Multi-Strategies Extraction of Chinese Synonyms

Natural Language Processing Lab, Nanjing Normal University

SONG Wenjie, GU Yanhui, ZHOU Junsheng

SUN Yujie, YAN Jie, QU Weiguang

Outline

- Introduction
- Strategies of Chinese Synonymy Extraction
- Experiment
- GKB_Synonym
- Conclusion

Introduction

- Synonym is an important relationship in many Natural Language Processing research issues
  - Information Retrieval(IR)
  - Machine Translation(MT)
  - Text Categorization(TC)
  - ……
- Synonym is also an important relationship in semantic dictionary
  - Tongyici Cilin(Cilin)
  - Chinese Concept Dictionary(CCD)
  - Grammatical Knowledge-base of Contemporary Chinese(GKB) ?

Introduction

WAN FANG DATA(2010-2014,Chinese)

Key Word	Journal	Degree	Conference
Synonym+IR	39	78	5
Synonym+MT	10	17	1
Synonym+TC	19	40	3

Introduction

WAN FANG DATA(2010-2014,Chinese)

Key Word	Journal	Degree	Conference
Cilin	127	97	10
CCD	90	30	5
GKB	15	8	2

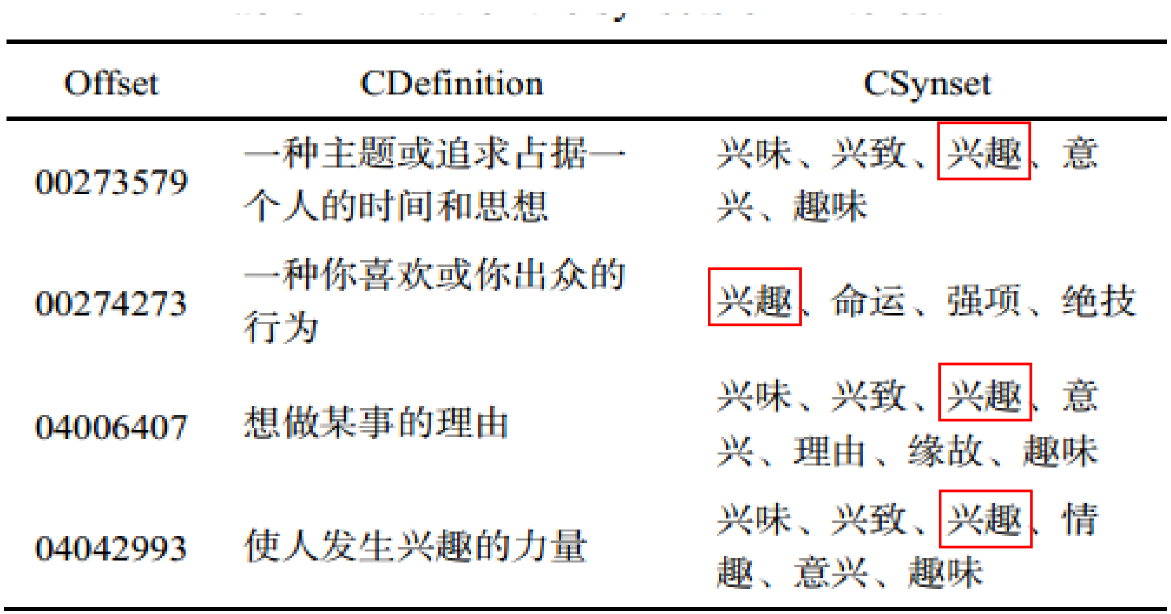
Introduction

- Both Cilin&CCD is built 10 years ago
    - the meaning of word in the dictionary may be changed
    - the semantic set needs to be expanded and updated
    - Internet resources are more popular over the years
- If we use Cilin+CCD+Internet resources to build semantic system for GKB?
    - contain semantic attributes
    - update synonym set for Cilin&CCD at the same time
    - a better dictionary for NLP research issues

Strategies of Chinese Synonymy Extraction

- Dictionary based method
    - Cilin
    - CCD
- HTML tag based method
    - Infobox in baidubaike
    - HTML tag in Zdic
- DIPRE based method

