The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 139-152

Sketching the Dependency Relations of Words in Chinese

Meng-Hsien Shih* and Shu-Kai Hsieh*

Abstract

We proposes a language resource by automatically sketching grammatical relations of words based on dependency parses from untagged texts. The advantage of word sketch based on parsed corpora is, compared to Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004), to provide more details about the different usage of each word such as various types of modification, which is also important in language pedagogy. Although some language resources of other languages have attempted to sketch words based on parsed data, in Chinese we have not seen a resource for dependency sketch of words in customized texts. Therefore, we propose such a resource and evaluate with Chinese Sketch Engine (Huang et al., 2005) in terms of corresponding thesaurus function.

Keywords: Dependency grammar, Grammatical relation, NLP tools/resources.

1. Introduction

Syntagmatic relational information has been the focus of the interface studies of syntax and semantics. With the rapid development of corpora, various corpus query, profiling and visualization tools have emerged quickly over the past years. Among these tools, Word Sketch Engine (Kilgarriff et al., 2004; Huang et al., 2005) has provided an effective approach to quantitatively summarize grammatical and collocation behavior[1]. The provided functions include Concordancer, Word Sketch, Sketch Diff, Thesaurus, and other web corpus crawling and processing tools.

Previous literatures have revealed that corpus linguistics has benefited greatly from Chinese Sketch Engine (Hong & Huang, 2006). Although proprietary, Word Sketch Engine system is popular among corpus linguists and language teachers because of its functions in language analysis. However, the construction of Sketch Engine is time-consuming due to the

* Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
E-mail: {simon.xian, shukai}@gmail.com

[1]http://www.sketchengine.co.uk

manually edited sketch grammar. Here we propose an alternative approach to sketch the grammar profile of words automatically from a text corpus.

The paper is organized as follows: Section 2 reviews the current design of related language resources. Section 3 describes the proposed method of sketching words in a parsed corpus. Section 4 presents the results from the proposed approach and evaluation. In the final section, we have a brief conclusion for this paper.

2. Review

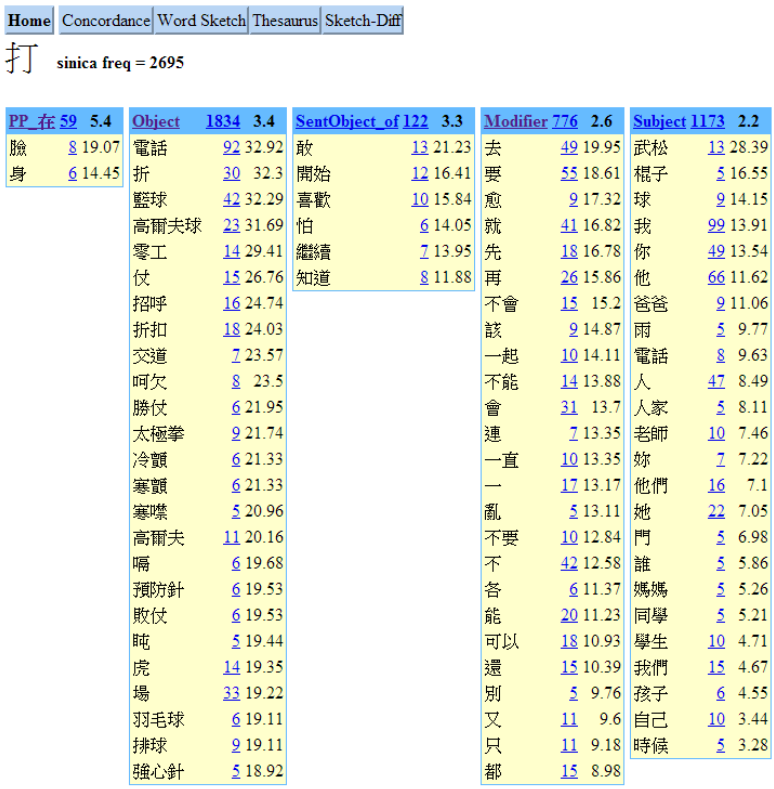
Word Sketch Engine (WSE) provides a set of corpus query tools that aims to help users reveal linguistic patterns in language use. Among these tools, word sketch function gains the most popularity and has widely applied in the studies of corpus linguistics and language pedagogy (Kilgarriff, 2007).

