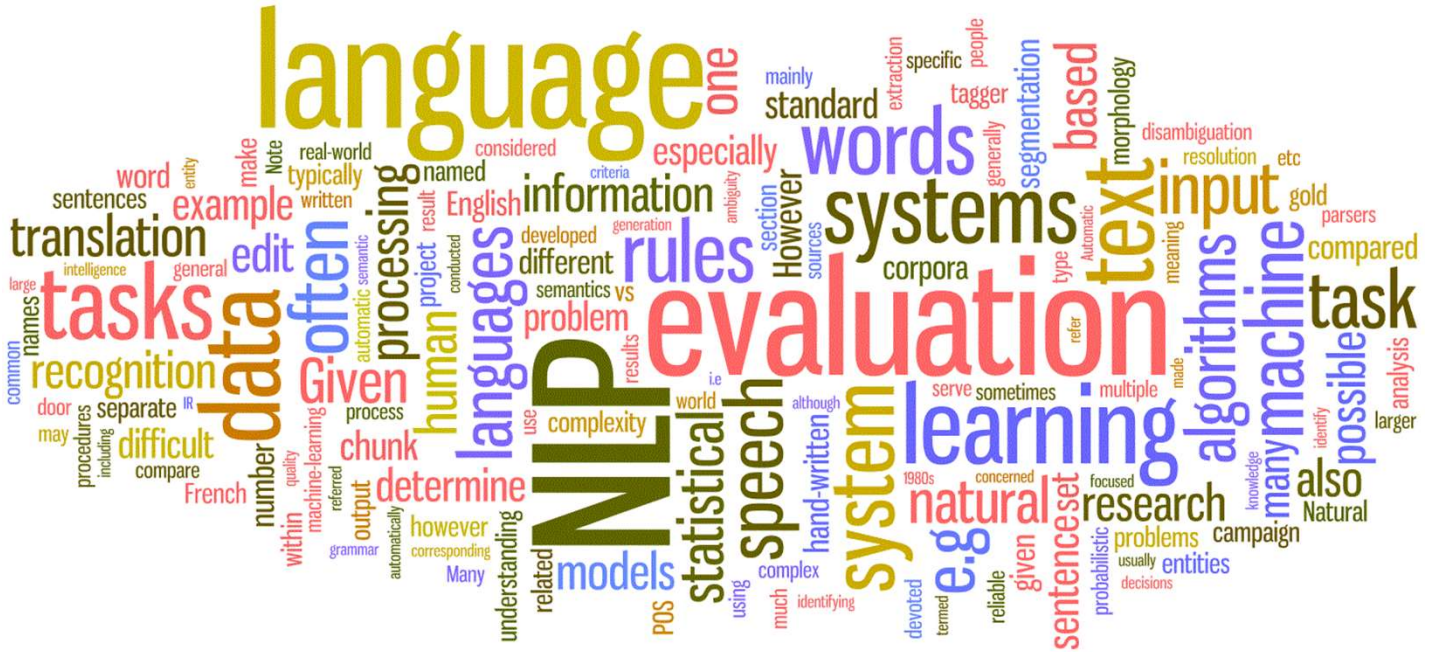


Lecture 7: Word Meanings

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Today's Class

- Word Meanings
- Computer Expression of Word Meanings
- English Semantic Resources
  - WordNet
- Chinese Semantic Resources
  - TongYiCiCiLin
  - HowNet
  - Chinese Concept Dictionary: CCD
- Applications of Semantic Resources

Word Meaning/Sense

- In linguistics, a word sense is one of the meanings of a word
- In each sentence, we associate a different meaning of the observing word based on hints the rest of the sentence
  - We went to see the **play** *Romeo and Juliet* at the theater
  - The children went out to **play** in the park
- Semantics: Study of word meanings

Expression and organization of Word Sense

- A word meaning is expressed by a description string in a dictionary
- Word Entries in different dictionary
  - 翻腾
  - 现代汉语词典
    - Sense 1: 上下滚动: 波浪~
    - Sense 2: 翻动: 几个柜子都~到了
  - Description string
  - Examples
  - Glosses: Definitions and/or example sentences

Chinese Dictionary

- 字典and词典
- Organization of Chinese dictionaries
  - By sound
    - Entries are ordered alphabetically using 拼音*pinyin* with its 25 letters
    - 倒序词典Reverse order dictionary
      - 腾　　翻腾 折腾 慢腾腾 闹腾
  - By graphic shape
    - graphic components known as radicals.部首
    - graphic-based look-up method四角号码
  - By meaning
    - TongYiCiCiLin- CiLin 同义词词林

Computer Expression of Word Meaning

- Described by Sememe 义元:
  - the smallest unit of meaning and is not further divisible
  - sememe <move> is shared by the various action morphemes *walk, run, roll, jump, etc*
- Described by definition
  - 翻腾
    - DEF=exercise|锻练,sport|体育
    - DEF=roll|滚
    - DEF=turn|扭转

- Described by Synonyms Set with glosses
  - WordNet

    good, right, ripe

    (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")
  - CiLin
    - 翻腾 叨登 捣腾 倒腾 翻倒 倾倒 扇风

Terminology

- **Homonyms** 同音
  - one of a group of words that share the same spelling and the same pronunciation but have different meanings
  - *Bank* a financial institution / river bank.
- **Polyseme** 多义:
  - words or phrases with multiple, related meanings
  - *Bed* in “river bed” / a household furniture.
  - Bank: financial institute, building of the financial
  - institute, storage of blood (blood bank)

- **Synonyms**同义
  - different words having similar or identical meanings
  - *student / pupil; buy / purchase*
- **Synset** 同义词集
  - a set of one or more synonyms
- **Antonyms** 反义
  - different words having contradictory or contrary meanings
  - *Perfect/imperfect big/small*

- **Hypernymy** 上位
  - the semantic relation of being super-ordinate or belonging to a higher rank or class
  - ‘living things’ is a hypernym of ‘human’
- **Hyponomy** 下位
  - the semantic relation of being sub-ordinate or belonging to a lower rank or class
  - 'adult' is a hyponom of 'human'.

- **Holonymy**整体
  - a word that defines the relationship between a term denoting the whole and a term denoting **a part of**, or **a member of**, the whole
  - *'tree'* is a holonym of *'bark*树皮*'*, *'trunk*树干*'*
- **Meronym**部份
  - a word that names **a part of** a larger whole
  - *'finger'* is a meronym of *'hand'*
  - *'wheel'* is a meronym of *'automobile'*.

- **Metonymy** 转指
  - a figure of speech in which a concept is referred to by the name of something closely associated with that concept
  - *White House* vs. *President* *Beijing* vs. *China*
- **Proposition**
  - it refers to the meaning of a statement
  - *All men are created equal* is a proposition.

WordNet

- WordNet is a lexical database for the English language
- Groups English words into synsets, provides short, general definitions, and records the various semantic relations between these synonym sets
- Purposes:
  - Produce a combination of dictionary and thesaurus that is more intuitively usable
  - Support automatic text analysis and artificial intelligence applications

WordNet Information

- Development began in 1985 at the Cognitive Science Laboratory of Princeton University
- WordNet's latest version is 3.0.
- The database contains 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs
- In compressed form, it is about 12 megabytes in size

WordNet: Design

- Organize the lexical knowledge according to word senses rather word form
- Description Target in WordNet
  - compound复合词
  - phrasal verb短语 动词
  - collocation搭配词
  - idiomatic phrase成语
  - word单词
- Word is the minimum unit
  - No morpheme
  - No frames

WordNet: Lexical Knowledge and World Knowledge

- Lexical knowledge:
  - Maintained in dictionary
    - ‘hit’ : a hard blow delivered with the hand or something held in it
    - ‘hit’ is verb
    - ‘hit’ is followed by the object
- World Knowledge
  - Maintained in Encyclopedia
    - a violent impact between things
    - Object should be solid / no gas
- WordNet focus on Lexical Knowledge
  - In some entries, combine lexical knowledge and world knowledge

WorNet Organization

- WordNet distinguishes
  - nouns, verbs, adjectives and adverbs
  - follow different grammatical rules
- Every synset contains a group of synonymous words or collocations
- Different senses of a word are in different synsets
- The meaning of the synsets is further clarified with short defining glosses
- Synsets are connected to other synsets via a number of semantic relations

WorNet: Sysnet Example

- Synset: Good, right, ripe
- Glosses: most suitable or right for a particular purpose
- Example collocations:
  - "a good time to plant tomatoes“
  - "the right time to act“
  - "the time is ripe for great sociological changes“
- Different meanings are in the different synsets

WorNet: Polysemes

POS	Monosemy and Senses	Polysemes	Senses of Polysemes
Noun	94685	14510	40002
Verb	5920	5168	18221
Adjectives	15981	5479	15175
Adverb	3820	787	1900
Total	120406	25944	75298

POS	Senses (including monosemy)	Senses
Noun	1.23	2.75
Verb	2.17	3.52
Adjectives	1.45	2.76
Adverb	1.24	2.41

Format of WordNet Entries

The noun "bass" has 8 senses in WordNet.

1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective "bass" has 1 sense in WordNet.

1. bass[1], deep[6] - (having or denoting a low vocal or instrumental range)
   "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

WorNet: Semantic Relations - Noun

- Hypernyms: Y is a hypernym of X if every X is a (kind of) Y
  - Canine is a hypernym of dog
- Hyponyms: Y is a hyponym of X if every Y is a (kind of) X
  - dog is a hyponym of canine
- Coordinate terms: Y is a coordinate term of X if X and Y share a hypernym
  - wolf is a coordinate term of dog
- Holonym 整体: Y is a holonym of X if X is a part of Y
  - building is a holonym of window
- Meronym 部分: Y is a meronym of X if Y is a part of X
  - window is a meronym of building

WorNet: Knowledge Structure

- Both nouns and verbs are organized into hierarchies, defined by hypernym

dog, domestic dog, Canis familiaris

=> canine, canid

=> carnivore

=> placental, placental mammal, eutherian, eutherian mammal

=> mammal

=> vertebrate, craniate

=> chordate

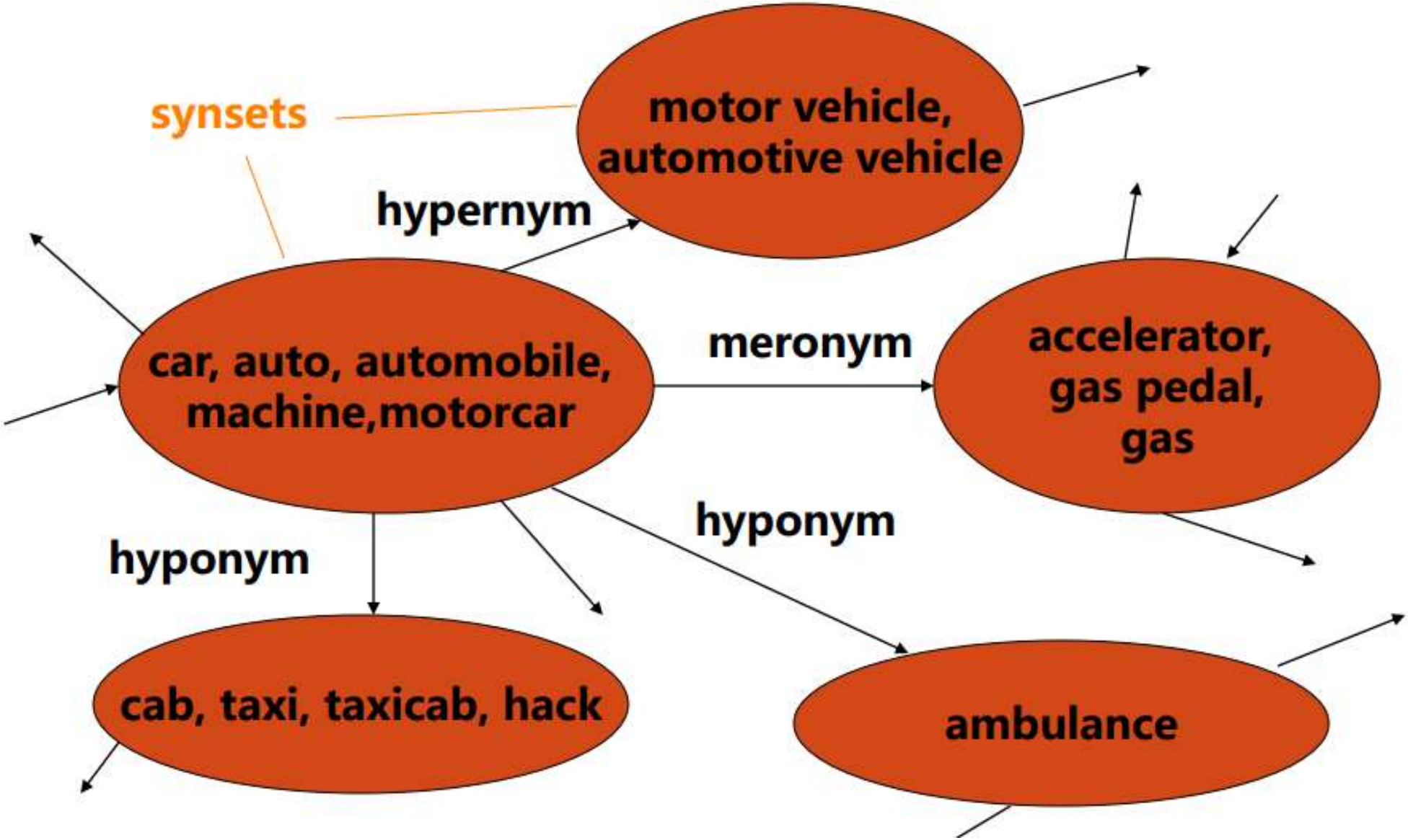
=> animal, animate being, beast, brute, creature, fauna

