# A Data Augmentation Method for Building Sememe Knowledge Base via Reconstructing Dictionary Definitions

Xiaoran Li Toshiaki Takano
Adaptive Systems Lab, Shizuoka Institute of Science and Technology
{2121026.rs,takano.toshiaki}@sist.ac.jp

## Abstract

A sememe is a semantic language unit of meaning; it is indivisible. Sememe knowledge bases (SKBs), which contain words annotated with sememes, have been successfully applied to many natural language processing tasks. Some extant construction methods for sememe knowledge bases are performed in a limited lexicon with fixed-size sememe annotations. However, the obtained sememe annotation is challenging to extend to more words from other lexicons. In this paper, we proposed a method via reconstructing word definitions for expanding the lexicon. Moreover, we presented an evaluation sememe method utilizing graph embedding techniques and performed many experiments to prove effective.[1]

## 1 Introduction

A sememe is a semantic language unit of meaning[1]; it is indivisible. However, people usually employ words as the minimum semantic unit because words as semantic representations are available for writing, yet sememe is only a semantic concept. Usually, there is a sense semantic unit between the sememe and the word (As shown in Figure 1). Moreover, linguists believe that all languages have the same limited sememe space[2] (e.g. HowNet SKB[3], which uses about 2,000 language-independent sememes to manual annotate senses of over 100 thousand Chinese and English words). Moreover, cooperate with the multilingual encyclopedic dictionary as BabelNet[4] to build a multilingual SKB as [5]. Furthermore, Sememe can synthesize words and represent the essential meaning was successfully applied to Neural Networks[6][7], Reverse Dictionaries[8] and Textual Adversarial Attacking[9], etc. However, manual annotation is flawed because the meaning of words is incremental with the amount of information, and it is impractical to add and modify SKB artificially. Some studies on automatic SKB construction have emerged recently. We previously proposed a method[10] based on deep clustering networks to learn sememe. Specifically, this method is based on an Auto-encoder to achieve minimal semantic clustering. Instead of generating specific language-independent sememes, it generates word embeddings with approximate sememe meaning, Regrettably, which is imprecise. Moreover, (Qi et al. 2021)[11] explored an automatic way to build an SKB via dictionaries with a Controlled Defining Vocabulary (CDV)[12], and demonstrate the effectiveness of this method; it is even superior to the most widely used HowNet SKB. The method is to extract the CDV as sememes in the dictionary definition. However A CDV is composed of high-frequency words. If the dictionary is large enough, it does not cover all words perfectly, which means some complex words can not acquire sememe.

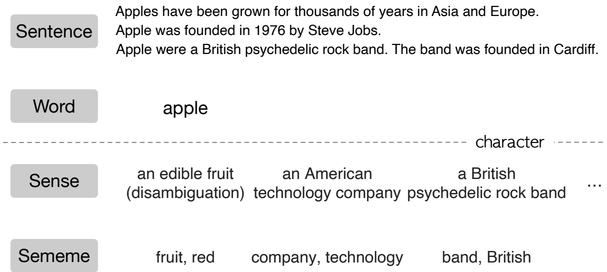
Sentence
Apples have been grown for thousands of years in Asia and Europe.
Apple was founded in 1976 by Steve Jobs.
Apple were a British psychedelic rock band. The band was founded in Cardiff.
Word
apple
character
Sense
an edible fruit (disambiguation)
an American technology company
a British psychedelic rock band
…
Sememe
fruit, red
company, technology
band, British

**Figure 1** Word-based semantic units (e.g. “*apple*”. It has very many senses. However, No matter how a word changes its sense, it is always composed of several primary and single sememes.)