:hod

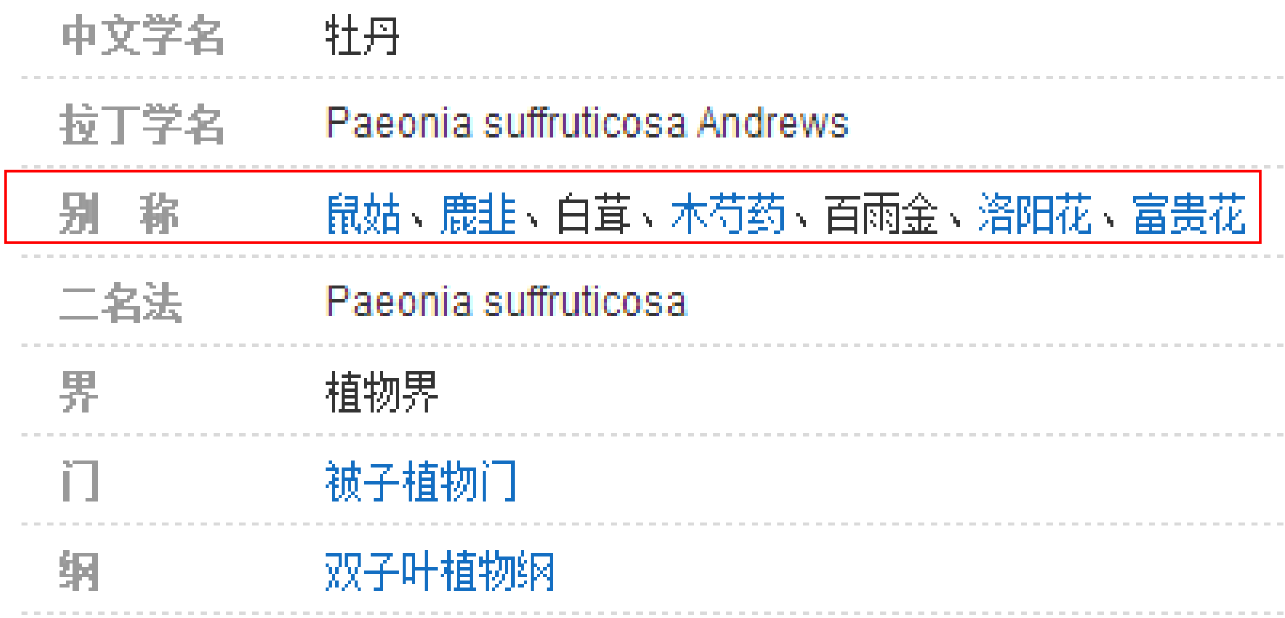


Offset	CDefinition	CSynset
00273579	一种主题或追求占据一个人的时间和思想	兴味、兴致、兴趣、意兴、趣味
00274273	一种你喜欢或你出众的行为	兴趣、命运、强项、绝技
04006407	想做某事的理由	兴味、兴致、兴趣、意兴、理由、缘故、趣味
04042993	使人发生兴趣的力量	兴味、兴致、兴趣、情趣、意兴、趣味

- CCD
  - one word in multiple CSynsets
  - not all words in CSynsets are synonym
  - use typical synonym filtering method*
  - {兴味、 兴致、 意兴、 趣味、 情趣 }

*孙玉霞 , 狄颖 , 曹冉, 等. 中文同义词自动抽取研究. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012testpaper.html

HTML tag based method



中文学名	牡丹
拉丁学名	Paeonia suffruticosa Andrews
别　称	鼠姑、鹿韭、白茸、木芍药、百雨金、洛阳花、富贵花
二名法	Paeonia suffruticosa
界	植物界
门	被子植物门
纲	双子叶植物纲