- At the top level, these hierarchies are organized into base types
  - 25 primitive groups for nouns, and 15 for verbs

Primitive groups for nouns

{act, action, activity}
{animal, fauna}
{artifact}
{attribute, property}
{body, corpus}
{cognition, knowledge}
{communication}
{event, happening}
{feeling, emotion}
{food}
{group, collection}
{location, place}
{motive}
{natural object}
{natural phenomenon}
{person, human being}
{plant, flora}
{possession}
{process}
{quantity, amount}
{relation}
{shape}
{state, condition}
{substance}
{time}

A WordNet Snapshot



WorNet: Semantic Relations - Verb

- Hypernym: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y
  - to perceive 感觉is an hypernym of to listen
- Troponym方式: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner
  - to lisp 发音 is a troponym of to talk
- Entailment蕴涵: the verb Y is entailed by X if by doing X you must be doing Y
  - to sleep is entailed by to snore
- Coordinate terms: those verbs sharing a common hypernym
  - to lisp and to yell (叫喊)

Primitive groups for Verbs

1) 身体动作动词（Verbs of Bodily Care and Functions）
2) 变化动词（Verbs of Change）
3) 通信动词（Verbs of Communication）
4) 竞争动词（Competition Verbs）
5) 消费动词（Consumption Verbs）
6) 接触动词（Contact Verbs）
7) 认知心理动词（Cognition Verbs）
8) 创造动词（Creation Verbs）
9) 运动动词（Motion Verbs）
10) 情感心理动词（Emotion or Psych Verbs）
11) 状态动词（Stative Verbs）
12) 感知动词（Perception Verbs）
13) 领属动词（Verbs of Possession）
14) 社会交互（Verbs of Social Interaction）
15) 气象动词（Weather Verbs）

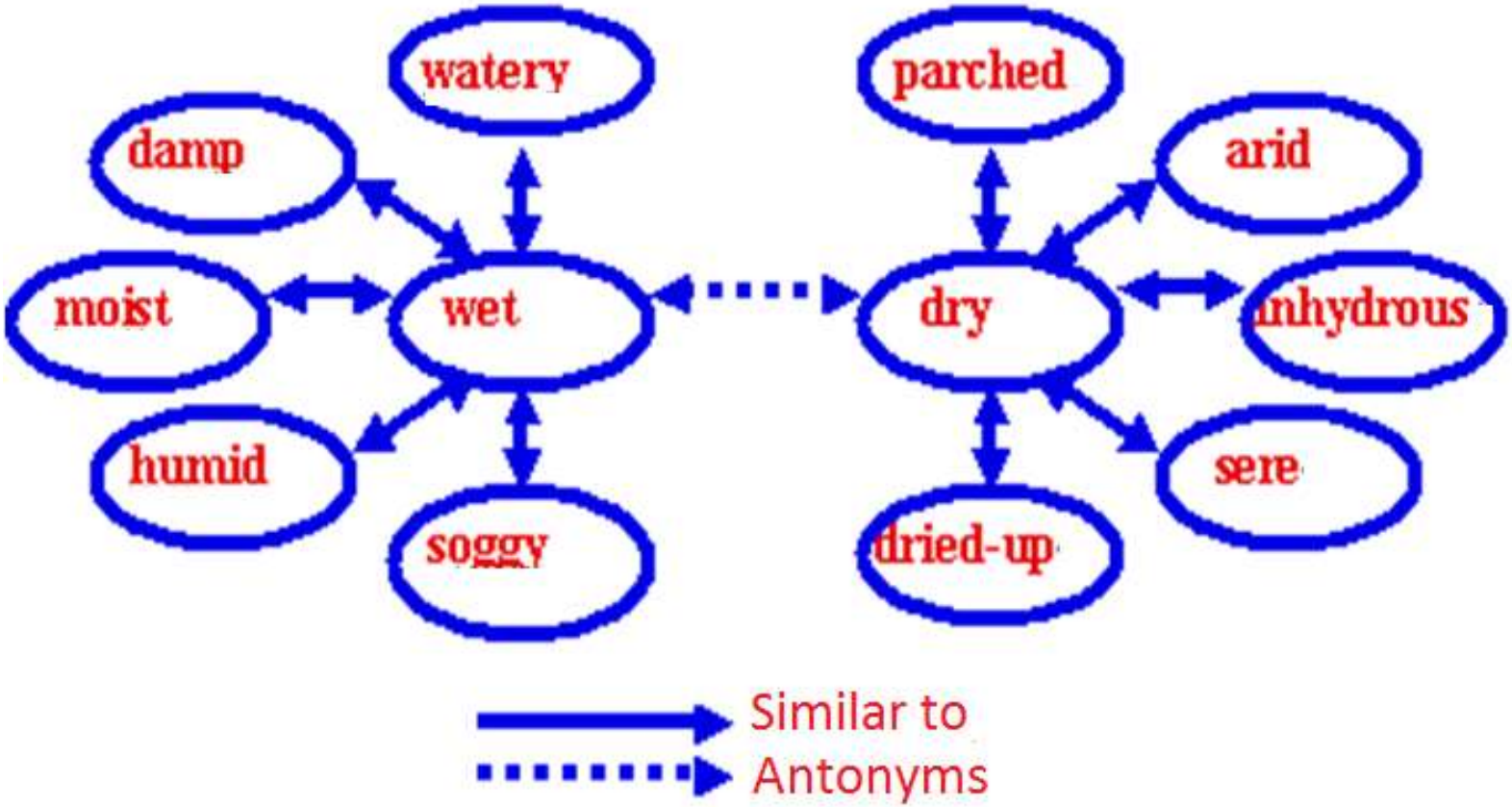
WorNet: Semantic Relations - Adjectives

- Described as N-D space
- Synset of adjective: Adjective Cluster
- Adjective clusters are connected through Antonyms relation

- Similar to relation 近似: ponderous/heavy
- Direct Antonyms: light/heavy
- Indirect Antonyms: light/ ponderous

Sense 2
ponderous -- (having great mass and weight and unwieldiness; '
burden"; "ponderous weapons")
=> heavy (vs. light) -- (of comparatively great physical we
is a heavy metal"; "heavy mahogony furniture")

WordNet: Adjective Cluster



WorNet as Thesaurs

- WordNet is organized based on Synset
- Synset is not the only expression for concept
- Synet is connected by semantic relations
  - Hypernyms
  - Hyponyms
  - Coordinate terms
  - Holonym
  - Meronym 部分
  - …

WorNet as Ontology

Ref

- Lexical Ontology
  - The hypernym/hyponym relationships among the noun synsets
  - Must be corrected before being used since hundreds of basic semantic inconsistencies
  - transforming WordNet into a lexical ontology
    - distinguishing the specialization relations into subtypeOf and instanceOf relations
    - associating intuitive unique identifiers to each category
- Part of the integration of WordNet 1.7 into the knowledge base of WebKB-2

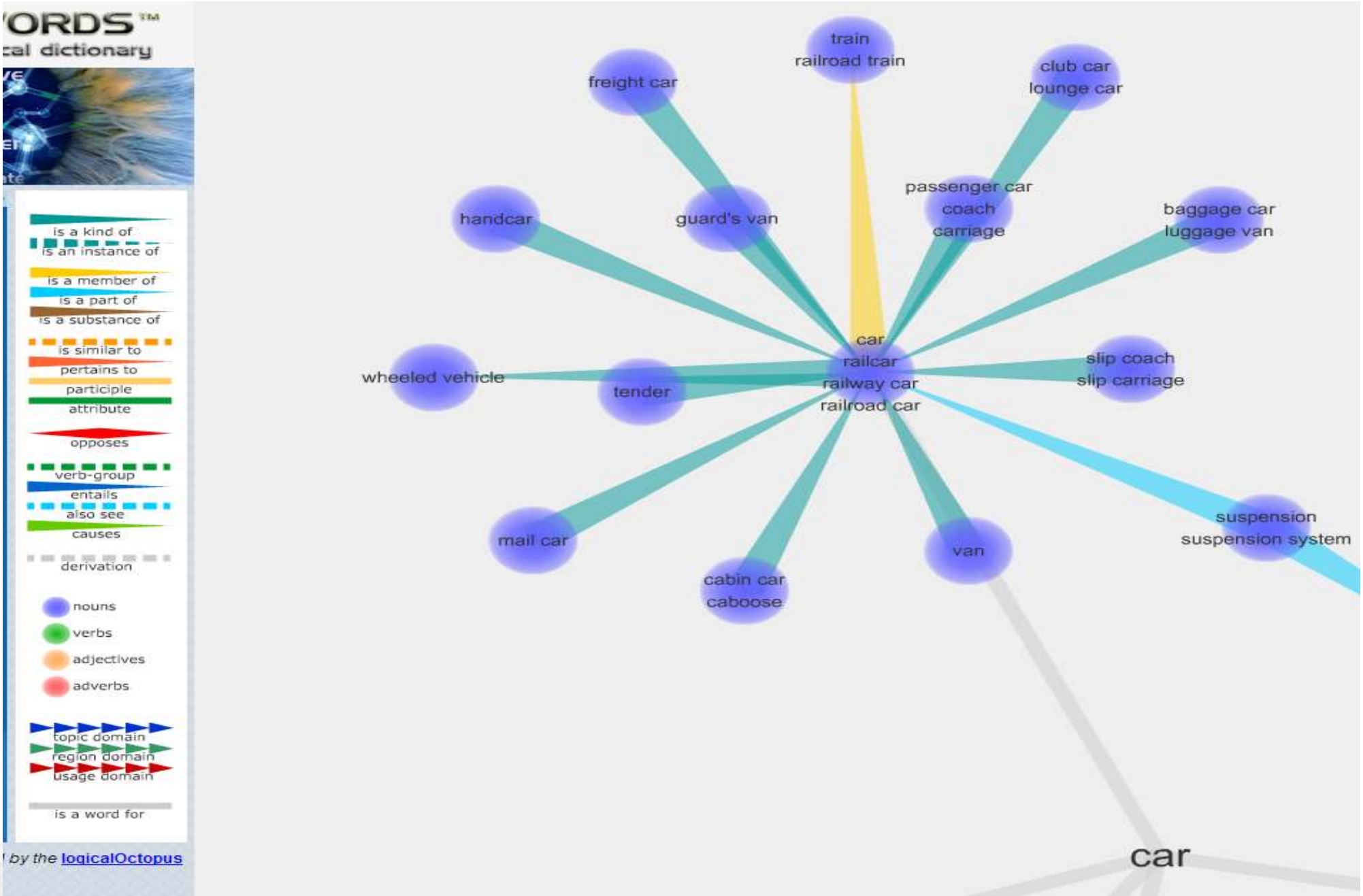
WordNet 3.0 Statistics

Ref

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117,798	82,115	146,312
Verb	11,529	13,767	25,047
Adjective	21,479	18,156	30,002
Adverb	4,481	3,621	5,580
Totals	155,287	117,659	206,941

Visualization: www.visuwords.com

Ref



Word Similarity Estimation based on WordNet

- Basic Idea: Two words(synset) which are more adjacent or have common hierarchy in WordNet, are more similar.

$$\text{sim}_{\text{path}}(c_1, c_2) = -\log \text{pathlen}(c_1, c_2)$$

$$\text{sim}_{\text{Resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2))$$

