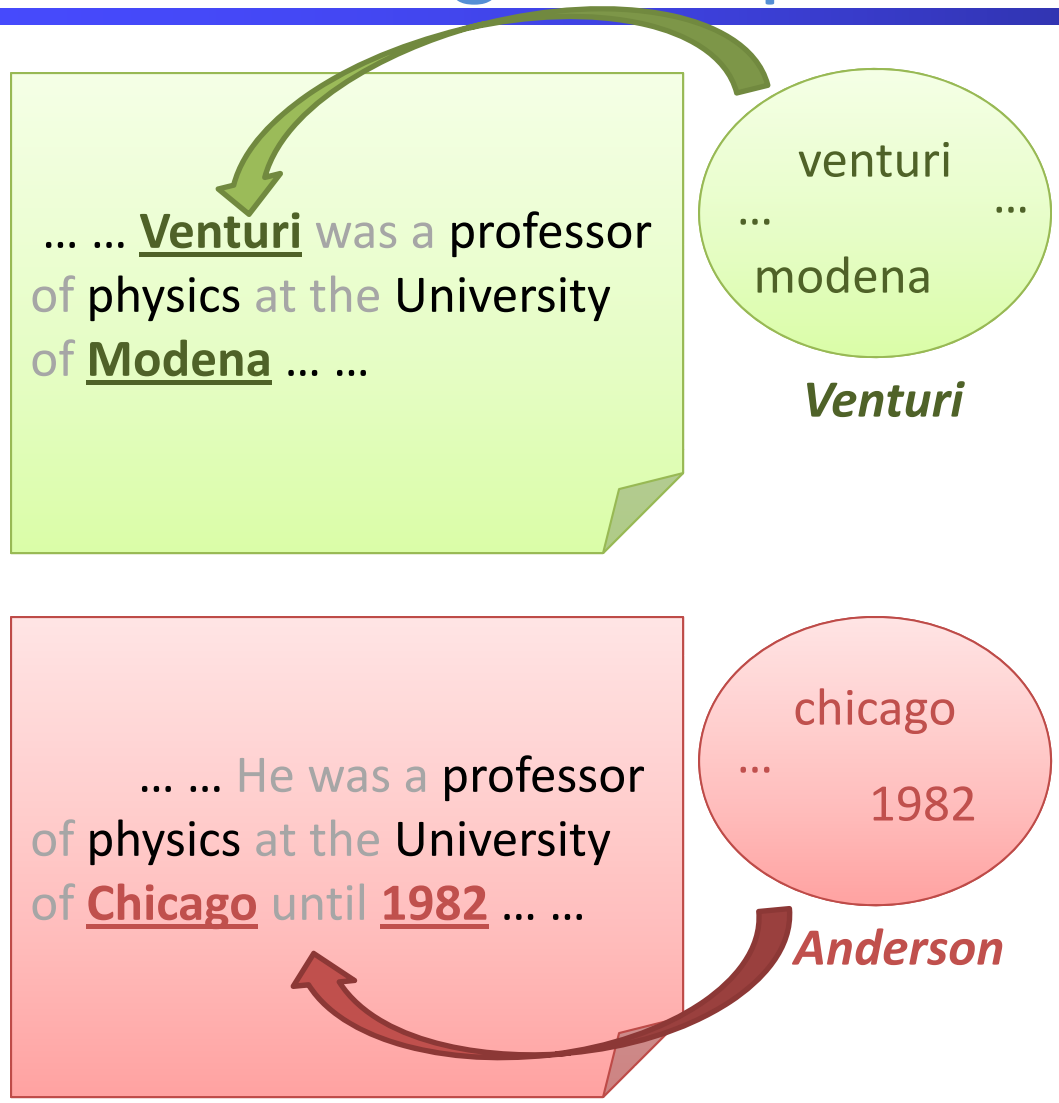
… … He was a professor of physics at the University of Chicago until 1982 … …

