위키피디아 기반 의미식별 연구동향

InSciTe 서비스 제작을 통한 경험을 중심으로

2013. 06. 26.

한국과학기술정보연구원

황명권

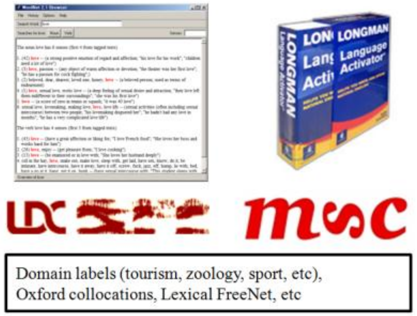
Intro – Word Sense Disambiguation

- WSD
  - 문맥의 의미를 파악하여 문맥 내의 단어들의 정확한 의미를 파악하는 것

- WSD의 중요성
  - 이미지, 비디오, 텍스트, 그래프, 소리 등 모든 개체를 상호 분석하기 위한 필수
    빅 데이터, 데이터 큐레이션 등

Intro – Historic Approach

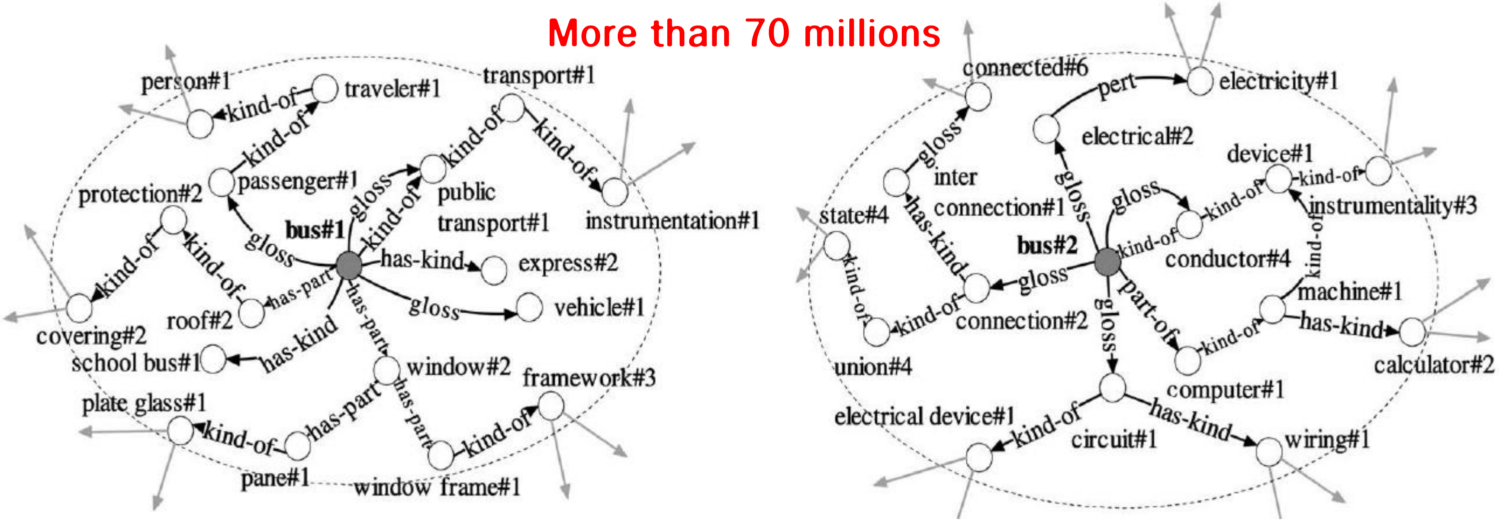
- WordNet based WSD – SSI Algorithm



Domain labels (tourism, zoology, sport, etc), Oxford collocations, Lexical FreeNet, etc

Resources	Explanations
WordNet 2.0	110,00 concepts
Annotated corpora	SemCor, LDC–DSO, WordNet glosses, WordNet usage examples
Dictionaries of collocations	Oxford Collocations, Longman Language Activator, Lexical FreeNet
Etc	Domain labels

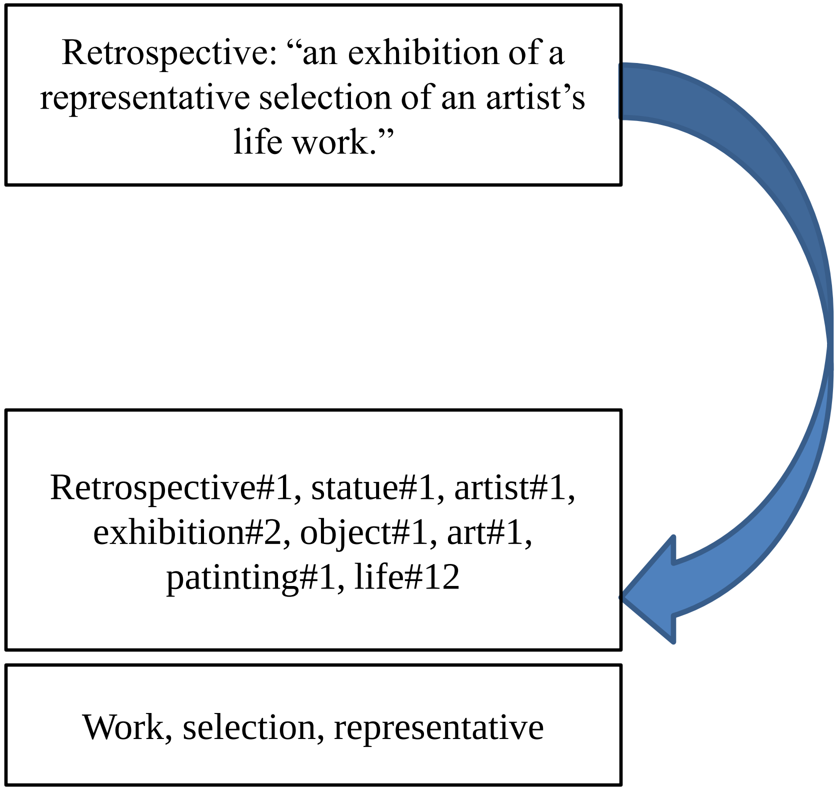
More than 70 millions



R. Navigli, et. al, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation," IEEE PAMI, Vol. 27, No. 7, pp. 1075-1086, 2005.

Intro – Historic Approach

- WordNet based WSD – SSI Algorithm



Retrospective: "an exhibition of a representative selection of an artist's life work."

Retrospective#1, statue#1, artist#1, exhibition#2, object#1, art#1, patinting#1, life#12

Work, selection, representative

$$T = [retrospective, work, object, exhibition, life, statue,$$
$$artist, selection, representative, painting, art]$$
$$I = [retrospective\#1, -, -, -, -, -, -, -, -, -, -]$$
$$P = [work, object, exhibition, life, statue, artist, selection,$$
$$representative, painting, art].$$

$$I = [retrospective\#1, statue\#1, artist\#1]$$
$$P = [work, object, exhibition, life, selection, representative,$$
$$painting, art]$$

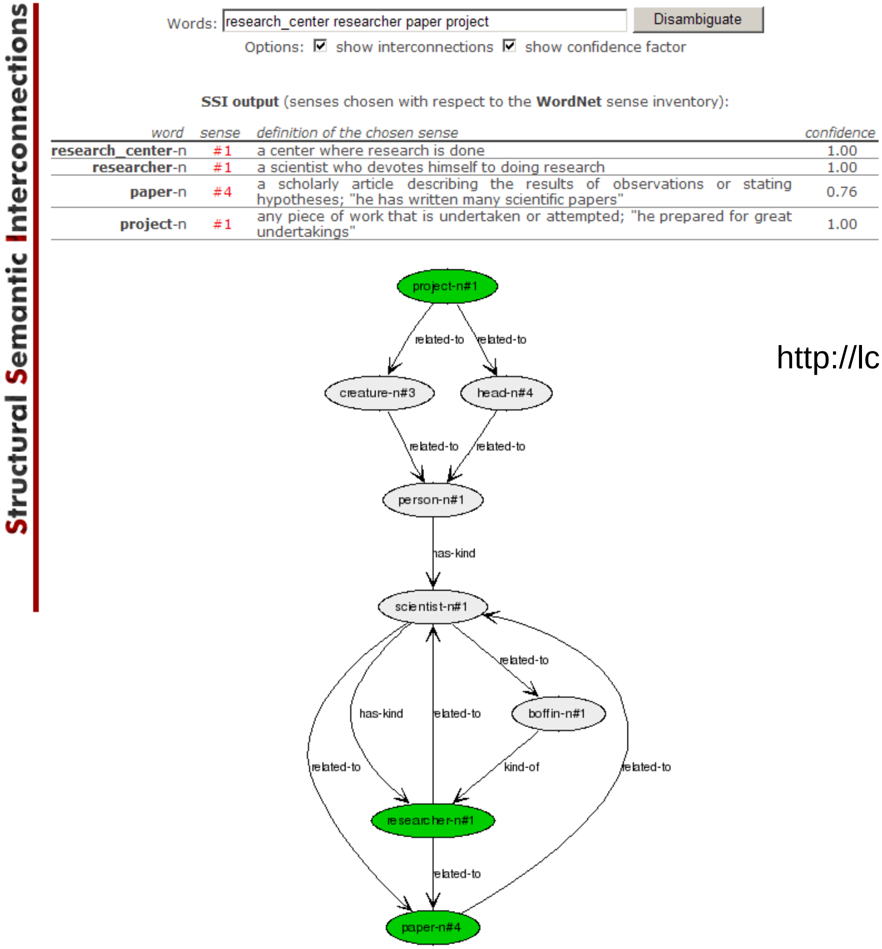
$$retrospective\#1 \xrightarrow{kind\text{-}of^2} exhibition\#2,$$
$$statue\#1 \xrightarrow{kind\text{-}of^3} art\#1 \text{ and } statue\#1 \xrightarrow{kind\text{-}of^6} object\#1.$$

$$I = [retrospective\#1, statue\#1, artist\#1, exhibition\#2,$$
$$object\#1, art\#1]$$
$$P = [work, life, selection, representative, painting].$$

R. Navigli, et. al, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation," IEEE PAMI, Vol. 27, No. 7, pp. 1075-1086, 2005.