Given the preprocessed corpus data, the available WSE system in most languages makes use of regular expressions to extract grammatical information from a POS-tagged corpus. The so-called *sketch grammars*, mostly manually crafted by linguists, describe the relation between a target word and its dependent, constrained on the surrounding context. In its design of grammar engineering, the sketch grammars are used for finite-state shallow parsing to extract the different grammatical relations [2]. Typical relations in English WSE include: [OBJECT_OF], [ADJ_MODIFIER], [NOUN_MODIFIER], [MODIFIES], [AND/OR], [PP_INTO], etc.

In terms of corpus linguistics, the *sketch* for a word presents a candidate set of its *collocates* organized by their grammatical relations they stand in to the target word. These collocates are sorted according to certain statistic measure of co-occurrence, as illustrated in the case of 打"hit"[3]:

[2] http://www.sketchengine.co.uk/documentation/wiki/SkE/Help/CreateCorpus

[3] http://wordsketch.ling.sinica.edu.tw



Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

打 sinica freq = 2695

PP_在 59 5.4		Object 1834 3.4		SentObject_of 122 3.3		Modifier 776 2.6		Subject 1173 2.2	
臉	8 19.07	電話	92 32.92	敢	13 21.23	去	49 19.95	武松	13 28.39
身	6 14.45	折	30 32.3	開始	12 16.41	要	55 18.61	棍子	5 16.55
		籃球	42 32.29	喜歡	10 15.84	愈	9 17.32	球	9 14.15
		高爾夫球	23 31.69	怕	6 14.05	就	41 16.82	我	99 13.91
		零工	14 29.41	繼續	7 13.95	先	18 16.78	你	49 13.54
		仗	15 26.76	知道	8 11.88	再	26 15.86	他	66 11.62
		招呼	16 24.74			不會	15 15.2	爸爸	9 11.06
		折扣	18 24.03			該	9 14.87	雨	5 9.77
		交道	7 23.57			一起	10 14.11	電話	8 9.63
		呵欠	8 23.5			不能	14 13.88	人	47 8.49
		勝仗	6 21.95			會	31 13.7	人家	5 8.11
		太極拳	9 21.74			連	7 13.35	老師	10 7.46
		冷顫	6 21.33			一直	10 13.35	妳	7 7.22
		寒顫	6 21.33			一	17 13.17	他們	16 7.1
		寒噤	5 20.96			亂	5 13.11	她	22 7.05
		高爾夫	11 20.16			不要	10 12.84	門	5 6.98
		嗝	6 19.68			不	42 12.58	誰	5 5.86
		預防針	6 19.53			各	6 11.37	媽媽	5 5.26
		敗仗	6 19.53			能	20 11.23	同學	5 5.21
		盹	5 19.44			可以	18 10.93	學生	10 4.71
		虎	14 19.35			還	15 10.39	我們	15 4.67
		場	33 19.22			別	5 9.76	孩子	6 4.55
		羽毛球	6 19.11			又	11 9.6	自己	10 3.44
		排球	9 19.11			只	11 9.18	時候	5 3.28
		強心針	5 18.92			都	15 8.98		

Figure 1. Word sketch of 打"hit"

The core component in WSE system is the *sketch grammar*, which defines the linear patterns with regular expression for the system to automatically identify possible relations to the target word. For instance, one of the sketch grammar rules defined in the huge Chinese corpus (zhTenTen11, with 2.1 billion tokens) provided by WSE are concerned with modification. That is, we can identify the cases of modification relation where the target word (indicated by the prefix "1:") can be any noun followed by non-nouns. And the collocates, i.e., that words we want to capture (marked with the prefix "2:") is taken to be any verb followed by a word 的:

**DUAL*

=A_Modifier/Modifies

2:[tag="V.*"] [word="的"] [tag="N.*"]{0,**1:**[tag="N.*"] [tag!="N.*"]

The sketch grammar can be more complicated with the granularity of POS. The following grammar shows the classification relation developed by Huang et al. (2005) and implemented in the Chinese WordSketch system[4], i.e., the target word can be a noun preceded by a measure word (tagged by Nf):

=Measure

2:"Nf.*" ("A"|"VH11"|"VH13"|"VH21"|"V.*" "DE") [tag="N[abcd].*" & tag!="Ncd"]
1:[tag="N[abcdhf].*" & tag!="Nbc.*" & tag!="Ncd.*" & word!="者" & word!="們"]
[tag!="N[abcdhef].*"|tag="Nbc.*"|tag="Ncd.*"]

However, the writing of grammar is time-consuming, running risk of 'low recall', so we turn to exploit the dependency parser for enriching the relational information. Unlike phrase-structure grammar, dependency grammar concentrates on the *typed dependency* between words, rather than constituent information. It is highly advantageous to our study, for it is linguistically-rich - capturing not only syntactic information such as *nsubj* (nominal subject) but also abstract semantic ones such as *loc* (localizer) - and can be further applied to other syntactic-semantic interface tasks (Chang, Tseng, Jurafsky, & Manning, 2009).

