Integrating heterogeneous thesauruses for Chinese synonyms

Jianbing ZHANG, Peng WU, Yingjie ZHANG, Shujian HUANG, Xinyu DAI, Jiajun CHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-020-9253-3

Problems & Ideas

- Can we build a better Chinese synonym resource by integrating the heterogeneous Chinese Concept Dictionary (CCD) and Tongyici-Cilin (Cilin) together?
  - CCD: cover rich semantic relations but miss many Chinese-specific words
  - Cilin: cover many Chinese words but miss clear relations between them

Table 1: Statistics of CCD Categories and Cilin Classes for each part-of-speech (POS)

POS	Cilin			CCD
	Class	Medium Class	Subdivision	Category
Noun	A B C D	48 –	588	26
Verb	F G H I J	34	567	15
Adjective	E	6 –	179	3
Adverb	K	1	35	1
Pronoun	A B C E	5 –	8	0
Others	K L	6	48	0

- Ideas: integrate CCD and Cilin by
  - direct mapping that ignores the inner structure of CCD and Cilin
  - hierarchical mapping that considers their structure

Main Experimental Results

Table 2: A Comparison of some metrics between direct and hierarchical mappings

	Direct	Hierarchical
# of synonym set	66,815	125,146
# of synonym set except single word	38,462	67,540
average # of words in synonym set	5.98	2.21
# of shared synonym set	58,332	
# of unique synonym set	8,517	66,814
average distance of unique synonym set	2.9868	2.1643

- Hierarchical mapping gets more synonym sets, because it keeps the synonym sets which contain the same word in mind, and meanwhile has different semantics via using the high-level words.
- The sets generated by direct mapping are larger since this method is straightforward and simple.
- For the unique set, hierarchical mapping can get closer distance in synonym set than direct mapping.