{牡丹、鼠姑、鹿韭、白茸、木芍药、百雨金、洛阳花、富贵花}

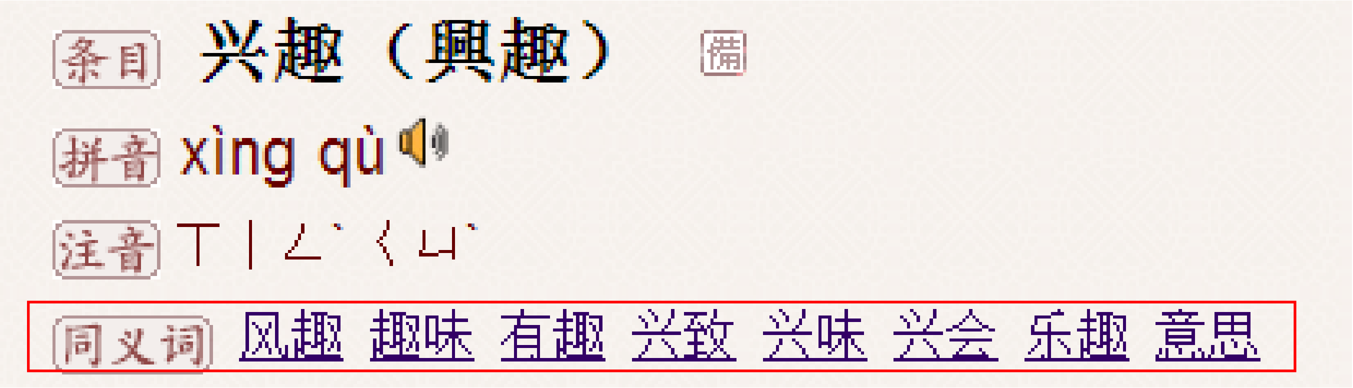
HTML tag based method

- Infobox in baidubaike
  - get baidubaike HTML page
    - http://baike.baidu.com/taglist?tag=XXX
    - XXX is the GKB coding of word
  - retrieve the position of the Infobox and get synonym
    - HTML tag:

      HT1=<div class=“biContent”>

      HT2=</div></div></div><div class="biItem">
    - if string contains key words, such as “别称”(another name)、“同义词”(synonym)、 “别名”(alias)…,get string from HT1 to last HT2 ,use a regular type matching the string to get synonym

HTML tag based method



条目 兴趣（興趣）

拼音 xìng qù

注音 ㄒㄧㄥˋ ㄑㄩˋ

同义词 风趣 趣味 有趣 兴致 兴味 兴会 乐趣 意思

{兴趣、风趣、趣味、有趣、兴致、兴味、兴会、乐趣、意思}

HTML tag based method

- HTML tag in Zdic
  - get Zdic HTML page
    - crawl web page of www.zdic.net
  - get synonym
    - HTML tag:

      HT1:<p><span class="dicty"><imgsrc="/images/c_i_tyc.gif" align="absmiddle">”

      HT2:</span></p>
    - if string contains key word “同义词”(synonym),get string from HT1 to HT2 ,use a regular type matching the string to get synonym

DIPRE based method

- DIPRE (Dual Iterative Pattern Relation Expansion) is applied to discover high credible patterns and synonymous instances in encyclopedia corpora
- use synonym sets extracted from Cilin&CCD as seed instances *i*={x,y} and get new pattern P from sentences in corpora which contain *i*
- *r*(*p*) means the reliability of P

$$r(P) = \frac{\sum_{i \in I} \frac{\mathrm{pmi}(i, p)}{\max_{\mathrm{pmi}}} \times r_i(i)}{|I|}, \qquad (1)$$

DIPRE based method

- pmi(*i*,*p*) means the mutual information of instance and pattern

$$\mathrm{pmi}(i,p)=\log\frac{|x,p,y|}{|x,*,y||*,p,*|} \qquad (2)$$

- $r_l(i)$ means the reliability of instance extracted from sentence

$$r_l(i)=\frac{\sum_{p\in P'}\frac{\mathrm{pmi}(i,p)}{\max_{\mathrm{pmi}}}\times r(p)}{|P|}, \qquad (3)$$