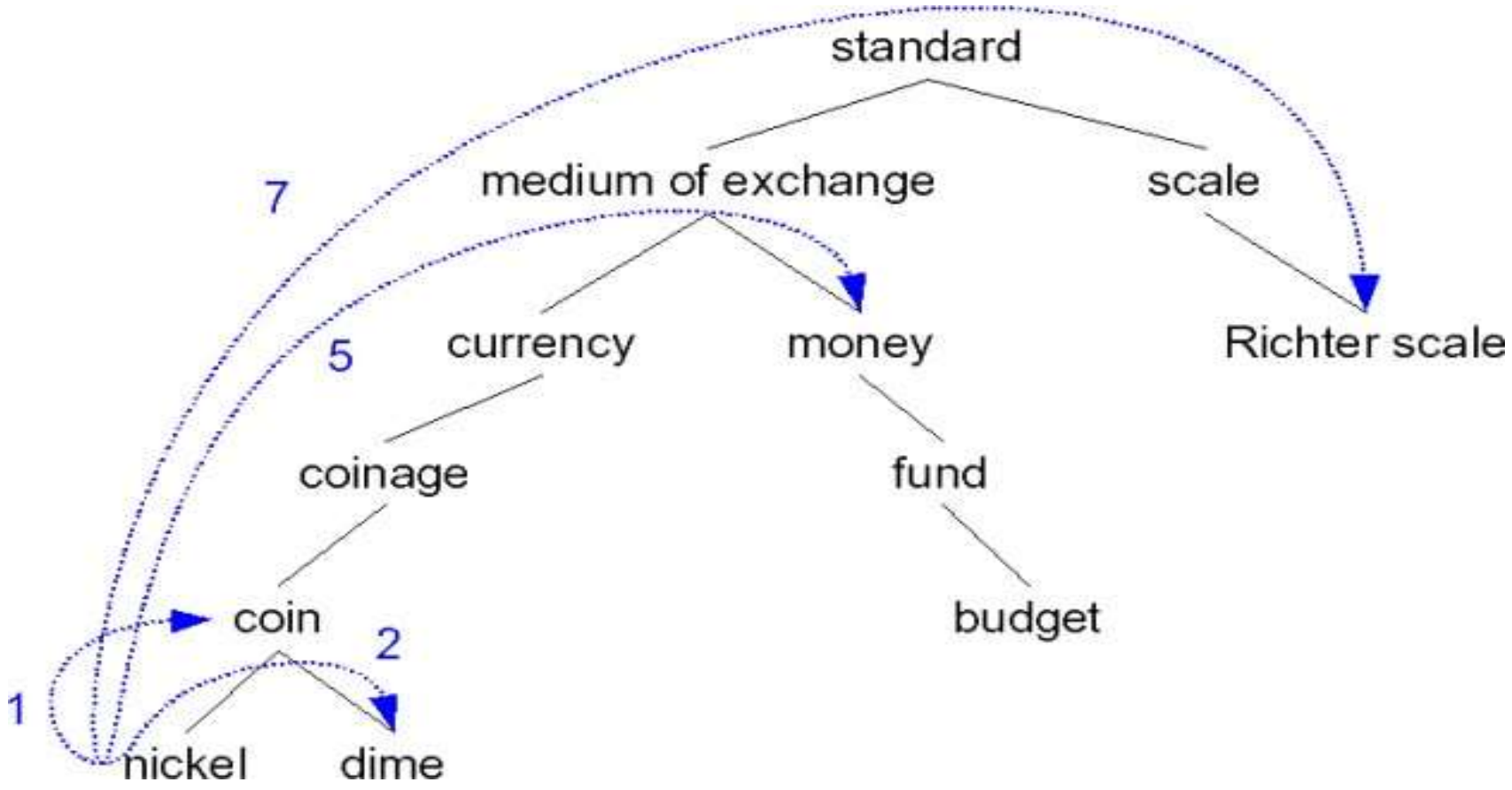
$$\text{sim}_{\text{Lin}}(c_1, c_2) = \frac{2 \times \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

$$\text{sim}_{\text{jc}}(c_1, c_2) = \frac{1}{2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))}$$

$$\text{sim}_{\text{eLesk}}(c_1, c_2) = \sum_{r,q \in \text{RELS}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$

Example 1: Path based similarity

- The shorter path in WordNet hierarchy, the greater similarity.



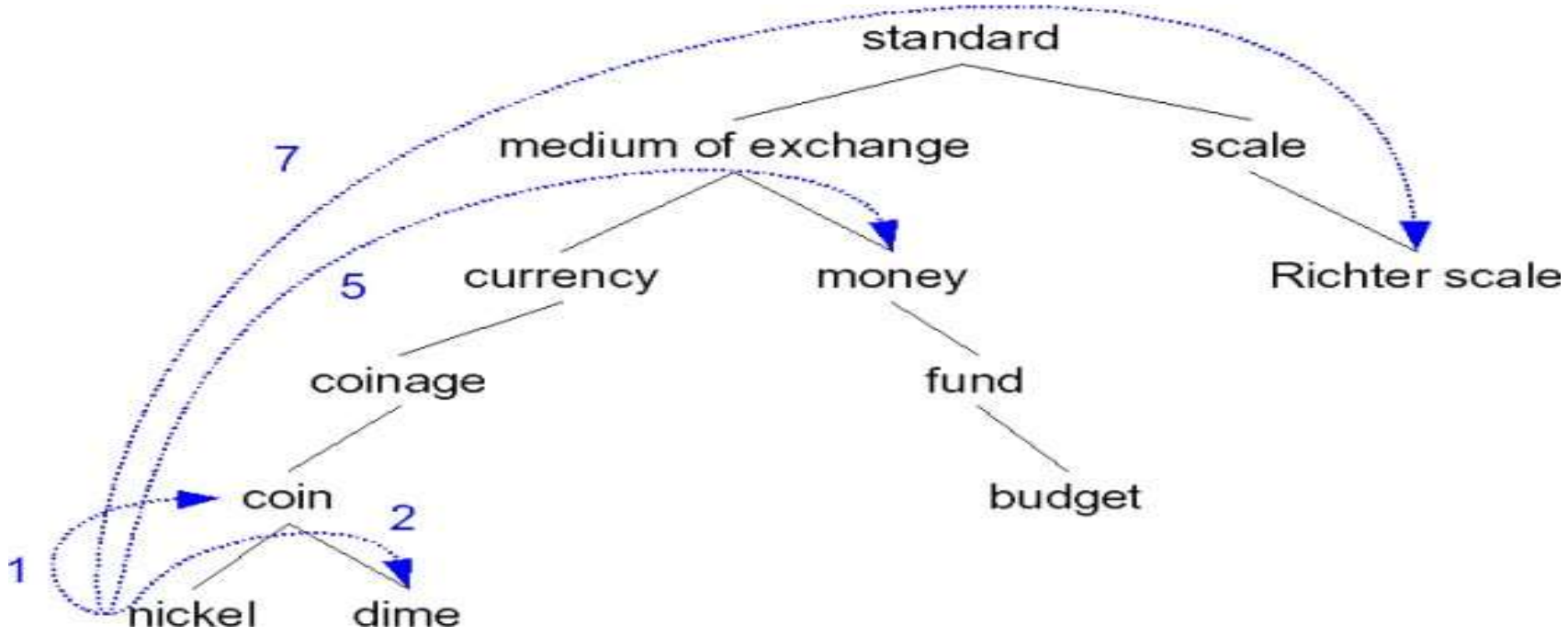
Example 1: Improvements of Path based similarity

- pathlen(c1,c2) = the number of edge in the the shortest path between c1 and c2
- simpath(c1,c2) = -log pathlen(c1,c2)
- wordsim(w1, w2) =

$$\max_{c1 \in senses(w1), c2 \in senses(w2)} sim(c1, c2)$$

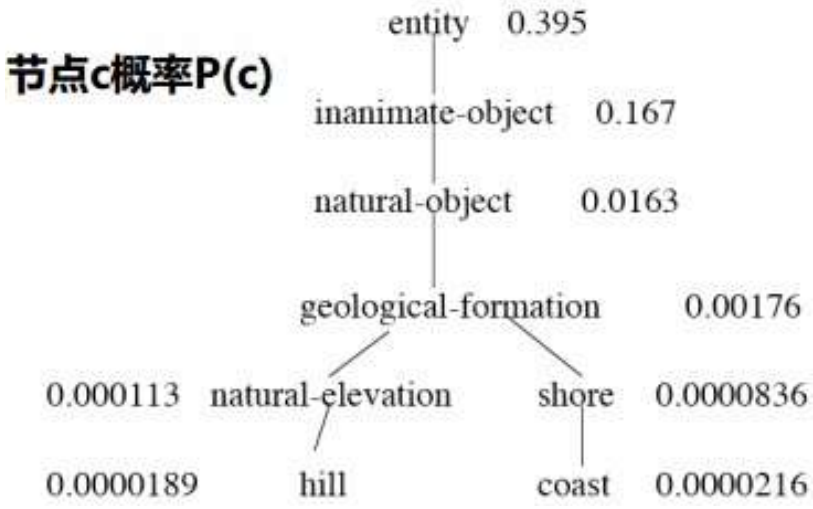
Example 1: Problem of Path based similarity

- Assuming each link(edge) represents the same distance
- Nickel to money and nickel to standard have the same similarity?



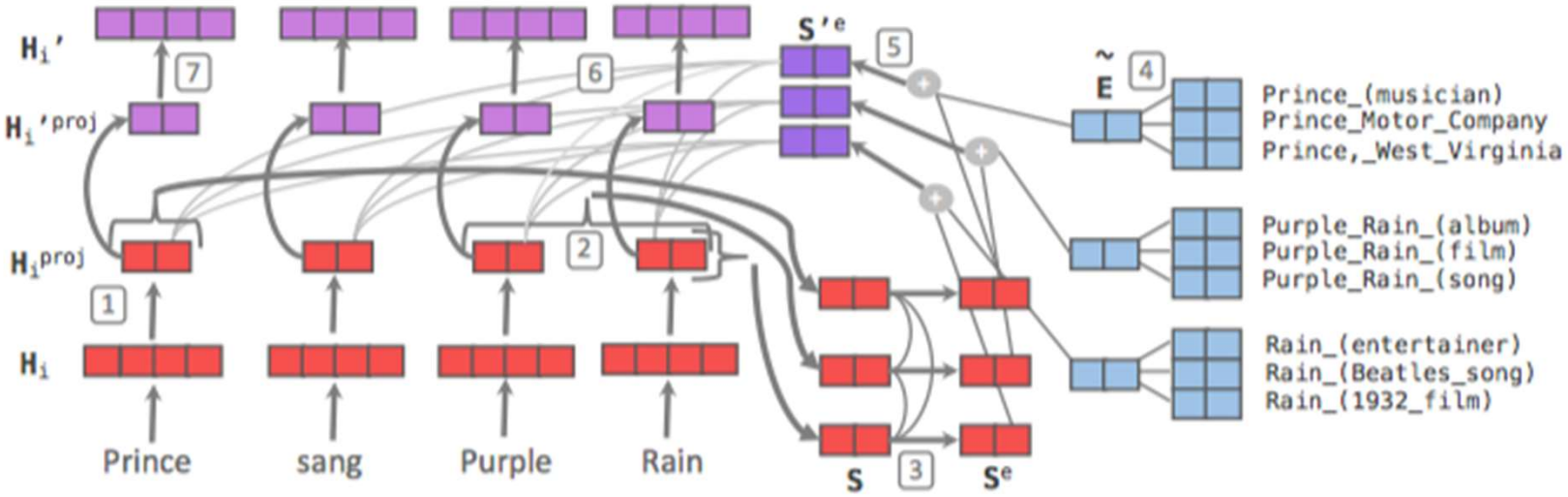
Example 2: Resnik method

- Use the information content of the lowest common ancestor node to measure the similarity.
- $sim_{resnik}(c1,c2) = -\log P(LCS(c1,c2))$
- The lower common ancestor node, the higher similarity.