To solve this problem, we proposed a way of reconstructing word definitions to increase the lexical coverage of CDV. We hypothesized that if the sememe can represent the basic meaning of a word, then if replacing the word with its sememes does not change its original meaning. More specifically, we solve the problem of CDV coverage by replacing the definitions of some words that CDV

[1] Codes are available at https://github.com/SauronLee/SKB-DA

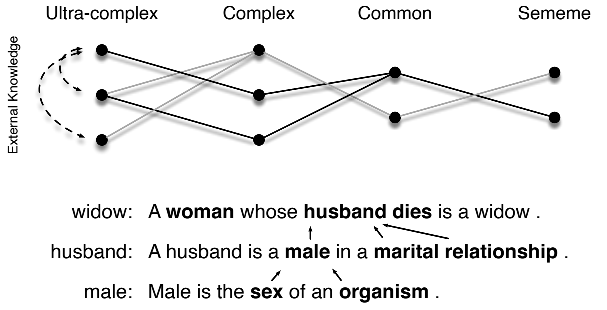
Ultra-complex
Complex
Common
Sememe
External Knowledge
widow: A woman whose husband dies is a widow .
husband: A husband is a male in a marital relationship .
male: Male is the sex of an organism .

**Figure 2** If "*widow*" is a complex word and "*husband*" is not in sememes, we can use the definition instead of "*husband*".

does not cover with definitions consisting of sememe (As shown in Figure 2). Finally, we employed the sememe internal evaluation criteria defined in [13] for evaluation. Moreover, we proposed a novel method to evaluate sememe by constructing a sememe graph. Because we consider the weight of sememe when constructing the sememe graph, it performs excellently in both evaluation methods. We shown the results in section 3.

## 2 Methodology

This section will detail the SKB construction method based on reconstructed word definitions and how to build a sememe graph for evaluation.

### 2.1 Building Sememe Knowledge Base

We employed the sememe search strategy of DictSKB[11]. It is intuitive to operate. Because a good sememe can represent the essential meaning of a word, it is most straightforward to extract the sememe from the word's definition. This method starts with finding the highest frequency $m$ words in the dictionary definition to cover as many dictionary words as possible (Previous experience: $m \approx 2k$). Unlike the specialized dictionaries employed by DictSKB, we utilized WordNet[2] and Wikipedia[3] as the base corpus for building SKB. WordNet contains 0.2 *million* pairs of word senses and definitions, while Wikipedia contains 6 *million* pairs of words and detailed explanations. Our approach is divided into two modules. Since the word definitions in WordNet are shorter and more precise, we first find the initial Sememe from WordNet and construct a WordNet-based SKB by matching the annotated words in WordNet. Then we expanded it to Wikipedia based on the WordNet-based SKB (Like Figure 3).

**Table 1** Statistics of WordNetSKB, WikiSKB & WikiSKB-DA. WikiSKB-DA is a data augmented version of WikiSKB, and compared with HowNet and DictSKB. The gray font represents the previous SKB results.[6] #AvgSem denotes the average Sememe number per sense, and "+" represents this average over four.

| SKB | #Word | #Sense | #Sememe | #AvgSem |
|---|---|---|---|---|
| HowNet | 50,879 | 111,519 | 2,187 | 2.26 |
| DictSKB+ | 70.218 | 105.160 | 2,046 | 6.03 |
| DictSKB | 70.218 | 105.160 | 1,682 | 2.04 |
| WordNetSKB | 121,697 | 163,340 | 2,000 | 1.83 |
| WikiSKB | 385,336 | 423,249 | 1,807 | 2.12 |
| WikiSKB-DA | 697,754 | 800,458 | 1,992 | 3.46 |
| WikiSKB-DA+ | 697,754 | 800,458 | 1,992 | 5.73 |

#### 2.1.1 WordNet-based SKB

First, we remove the stop words[4] and meaningless characters and used the entity linking tool TAGME[5] to find the $2k$ most frequent topic words from HowNet word definitions as the base sememes. TAGME can identify meaningful short phrases in an unstructured text and link them to a relevant Wikipedia page. This annotation process has implications that go far beyond the enrichment of the text with explanatory links because it concerns contextualization and, in some way, the understanding of the text. A case in point, the definition "*A husband is a male in a marital relationship.*" is semantically enriched by the relations with the entities "*husband*", "*male*" and "*marital relationship*". We then used the *link_probability* parameter provided by TAGME to rank the entities in the word's definition and kept only the first four entities with the highest probability as sememes. We compared the filtered results with DicSKB and HowNet (In Table 1).

#### 2.1.2 Wikipedia-based SKB

The Wikipedia-based SKB construction is roughly the same as WordNetSKB. The difference is that Wikipedia's explanation is too detailed, and in addition, Wikipedia does not have semantic sense concepts, which significantly increases the noise of constructing SKB. To solve these problems, we took the following trick,

- Only the first two sentences of each entry in Wikipedia are adopted.