Intro – Historic Approach

- WordNet based WSD – SSI Algorithm



Structural Semantic Interconnections

Words: research_center researcher paper project Disambiguate

Options: ☑ show interconnections ☑ show confidence factor

SSI output (senses chosen with respect to the **WordNet** sense inventory):

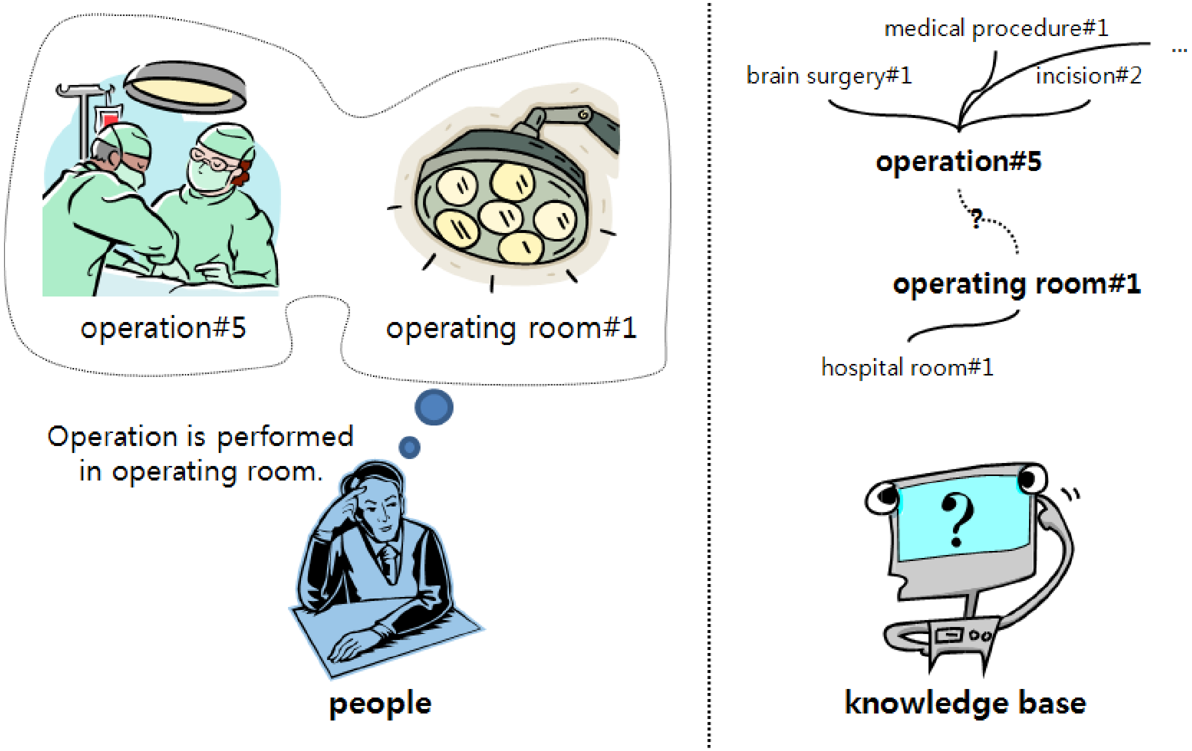
word	*sense*	*definition of the chosen sense*	*confidence*
research_center-n	#1	a center where research is done	1.00
researcher-n	#1	a scientist who devotes himself to doing research	1.00
paper-n	#4	a scholarly article describing the results of observations or stating hypotheses; "he has written many scientific papers"	0.76
project-n	#1	any piece of work that is undertaken or attempted; "he prepared for great undertakings"	1.00

http://lcl.uniroma1.it/ssi/

R. Navigli, et. al, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation," IEEE PAMI, Vol. 27, No. 7, pp. 1075-1086, 2005.

Intro – Historic Approach

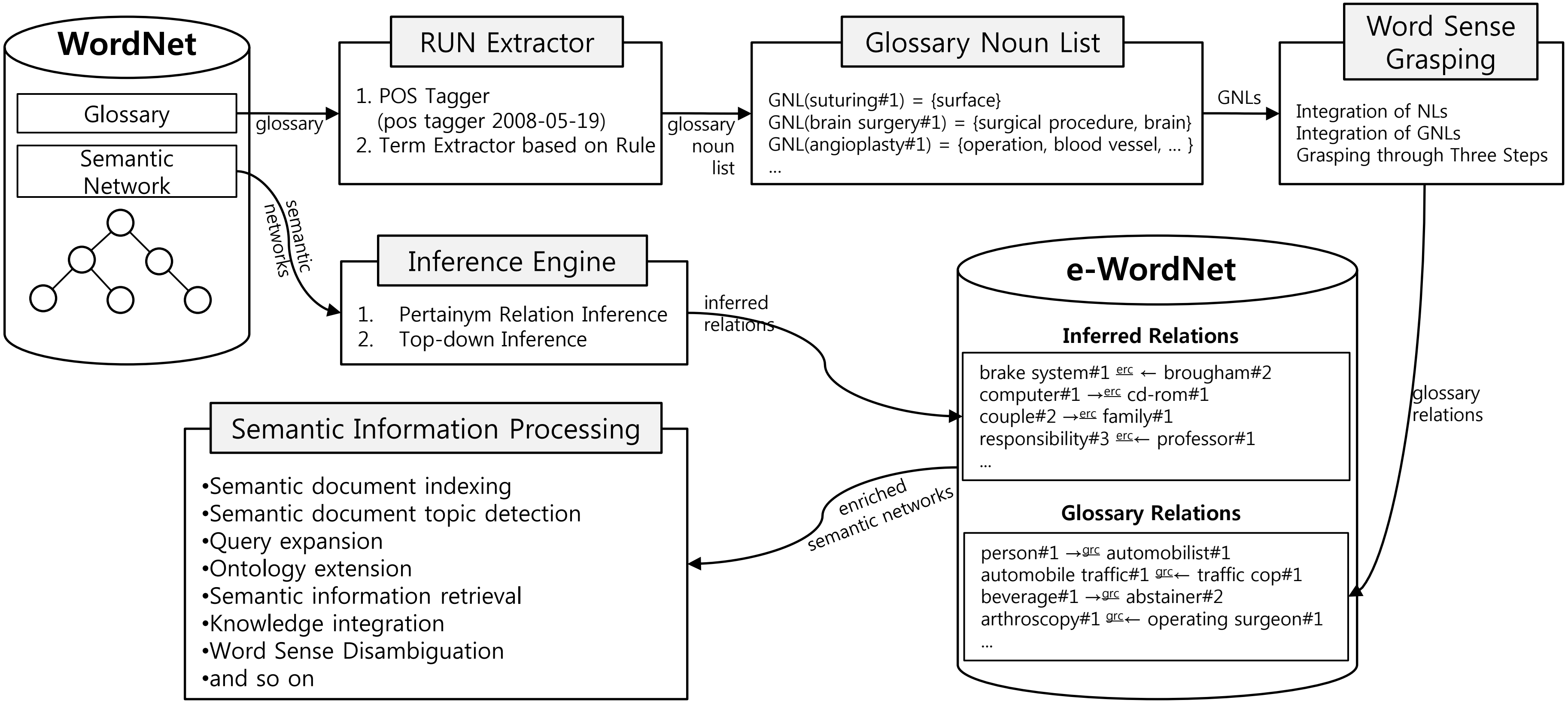
- **WordNet based WSD – WSD SemNet**



M. Hwang, et. al, "Automatic enrichment of semantic relation network and its application to word sense disambiguation," IEEE TKDE, Vol. 23, No. 6, pp. 845-858, 2011.

Intro – Historic Approach

- **WordNet based WSD – WSD SemNet**



WordNet

Glossary

Semantic Network

glossary

semantic networks

RUN Extractor

1. POS Tagger (pos tagger 2008-05-19)
2. Term Extractor based on Rule

glossary noun list

Glossary Noun List

GNL(suturing#1) = {surface}
GNL(brain surgery#1) = {surgical procedure, brain}
GNL(angioplasty#1) = {operation, blood vessel, … }
…

GNLs

Word Sense Grasping

Integration of NLs
Integration of GNLs
Grasping through Three Steps

glossary relations

Inference Engine

1. Pertainym Relation Inference
2. Top-down Inference

inferred relations

e-WordNet

Inferred Relations

brake system#1 $\xleftarrow{erc}$ brougham#2
computer#1 $\xrightarrow{erc}$ cd-rom#1
couple#2 $\xrightarrow{erc}$ family#1
responsibility#3 $\xleftarrow{erc}$ professor#1
…

Glossary Relations

person#1 $\xrightarrow{grc}$ automobilist#1
automobile traffic#1 $\xleftarrow{grc}$ traffic cop#1
beverage#1 $\xrightarrow{grc}$ abstainer#2
arthroscopy#1 $\xleftarrow{grc}$ operating surgeon#1
…

enriched semantic networks

Semantic Information Processing

- Semantic document indexing
- Semantic document topic detection
- Query expansion
- Ontology extension
- Semantic information retrieval
- Knowledge integration
- Word Sense Disambiguation
- and so on

More than 1.7 millions

M. Hwang, et. al, "Automatic enrichment of semantic relation network and its application to word sense disambiguation," IEEE TKDE, Vol. 23, No. 6, pp. 845-858, 2011.