Applications of WordNet

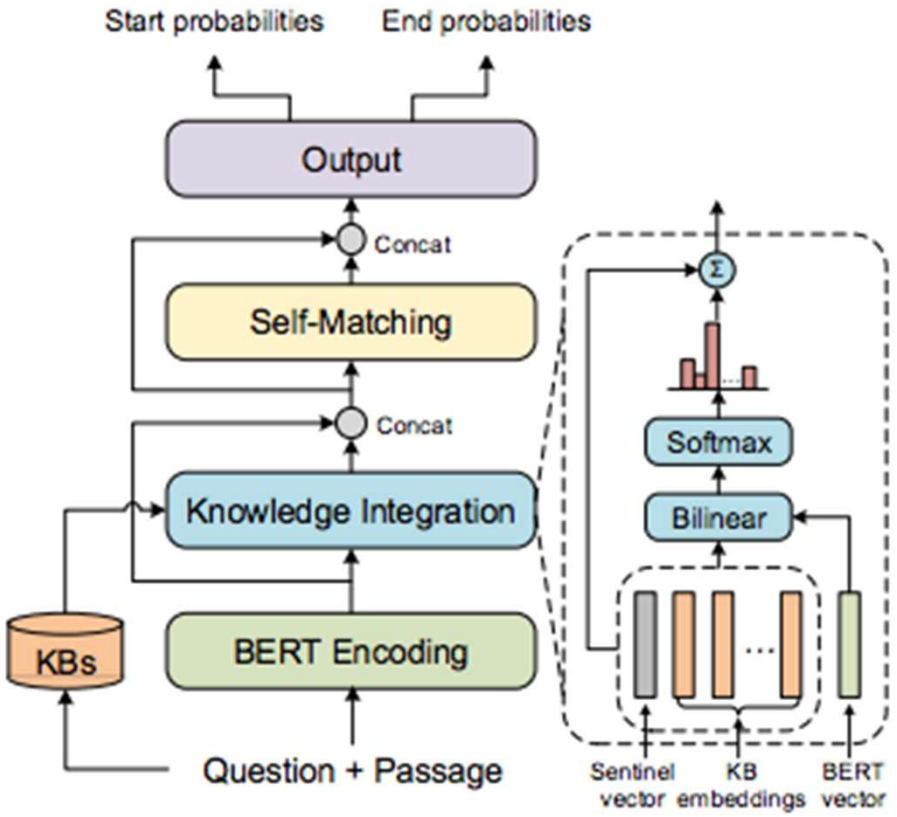
- Knowledge Enhanced Contextual Word Representations
  - [Peters et al. 2019] propose a general method to embed multiple knowledge bases (KBs) into large scale models, and thereby enhance their representations with structured, human-curated knowledge.



Peters, M.E., Neumann, M., Logan IV, R.L., Schwartz, R., Joshi, V., Singh, S. and Smith, N.A., 2019. Knowledge enhanced contextual word representations.

Applications of WordNet

- Machine reading comprehension
  - [Yang et al. 2019] investigate the potential of leveraging external knowledge bases (KBs) to further improve BERT for MRC.



Yang, A., Wang, Q., Liu, J., Liu, K., Lyu, Y., Wu, H., She, Q. and Li, S., 2019, July. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension.

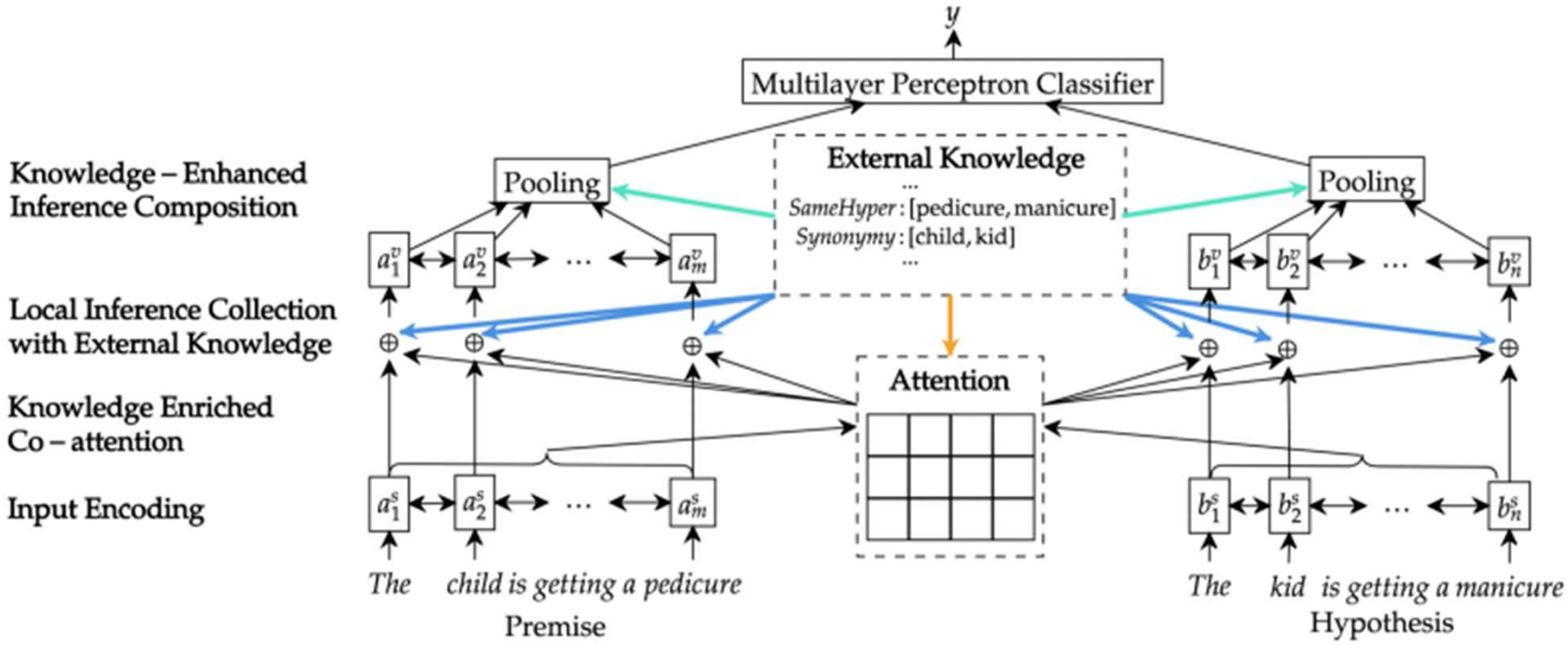
Applications of WordNet

- Data Augmentation
  - [Wei et al. 2019] present EDA: easy data augmentation techniques for boosting performance on text classification tasks.
    - Synonym Replacement (SR) → WordNet
    - Random Insertion (RI)
    - Random Swap (RS):
    - Random Deletion (RD)

Wei, J. and Zou, K., 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

Applications of WordNet

- Neural Natural Language Inference
  - [Chen et al. 2018] enrich neural natural language inference models with external knowledge.

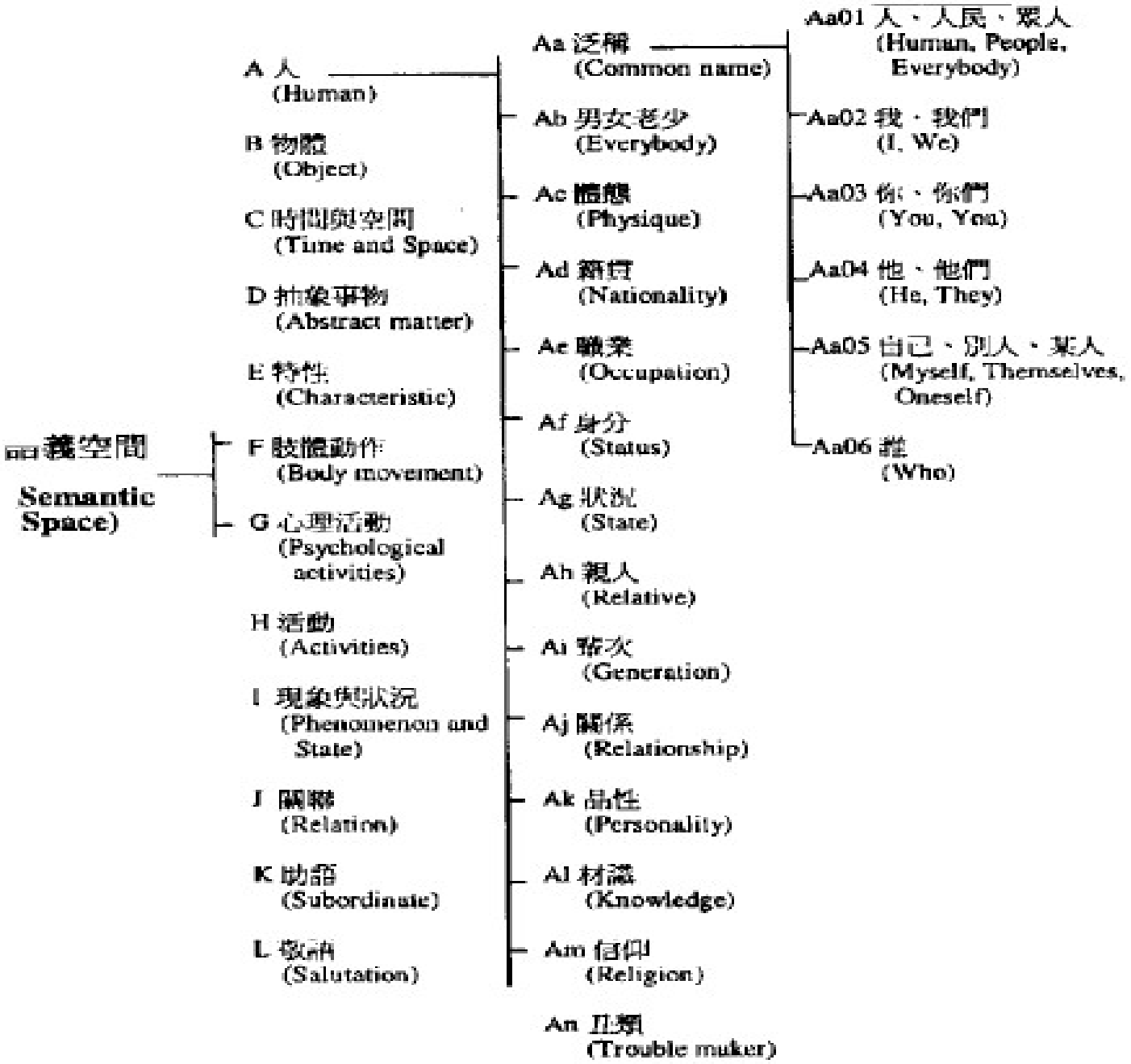


Chen, Q., Zhu, X., Ling, Z.H., Inkpen, D. and Wei, S., 2017. Neural natural language inference models enhanced with external knowledge.

TongYiCiCiLin- CILIN

- A contemporary Chinese thesaurus
- Developed for translation and writing purposes
- CILIN contains approximately 63,600 entries for about 52,500 Chinese words
- Classifies commonly used Chinese words in a three-level hierarchical tree structure

- This hierarchical structure reflects the semantic relationship between words
- Defined by 12 major (top level), 95 medium (middle level), and 1428 minor (bottom level) semantic clusters
- Each minor semantic cluster consists of a set of words.
- Words under the same minor semantic cluster share the concept of this class


義空間
Semantic Space)
A 人 (Human)
B 物體 (Object)
C 時間與空間 (Time and Space)
D 抽象事物 (Abstract matter)
E 特性 (Characteristic)
F 肢體動作 (Body movement)
G 心理活動 (Psychological activities)
H 活動 (Activities)
I 現象與狀況 (Phenomenon and State)
J 關聯 (Relation)
K 助語 (Subordinate)
L 敬詞 (Salutation)
Aa 泛稱 (Common name)
Ab 男女老少 (Everybody)
Ac 體態 (Physique)
Ad 籍貫 (Nationality)
Ae 職業 (Occupation)
Af 身分 (Status)
Ag 狀況 (State)
Ah 親人 (Relative)
Ai 輩次 (Generation)
Aj 關係 (Relationship)
Ak 品性 (Personality)
Al 材識 (Knowledge)
Am 信仰 (Religion)
An 卫類 (Trouble muker)
Aa01 人、人民、眾人 (Human, People, Everybody)
Aa02 我、我們 (I, We)
Aa03 你、你們 (You, You)
Aa04 他、他們 (He, They)
Aa05 自己、別人、某人 (Myself, Themselves, Oneself)
Aa06 誰 (Who)

- CILIN is a valuable knowledge source for Chinese NLP
- Lua use the conventional simple co-occurrence approach for extracting the semantic knowledge from the CILIN
- Wong proposed an approach toextract partial semantic information from the CILIN

- New version of CILIN

HowNet 知网

- Developed by Dong Zhendong and Dong Qiang
  - http://www.keenage.com/
- HowNet is a extra-linguistic common-sense knowledge system for the computation of meaning in human language technology
- Unveils inter-concept relations and inter-attribute relations of concepts co-annoted in its Chinese-English lexicon

HowNet : Features

- The concept definition (DEF) in HowNet is based on sememes
  - Sememes are in a structured markup language.
    - 监狱 (prison)
    - DEF={InstitutePlace|场所:domain={police|警}, {detain|扣住:location={~}, patient={human|人:modifier={guilty|有罪}}}, {punish|处罚:location={~}, patient={human|人:modifier={guity|有罪}}},
    - “A prison is an institutional place, which is the location where guilty people are detained and punished. The place belongs to the domain of police and law”.