[2] https://wordnet.princeton.edu/
[3] https://dumps.wikimedia.org/
[4] https://code.google.com/archive/p/stop-words/
[5] https://sobigdata.d4science.org/group/tagme/
[6] The same is true for the following tables.

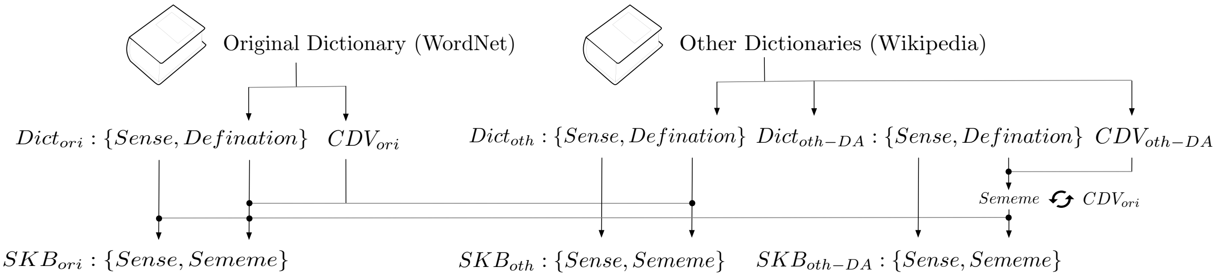
Original Dictionary (WordNet)
Other Dictionaries (Wikipedia)
$Dict_{ori} : \{Sense, Defination\}$
$CDV_{ori}$
$Dict_{oth} : \{Sense, Defination\}$
$Dict_{oth-DA} : \{Sense, Defination\}$
$CDV_{oth-DA}$
Sememe ⟲ $CDV_{ori}$
$SKB_{ori} : \{Sense, Sememe\}$
$SKB_{oth} : \{Sense, Sememe\}$
$SKB_{oth-DA} : \{Sense, Sememe\}$

**Figure 3** SKB Expansion Flow Chart: We use WordNet as the original dictionary and lexical expansion through Wikipedia, sharing a sememe set (As $CDV_{ori}$) between them. From the right side of the illustration, we used the look-up table method to replaced the $Definition$ of $Dict_{oth-DA}$ with the $Sememe$ of $SKB_{ori}$, which is a straightforward operation.

- Traverse each word in the Wikipedia entry and use NLTK to terms lemmatization.
- Use TF-IDF to remove words below the threshold from the definition. (We set the lowe bound = 4)
- Use the polysemantic annotations in Wikipedia as the sense.
- Delete the senses of polysemous words containing the meaning associated with “*film*” , “*novel*” , “*album*” , “*song*” , “*band*” , “*name*” , “*ep*” , “*game*” , “*surname*” and “*tv series*” .
- Keep only Wikipedia entries with a one-word title.

The WikiSKB was then constructed using the same $2k$ sememes as WordNetSKB. The statistics in Table 1. Since WordNet-based sememes do not cover the Wikipedia entry definitions perfectly, we reconstructed the entries that were not covered. In detail, we first used TAGME to extract the entities of each entry and adopted the $2k$ most frequent entities as sememes according to the same method as WordNetSKB. Then we found these words with the same entities from WordNetSKB and replaced them with the sememes of these words. We also put the statistics of the augmented version for WikiSKB in Table 1.

### 2.1.3 Sememe-based Graph Embedding

Our idea about the evaluation of sememe is to construct a bipartite graph by linking words with sememe to learn the embedding representation of words and then evaluate the quality of sememe by assessing the quality of word embedding (Like Figure 4). We employ second-order similarity of LINE[14] to train the node vectors regarding the graph embedding model, which is fast and intuitive. In detail, the probability of generating a neighbor node $v_j$ given a node $v_i$ can be expressed in the following form:

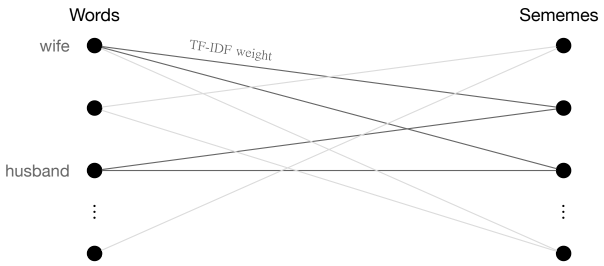

**Figure 4** Sememe based bipartite graph: We only learn the embedding representation of words by sememe, which means there is no line between words; we use TF-IDF for the weights of edges. The figure shows that “$wife$” and “$husband$” contain the same sememe, which should have approximate embedding representations. Note that here we do not consider the embedding representation of the sememe, so we use index instead of the word of the sememe itself.

$$p\left(v_j \mid v_i\right) = \frac{\exp\left(\vec{u}_j^{\prime T} \cdot \vec{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(\vec{u}_k^{\prime T} \cdot \vec{u}_i\right)} \tag{1}$$

Where $\vec{u}$ and $\vec{u}'$ are denoted as the vector of $v$ itself and $v$ when it is a neighbor, respectively. The empirical probability as $\hat{p}\left(v_j \mid v_i\right) = \frac{w_{ij}}{d_i}$, where $w_{ij}$ is the weight of the edge $i, j$; and $d_i$ is the degree of vertex $i$. If we define the importance factor $d_i$ of the node, the loss function can be defined as

$$-\sum_{i \in V} d_i KL_{divergence}\left(\hat{p}\left(\cdot \mid v_i\right), p\left(\cdot \mid v_i\right)\right). \tag{2}$$

## 3 Evaluations

This section uses two methods to evaluate our SKB: a collaborative filtering method[15] in subsection 3.2 and an approach based on constructing sememe graphs in subsection 3.2.

**Table 2** The results on Consistency Check of Sememe Annotations: MAP score of WordNetSKB exceeds the DictSKB, which we indicated in **boldface**.

| SKB | MAP | F1 |
|---|---|---|
| HowNet | 0.93 | 0.91 |
| DictSKB+ | 0.88 | 0.86 |
| DictSKB | 0.95 | 0.91 |
| WordNetSKB | **0.96** | 0.87 |
| WikiSKB | 0.95 | 0.86 |
| WikiSKB-AD | 0.93 | 0.90 |

**Table 3** The results on the Sememe graph: We merged WordNetSKB and WikiSKB. These tasks provide human scoring of the relationship between two words, thus assessing the degree of positive word relatedness. The method is to first calculate the cosine similarity of the two words and then compare them with the manual tags to calculate the Spearman correlation coefficient[8] for scoring.

| Similarity Tasks | WordNet&WikiSKB | WordNet&WikiSKB+ |
|---|---|---|
| WS-353-ALL[16] | 36.23 | 40.00 |
| WS-353-SIM | 41.66 | 45.12 |
| WS-353-REL | 22.70 | 30.26 |
| MC-30[17] | 26.32 | 31.19 |
| RG-65[18] | 28.71 | 32.57 |

### 3.1 Evaluate on Consistency Check of Sememe Annotations

This method is motivated by the idea that semantically close senses should have similar sememes. It actually implements a sememe prediction process that predicts sememes for a small proportion of senses according to the sememe annotations of the other senses. We have evaluated our SKB using open source code[7]. For hyperparameters, we set the same hyperparameters as the original paper[11], the number of evaluating epochs is 10, the threshold is 0.8, and the descending confidence factor is 0.93. The evaluation results are in Table 2. We discovered that the accuracy decreases with the increase of the dictionary lexicon. It is intuitive that the dictionary has many synonyms, while the sememe is static.

### 3.2 Evaluate on Sememe Graph

Our ultimate goal is to build a large SKB, so we combined the knowledge of both SKBs and evaluated them on some test sets (Shown in Table3). We first performed node embedding training using LINE[9], where the size of the node embedding is 200 dimensions, the total number of training samples is 100 million, the starting value of the learning rate is 0.025, the number of negative samples is 5, and we only used second-order proximity for training. SKB is sense-based, but each word in Word Similarity Tasks has only one meaning. Since the sense of words in Wikipedia is huge, we keep only *disambiguation* and *noun* sense for each word, and found that a higher number of sememe for a sense is a more accurate representation of the meaning. It is undoubtedly. Since we use only about four sememes to represent the meaning of a sense, it may lose some semantics, but the essential meaning can be kept. We released the data to reproduce the results.[10] Note that since the previous SKBs (As HowNet&DictSKB) did not provide a weight parameter for each sememe, we cannot compare the previous SKBs. However, to demonstrate that our SKB can better represent specialized words we extracted some specialized words in the Ohsumed dataset[11] for comparison (Shown in Table 4). Note that we boost the number of sememe to 5,000 to maximize the lexicon, other parameters are the same as before, and the final lexicon size of our SKB-DA is 910,369.