DIPRE based method

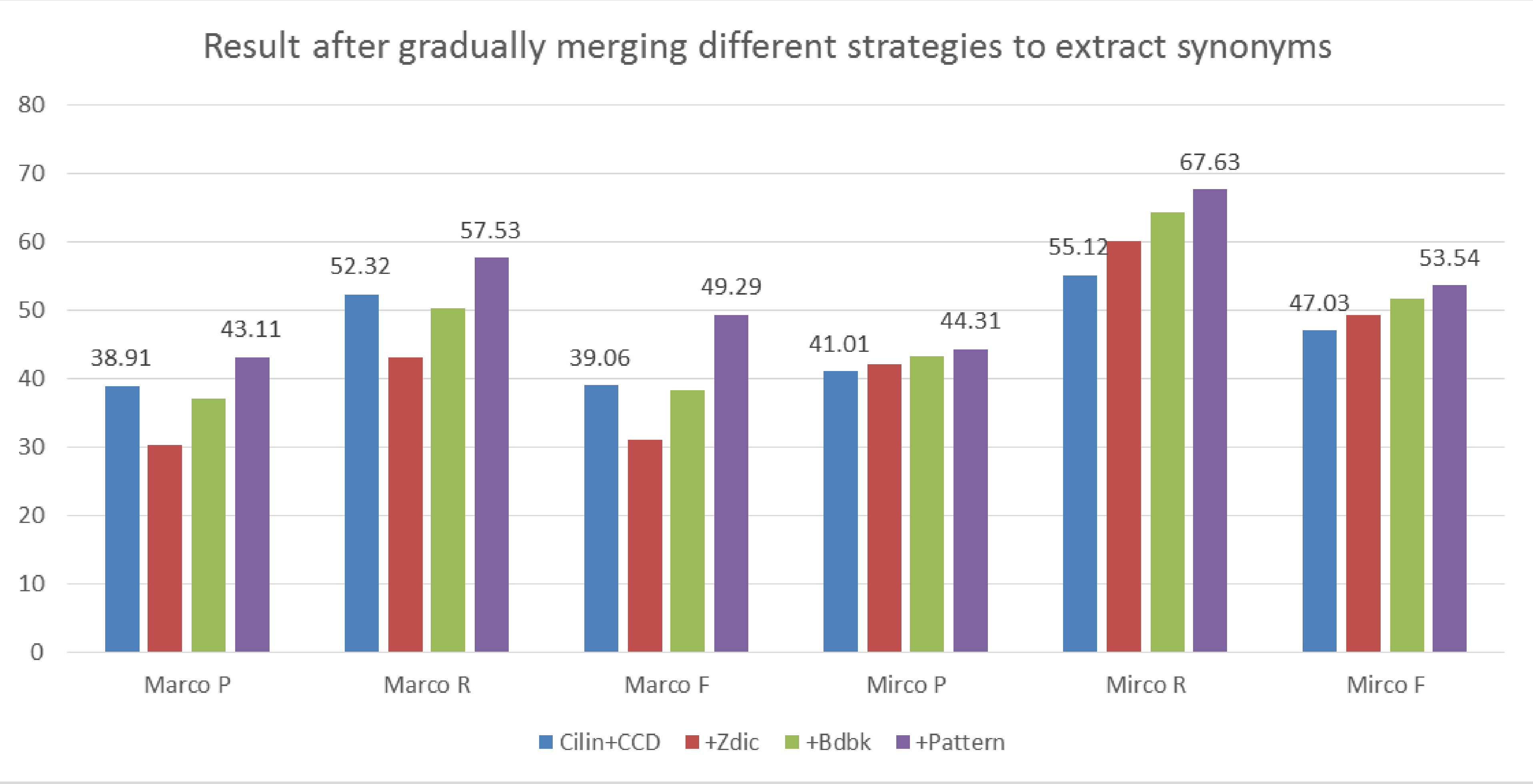
- iterate until no new reliable pattern or instance can be obtained
- when $r(p) \geq \alpha$, pattern p is reliable
- when $r_l(i) \geq \beta$, instance i is reliable
- adjust α and β according to Precision and Recall

