

AI Tutorial

Sense Alignment of Wikipedia and WordNet

杜海倫 · 國立清華大學資訊工程系 · 自然語言研究小組

https://github.com/NTHU-NLPLAB/AI_Tutorial



Motivation

- Word sense disambiguation (WSD)
 - A long-standing but open problem in NLP
 - Semi-supervised systems have been constantly observed as leading to the highest performance
 - The sense-tagged data bottleneck
- [Rada Mihalcea 2007]: Generating sense-tagged data using Wikipedia as a source of sense annotations

Dataset	#s	#ex	baselines		word sense disambig.
			MFS	LeskC	
SENSEVAL	4.60	226	51.53%	58.33%	68.13%
WIKIPEDIA	3.31	316	72.58%	78.02%	84.65%

Procedure (cont')

1. Develop Wiki sense inventory (WikiSI)

- Using [hyperlink references](#) in Wikipedia:
 - `[[political party|party]], [[party]], [[party \(law\)|party]]`

Comparison with pledged delegates [\[edit \]](#)

Democratic Party rules distinguish pledged and unpledged delegates. Pledged delegates are selected based on their announced preferences in the contest for the presidential nomination. In the [party primary elections](#) and [caucuses](#) in each U.S. state, voters [express the](#) [States](#). Pledged delegates su
They fall into three categories delegates, and pledged PLEC
legally required to support the state (Rule 12.J): "Delegates
conscience reflect the sentime

A **political party** is an organized group of people who have the same ideology, or who otherwise have the same political positions, and who field candidates for elections, in an attempt to get them elected and thereby implement the party's agenda.

nomination for [President of the United](#)
ratio to their candidate's share of the vote.
[ssional districts](#)), at-large pledged
a minority of the states, delegates are
to the states' requirements, the party rules
idential candidate shall in all good

Procedure

2. Disambiguate and align Wikipedia hyperlinked articles to WordNet senses
 - [[political party|party]] >= party.n.01
 - [[party]] >= party.n.04
 - [[party (law)|party]] >= party.n.05
3. Extract Wikipedia sentences with anchor | Wiki page | WN sense
 - party | Political_party | party.n.01 (865)
 - party | Party | party.n.04 (320)
 - party | Party (law) | party.n.05 (123)

Example

- In Wiki page “Superdelegate” (<https://en.wikipedia.org/wiki/Superdelegate>)
 - We extract the sentence for party.n.05 : In the party primary elections and caucuses in each U.S. state,

Comparison with pledged delegates [\[edit \]](#)

Democratic Party rules distinguish pledged and unpledged delegates. Pledged delegates are selected based on their announced preferences in the contest for the presidential nomination. In the party primary elections and caucuses in each U.S. state, voters express the States. Pledged delegates su They fall into three categories delegates, and pledged PLEC legally required to support the state (Rule 12.J): "Delegates conscience reflect the sentime

A **political party** is an organized group of people who have the same ideology, or who otherwise have the same political positions, and who field candidates for elections, in an attempt to get them elected and thereby implement the party's agenda.

omination for President of the United ratio to their candidate's share of the vote. ssional districts), at-large pledged a minority of the states, delegates are to the states' requirements, the party rules identical candidate shall in all good

Hands-on Activity

- Preprocessing
 - Download Wiki Data from **Wikimedia Downloads** (<https://dumps.wikimedia.org>)
 - Extract useful information

Manual Alignment Guideline

- Align a WordNet sense with Wiki page if they are:
 - WordNet lemma is identical to page Identical
 - **bank.n.01** (*WordNet*): sloping land (especially the slope beside a body of water)
 - **Bank (geography)** (*Wikipedia*)
 - Else if a sister term is identical to page title
 - **disk.n.01**: something with a round shape resembling a flat circular plate
 - **Ball**
 - Else a daughter term is identical to page title
 - **party.n.05**: a person involved in legal proceedings
 - **Plaintiff**
- Exclude
 - Name entity and modifier
 - **Arms** industry

Evaluation

- Development set: 30 words in 158 senses and about 5,000 sentences
- Full set

	Unique Words	Synsets	Word-Sense Pairs	Ambig. words	(# of Sentences)
Wiki (2019/7/2)	41,615	-	(En) 88,571 (Zh) 22,781	12,159	3,741,348
WordNet	117,798	82,115	146,312	15,935	48,247