- HowNet constructs a graph structure of its knowledge base on the inter-concept relations and inter-attribute relations
  - Different from tree-structure lexical databases
  - Reveals concept relations within the same parts-of-speech (POS) categories and across POS categories
  - Especially the semantic-role relations between nouns and verbs.
- The representation is based on concepts denoted by words and expressions in both Chinese and English

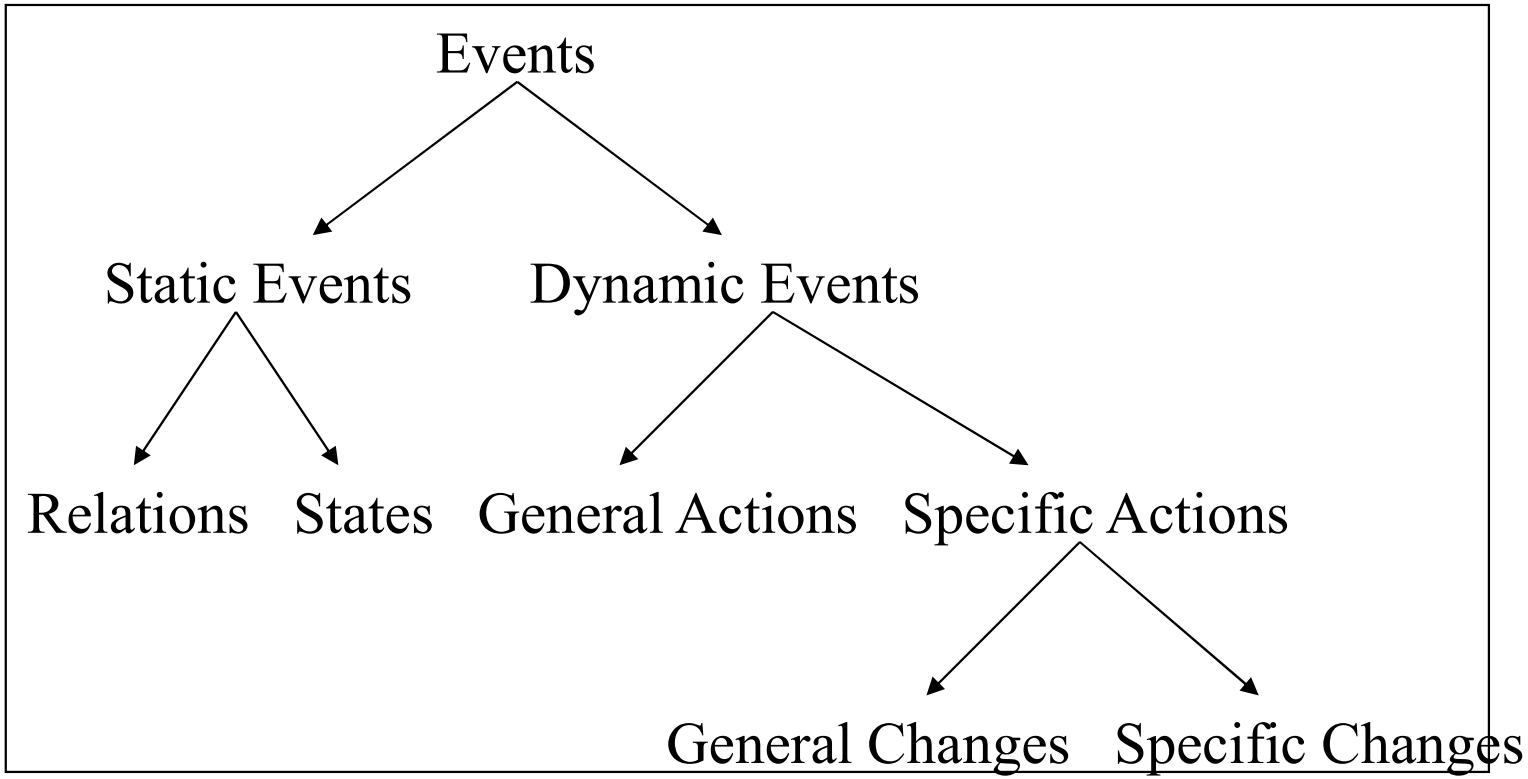
HowNet: Sememes

- Hypothesizes that all concepts can be reduced to the relevant sememes
- Defines a close set of sememes from which an open set of concepts can be composed
- Sememes in HowNet are obtained by examining about 4,000 frequently used characters and their senses.
- This large-scale linguistic engineering process takes about three years of annotation and modification.
- 2,088 sememes in HowNet

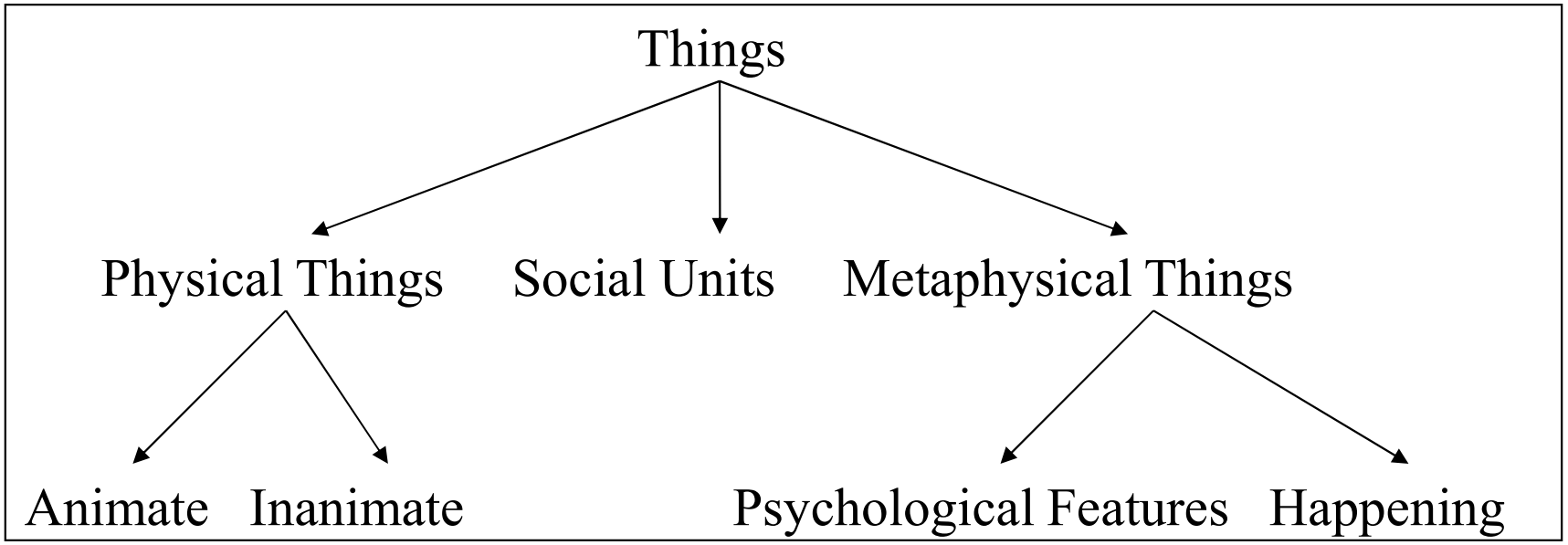
HowNet: Taxonomy

- Sememes are classified in a hierarchical structure, called taxonomy
- Taxonomy mainly provides the hypernym-hyponym relations of concepts
- 4 taxonomies in HowNet organizes its
  - taxonomy of events: 805 event sememes
  - taxonomy of entity, including things, parts, time and space: 152 entity sememes
  - taxonomy of attributes: 245 attribute sememes
  - taxonomy of attribute-values: 886 attribute-value sememes.