**Table 4** Sememe comparison on the Ohsumed dataset. Where "*non*" means no sememe of the word, we have merged WordNetSKB and WikiSKB+ as SKB-DA. More examples can be found in Appendix Table5.

| Word | SKB | Sememe |
|---|---|---|
| clostridium | DictSKB | {cause, illness} |
| | SKB-DA | {bacterial, cell, swollen} |
| colitis | DictSKB | {cause, illness} |
| | SKB-DA | {colon, inflammation} |
| pediatric | DictSKB | non |
| | SKB-DA | {care, child, medical} |

## 4 Conclusion

This study focuses on the lexical expansion of SKB, which expands the vocabulary and dramatically increases the lexical sense. It helps in the semantic understanding of particular domain text classification and some downstream tasks. This paper's proposed method of reconstructing dictionary definitions can effectively expand the original SKB and have promising results on some internal evaluation metrics. In future research, we will use sememe to perform short text classification tasks or use SKB knowledge to apply to Abstract Meaning Representation to improve the accuracy of downstream tasks.

[7] https://github.com/thunlp/DictSKB/tree/main/ConsistencyCheck

[8] https://github.com/scipy/scipy/blob/v1.7.1/scipy/stats/stats.py#L4343-L4525

[9] https://github.com/tangjianpku/LINE

[10] https://github.com/SauronLee/SKB-DA

[11] Ohsumed dataset: it includes medical abstracts from the cardiovascular diseases. http://disi.unitn.it/moschitti/corpora.htm

# References

[1] Leonard Bloomfield. A set of postulates for the science of language. **Language**, Vol. 2, No. 3, pp. 153–164, 1926.

[2] Anna Wierzbicka. **Semantics: Primes and universals: Primes and universals**. Oxford University Press, UK, 1996.

[3] Zhendong Dong and Qiang Dong. **Hownet and the computation of meaning (with Cd-rom)**. World Scientific, 2006.

[4] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In **Proceedings of the 48th annual meeting of the association for computational linguistics**, pp. 216–225, 2010.

[5] Fanchao Qi, Liang Chang, Maosong Sun, Sicong Ouyang, and Zhiyuan Liu. Towards building a multilingual sememe knowledge base: Predicting sememes for babelnet synsets. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 8624–8631, 2020.

[6] Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. Language modeling with sparse product of sememe experts. **arXiv preprint arXiv:1810.12387**, 2018.

[7] Fanchao Qi, Junjie Huang, Chenghao Yang, Zhiyuan Liu, Xiao Chen, Qun Liu, and Maosong Sun. Modeling semantic compositionality with sememe knowledge. **arXiv preprint arXiv:1907.04744**, 2019.

[8] Lei Zheng, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. Multi-channel reverse dictionary model. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 312–319, 2020.

[9] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. **arXiv preprint arXiv:1910.12196**, 2019.

[10] 李笑然, 高野敏明. The analysis about building cross-lingual sememe knowledge base based on deep clustering network. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 92 回 (2021/9), p. 06. 一般社団法人 人工知能学会, 2021.

[11] Fanchao Qi, Yangyi Chen, Fengyu Wang, Zhiyuan Liu, Xiao Chen, and Maosong Sun. Automatic construction of sememe knowledge bases via dictionaries. **arXiv preprint arXiv:2105.12585**, 2021.

[12] Sidney I Landau. Bt sue atkins and michael rundell. the oxford guide to practical lexicography., 2009.

[13] LIU Zhiyuan SUN Maosong LIU Yangguang, QI Fanchao. Research on consistency check of sememe annotations in hownet. Vol. 35, No. 4, p. 23, 2021.