Alignment Methods

- Lemma
- Bilingual alignment
- # of transitive closures (hypos/hypers/sisterms...) of a synset in the text/hyperlink of the Wiki page
- Category
-

Download

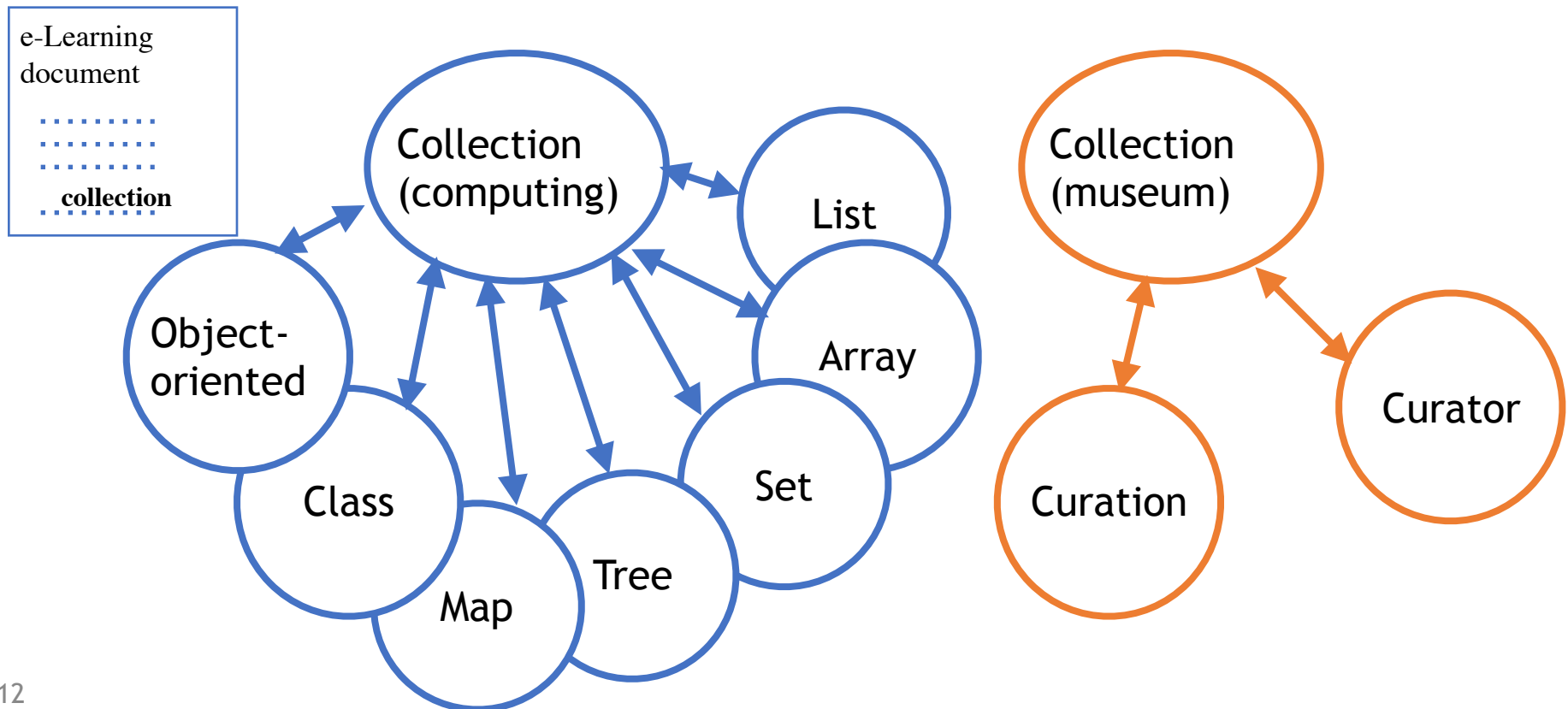
- Source code
 - https://github.com/NTHU-NLPLAB/AI_Tutorial
- TextNet
 - <http://textnet.nlplab.cc>
- Example Sentences
 - <https://tinyurl.com/y4jjsuej>

AI Tutorial

Graph Based Word Sense Disambiguation

WSD by Wiki Link Graph (cont')