HowNet: Taxonomy - Event



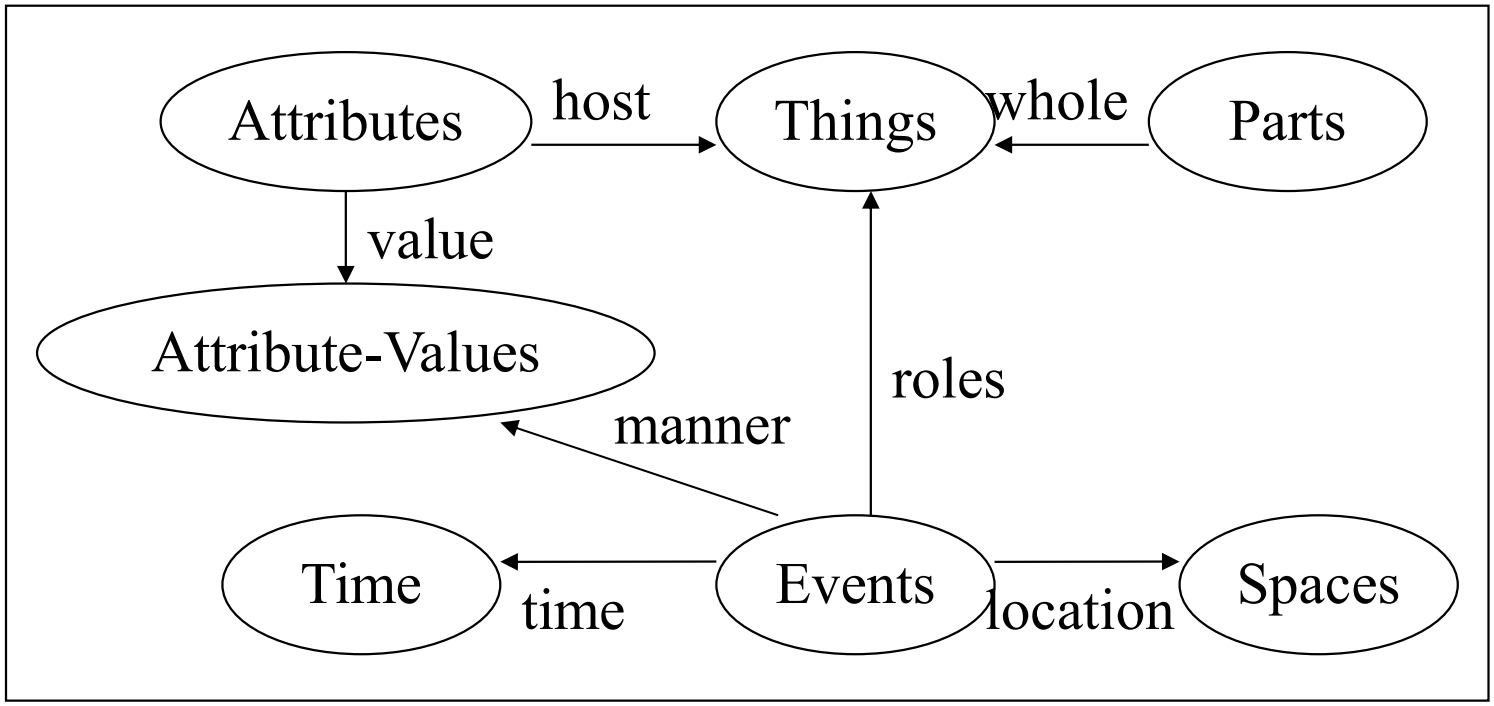
HowNet: Taxonomy - Thing



HowNet: Relations

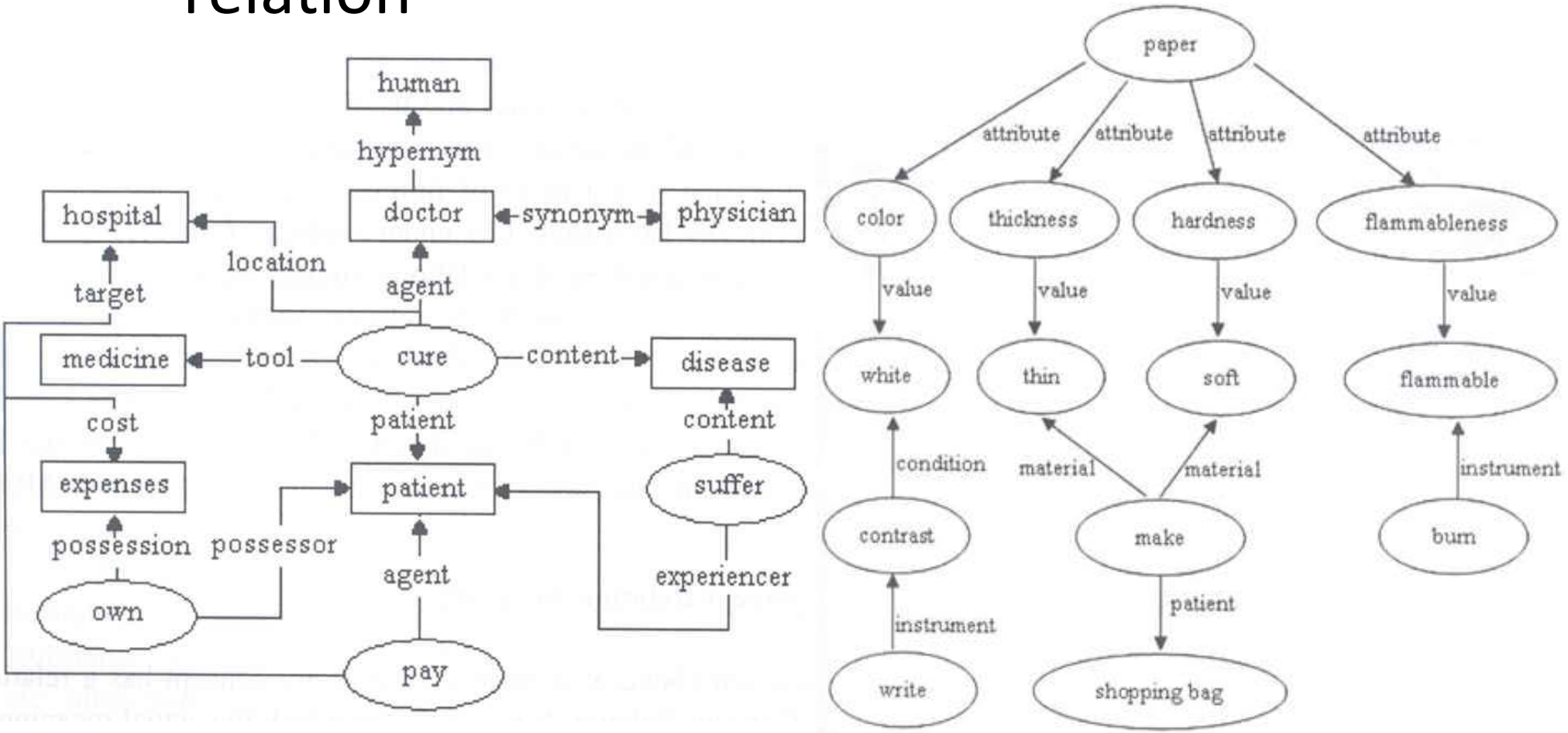
- Relations are the soul of HowNet.
- Classified into explicit relations and implicate relations
  - 11 types of explicit relation in HowNet.
  - (1) synonym (), (2) synclass, (3) antonym, (4) converse, (5) hypernym, (6) hyponym, (7) part-whole, (8) value-attribute, (9) attribute-host, (10) cognate role-frame, and (11) semantic roles-event.
  - the last four types are cross POS relations.

HowNet: Relation of Top-level Sememes



HowNet: Relations

- Divided into concept relation and attribute relation



HowNet: Relations

(a) 上下位关系（由概念的主要特征体现，请参看《知网管理工具》）

(b) 同义关系（可通过《同义、反义以及对义组的形成》获得）

(c) 反义关系（可通过《同义、反义以及对义组的形成》获得）

(d) 对义关系（可通过《同义、反义以及对义组的形成》获得）

(e) 部件-整体关系（由在整体前标注 % 体现，如"心"，"CPU"等）

(f) 属性-宿主关系（由在宿主前标注 & 体现，如"颜色"，"速度"等）

(g) 材料-成品关系（由在成品前标注 ？体现，如"布"，"面粉"等）

(h) 施事/经验者/关系主体-事件关系（由在事件前标注 * 体现，如"医生"，"雇主"等）

(i) 受事/内容/领属物等-事件关系（由在事件前标注 $ 体现，如"患者"，"雇员"等）

(j) 工具-事件关系（由在事件前标注 * 体现，如"手表"，"计算机"等）

(k) 场所-事件关系（由在事件前标注 @ 体现，如"银行"，"医院"等）

(l) 时间-事件关系（由在事件前标注 @ 体现，如"假日"，"孕期"等）

(m) 值-属性关系（直接标注无须借助标识符，如"蓝"，"慢"等）

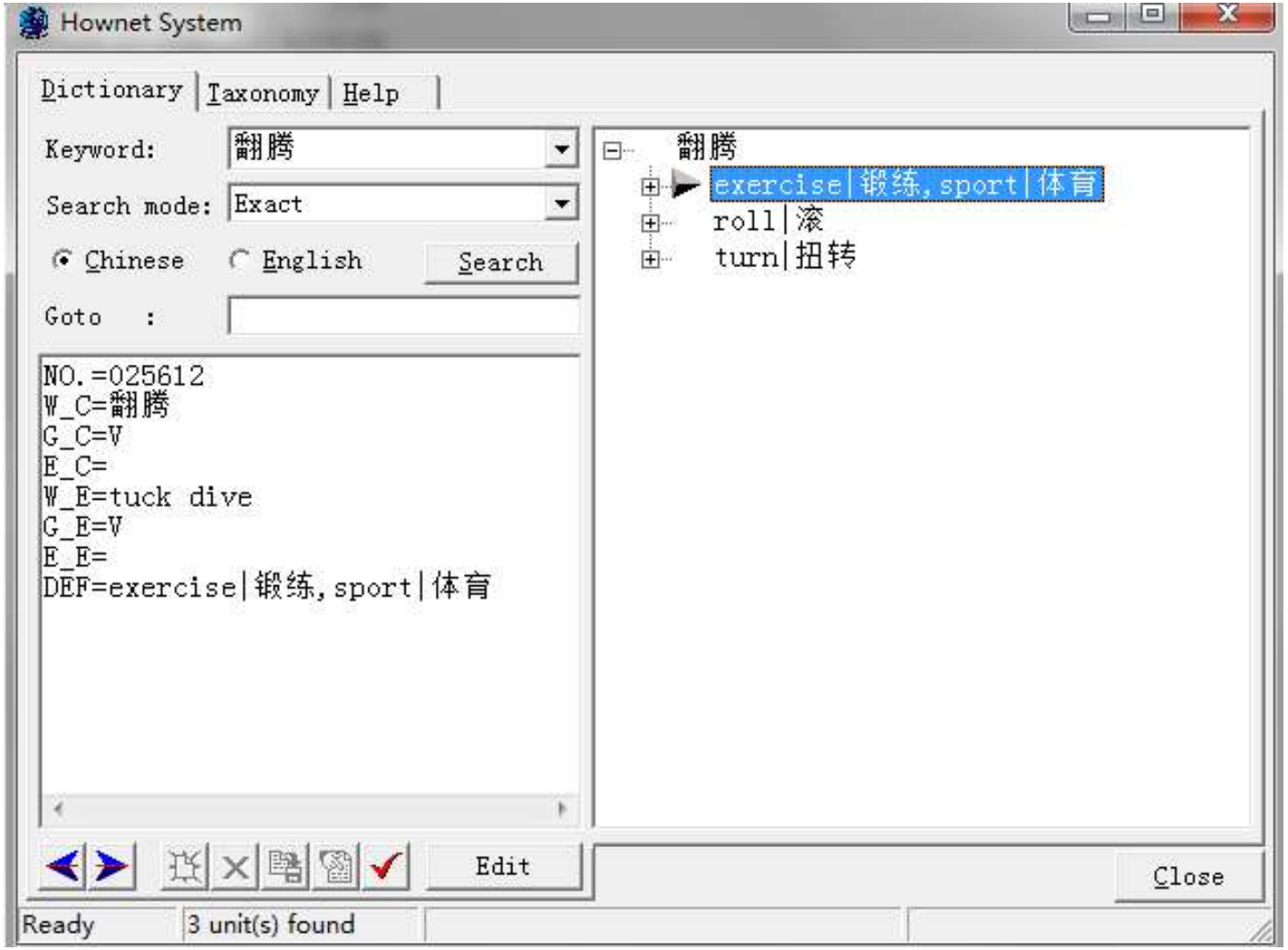
(n) 实体-值关系（直接标注无须借助标识符，如"矮子"，"傻瓜"等）

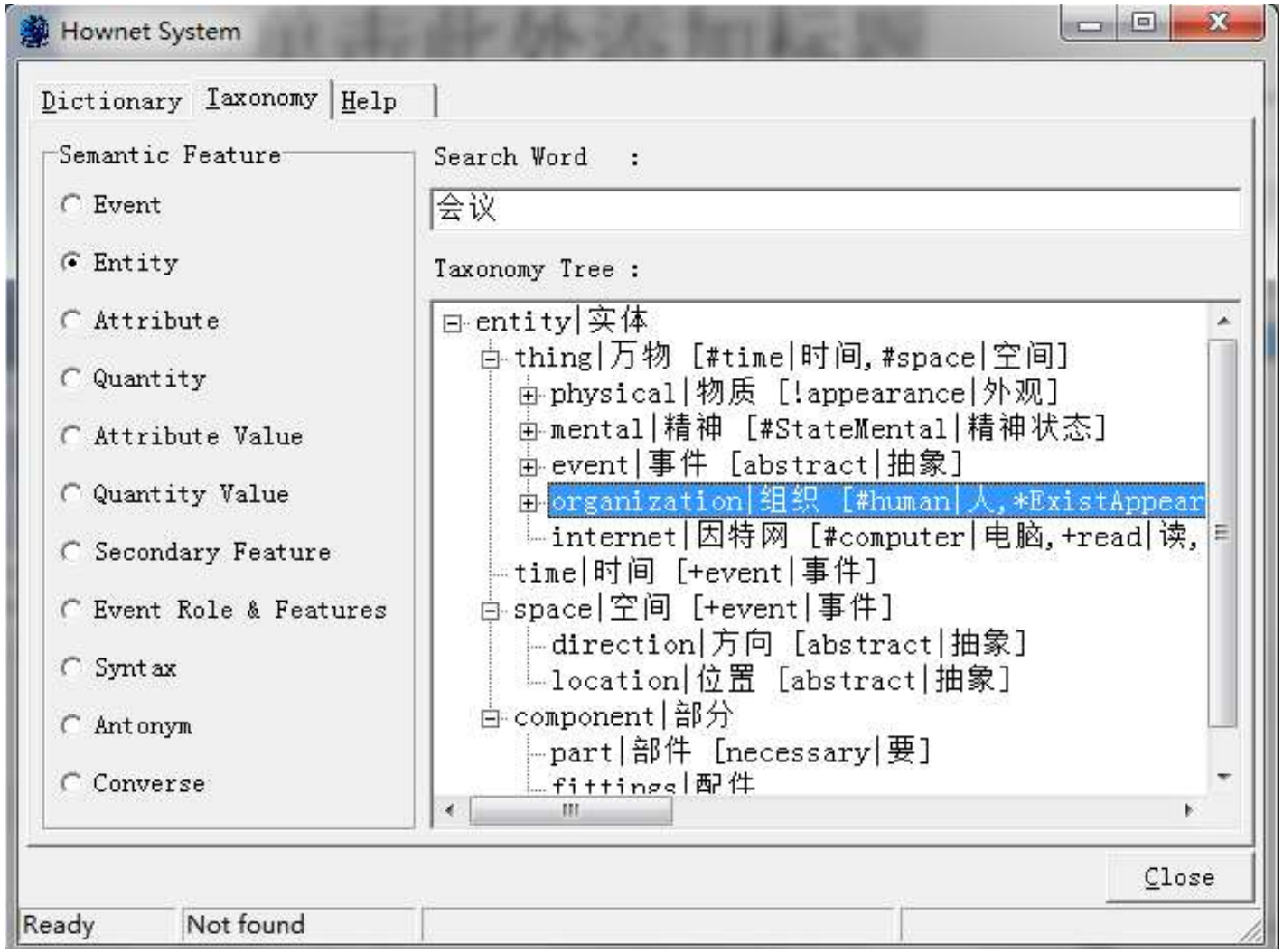
(o) 事件-角色关系（由加角色名体现，如"购物"，"盗墓"等）

(p) 相关关系（由在相关概念前标注 # 体现，如"谷物"，"煤田"等）

HowNet: Interface

Ref




Hownet System
Dictionary
Taxonomy
Help
Semantic Feature
Event
Entity
Attribute
Quantity
Attribute Value
Quantity Value
Secondary Feature
Event Role & Features
Syntax
Antonym
Converse
Search Word :
会议
Taxonomy Tree :
entity|实体
thing|万物 [#time|时间,#space|空间]
physical|物质 [!appearance|外观]
mental|精神 [#StateMental|精神状态]
event|事件 [abstract|抽象]
organization|组织 [#human|人,*ExistAppear
internet|因特网 [#computer|电脑,+read|读,
time|时间 [+event|事件]
space|空间 [+event|事件]
direction|方向 [abstract|抽象]
location|位置 [abstract|抽象]
component|部分
part|部件 [necessary|要]
Close
Ready
Not found

HowNet: Word Similarity Estimation

- For two Chinese words W_1 and W_2 in HowNet, if W_1 has concept: S_{11}, S_{12}, ... , S_{1n}, W_2 has concept: S_{21}, S_{22}, ... , S_{2m},

  the similarity between W_1 and W_2 is equals to the maximum similarity between two concepts

$$Sim(W_1, W_2) = \max_{i=1..n, j=1...m} Sim(S_{1i}, S_{2j})$$

HowNet: Word Similarity Computing

•Different from traditional semantic dictionary, each concept can **not** project to a concept tree node in HowNet, it is described by a series of primitive (义原).

similarity	function word concept (虚词概念)	notional word concept (实词概念)
function word concept	similarity of corresponding syntactic primitive(句法义原) or relation primitive(关系义原)	0
notional word concept	0	sum of 4 parts: 1. 第一独立义原描述式 $Sim_1(S_1,S_2)$ 2. 其他独立义原描述式 $Sim_2(S_1,S_2)$ 3. 关系义原描述式 $Sim_3(S_1,S_2)$ 4. 符号义原描述式 $Sim_4(S_1,S_2)$

HowNet: Word Similarity Computing

- The semantic distance between tow primitive is defined as:

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

  - d is the length of the path between p1 and p2 in the primitive hierarchy of HowNet.
  - α is an adjustable parameter.
- $Sim_1(S_1, S_2)$ can be calculated according to the above formula, the other three are different.