The Stanford lexicalized probabilistic parser (Levy & Manning, 2003) works out the grammatical structure of sentences with a factored product model efficiently combing preferences of PCFG phrase structure and lexical dependency experts. In addition to phrase structure tree, the parser also provides Stanford Dependencies (SD)[5] that are known as grammatical relations between words in a sentence. Take the following Chinese sentence for example: 我 很 喜歡 兩 則 惜福 與 惜緣 的 故事。 The *head* 喜歡 has a *dependent* of 我 as its nominal subject, and another dependent of 故事 as direct object (Fig. 2).

[4] http://wordsketch.ling.sinica.edu.tw

[5] http://nlp.stanford.edu/software/stanford-dependencies.shtml

(ROOT (IP (NP (PN 我)) (VP (ADVP (AD 很)) (VP (VV 喜歡) (NP (DNP (NP (NP (NR 兩)) (NP (NN 則) (NN 惜福) (NN 與) (NN 惜緣))) (DEG 的)) (NP (NN 故事))))) (PU 。)))	nsubj(喜歡-3, 我-1) advmod(喜歡-3, 很-2) root(ROOT-0, 喜歡-3) nn(惜緣-8, 兩-4) nn(惜緣-8, 則-5) nn(惜緣-8, 惜福-6) nn(惜緣-8, 與-7) assmod(故事-10, 惜緣-8) assm(惜緣-8, 的-9) dobj(喜歡-3, 故事-10)

Figure 2. Dependencies in a Chinese sentence with PCFG: 我很喜歡兩則惜福與惜緣的故事。

The SD has been widely used in NLP-related fields such as sentiment analysis (Meena & Prabhakar, 2007), textual entailment (Androutsopoulos & Malakasiotis, 2010). The Chinese version of SD (Chang et al., 2009) is also available on the Stanford Dependencies page[6]. The SD can even distinguish 45 typed dependencies among Chinese words, as shown in Table 1.

[6] http://nlp.stanford.edu/software/stanford-dependencies.shtml#Chinese

Table 1. Chinese dependency relations (Chang et al., 2009)