Intro – Historic Approach

- WordNet based WSD – WSD SemNet

Coverage Test

KB	Basic KB	Light KB	Heavy KB	SSI
Coverage (%)	74.25 444/598	82.11 491/598	**89.13** **533/598**	85.45 511/598

Concept-pairs	The basic	The light	The heavy	SSI KB
love#1-family#1	X	O	O	O
book#1-desk#1	X	O	O	O
travel#1-tour_guide#1	X	X	O	O
police#1-traffic#1	X	X	X	O
brougham#2-fan_belt#1	X	X	O	X
fire_engine#1-fireman#4	X	X	O	X
fireman#4-fire#1	X	X	O	O
home#1-marriage#1	O	O	O	X
crop#1-growing_season#1	X	O	O	X
emergency#1-fire#1	O	O	O	X
bank#1-money#1	X	O	O	O
buddhist#1-India#1	X	O	O	X
education#1-school#1	X	O	O	O
captain#3-ship#1	X	O	O	O
liquid#1-water#1	X	O	O	O
parent#1-love#1	X	X	X	O

WSD based on SemCor

Method	Total Count of Concept Pairs	Pre.	Recall	F1
SSI	70,005,325	56.21	**95.08**	70.65
WSD-SemNet with the Basic KB	203,760	56.20	82.33	66.24
WSD-SemNet with the Light KB	318,160	**59.82**	86.84	70.84
WSD-SemNet with the Heavy KB	1,748,627	57.94	92.69	**71.31**

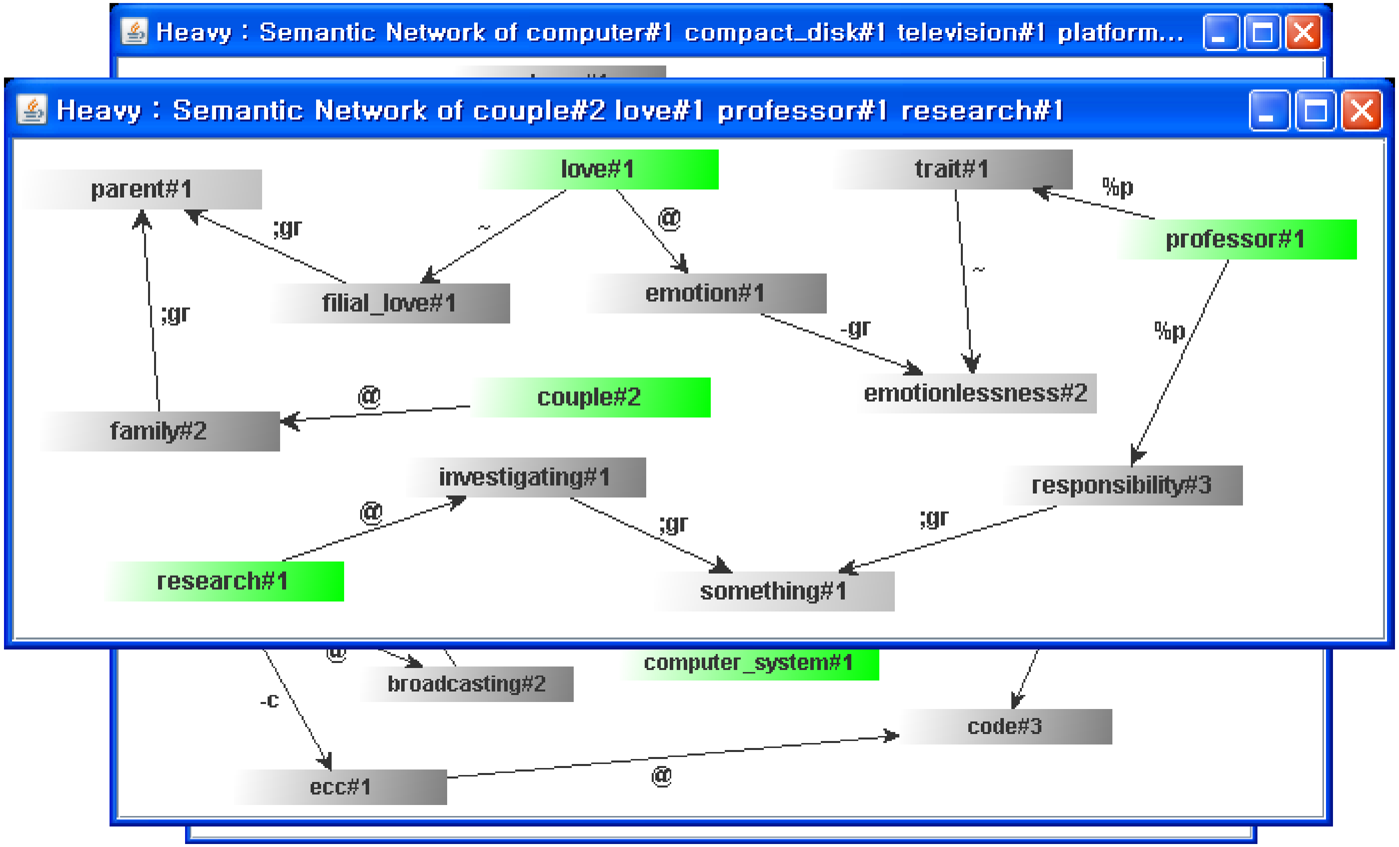
WSD based on Senseval-3

KB	Pre.	Recall	F1
The basic	70.3	74.3	72.2
The light	75.7	78.9	77.2
The heavy	71.3	85.2	77.7
The heavy without 'gr'	71.2	83.3	76.8

M. Hwang, et. al, "Automatic enrichment of semantic relation network and its application to word sense disambiguation," IEEE TKDE, Vol. 23, No. 6, pp. 845-858, 2011.

Intro – Historic Approach

- WordNet based WSD – WSD SemNet



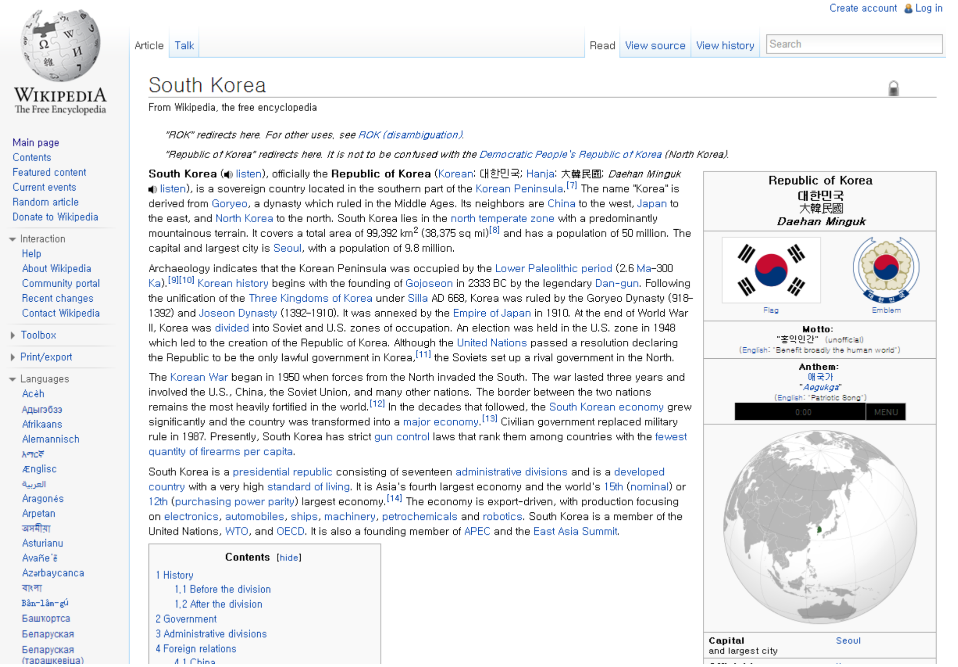
M. Hwang, et. al, “Automatic enrichment of semantic relation network and its application to word sense disambiguation,” IEEE TKDE, Vol. 23, No. 6, pp. 845-858, 2011.

Intro – Wikipedia

- 강점
  - 전세계 모든 개념 커버, 시맨틱 네트워크 형성
  - 지속적인 개념 증가
  - 신뢰성 보장

Intro – Wikipedia

- Components
  - Title (concept)
  - Abstract
  - Info box
  - Contents
  - In/out link
  - Article–category
  - …



South Korea

From Wikipedia, the free encyclopedia

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

"Republic of Korea" redirects here. It is not to be confused with the Democratic People's Republic of Korea (North Korea).

South Korea (listen), officially the **Republic of Korea** (Korean: 대한민국; Hanja: 大韓民國; *Daehan Minguk* listen), is a sovereign country located in the southern part of the Korean Peninsula.[7] The name "Korea" is derived from Goryeo, a dynasty which ruled in the Middle Ages. Its neighbors are China to the west, Japan to the east, and North Korea to the north. South Korea lies in the north temperate zone with a predominantly mountainous terrain. It covers a total area of 99,392 km^2 (38,375 sq mi)[8] and has a population of 50 million. The capital and largest city is Seoul, with a population of 9.8 million.