Motivating Example

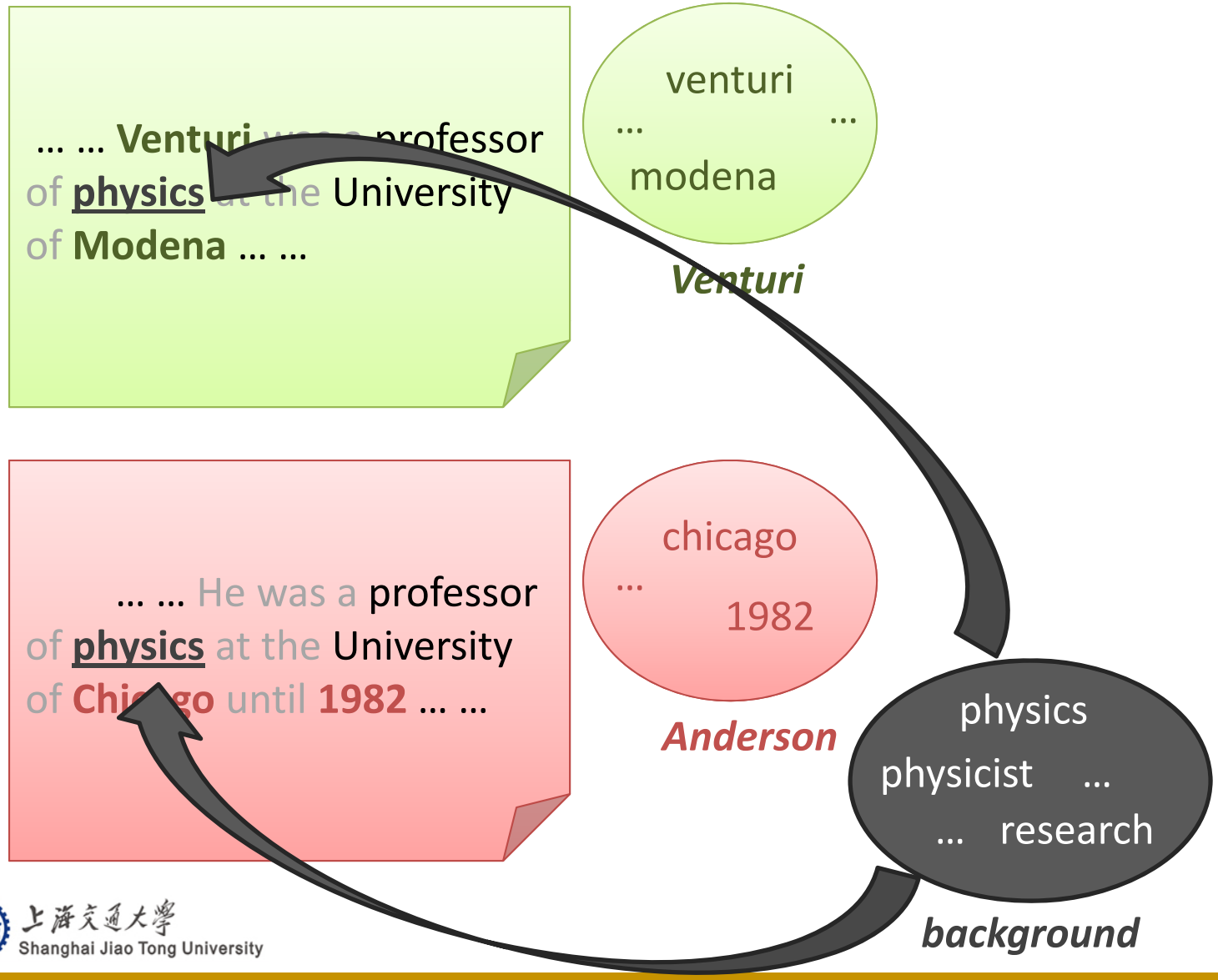
… … Venturi was a professor of physics at the University of Modena … …