abbreviation	short description	Chinese example	typed dependency	counts	percentage
nn	noun compound modifier	服务 中心	nn(中心, 服务)	13278	15.48%
punct	punctuation	海关 统计 表明 ，	punct(表明, ，)	10896	12.71%
nsubj	nominal subject	梅花 盛开	nsubj(盛开, 梅花)	5893	6.87%
conj	conjunct (links two conjuncts)	设备 和 原材料	conj(原材料, 设备)	5438	6.34%
dobj	direct object	浦东 颁布 了 七十一 件 文件	dobj(颁布, 文件)	5221	6.09%
advmod	adverbial modifier	部门 先 送上 文件	advmod(送上, 先)	4231	4.93%
prep	prepositional modifier	在 实践 中 逐步 完善	prep(完善, 在)	3138	3.66%
nummod	number modifier	七十一 件 文件	nummod(件, 七十一)	2885	3.36%
amod	adjectival modifier	跨世纪 工程	amod(工程, 跨世纪)	2691	3.14%
pobj	prepositional object	根据 有关 规定	pobj(根据, 规定)	2417	2.82%
rcmod	relative clause modifier	不 曾 遇到 过 的 情况	rcmod(情况, 遇到)	2348	2.74%
cpm	complementizer	开发 浦东 的 经济 活动	cpm(开发, 的)	2013	2.35%
assm	associative marker	企业 的 商品	assm(企业, 的)	1969	2.30%
assmod	associative modifier	企业 的 商品	assmod(商品, 企业)	1941	2.26%
cc	coordinating conjunction	设备 和 原材料	cc(原材料, 和)	1763	2.06%
clf	classifier modifier	七十一 件 文件	clf(文件, 件)	1558	1.82%
ccomp	clausal complement	银行 决定 先 取得 信用 评级	ccomp(决定, 取得)	1113	1.30%
det	determiner	这些 经济 活动	det(活动, 这些)	1113	1.30%
lobj	localizer object	近年 来	lobj(来, 近年)	1010	1.18%
range	dative object that is a quantifier phrase	成交 药品 一亿多 元	range(成交, 元)	891	1.04%
asp	aspect marker	发挥 了 作用	asp(发挥, 了)	857	1.00%
tmod	temporal modifier	以前 不 曾 遇到 过	tmod(遇到, 以前)	679	0.79%
plmod	localizer modifier of a preposition	在 这 片 热土 上	plmod(在, 上)	630	0.73%
attr	attributive	贸易额 为 二百亿 美元	attr(为, 美元)	534	0.62%
mmod	modal verb modifier	利益 能 得到 保障	mmod(得到, 能)	497	0.58%
loc	localizer	占 九成 以上	loc(占, 以上)	428	0.50%
top	topic	建筑 是 主要 活动	top(是, 建筑)	380	0.44%
pccomp	clausal complement of a preposition	据 有关 部门 介绍	pccomp(据, 介绍)	374	0.44%
etc	etc modifier	科技 、 文教 等 领域	etc(文教, 等)	295	0.34%
lccomp	clausal complement of a localizer	中国 对 外 开放 中 升起 的 明星	lccomp(中, 开放)	207	0.24%
ordmod	ordinal number modifier	第七 个 机构	ordmod(个, 第七)	199	0.23%
xsubj	controlling subject	银行 决定 先 取得 信用 评级	xsubj(取得, 银行)	192	0.22%
neg	negative modifier	以前 不 曾 遇到 过	neg(遇到, 不)	186	0.22%
rcomp	resultative complement	研究 成功	rcomp(研究, 成功)	176	0.21%
comod	coordinated verb compound modifier	颁布 实行	comod(颁布, 实行)	150	0.17%
vmod	verb modifier	其 在 支持 外商 企业 方面 的 作用	vmod(方面, 支持)	133	0.16%
prtmod	particles such as 所, 以, 来, 而	在 产业化 所 取得 的 成就	prtmod(取得, 所)	124	0.14%
ba	"ba" construction	把 注意力 转向 市场	ba(转向, 把)	95	0.11%
dvpm	manner DE(地) modifier	有效 地 防止 流失	dvpm(有效, 地)	73	0.09%
dvpmod	a "XP+DEV(地)" phrase that modifies VP	有效 地 防止 流失	dvpmod(防止, 有效)	69	0.08%
prnmod	parenthetical modifier	八五 期间 （ 1990 – 1995 ）	prnmod(期间, 1995)	67	0.08%
cop	copular	原 是 自给自足 的 经济	cop(自给自足, 是)	59	0.07%
pass	passive marker	被 认定 为 高 技术 产业	pass(认定, 被)	53	0.06%
nsubjpass	nominal passive subject	镍 被 称作 现代 工业 的 维生素	nsubjpass(称作, 镍)	14	0.02%

On the other hand, most semantic resources like PropBank (Palmer, Gildea, & Kingsbury, 2005) and FrameNet (Baker, Fillmore, & Lowe, 1998) either provide coarse-grained information or with limited coverage. In this paper, we propose a lexical resource tool to describe more detailed information for all words in a text corpus. We choose Sinica Corpus (Chen, Huang, Chang, & Hsu, 1996) as our texts and evaluate the results with Chinese Sketch Engine in terms of corresponding thesaurus function.

3. Method

In this case study, untagged texts of 567,702 sentences from Sinica Corpus 3.0[7] were parsed with dependency relations by the Stanford Parser (Chang et al., 2009). We obtained 574,552 dependency relations (of 23 types) between 44,257 words.