Archaeology indicates that the Korean Peninsula was occupied by the Lower Paleolithic period (2.6 Ma–300 Ka).[9][10] Korean history begins with the founding of Gojoseon in 2333 BC by the legendary Dan-gun. Following the unification of the Three Kingdoms of Korea under Silla AD 668, Korea was ruled by the Goryeo Dynasty (918–1392) and Joseon Dynasty (1392–1910). It was annexed by the Empire of Japan in 1910. At the end of World War II, Korea was divided into Soviet and U.S. zones of occupation. An election was held in the U.S. zone in 1948 which led to the creation of the Republic of Korea. Although the United Nations passed a resolution declaring the Republic to be the only lawful government in Korea,[11] the Soviets set up a rival government in the North.

The Korean War began in 1950 when forces from the North invaded the South. The war lasted three years and involved the U.S., China, the Soviet Union, and many other nations. The border between the two nations remains the most heavily fortified in the world.[12] In the decades that followed, the South Korean economy grew significantly and the country was transformed into a major economy.[13] Civilian government replaced military rule in 1987. Presently, South Korea has strict gun control laws that rank them among countries with the fewest quantity of firearms per capita.

South Korea is a presidential republic consisting of seventeen administrative divisions and is a developed country with a very high standard of living. It is Asia's fourth largest economy and the world's 15th (nominal) or 12th (purchasing power parity) largest economy.[14] The economy is export-driven, with production focusing on electronics, automobiles, ships, machinery, petrochemicals and robotics. South Korea is a member of the United Nations, WTO, and OECD. It is also a founding member of APEC and the East Asia Summit.

Contents [hide]

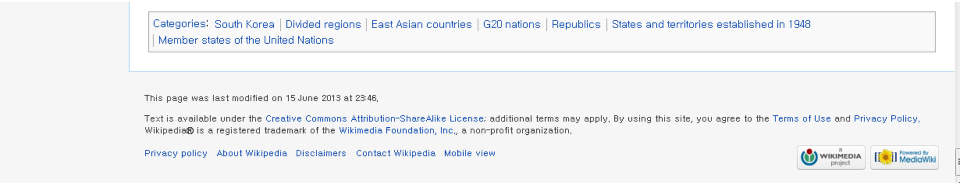
Republic of Korea
대한민국
大韓民國
Daehan Minguk

Motto:
"홍익인간" (unofficial)
(English: "Benefit broadly the human world")

Anthem:
애국가
"*Aegukga*"
(English: "Patriotic Song")

Capital and largest city — Seoul

…

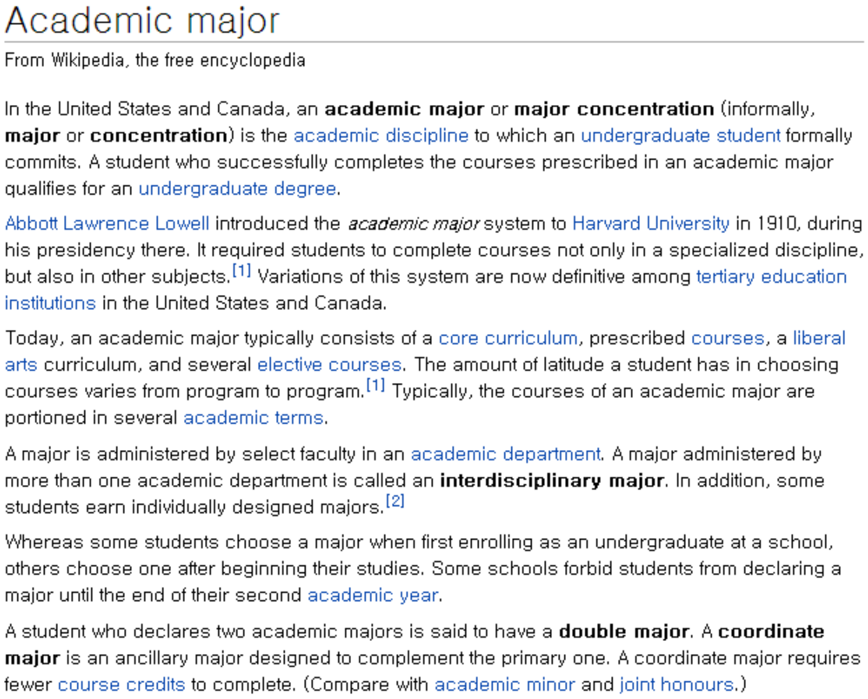


Categories: South Korea | Divided regions | East Asian countries | G20 nations | Republics | States and territories established in 1948 | Member states of the United Nations

This page was last modified on 15 June 2013 at 23:46.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Context based WSD



Academic major

From Wikipedia, the free encyclopedia

In the United States and Canada, an **academic major** or **major concentration** (informally, **major** or **concentration**) is the academic discipline to which an undergraduate student formally commits. A student who successfully completes the courses prescribed in an academic major qualifies for an undergraduate degree.

Abbott Lawrence Lowell introduced the *academic major* system to Harvard University in 1910, during his presidency there. It required students to complete courses not only in a specialized discipline, but also in other subjects.[1] Variations of this system are now definitive among tertiary education institutions in the United States and Canada.

Today, an academic major typically consists of a core curriculum, prescribed courses, a liberal arts curriculum, and several elective courses. The amount of latitude a student has in choosing courses varies from program to program.[1] Typically, the courses of an academic major are portioned in several academic terms.

A major is administered by select faculty in an academic department. A major administered by more than one academic department is called an **interdisciplinary major**. In addition, some students earn individually designed majors.[2]

Whereas some students choose a major when first enrolling as an undergraduate at a school, others choose one after beginning their studies. Some schools forbid students from declaring a major until the end of their second academic year.

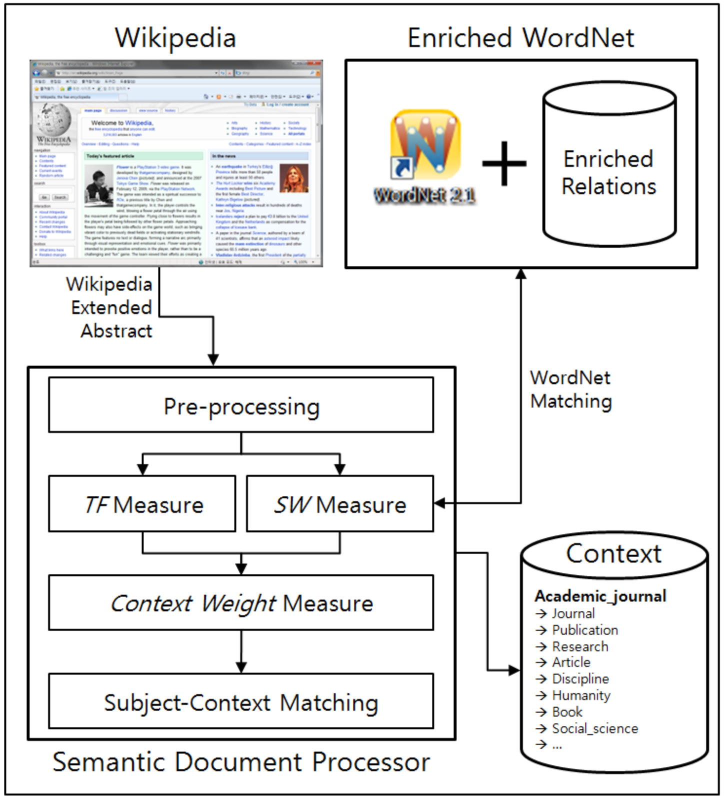
A student who declares two academic majors is said to have a **double major**. A **coordinate major** is an ancillary major designed to complement the primary one. A coordinate major requires fewer course credits to complete. (Compare with academic minor and joint honours.)

Noun (*tf*)	Student(0.12), university(0.08), major(0.08), core(0.06), field(0.04), study(0.04), education(0.04), college(0.04), term(0.02), portion(0.04), year(0.04), number(0.04), curriculum(0.04), concentration(0.04), art(0.02), program(0.02), framework(0.02), …

	Context Word#Sense	Context_CW
Context Weight	student#1	0.654
	university#2	0.344
	major#4	0.312
	core#4	0.271
	field#4	0.259
	study#6	0.259
	education#1	0.222
	college#1	0.181
	curriculum#1	0.147
	…	…

D. Choi, et. al, "Semantic Context Extraction from Wikipedia Document," In Proceedings of The 2010 International Conference on Semantic Web & Web Services, pp. 38-41, 2010.

Context based WSD



Wikipedia

Enriched WordNet

WordNet 2.1

+

Enriched Relations

Wikipedia Extended Abstract

WordNet Matching

Pre-processing

TF Measure

SW Measure

Context Weight Measure

Subject-Context Matching

Semantic Document Processor

Context

Academic_journal
→ Journal
→ Publication
→ Research
→ Article
→ Discipline
→ Humanity
→ Book
→ Social_science
→ ...

D. Choi, et. al, "Semantic Context Extraction from Wikipedia Document," In Proceedings of The 2010 International Conference on Semantic Web & Web Services, pp. 38-41, 2010.