Experiment

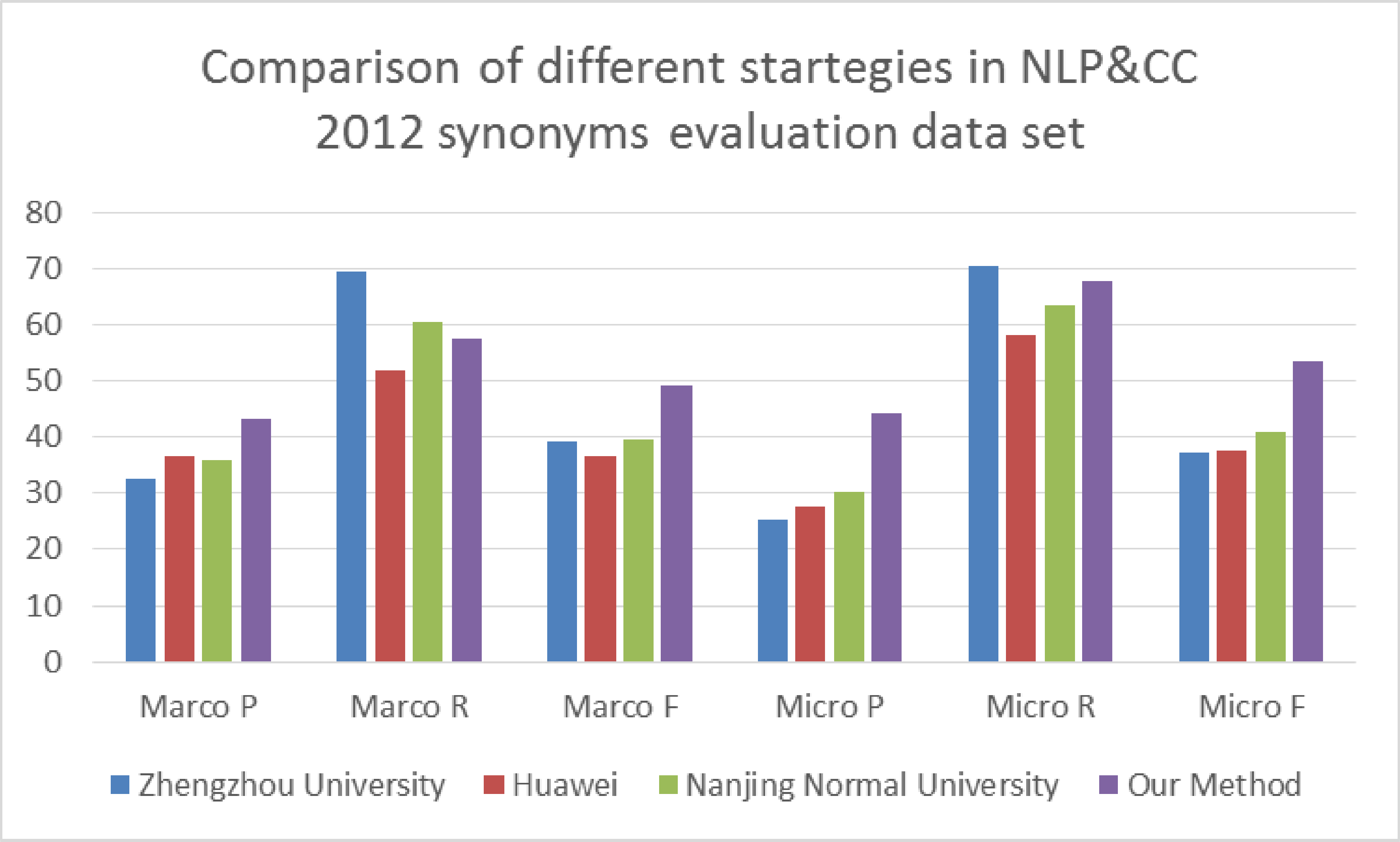
- evaluation standard:
  - NLP&CC 2012 Words Relations Extraction Evaluation*
  - Precision、Recall、F-measure(Macro&Micro)
- test set:
  - 778 words
  - 5890 synonyms

*http://tcci.ccf.org.cn/conference/2012/dldoc/评测大纲-词义.pdf

Experiment

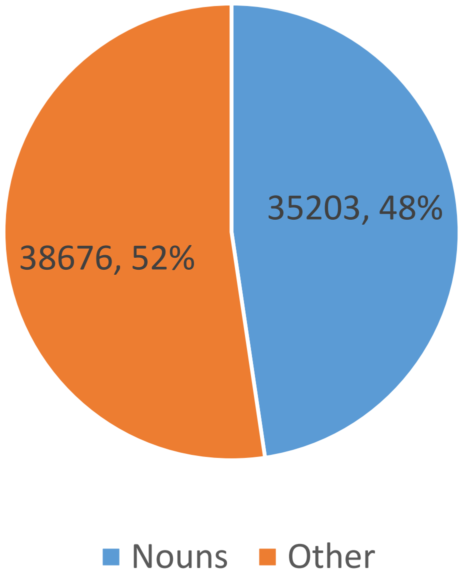

Result after gradually merging different strategies to extract synonyms
80
70
60
50
40
30
20
10
0
38.91
43.11
52.32
57.53
39.06
49.29
41.01
44.31
55.12
67.63
47.03
53.54
Marco P
Marco R
Marco F
Mirco P
Mirco R
Mirco F
Cilin+CCD
+Zdic
+Bdbk
+Pattern