Applications of HowNet

- Word Representation Learning
  - [Niu et al. 2017] present that, word sememe information can improve word representation learning (WRL), which maps words into a low-dimensional semantic space and serves as a fundamental step for many NLP tasks.

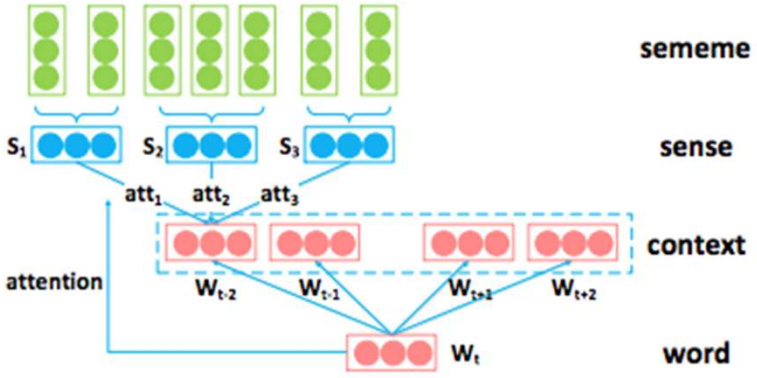


Figure 2: Sememe Attention over Context Model.

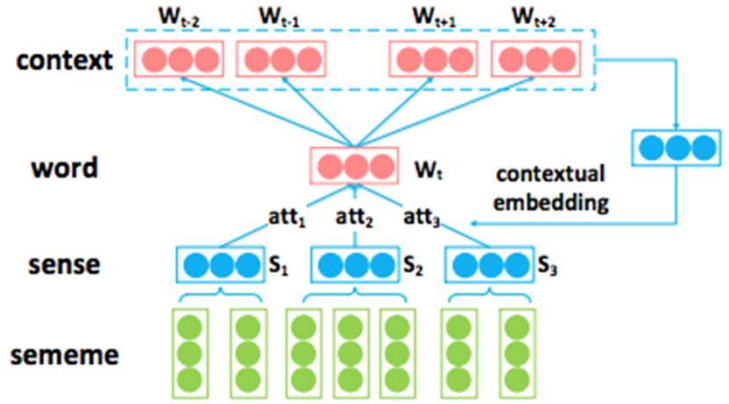


Figure 3: Sememe Attention over Target Model.

Niu, Y., Xie, R., Liu, Z. and Sun, M., 2017, July. Improved word representation learning with sememes.

Applications of HowNet

- Event Detection
  - [Ding et al. 2019] propose the Trigger-aware Lattice Neural Network (TLNN).
  - For polysemous characters and words, this method model all senses of them with the help of HowNet, so as to alleviate the problem of ambiguous triggers.

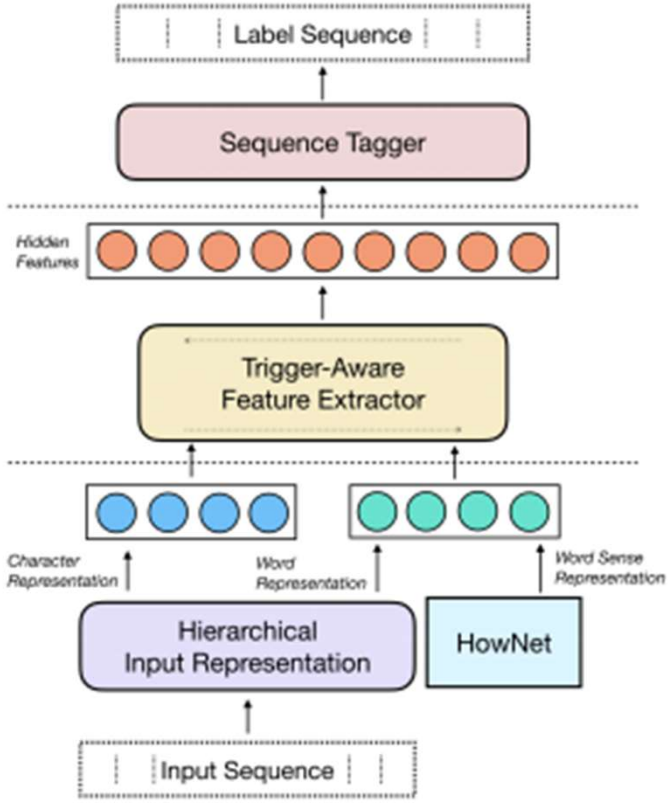
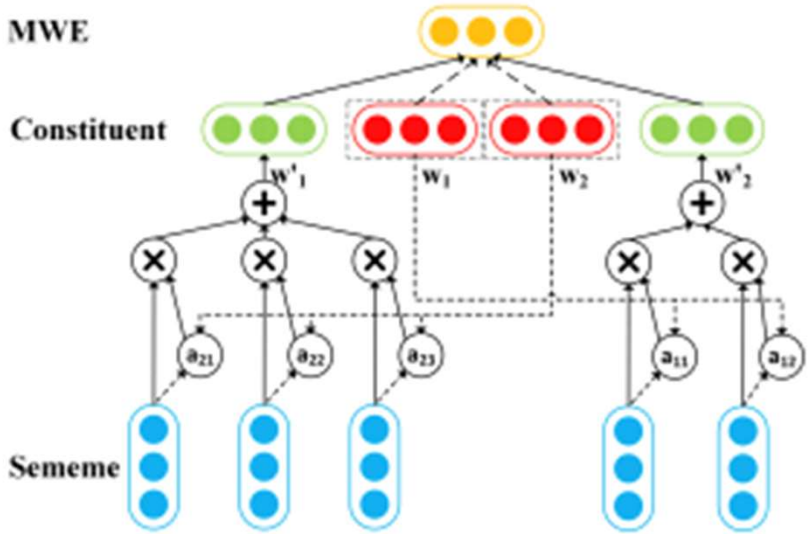


Figure 2: The architecture of TLNN.

Ding, N., Li, Z., Liu, Z., Zheng, H. and Lin, Z., 2019, November. Event detection with trigger-aware lattice neural network.

Applications of HowNet

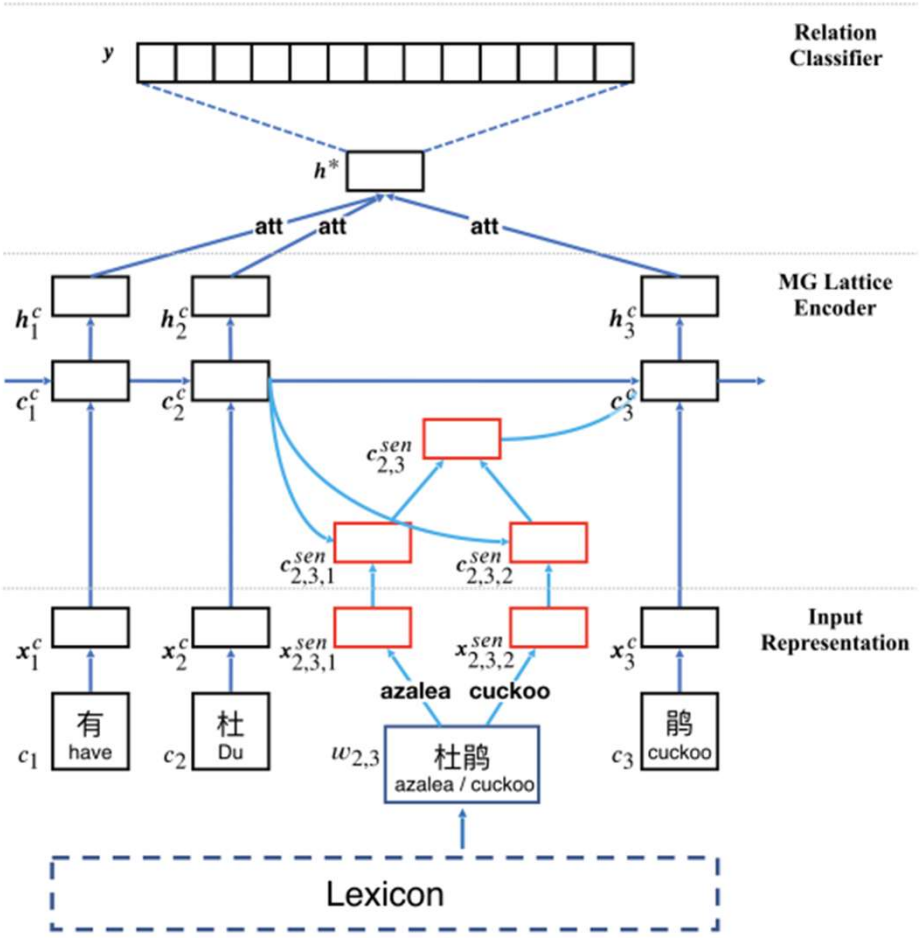
- Semantic compositionality
  - [Qi et al. 2019] verify the effectiveness of sememes, the minimum semantic units of human languages, in modeling SC by a confirmatory experiment.



Qi, F., Huang, J., Yang, C., Liu, Z., Chen, X., Liu, Q. and Sun, M., 2019. Modeling semantic compositionality with sememe knowledge.

Applications of HowNet

- Relation Extraction
  - [Li et al. 2019] propose a multi-grained lattice framework (MG lattice) for Chinese relation extraction to take advantage of multi-grained language information and external linguistic knowledge.



Li, Z., Ding, N., Liu, Z., Zheng, H. and Shen, Y., 2019, July. Chinese relation extraction with multi-grained information and external linguistic knowledge.

CCD: Chinese Concept Dictionary 中文概念词典

- Developed in ICT, Peking University
- A bilingual Chinese-English WordNet-like Lexicon follow the framework of WordNet
- Compatible with WordNet structurally
- CCD records
  - synonyms, antonyms
  - hypernyms-hyponyms
  - meronyms-holonyms
  - collocations of wordsas
  - major POS
  - nouns, verbs, adjectives and adverbs

CCD: Characteristics

- Nouns in CCD include
  - words of time, e.g. 下午 (afternoon),
  - location, e.g. 西部 (Western)
  - direction-position, e.g. 上 (up)
  - numerals, e.g. 甲 (first)
  - quantifiers, e.g. 批 (lots),
  - Pronouns
  - differentiation, e.g. 男/女 (male/female)
  - suffixes, e.g.性 (gender), 器 (apparatus)
  - idioms, e.g. 八拜之交, 铜墙铁壁
  - semi-idioms (木头疙瘩、光杆司令)
  - abbreviated forms (政协)

- The intension and extension of verbs, adjectives and adverbs in CCD are different from those in WordNet
- Semantically, concepts from different POSs may be related
  - 战争/n(battle) and打仗 /v(fighting in a war) 战斗/v (fight) are associated
- CCD alleviates this restriction in WordNet.
- An attribute of association is appended to the feature structure of a concept.

- Relations described in CCD are much finer than WordNet
  - Synonyms relation is divided into equivalence and similarity
  - Antonyms relation is divided into essential antonyms and nonessential ones
  - Three new types of meronyms-holonyms relations are introduced
- Examples of collocations are selected from real corpora with quantitative descriptions.