Context based WSD

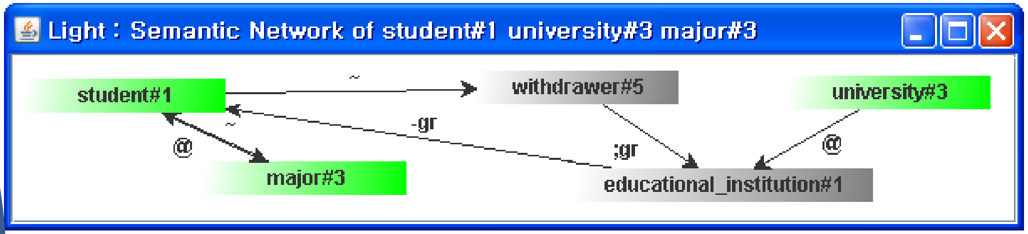
$$relatedness(s_{ia}, s_{jb}) = \frac{1}{\underset{s_{ia} \in SL_i, s_{jb} \in SL_j}{\arg\ \min}\left(dist(s_{ia}, s_{jb})\right)}, i \neq j$$

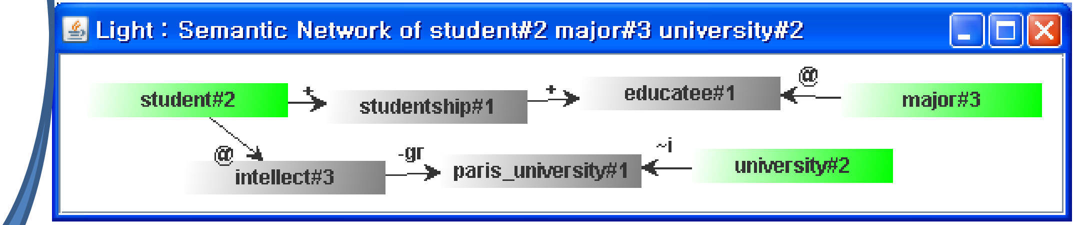
$$sw(s_{ia}) = \sum_{j=1}^{n} \underset{s_{jb} \in S_j}{\arg\ \max}\left(relatedness(s_{ia}, s_{jb})\right), i \neq j$$

$$sw(t_i) = \underset{s_{ia} \in S_i}{\arg\ \max}\left(sw_{ia}\right)$$

$$cw(t_i) = tf(t_i) \times \left(sw(t_i) + 1\right)$$

Student, major, university


Light : Semantic Network of student#1 university#3 major#3
student#1
~
withdrawer#5
university#3
~
-gr
@
;gr
@
major#3
educational_institution#1



Academic major

Context Word#Sense	Context_CW
student#1	0.654
university#2	0.344
major#4	0.312
core#4	0.271
field#4	0.259
study#6	0.259
education#1	0.222
college#1	0.181
curriculum#1	0.147
…	…

relatedness(student#1, major#3) = 0.5
relatedness(student#1, university#3) = 0.33
***cw(student#1)* = 0.83**

relatedness(student#2, major#3) = 0.25
relatedness(student#2, university#2) = 0.25
cw(student#2) = 0.5

D. Choi, et. al, "Semantic Context Extraction from Wikipedia Document," In Proceedings of The 2010 International Conference on Semantic Web & Web Services, pp. 38-41, 2010.

Article/Category based WSD (Similarity)

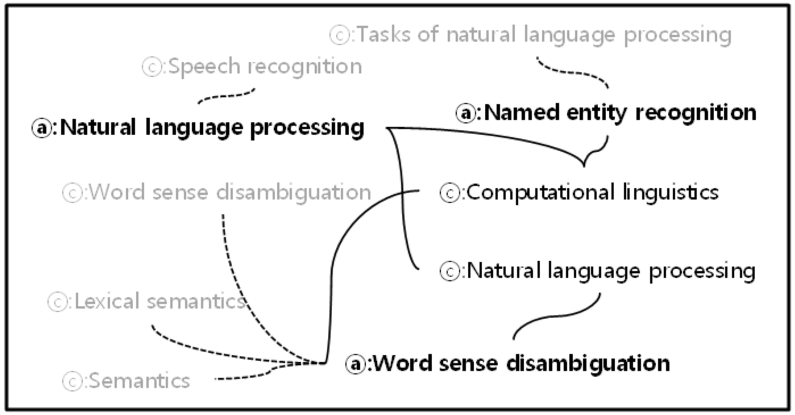
Similarity (natural language processing, word sense disambiguation)

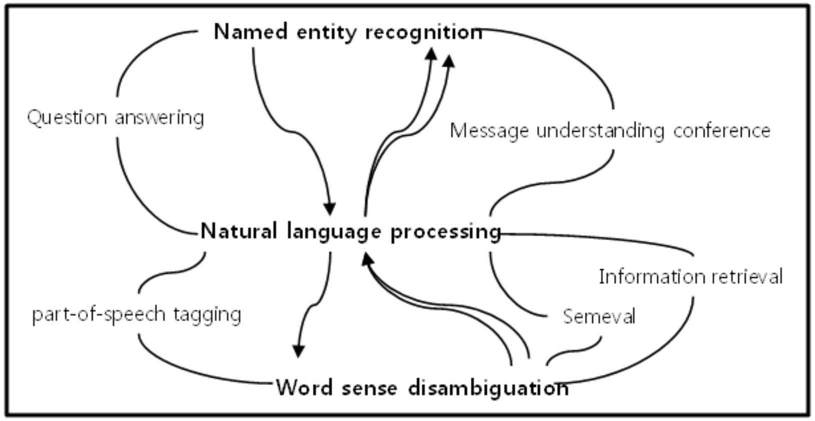
Similarity (natural language processing, named entity recognition)

Pair	Method	Similarity measure
NLP, NER	sim_C	(2*1)/(3+2) = 0.400
	sim_I	(3*3)/(2*(212+152)+(2*43)/(3*(212+152) = 0.041
	sim_H	(0.4+0.041)/2 = 0.221
NLP, WSD	sim_C	(2*2)/(3+5) = 0.500
	sim_I	(2*3)/(2*(212+26)+(2*10)/(3*(212+26)) = 0.= 0.087
	sim_H	(0.5+0.087)/2 = 0.294

M. Hwang, et. al, “Measuring Similarities between Technical Terms Based on Wikipedia,” In Proceedings of IEEE International Conferences on Things, and Cyber, Physical and Social Computing, pp. 533-536, 2011.

Article/Category based WSD


ⓒ:Tasks of natural language processing
ⓒ:Speech recognition
ⓐ:Natural language processing
ⓐ:Named entity recognition
ⓒ:Word sense disambiguation
ⓒ:Computational linguistics
ⓒ:Natural language processing
ⓒ:Lexical semantics
ⓐ:Word sense disambiguation
ⓒ:Semantics



$$sim_C(a,b) = \frac{2 \times |C_a \cap C_b|}{(|C_a| + |C_b|)}$$

$$n\_sim_C(a,b) = \frac{sim_C(a,b)}{\max \arg_{art \in cat_a, art \neq a} sim_C(a, art)}$$

art means an article and cat_a is a category of an article *a*.

SN_A: 'NLP' – 'NER'
SN_B:'NLP' – 'WSD' based on Wikipeida interlink
Actually 'NLP,' 'NER,' and 'WSD' has 212, 152, and 26 interlinks
(DL, IL) of SN_A and SN_B is (3, 43) and (3, 10) respectively in Wikipedia.

$$sim_I(a,b) = \frac{2 \times |DL(a,b)|}{2 \times (|\vec{a}| + |\vec{b}|)} + \frac{2 \times |IL(a,b)|}{3 \times (|\vec{a}| + |\vec{b}|)}$$

$\vec{a}$ and $\vec{b}$ is out-link of the article. The results are also normalized by the maximum value. Hereafter, *sim* means the normalized similarity for each measure.

M. Hwang, et. al, "Measuring Similarities between Technical Terms Based on Wikipedia," In Proceedings of IEEE International Conferences on Things, and Cyber, Physical and Social Computing, pp. 533-536, 2011.