To sketch a word, we make use of the dependency tuples from the parsed corpus (see the right panel of Fig. 2) to extract the relations of each word with its dependents, and obtain the sketch of words such as 打 "hit" shown below:

Table 2. Dependency sketch of 打 "hit"

(Matches with Chinese Sketch Engine are marked in red bold face)

prep	dobj	advmod/mmod	nsubj	asp	conj
在	**電話**	**去**	**武松**	了	重建
到	**折**	**要**	**棍子**	著	是
自	**籃球**	**就**	**球**		鬧
	高爾夫球	**先**	**我**		
	硬仗	**不會**	**你**		
	招呼	**該**	**他**		
	折扣	**一起**	**爸爸**		
	哈欠	**會**	**雨**		
	太極拳	連續	**人**		
	麻藥針	**一**	**老師**		
	盹兒	**能**	**他們**		
	虎	**可以**	**她**		
	羽毛球	還要	**學生**		
	排球	**都**	**自己**		
	蛇	雖然	湖人		
	起來	仍然	來		
	秋千	而	政		

[7] www.sinica.edu.tw/SinicaCorpus

Since the Stanford Parser still suffers from parsing difficulty in Chinese (Levy & Manning, 2003), the grammatical relations automatically required, though impressive, may contain heterogeneous errors originating from mistagging errors[8], syntactic ambiguities and other dependency parsing issues, so we have observed some minor sketch errors in the result. However, it's hard to evaluate the results in an automatic way as conventionalized in the field of NLP. The main reasons are:

[1]. Currently, there is no gold-standard (in Chinese). It is particularly hard to measure recall for the set of 'correct answer' is not available.

[2]. An overall evaluation of the sketch performance will have to rely on the assessment of each module (word segmentation, POS tagging, sketch grammar and/or dependency parsing, etc.) separately. A comparative table is shown in Table 3.

Table 3. Comparison of Different Word Sketch Systems

Word Sketch System	word segmentation	pos tagging/tagset	sketch grammar	dependency parser
CWSE.sinica	CKIP	CKIP/ASBC	hand-crafted rules	*
zhTenTen.11	Stanford Chinese Word Segmenter	Stanford Log-linear Part-Of-Speech Tagger / Chinese Penn Treebank standard	hand-crafted (2 rules)	*
Proposed	Stanford Chinese Word Segmenter	*	*	Stanford dependencies

[8] In this study, since Stanford Parser takes manully tokenized input from Sinica Corpus, the segmentation error may be less than that from an automatic segmenter and be omitted here.

In addition, from the perspective of language resources construction as well as applied lexicography, as the system aims to identify highly salient candidate patterns, the noisy data should not constitute a serious problem for the task. The position is also well-articulated and proposed in (Kilgarriff, Kovář, Krek, Srdanović, & Tiberius, 2010), where a variant of evaluation paradigm (user/developer-oriented paradigm) is required.

Different from Ambati, Reddy, and Kilgarriff (2012) and Reddy, Klapaftis, McCarthy, and Manandhar (2011) where an external evaluation task such as *topic coherence* or *semantic composition* were adopted, we evaluated the proposed method with the task of automatic construction of thesaurus, for our main concern is the construction of language resource rather than NLP system performance.

The thesaurus in WSE is called **distributional thesaurus**, and can be built for any language if the word sketches data of the language is available. The thesaurus is constructed by computing the similarity between words based upon the overlapping rate of their word sketches. Our method instead, follows the **distributional semantic model** (Dinu, Pham, & Baroni, 2013; Turney & Pantel, 2010) and anchors on two manually constructed resources of the **Chinese Wordnet**[9] and Chilin (Chao & Chung, 2013)[10].

4. Evaluation

The dependency data of five selected synonym sets (經常, 原因, 按照, 相當, and 快樂) from Chinese Wordnet were converted into multi-dimensional (to avoid sparseness, only dependents shared by both synonyms were included) in order to calculate distributional similarity between synonyms. Five synonym sets from Chinese Wordnet were examined. For example, the dependency data of 高興 and 快樂 are converted as follows (disregarding the dependency type):

	不	也	了	他	可以
高興	7	1	5	5	1
快樂	1	4	1	3	2

[9] http://lope.linguistics.ntu.edu.tw/cwn2

[10] http://code.google.com/p/tw-synonyms-chilin

Then we adopt one of the common measures for similarity in distributional models, *cosine similarity*, to calculate the similarity between two words (e.g., 高興 and 快樂). The meaning of a word is determined by its collocation, and represented as a vector of its co-occurrence with other words in multiple dimensions. In this model, the similarity between two word vectors, w_1 and w_2, can be calculated by their cosine value:

$$CosSimilarity(w_1, w_2) = \frac{w_1 \cdot w_2}{|w_1||w_2|} \quad (1)$$

To illustrate, consider only the first two dimensions of 高興 and 快樂, the cosine similarity between the two words would be $(7,1)\cdot(1,4)/\sqrt{}(7^2+1^2)/\sqrt{}(1^2+4^2)=0.377$, and the calculation can be extended to even more dimensions. If two words have similar collocation with other words, the value of cosine similarity will approach the upper bound of 1.0 and could be considered a pair of synonyms.