… … He was a professor of physics at the University of Chicago until 1982 … …

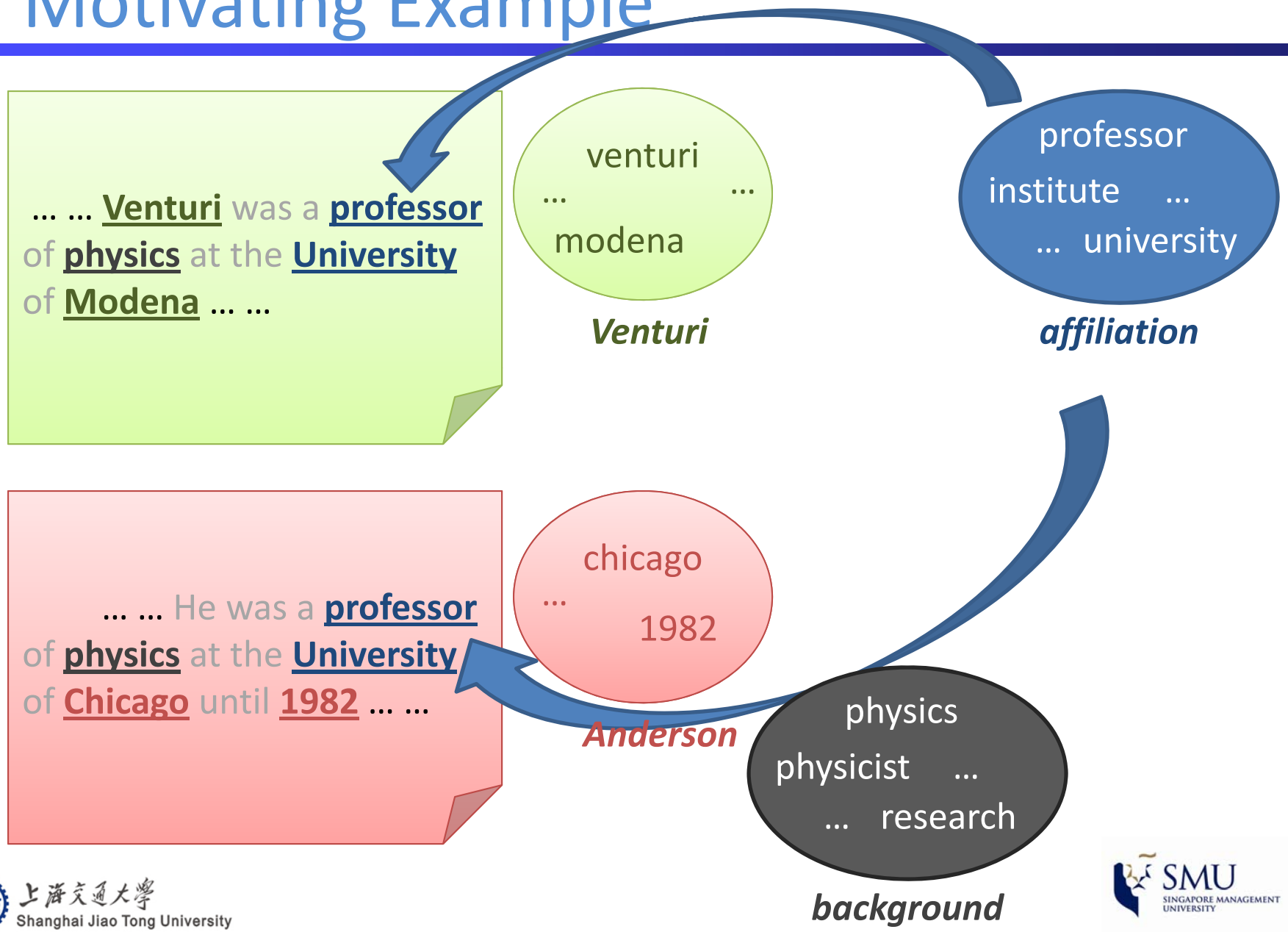
Motivating Example



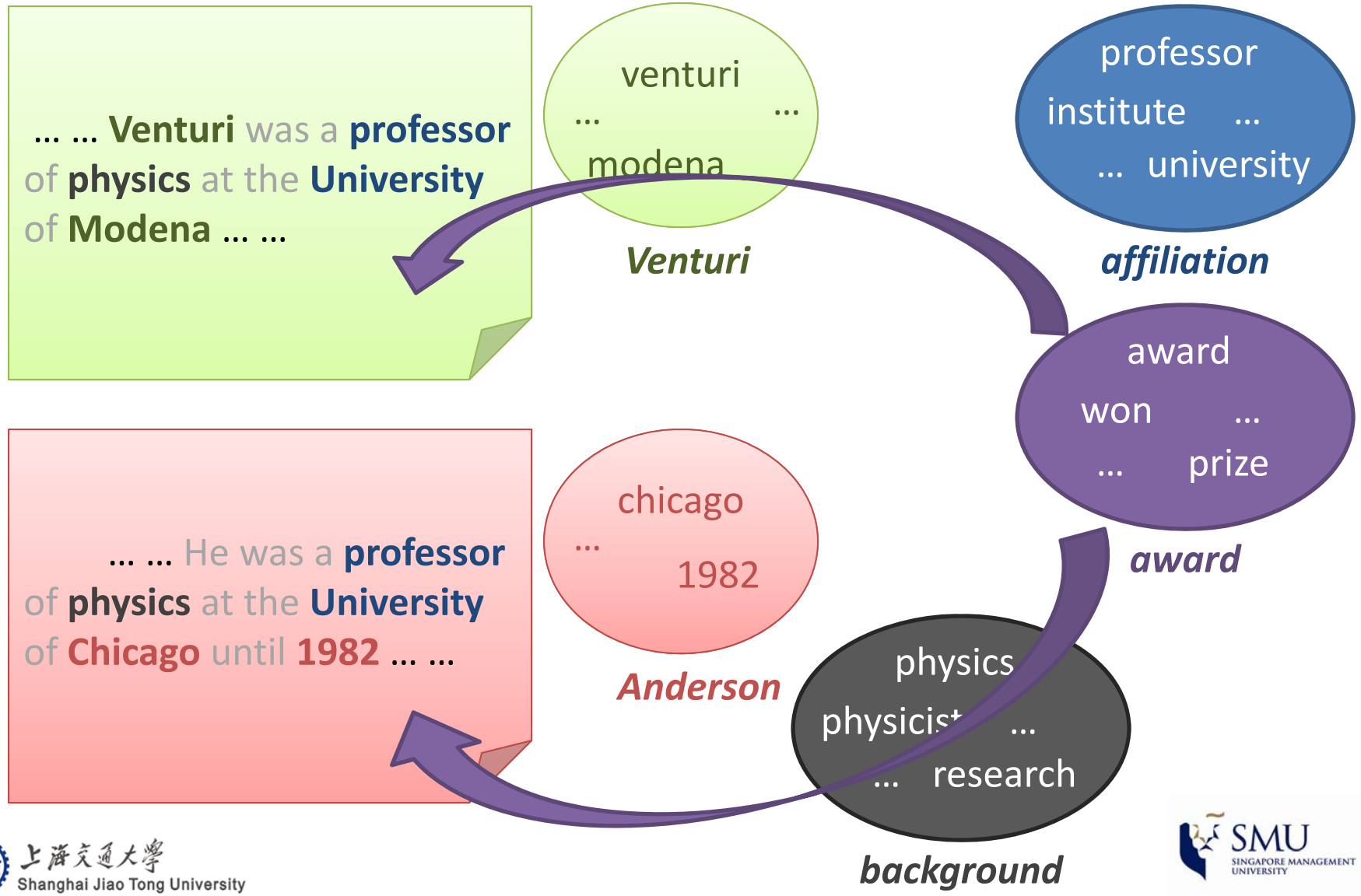
Motivating Example



Motivating Example