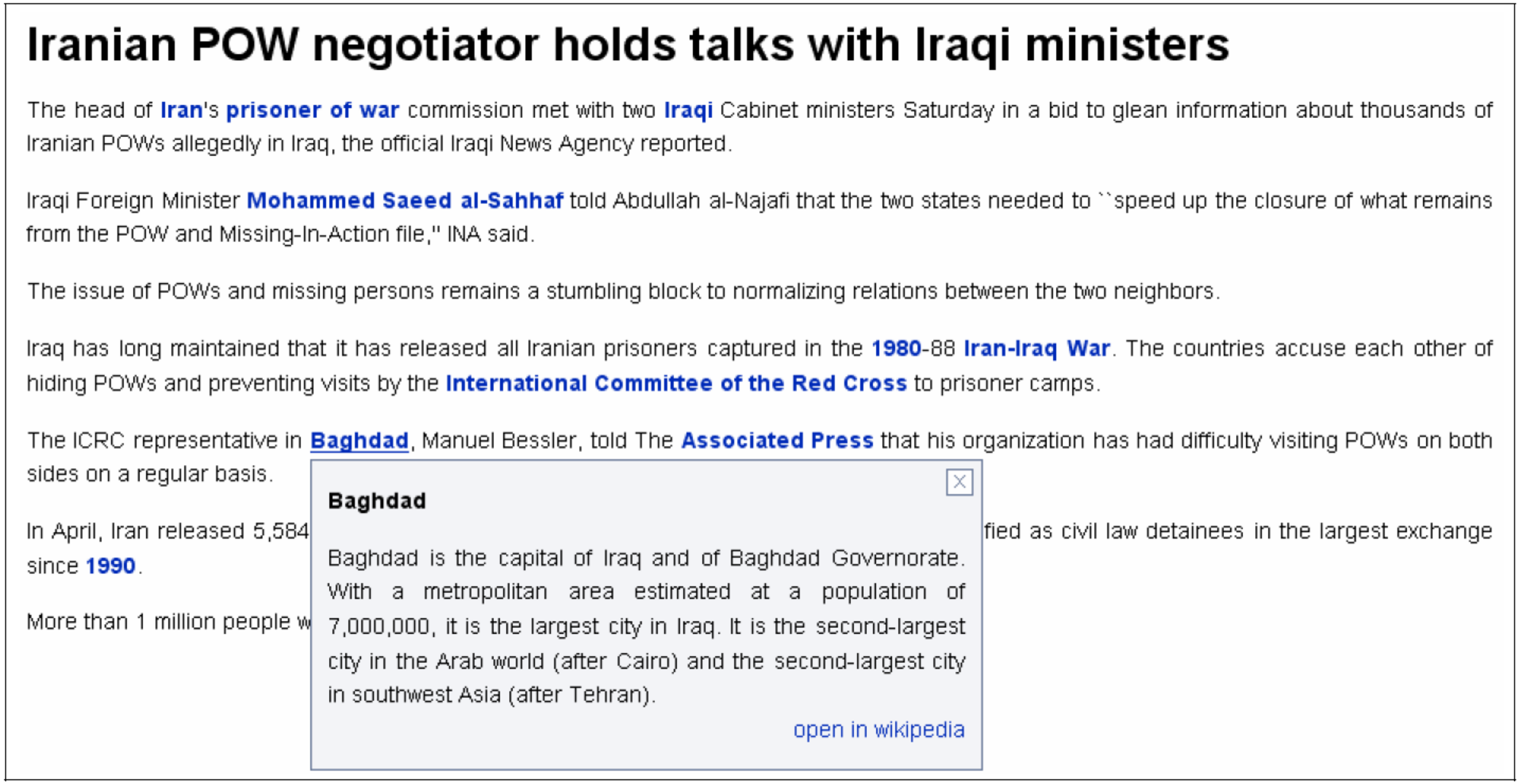
Article/Category based WSD

$$sim_H(a,b) = (1-\alpha) \times sim_C(a,b) + \alpha \times sim_I(a,b)$$

Pair	Method	Similarity measure
NLP, NER	sim_C	(2*1)/(3+2) = 0.400
	sim_I	(3*3)/(2*(212+152)+(2*43)/(3*(212+152) = 0.041
	sim_H	(0.4+0.041)/2 = 0.221
NLP, WSD	sim_C	(2*2)/(3+5) = 0.500
	sim_I	(2*3)/(2*(212+26)+(2*10)/(3*(212+26)) = 0.= 0.087
	sim_H	(0.5+0.087)/2 = 0.294

M. Hwang, et. al, "Measuring Similarities between Technical Terms Based on Wikipedia," In Proceedings of IEEE International Conferences on Things, and Cyber, Physical and Social Computing, pp. 533-536, 2011.

Wikipedia Link-based Measure



Iranian POW negotiator holds talks with Iraqi ministers

The head of **Iran**'s **prisoner of war** commission met with two **Iraqi** Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister **Mohammed Saeed al-Sahhaf** told Abdullah al-Najafi that the two states needed to ``speed up the closure of what remains from the POW and Missing-In-Action file,'' INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the **1980**-88 **Iran-Iraq War**. The countries accuse each other of hiding POWs and preventing visits by the **International Committee of the Red Cross** to prisoner camps.

The ICRC representative in **Baghdad**, Manuel Bessler, told The **Associated Press** that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 fied as civil law detainees in the largest exchange since **1990**.

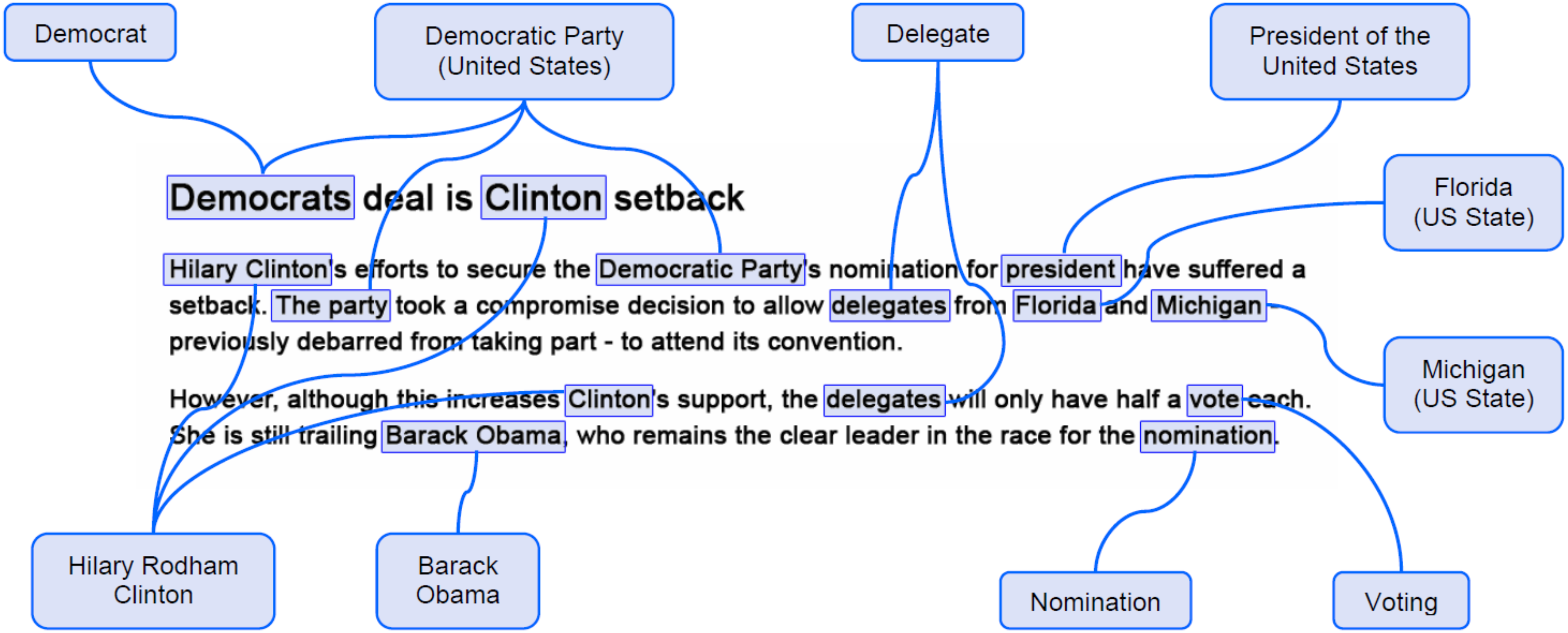
More than 1 million people w

Baghdad

Baghdad is the capital of Iraq and of Baghdad Governorate. With a metropolitan area estimated at a population of 7,000,000, it is the largest city in Iraq. It is the second-largest city in the Arab world (after Cairo) and the second-largest city in southwest Asia (after Tehran).

David Milne, "Learning to Link with Wikipedia," In Proceedings of CIKM 2008.