Finally, to obtain a synonym list, the dependents of a target word are ranked by their similarity with the target word, regardless of their dependency relations. The results for the selected five synonym sets in Chinese Wordnet and Chilin[11] are shown in Table 4.

Table 4. Comparison of the results with Sketch Engine

	快樂	經常	原因	按照	相當
Cilin	高興,愉快,樂,…	時常,常常,時時,常..	因,故,緣故,緣由,…	依照,比照,遵照,…	頂,相當於,…
Chinese Wordnet	樂,愉快,愉	時常,勤,常常,常,恆	關係,肇始,因, 故,導因,緣	按,依照,依據,根據,…	很,相當於,具體
Proposed Method	有趣,愉快,美好,…	時常	關係	按,依照	具體, …
Sketch Engine	愉快,美好,…	n/a	因素,背景,條件,環境,理由,..	n/a	莫大,重大,重要,直接,…

[11]Although WordNet is more used in natural language processing, Chilin is considered a more appropriate resource designed for thesaurus. Here we present the comparison with both.

For a brief look, we observed that the proposed method is capable of extracting more synonyms from a text corpus, which might be absent in the Sketch Engine, although we still cannot perform as accurately as does the manual *sketch grammar* of Sketch Engine.

We also built a web interface considering the friendly access for potential users from TCSL (Teaching Chinese as Second Language) and linguistics[12]. Figure 3 shows a snapshot of the prototype. The sketch page first shows the frequent roles of the query word ranked by their frequency, followed by the collocation in each role. The page also shows, as the classical Word Sketch does, an analysis of the types of words which the query word collocates with. For example, in Figure 3 we can know that 快樂 "happy" frequently serves as an associative modifier (14.3%) and modifies 笑容 "smile" twice in this corpus. We believe that such word sketch information is useful in TCSL application. The scripts and data has been put on Github[13] for open access and further collaboration.