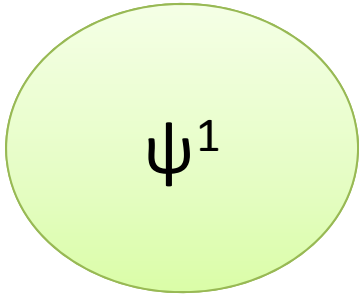


Motivating Example

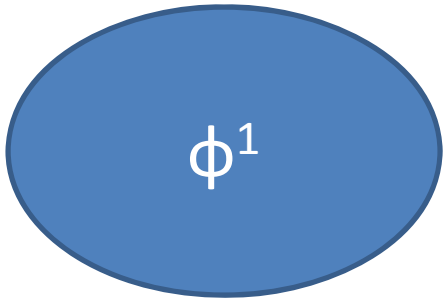


Motivating Example

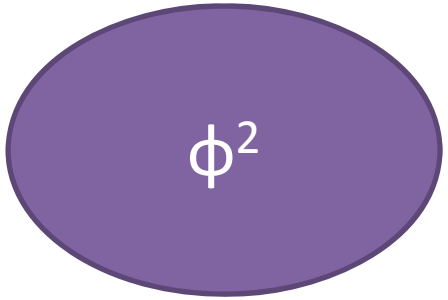
… … **Venturi** was a **professor** of **physics** at the **University** of **Modena** … …