Nouns in CCD

- Synonymy
  - CCD records the word formation rule of the nouns if applicable.
    - attribute+part+noun->noun (长 *long*+鹿*deer* + 颈 *neck* is a part of 鹿*deer*->长颈鹿*giraffe*)
  - CCD records the syntactic relations between the morphemes forming a noun.
    - 美人*beautiful girl* has 美*beautiful* and 人*person*.
  - CCD records the refined noun categorization information
    - Abbreviations 北大 vs. 北京大学 Peking University
    - Regular and irregular words 教师 vs. 孩子王
    - Sentiment polarity 教师 neutral vs. 臭老九negative

- Antonymy
  - Divided into essential / nonessential
  - Word *A* and word *B* is named essential antonym if under one specific condition, *A* is the only antonym choice.
    - 丈夫 *husband*-妻子*wife* is essential antonym in the relationship of husband-wife.
  - Word *A* and word *B* are named nonessential antonyms if under one specific condition, *A* is not the only antonym choice.
    - under the relationship of persons with ages, 老人 *elder* and 孩子*children*, 老人*elder* and年轻人*young man*

- Hyponymy
  - Hyponymy is transitive and asymmetrical

    树 tree--

    =>木本植物 woody plant--

    =>维管植物 vascular plan

    =>植物，植物生命体 plant

    =>生命形式，生物体，生物 life form

    =>实体 entity
  - Less than 10 levels

- Meronymy
  - a component part of individual *w'*. 车轮—车
  - a member of set (composed by individual *w'*). 树—森林
  - the stuff that *w'* is made from. 钢—钢板
  - a subarea of *w'*. 石家庄—河北
  - event *w* is a part of event *w'*. 开幕式—会议
  - time *w* is a part of time *w'*. 唐朝—古代
- Attribute

Verbs in CCD

- Synonymy
  - equivalent synonym/near synonyms
    死,逝世,去世,升天
  - Provide abbreviation information
    死 亡/ 死亡
  - Provide regular/irregular use
    死 去世 逝世/老 归西 上西天
  - Provide sentiment polarity of verbs
    死 亡 死亡
    去世 逝世 过世 老 仙逝 Positive
    完蛋 一命呜呼 见阎王 Negative

- Antonymy
- Hypernymy
- Entailment
  - The temporal interval (or point) of verb $\bar{\bar{U}}$ embeds in that of $\acute{U}$. 打鼾(*snore)* entails 睡觉 (*sleep*)
  - The temporal interval (or point) of verb $\bar{\bar{U}}$ is equal to that of $\acute{U}$. 跛行(*limp)* entails行走
  - The temporal interval of verb $\bar{\bar{U}}$ has nothing to do with that of $\acute{U}$. But inferring back from $\bar{\bar{U}}$, 成功 (*succeed)* entails尝试 (*try)*
  - The temporal interval of verb $\bar{\bar{U}}$ has nothing to do with that of $\acute{U}$. But inferring back from $\bar{\bar{U}}$, 给 (*give)* entails拥有(*own)*

- Sentence Frame
  - semantic roles of agent and patient

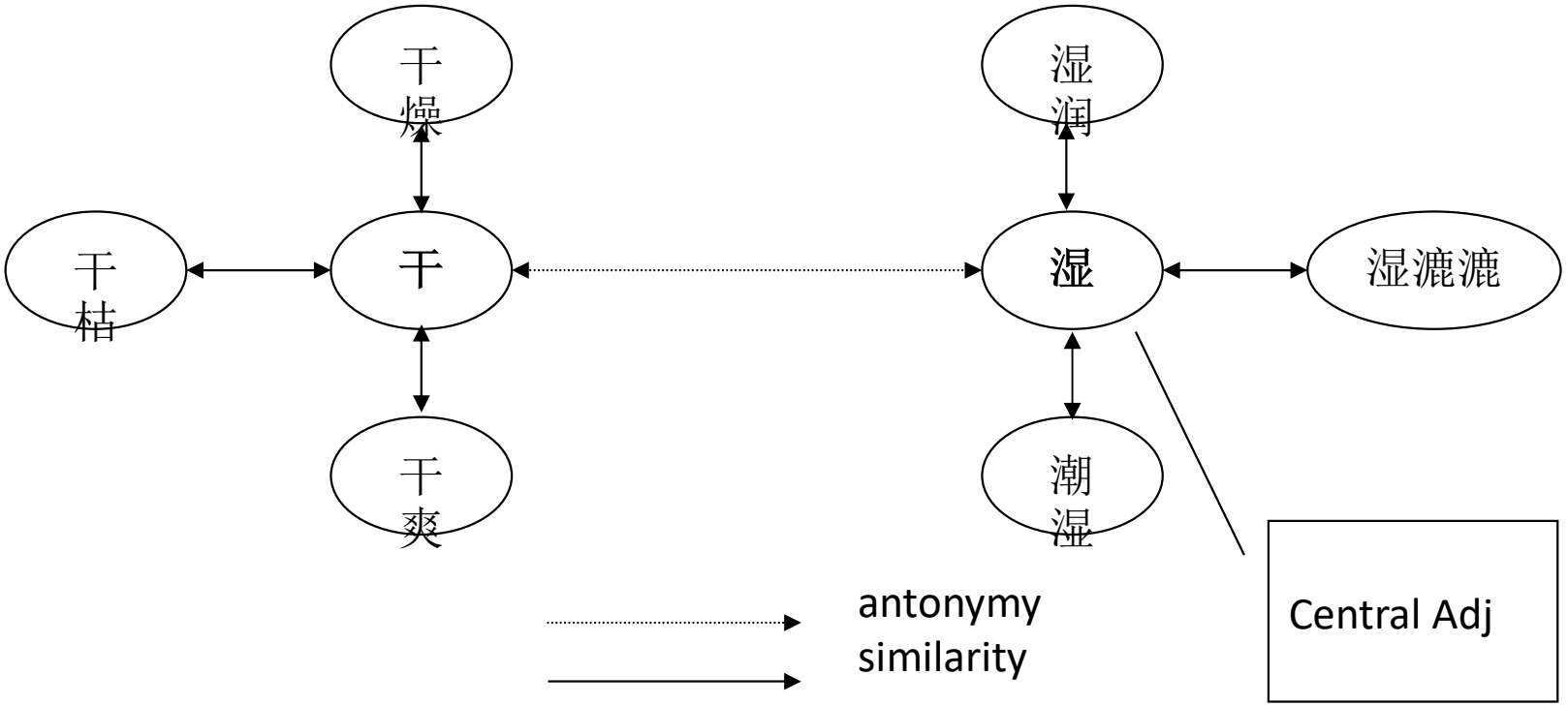
AGN Agent 施事	COS Course 经事
ESS Essive 当事	GOL Goal 向事
EXP Experience 感事	SCP Scope 范围
GEN Genitive 领事	RSN Reason 缘由
PAT Patient 受事	INT Intention 意图
IMP Impellee 致事	TIM Time 时间
RST Result 结果	SPA Space 空间
CON Content 内容	MAN Manner 方式
BRL Belongings 属事	INS Instrument 工具
PAR Part 分事	MAT Material 材料
CAT Category 类事	QNT Quantity 数量
COM Comitative 涉事	DUR Duration 历时
SOR Source 源事	FRQ Frequency 频次

Adjectives in CCD

- Classifies Chinese adjectives into common adjectives and special adjectives
- CCD particularly emphasizes syntactic-semantic differences

<table>
<tr><td>Common Adjectives
can be modified by 很 very</td><td colspan="3">Special Adjectives
Cannot be modified by 很 very</td></tr>
<tr><td rowspan="3">大
小
高
低
重</td><td>Serve as adverbial modifier</td><td colspan="2">Not serve as adverbial modifier</td></tr>
<tr><td rowspan="2">高速
大规模
小范围</td><td>Serve as attribute</td><td>Not serve as attribute</td></tr>
<tr><td>男
女</td><td>红彤彤
绿油油</td></tr>
</table>

- Synonymy
  - 干 (*dry)* , {干，干燥，干枯，干爽}
- Similarity
- Antonymy



Status of CCD

- CCD version 1.0 contains synsets and relations of
  - about 66,000 noun concepts
  - 12,000 verbal concepts
  - 21,000 adjectival and adverbial concepts
  - The annotated concepts can be aligned with about 100,000 concepts in English WordNet version 1.6.
- It plays an important role in the construction of Global WordNet.
- More information
  http://icl.pku.edu.cn/icl_groups/ccd.asp.

Applications of Semantic Resources

- Word Sense Disambiguation
- Word Similarity Estimation
- Sentence Polarity Determination
- Ontology Construction
- Semantic Field Construction
- General Knowledge Reasoning
  - Transfer WordNet descripts to semantic network
  - hungry（饿）-- food（食物）-- refrigerator（冰箱）

SenticNet

- SenticNet provides a set of semantics, sentics, and polarity associated with 200,000 natural language concepts.
  - Semantics
    Semantics define the denotative information associated with words and multiword expressions (i.e., semantically-related concepts)
  - Sentics
    Sentics define the connotative information associated with natural language concepts (i.e., emotion categorization values expressed in terms of four affective dimensions: Pleasantness, Attention, Sensitivity, and Aptitude)
  - Polarity
    Polarity is floating number between -1 and +1 (where -1 is extreme negativity and +1 is extreme positivity)
- Resources: sentic.net

Four sentic dimensions

- Emotion categorization values expressed in terms of four affective dimensions.
  - Pleasantness
  - Attention
  - Sensitivity
  - Aptitude

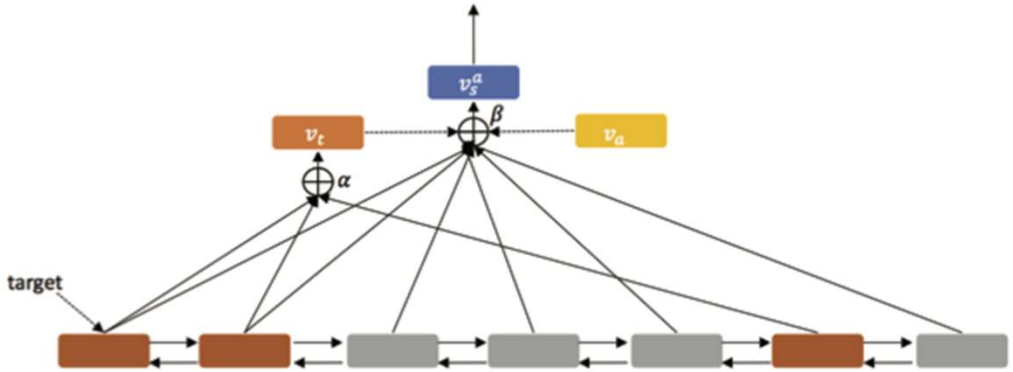
	愉悦 (Pleasantness)	专注 (Attention)	敏感 (Sensitivity)	倾向 (Aptitude)
+3	狂喜 (ecstasy)	警惕 (vigilance)	发怒 (rage)	钦佩 (admiration)
+2	快乐 (joy)	期待 (anticipation)	愤怒 (anger)	信任 (trust)
+1	宁静 (serenity)	兴趣 (interest)	烦恼 (annoyance)	接受 (acceptance)
0				
−1	沉思 (pensiveness)	分心 (distraction)	忧虑 (apprehension)	无聊 (boredom)
−2	悲伤 (sadness)	惊喜 (surprise)	害怕 (fear)	厌恶 (disgust)
−3	悲伤 (grief)	惊奇 (amazement)	恐怖 (terror)	极度讨厌 (loathing)

SenticNet

- For example:
- **Concepts**: a_little
- **Semantics**:
  - least
  - little
  - small_amount
  - shortage
  - scarce
- **Sentics**:
  - Pleasantness: -0.99
  - Attention: 0
  - Sensitivity: 0
  - Aptitude: -0.70
- **Moodtags**:
  - sadness
  - disgust
- **Polarity**:
  - negative
  - -0.84

Applications of SenticNet

- Targeted Aspect-Based Sentiment Analysis
  - [Ma et al. 2018] propose a novel solution to targeted aspect-based sentiment analysis, which tackles the challenges of both aspect-based sentiment analysis and targeted sentiment analysis by exploiting commonsense knowledge.
  - Commonsense knowledge of sentiment-related concepts is incorporated into the end-to-end training of a deep neural network for sentiment classification.



$$f_i = \sigma(W_f[x_i, h_{i-1}, \mu_i] + b_f)$$
$$I_i = \sigma(W_I[x_i, h_{i-1}, \mu_i] + b_I)$$
$$\widetilde{C}_i = tanh(W_C[x_i, h_{i-1}] + b_C)$$
$$C_i = f_i * C_{i-1} + I_i * \widetilde{C}_i$$
$$o_i = \sigma(W_o[x_i, h_{i-1}, \mu_i] + b_o)$$
$$o_i^c = \sigma(W_{co}[x_i, h_{i-1}, \mu_i] + b_{co})$$
$$h_i = o_i * tanh(C_i) + o_i^c * tanh(W_c\mu_i)$$

Ma, Y., Peng, H. and Cambria, E., 2018, February. Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM.

The Next Lecture

- Lecture 8

  Word Sense Disambiguation