Wikipedia Link-based Measure



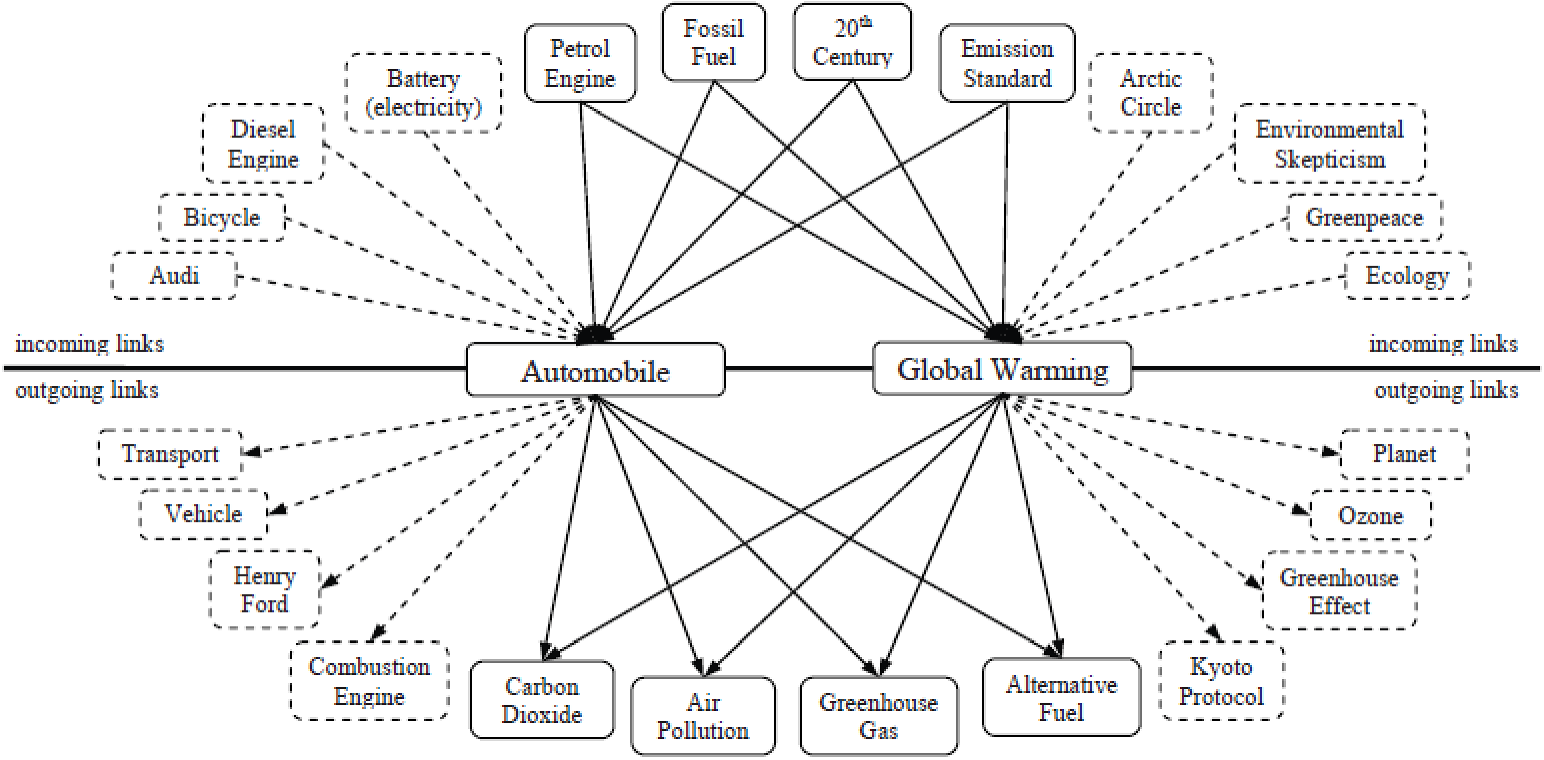
Democrats deal is Clinton setback

Hilary Clinton's efforts to secure the Democratic Party's nomination for president have suffered a setback. The party took a compromise decision to allow delegates from Florida and Michigan previously debarred from taking part - to attend its convention.

However, although this increases Clinton's support, the delegates will only have half a vote each. She is still trailing Barack Obama, who remains the clear leader in the race for the nomination.

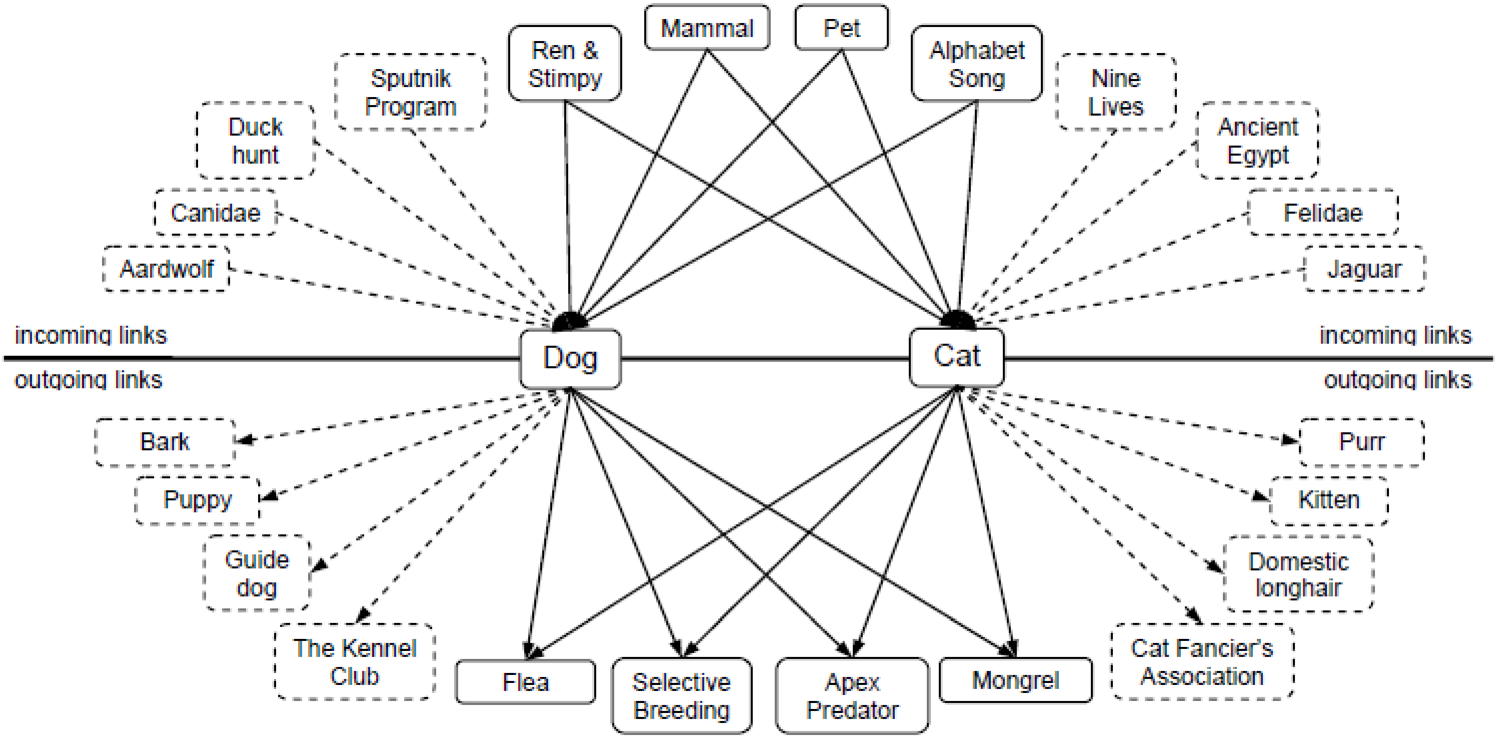
David Milne, "Learning to Link with Wikipedia," In Proceedings of CIKM 2008.

Wikipedia Link-based Measure



David Milne and Ian H. Witten, “An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links,” In Proceedings fo AAAI 2008

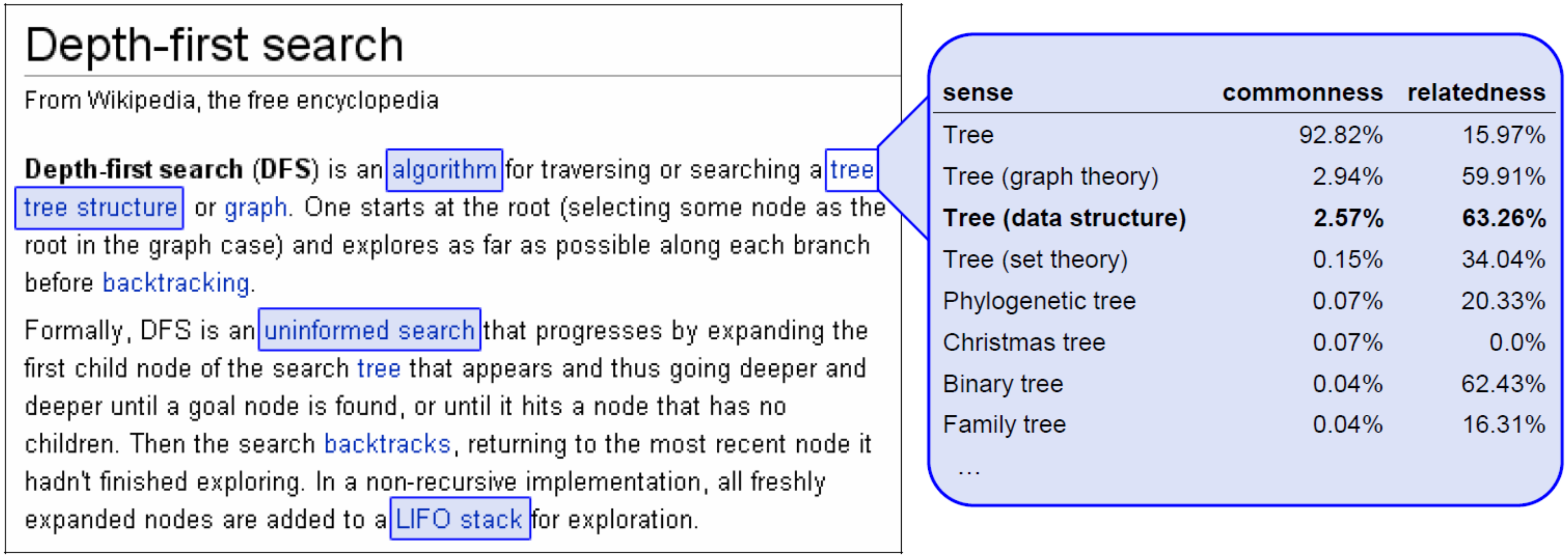
Wikipedia Link-based Measure



David Milne, "An Open-Source Toolkit for Mining Wikipedia," In Proceedings of New Zealand Computer Science Research Student Conference, 2009.