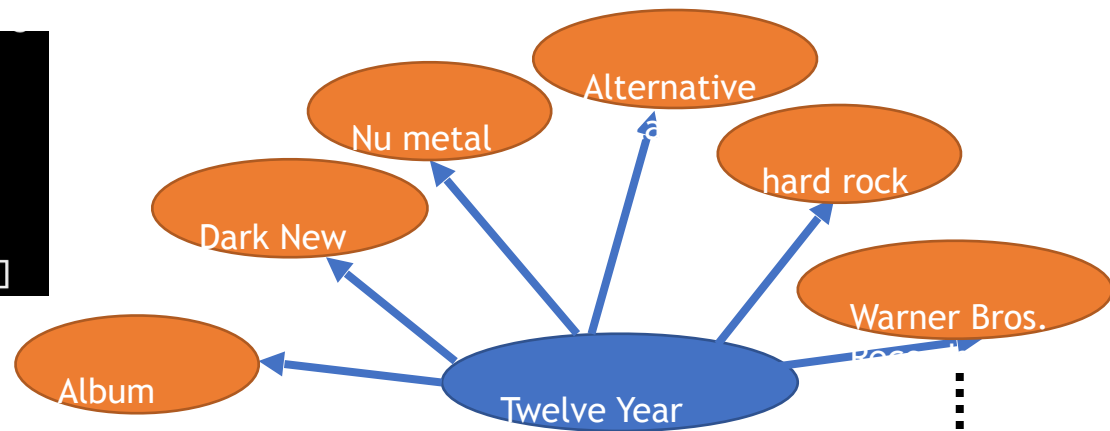
- [A Fogarolli 2009]: Word Sense Disambiguation based on Wikipedia Link Structure



WSD by Wiki Link Graph (cont')

- Wikipedia Link Graph
 - Node: each article
 - Edge: each link between articles

```
<title>Twelve Year Silence  
[[Album]]  
[[Dark New Day]]  
[[Nu metal]]  
[[alternative metal]]  
[[hard rock]]  
[[Warner Bros. Records|Warner Bros.]]
```



WSD by Wiki Link Graph (cont')

- Use Wikipedia link graph for WSD (without [[linkanchor]])

- 檢索輸入句 (rising senior) 相關文章，
- 在Wikipedia Link Graph, 經過較短且較多的路徑
- 連到解答的linkanchor (例如 Senior (education) 節點)

- 以“rising senior”中的senior(升大四生)為例

- “rising senior”中的 senior 不是有超連結的 anchor
- 無法直接連接到與senior較相關的wiki page

— Senior (education)

— Open class (track and field)

— Old age

```
senior Senior (education)      ___  518  Counter({'[[Senior (education)|senior]]': 158, '[[Senior (education)|senior]]': 8})
senior Open class (track and field)  ___  145  Counter({'[[senior (athletics)|senior]]': 60})
senior Old age 老年      58  Counter({'[[old age|senior]]': 58})
```

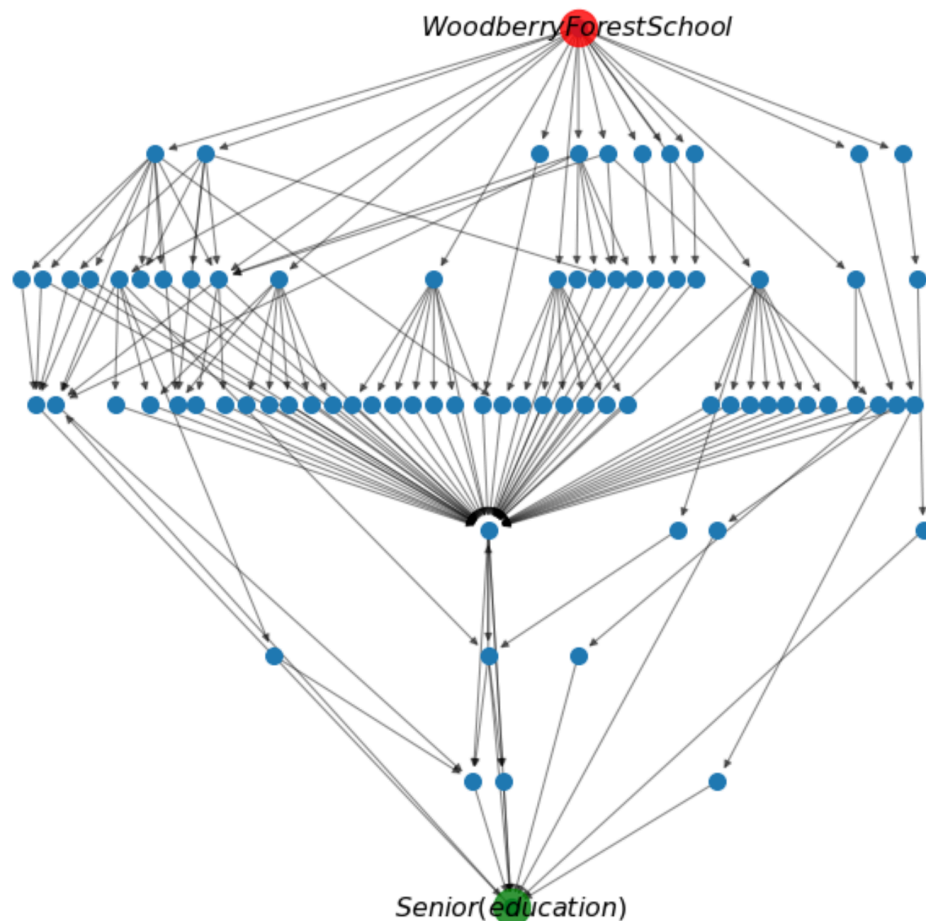
WSD by Wiki Link Graph (cont')