[12] http://140.112.147.131:8000/sketch

[13] http://github.com/mhshih/sketch

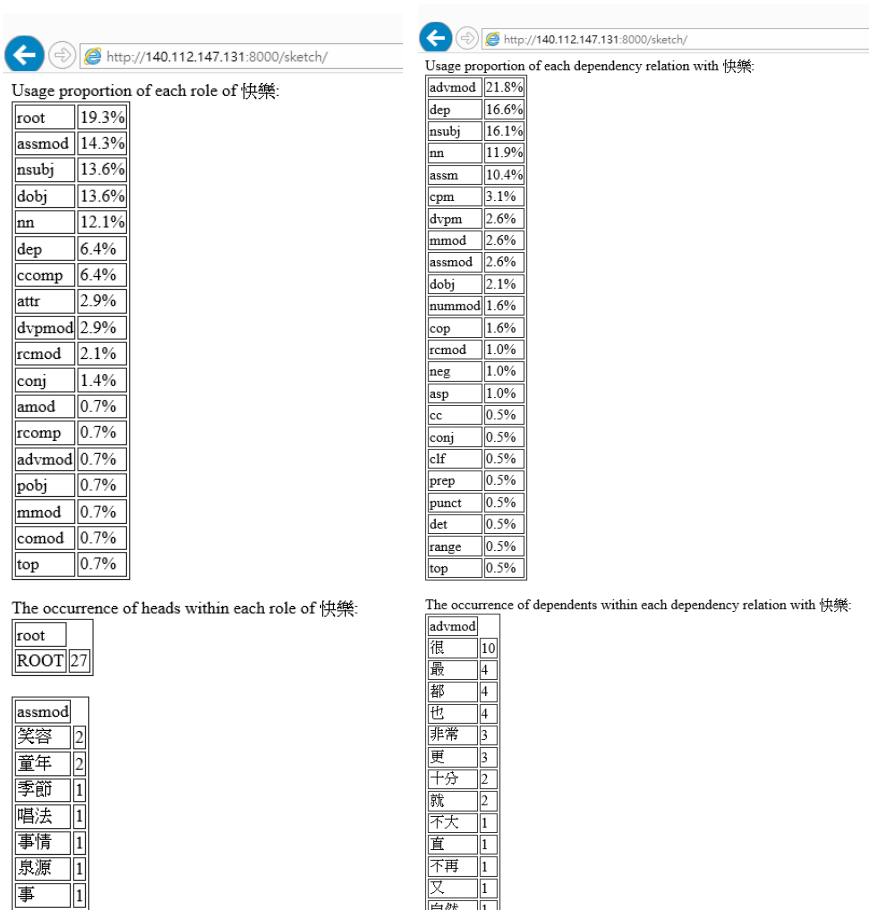


Figure 3. Snapshot of the sketch function

5. Conclusion

Word sketch is a corpus-based automatic summary of a word's grammatical and collocational behavior. Based on the hand-crafted finite-state sketch grammar over a POS-tagged corpus, word sketch system can identify the collocates with grammatical relations to the target word. However, the grammar engineering is time-consuming and requires experts, in this paper, we propose an alternative by leveraging an existing dependency parser. The results were evaluated based on the comparison of distributional thesaurus with significance.

This paper serves as the first attempt to create an open-sourced word sketch-like corpus profiling system for Chinese linguistics and Teaching Chinese as Second Language. The

proposed method is pipelined and can be applied to user-created corpus. The extracted relation triples <w1, R, w2> can be used to enrich our on-going Chinese BIGLEX database. Future works include exploring other dependency parsing algorithm, incorporating advanced statistics to single out salient collocations, and an open evaluation platform for the further improvement of the resource are in progress.

Acknowledgements

The authors would like to thank reviewers of ROCLING 2014 for their insightful comments that help improve the manuscript.

References

Ambati, B. R., Reddy, S., & Kilgarriff, A. (2012). *Word Sketches for Turkish.* Paper presented at the LREC.

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res., 38*(1), 135-187.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). *The Berkeley FrameNet Project.* Paper presented at the Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, Montreal, Quebec, Canada.

Chang, P.-C., Tseng, H., Jurafsky, D., & Manning, C. D. (2009). *Discriminative reordering with Chinese grammatical relations features*. Paper presented at the Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, Boulder, Colorado.

Chao, F. A., & Chung, S.-F. (2013). A Definition-based Shared-concept Extraction within Groups of Chinese Synonyms: A Study Utilizing the Extended Chinese Synonym Forest. *IJCLCLP, 18*(2).

Chen, K.-J., Huang, C.-R., Chang, L.-P., & Hsu, H.-L. (1996). *Sinica corpus: Design methodology for balanced corpora.* Paper presented at the The 11th Pacific Asia Conference on Language, Information and Computation (PACLIC-11).

Dinu, G., Pham, N., & Baroni, M. (2013). *DISSECT-DIStributional SEmantics Composition Toolkit.* Paper presented at the 51st Annual Meeting of the Association for Computational Linguistics.

Hong, J.-F., & Huang, C.-R. (2006). *Using chinese gigaword corpus and chinese word sketch in linguistic research.* Paper presented at the The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20).

Huang, C.-R., Kilgarriff, A., Wu, Y., Chiu, C.-M., Smith, S., Rychly, P., . . . Chen, K.-J. (2005). *Chinese Sketch Engine and the extraction of grammatical collocations.* Paper presented at the Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Kilgarriff, A. (2007). Using corpora in language learning: the Sketch Engine. *Optimizing the role of language in Technology-Enhanced Learning, 22*, 21.

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I., & Tiberius, C. (2010). *A quantitative evaluation of word sketches.* Paper presented at the Proceedings of the 14th EURALEX International Congress, Leeuwarden, The Netherlands.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology, 105*, 116.

Levy, R., & Manning, C. (2003). *Is it harder to parse Chinese, or the Chinese Treebank?* Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Sapporo, Japan.

Meena, A., & Prabhakar, T. V. (2007). *Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis*. Paper presented at the Proceedings of the 29th European conference on IR research, Rome, Italy.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics, 31*(1), 71-106. doi: 10.1162/0891201053630264

Reddy, S., Klapaftis, I. P., McCarthy, D., & Manandhar, S. (2011). *Dynamic and Static Prototype Vectors for Semantic Composition.* Paper presented at the IJCNLP.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res., 37*(1), 141-188.