Experiment


Comparison of different startegies in NLP&CC 2012 synonyms evaluation data set
80
70
60
50
40
30
20
10
0
Marco P
Marco R
Marco F
Micro P
Micro R
Micro F
Zhengzhou University
Huawei
Nanjing Normal University
Our Method

GKB_Synonym

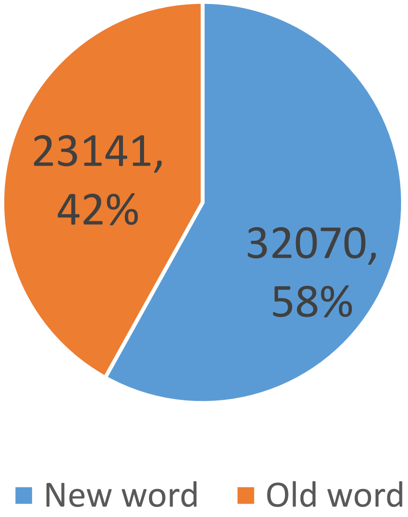
- GKB_Synonym: synonym system for Noun in GKB
  - synonym relation is more common between Nouns
  - Nouns account for 48% in all POS of GKB

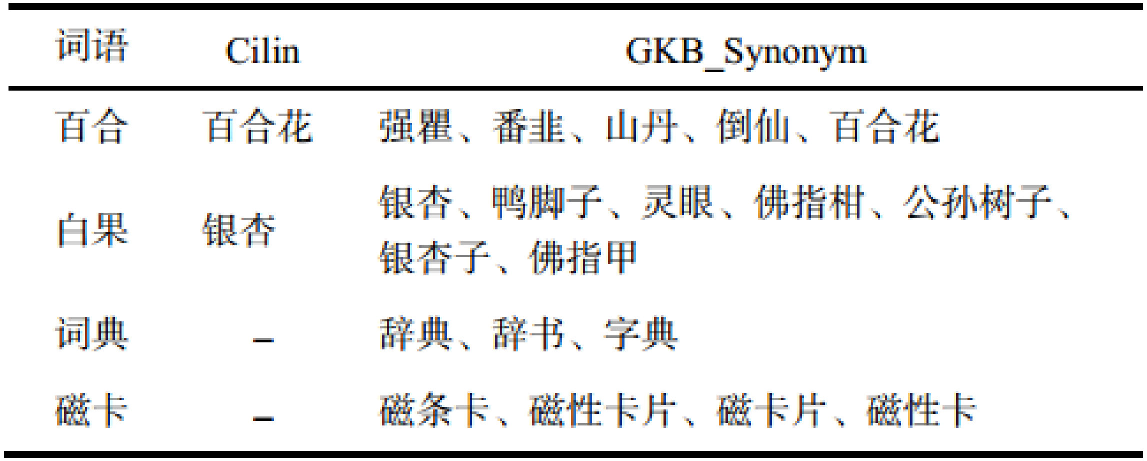


GKB_Synonym

- Artificial proofreading:10 people*30 days

GKB_Synonym	No synonym word	New Word	All Word
Before	6684	57586	87640
After	13597	32070	55211





词语	Cilin	GKB_Synonym
百合	百合花	强瞿、番韭、山丹、倒仙、百合花
白果	银杏	银杏、鸭脚子、灵眼、佛指柑、公孙树子、银杏子、佛指甲
词典	–	辞典、辞书、字典
磁卡	–	磁条卡、磁性卡片、磁卡片、磁性卡

Conclusion

- propose multi-strategies to extract Chinese synonyms
  - dictionary based method
  - HTML tag based method
  - DIPRE based method
- build GKB_Synonym and complete artificial proofreading using 10 people∗30 days
  - more comprehensive than Cilin
  - attempt to build semantic system for GKB
- future work
  - more strategies
  - automatic filtration
  - other POS of GKB

Thank You!