D_1

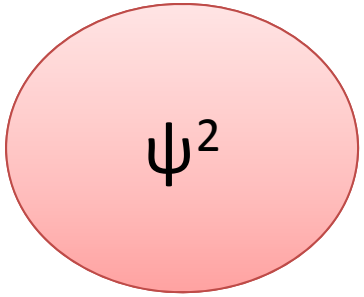


A_1

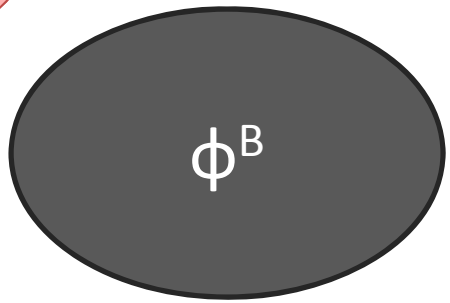


A_2

… … He was a **professor** of **physics** at the **University** of **Chicago** until **1982** … …

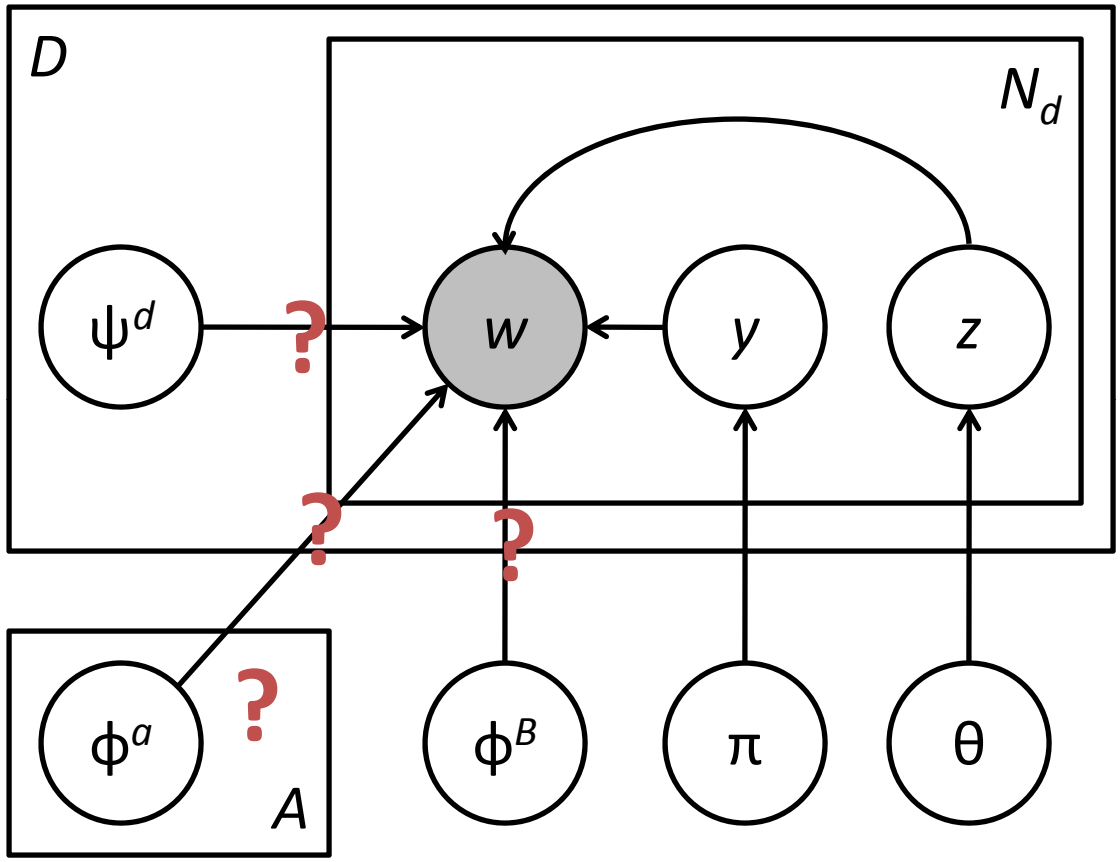


D_2



background