Wikipedia Link-based Measure

$$sr(a,b)=\frac{\log(\max(|A|,|B|))-\log(|A\cap B|)}{\log(|W|)-\log(\min(|A|,|B|))}$$



Depth-first search

From Wikipedia, the free encyclopedia

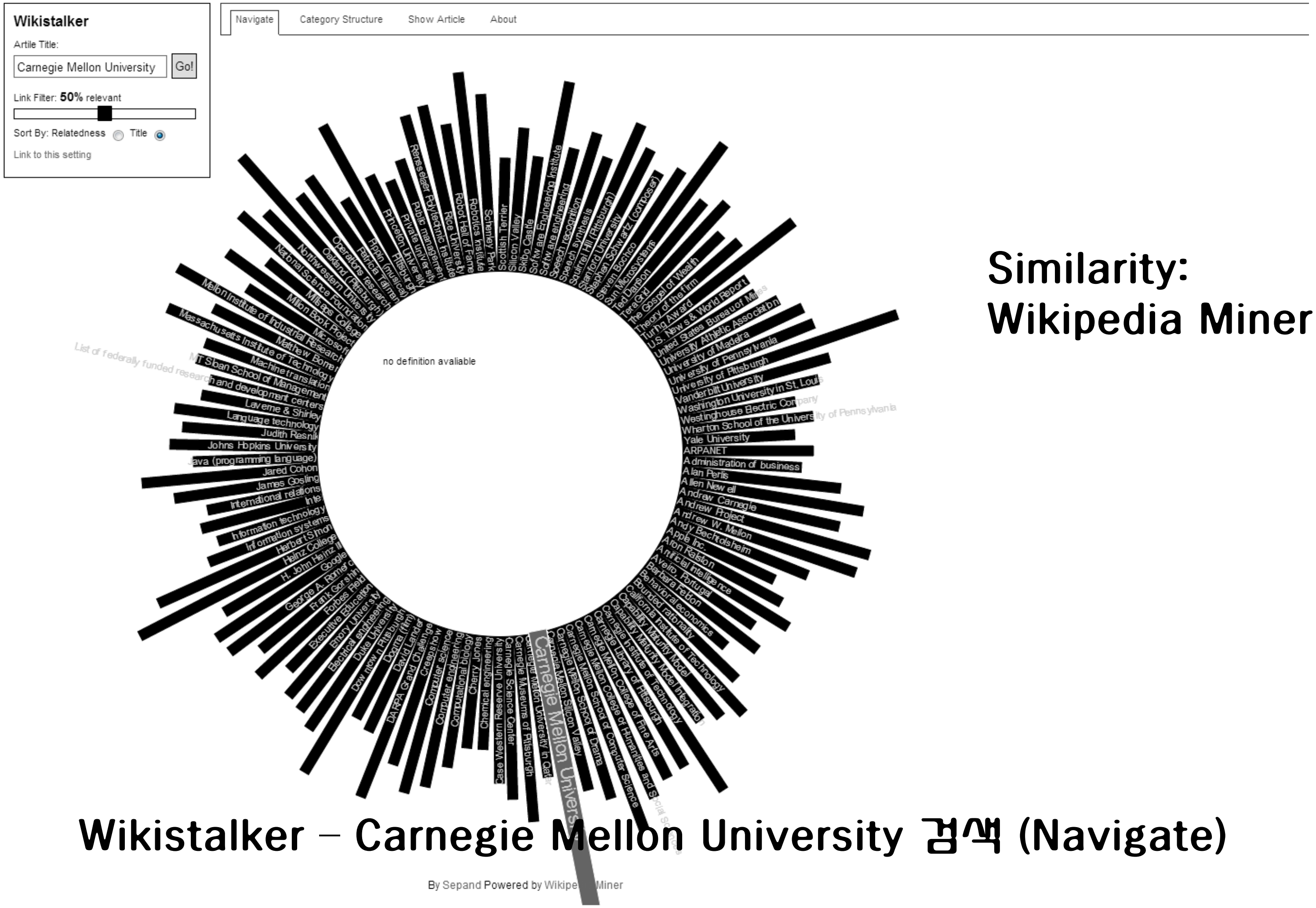
Depth-first search (**DFS**) is an algorithm for traversing or searching a tree, tree structure, or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

David Milne, "Learning to Link with Wikipedia," In Proceedings of CIKM 2008.

Wikistalker



Wikistalker

Artile Title: Carnegie Mellon University Go!

Link Filter: **50%** relevant

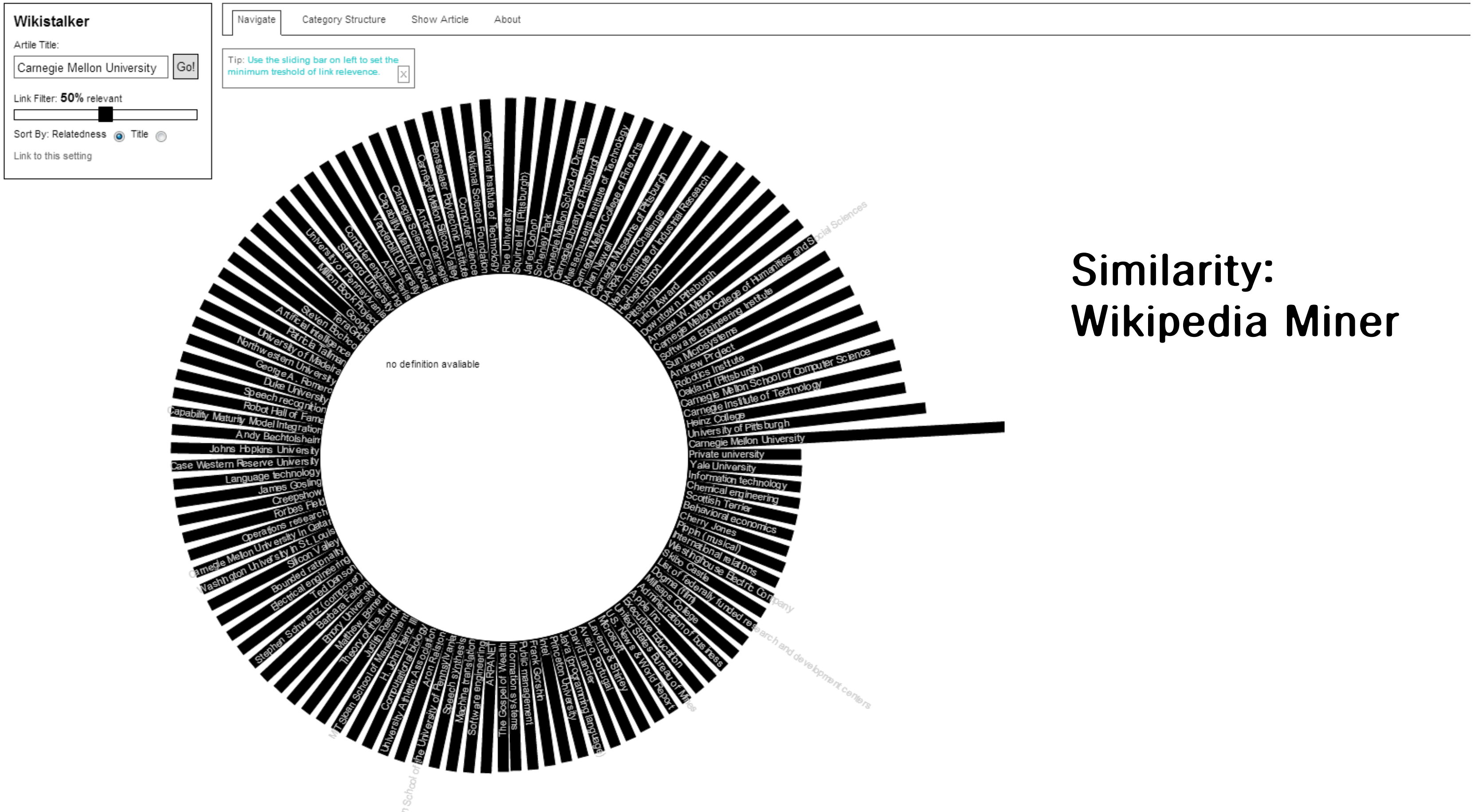
Sort By: Relatedness ○ Title ●

Link to this setting

Similarity:
Wikipedia Miner

Wikistalker – Carnegie Mellon University 검색 (Navigate)

Wikistalker



Wikistalker

Artile Title: Carnegie Mellon University Go!

Link Filter: **50%** relevant

Sort By: Relatedness ● Title ○

Link to this setting

Navigate | Category Structure | Show Article | About

Tip: Use the sliding bar on left to set the minimum treshold of link relevence. ☒

no definition avaliable

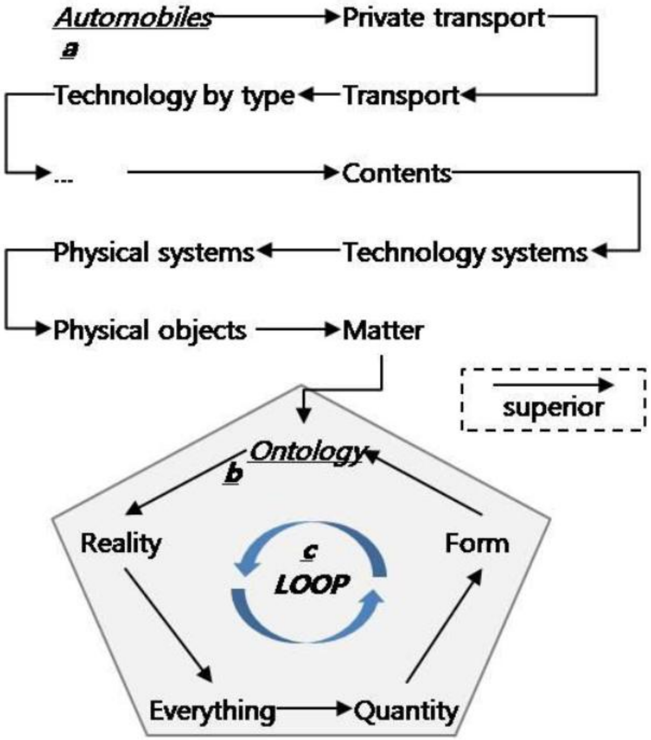
By Sepand Powered by Wikipedia Miner

Similarity:
Wikipedia Miner

Wikistalker – Carnegie Mellon University 검색 (Navigate)

Outro

- Limitations



Thank you for having me.

황명권

mgh@kisti.re.kr

http://johnnie.kisti.re.kr