[14] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In **Proceedings of the 24th international conference on world wide web**, pp. 1067–1077, 2015.

[15] Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, and Maosong Sun. Lexical sememe prediction via word embeddings and matrix factorization. In **IJCAI**, pp. 4200–4206, 2017.

[16] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. 2009.

[17] George A Miller and Walter G Charles. Contextual correlates of semantic similarity. **Language and cognitive processes**, Vol. 6, No. 1, pp. 1–28, 1991.

[18] Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. **Communications of the ACM**, Vol. 8, No. 10, pp. 627–633, 1965.

**Table 5** Sememe comparison on the Ohsumed dataset. Where "*non*" means no sememe of the word, we have merged WordNetSKB and WikiSKB+ as SKB-DA, note that the SKB-DA* with an asterisk indicates that this sememe is from WikiSKB. We extracted only the first 15 sentences of specialized words in the Ohsumed. When constructing SKB-DA, we did not purposely adjust the medical-related words, but it performed well. We found that DictSKB has insufficient vocabulary and words with similar meanings with the same sememes. e.g., "*clostridium*" and "*colitis*". In this way, it is questioning to classify words effectively in the downstream task of natural language processing.

| Word | SKB | Sememe |
|---|---|---|
| clostridium | DictSKB | {cause, illness} |
| | SKB-DA | {bacterial, cell, swollen} |
| colitis | DictSKB | {cause, illness} |
| | SKB-DA | {colon, inflammation} |
| pediatric | DictSKB | non |
| | SKB-DA | {care, child, medical} |
| thoracic | DictSKB | {neck, part} |
| | SKB-DA | {chest} |
| empyema | DictSKB | non |
| | SKB-DA | {body, cavity, lung, pu} |
| diagnosis | DictSKB | {wrong} |
| | SKB-DA | {identify, nature, phenomenon} |
| helicobacter | DictSKB | non |
| | SKB-DA | {bacteria, gram, negative, shape} |
| infection | DictSKB | {disease, someone} |
| | SKB-DA | {body, invasion, microorganism, pathogenic} |
| salmonella | DictSKB | {make} |
| | SKB-DA | {Gram-negative, bioweapon, fever, food, poisoning, rod-shaped} |
| cerebrospinal | DictSKB | non |
| | SKB-DA | {brain, cord, spinal} |
| rhesus | DictSKB | non |
| | SKB-DA | {Asia, medical, southern} |
| mangabey | DictSKB | non |
| | SKB-DA | {arboreal, eyelid, limb, monkey, tail, white} |
| splenic | DictSKB | non |
| | SKB-DA | {spleen} |
| tissue | DictSKB | Sense1: {nose, paper, piece} |
| | | Sense2: {paper, use, wrap} |
| | | Sense3: {cell, form} |
| | SKB-DA | Sense1: {cloth, cotton, fabric, interlace, piece, strand, wool} |
| | | Sense2: {paper, soft, translucent} |
| | | Sense3: {cell, function, organism, structure} |
| itraconazole | DictSKB | non |
| | SKB-DA* | {fungal, infection, medication, mouth, treat} |
| phaeohyphomycosis | DictSKB | non |
| | SKB-DA* | {cell,characteristic,diverse,fungi,infection,tissue,yeast} |
| dermatophyte | DictSKB | non |
| | SKB-DA* | {chlorophyll,evolution,feed,fungi,fungus,protective,sac,spore-bearing,unicellular,vascular} |
| percutaneous | DictSKB | non |
| | SKB-DA | {cream, form, medication, ointment, patch, skin} |
| venous | DictSKB | {carry} |
| | SKB-DA | {function, vein} |
| catheterization | DictSKB | non |
| | SKB-DA | {body, operation} |
| subspecialty | DictSKB | non |
| | SKB-DA* | {field, knowledge, medical, professional, skill, trade} |
| tinea | DictSKB | non |
| | SKB-DA | Sense1 {fungi, infection, nail, patch, skin} |
| | | Sense2 {genus, moth, type} |
| candidiasis | DictSKB | non |
| | SKB-DA | {fungi, genus, infection} |
| immunization | DictSKB | protect |
| | SKB-DA* | {agent, immune, infectious, process} |