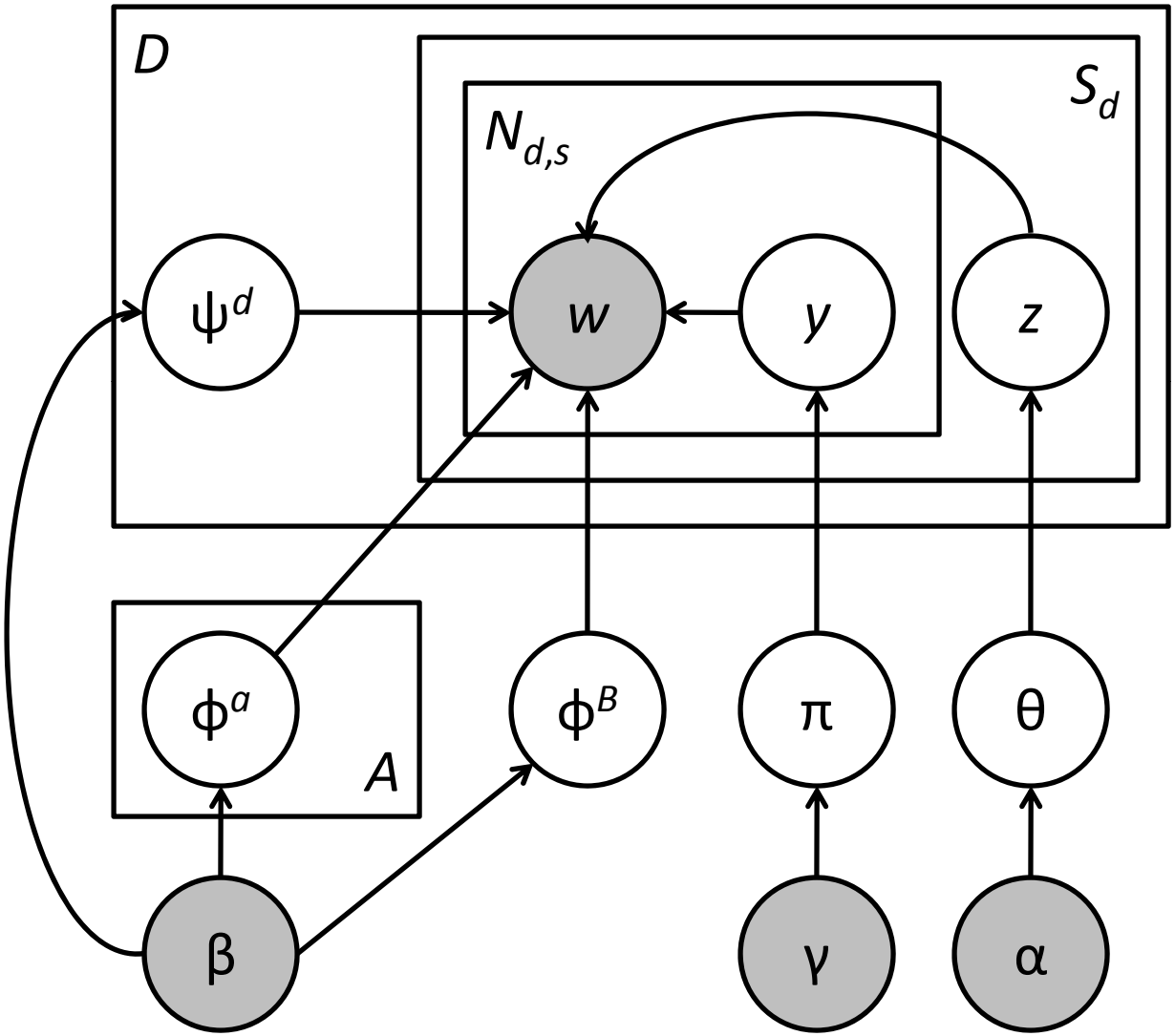
Entity-Aspect Model



$y \in \{1, 2, 3\}$

$z \in \{1, \ldots, A\}$

Entity-Aspect Model



$y \in \{1, 2, 3\}$

$z \in \{1, \ldots, A\}$

Model Inference

- Gibbs sampling

$$p(z_{d,s} = a | \boldsymbol{z}_{\neg\{d,s\}}, \boldsymbol{y}, \boldsymbol{w})$$