- 實驗步驟

- 找出所有含“rising senior”的wiki pages set RS
- 令set $S = \{\text{“Senior (education)”}, \text{“Open class (track and field)”}, \text{“Old age”}\}$
- 找出所有set RS 到 S 距離小於5的path, 以這些path建立 G 的subgraph G_{rs}
 - 範例圖為 RS 裡“Woodberry Forest School”到 S 裡“Senior (education)”的subgraph
- 計算 G_{rs} 的page rank

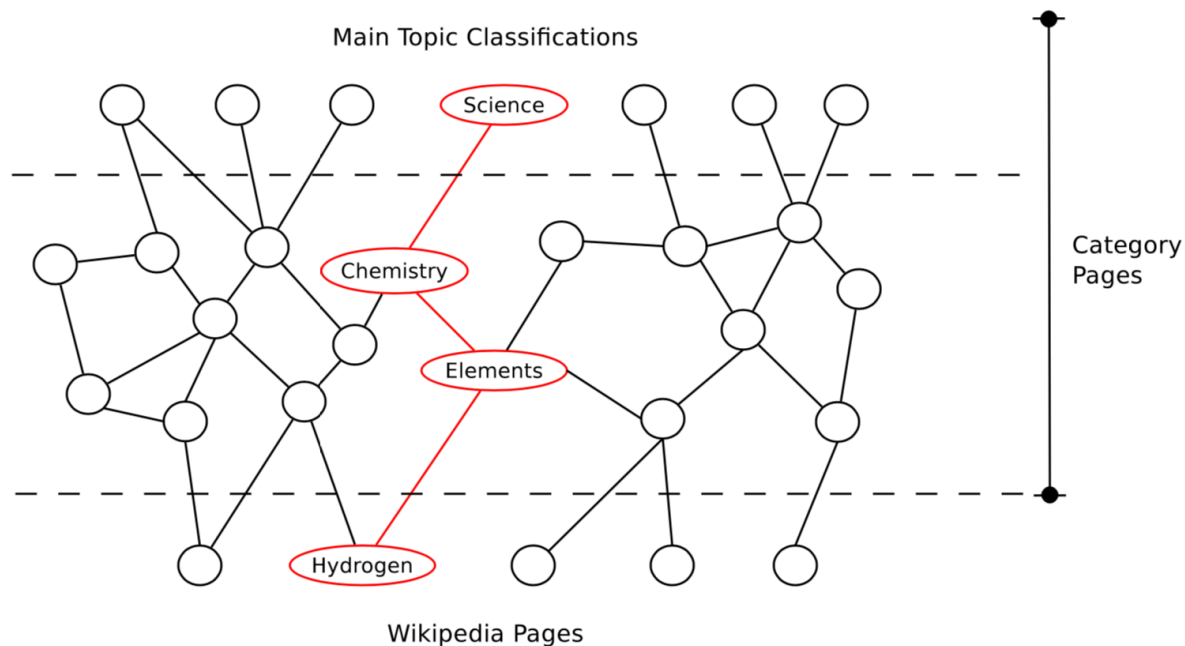
- 實驗結果

- Page rank 值:
 - Senior (education): 0.1037
 - Old age: 0.0083
- Open class (track and field)不在 G_{rs} 中



Hands-on Activity

- Category



Source: Christopher M. De Vries et al., 2010