$$\propto \frac{C^{\mathcal{A}}_{(a)} + \alpha}{C^{\mathcal{A}}_{(\cdot)} + A\alpha} \cdot \frac{\prod_{v=1}^{V} \prod_{i=0}^{E_{(v)}} (C^{a}_{(v)} + i + \beta)}{\prod_{i=0}^{E_{(\cdot)}} (C^{a}_{(\cdot)} + i + V\beta)}$$

$$p(y_{d,s,n} = 1 | \boldsymbol{z}, \boldsymbol{y}_{\neg\{d,s,n\}}) \propto \frac{C^{\pi}_{(1)} + \gamma}{C^{\pi}_{(\cdot)} + 3\gamma} \cdot \frac{C^{\mathcal{B}}_{(w_{d,s,n})} + \beta}{C^{\mathcal{B}}_{(\cdot)} + V\beta}$$

$$p(y_{d,s,n} = 2 | \boldsymbol{z}, \boldsymbol{y}_{\neg\{d,s,n\}}) \propto \frac{C^{\pi}_{(2)} + \gamma}{C^{\pi}_{(\cdot)} + 3\gamma} \cdot \frac{C^{d}_{(w_{d,s,n})} + \beta}{C^{d}_{(\cdot)} + V\beta}$$

$$p(y_{d,s,n} = 3 | \boldsymbol{z}, \boldsymbol{y}_{\neg\{d,s,n\}}) \propto \frac{C^{\pi}_{(3)} + \gamma}{C^{\pi}_{(\cdot)} + 3\gamma} \cdot \frac{C^{a}_{(w_{d,s,n})} + \beta}{C^{a}_{(\cdot)} + V\beta}$$

Clustered Sentences

Venturi/D was/S a/S **professor/A** of/S physics/B at/S the/S **University/A** of/S Modena/D ./S

He/S was/S a/S **professor/A** of/S physics/B at/S the/S **University/A** of/S Chicago/D until/S 1982/D ./S

… …

S: stop word
B: background word
A: aspect word
D: document word

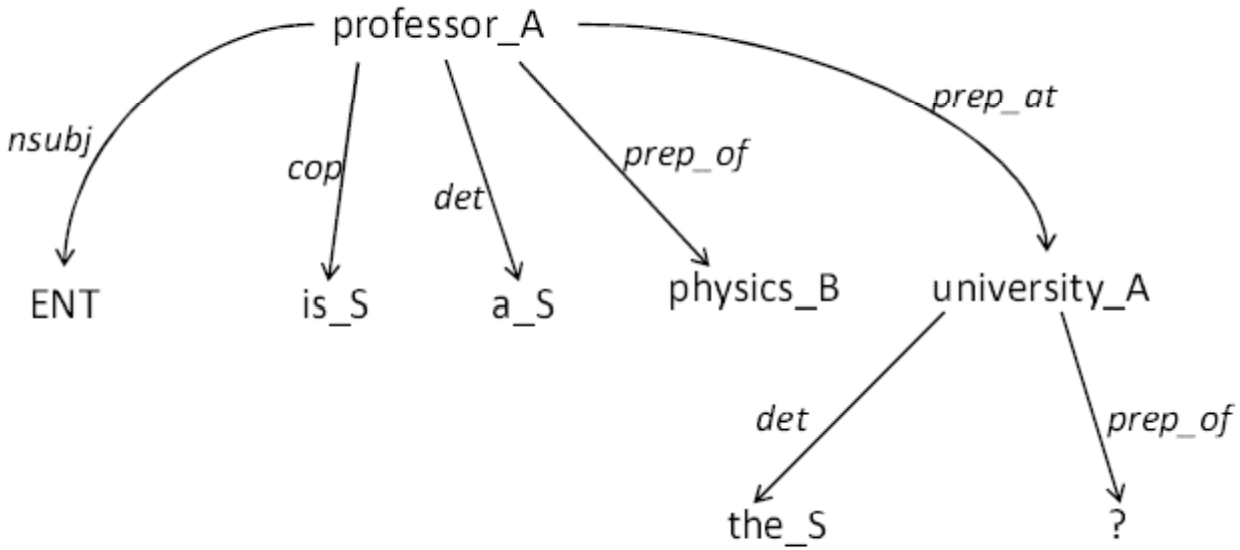
Outline

- Task and motivation
- Related work
- Our approach
  - Sentence clustering
  - **Pattern mining**
- Evaluation
- Conclusions

Pattern Mining

- Use heuristics to locate subject entities
  - Title of the Wikipedia article
  - Top 3 frequent subject noun phrases in the article
- Generate labeled dependency parse trees
  - ***Replace document words with "?"***
- Mine frequent subtree patterns
- Prune patterns
  - Remove patterns with no subject entity
  - ***Remove patterns with no aspect word***
- Convert subtree patterns to sentence patterns

A Labeled Dependency Parse Tree



Sample Aspects and Patterns

Aspect	Sample Sentence Patterns
1	X received his PhD from ___ University X studied ___ under ___ X earned his ___ in physics from University of ___
2	X was awarded the medal in ___ X won the ___ award X received the Nobel Prize in physics in ___
3	X was ___ director X was the head of ___ X worked for ___
4	X made contributions to ___ X is best known for work on ___ X is noted for ___

Outline

- Task and motivation
- Related work
- Our approach
  - Sentence clustering
  - Pattern mining
- **Evaluation**
- Conclusions

Data Set

- Wikipedia articles from 5 categories

Category	Documents	Sentences	Avg Sent/Doc
US Actress	407	1721	4
Physicist	697	4238	6
US CEO	179	1040	5
US Company	375	2477	6
Restaurant	152	1195	7

Quantitative Evaluation

- Sentence patterns
  - Manually judged whether each sentence pattern is meaningful for the given entity category
    - Allows us to compute *precision*, *recall* and *f1*
  - Baseline 1: pattern mining without entity-aspect model
  - Baseline 2: verb-based pruning (Filatova et al. 2006) on top of BL-1
- Aspect clusters
  - Manually grouped meaningful patterns into clusters
    - Allows us to compute *purity*

Quality of Sentence Patterns

Method		Category				
		US Actress	Physicist	US CEO	US Company	Restaurant
BL-1	precision	0.714	0.695	0.778	0.622	0.706
	recall	0.545	0.300	0.367	0.425	0.361
	f1	0.618	0.419	0.499	0.505	0.478
BL-2	precision	0.845	0.767	0.829	0.809	1.000
	recall	0.260	0.096	0.127	0.167	0.188
	f1	0.397	0.17	0.220	0.276	0.316
Ours	precision	0.544	0.607	0.586	0.450	0.560
	recall	0.710	0.785	0.712	0.618	0.701
	f1	0.616	0.684	0.643	0.520	0.624

- No consideration of aspect clusters.
- High precision of BL-1 and BL-2: They use a higher frequency threshold to select frequent sentence patterns.
- Low recall of BL-2: Many meaningful sentence patterns do not contain a non-auxiliary verb.
- BL-1 and BL-2 do not generate template slots.

Quality of Aspect Clusters

Category	B	Purity
US Actress	4	0.626
Physicist	6	0.714
US CEO	4	0.674
US Company	4	0.614
Restaurant	3	0.587

Sample Aspects and Their Representative Words

Method	Sample Aspects		
	1	2	3
Our entity-aspect model	university received ph.d. college degree	prize nobel physics awarded medal	academy sciences member national society
Standard LDA	physics american professor received university	nobel prize physicist awarded john	physics institute research member sciences
K-means	physics university institute work research	physicist american physics university nobel	physics academy sciences university new

Conclusions

- We proposed an LDA-based entity-aspect model to simultaneously cluster sentences and label words
- We used pattern mining to identify sentence patterns with slots
- We empirically evaluated our method and showed its effectiveness

Future Directions

- Use linguistic knowledge to further prune sentence patterns and improve their readability
- Automatic aspect labeling
- Automatic entity summary generation

THANK YOU!