Chengyu Dictionary Documentation

Xuefeng Luo Jingwen Li

Contents

List of Figures

List of Tables

1 Introduction

The Internet has promoted globalisation and along with it, the urge (or necessity) to be multilingual. Following this trend, Chinese learners worldwide have also increased in number. As an inseparable part of the Chinese language, *Chengyu*, or the so-called "four-character idioms", is on the other hand difficult to master. One reason behind this is that there are simply so many of them and it is hard for a learner to learn them in chunks. Most of the *Chengyu* are learned as they are encountered, which surely satisfies most learners practically, but not the hungry ones. To help the intermediate/advanced Chinese learners, we present this Chengyu Dictionary as an entry-point to tame these culturally rich beasts.

This paper is presented in the following structure:
Section 2 talks about the existing tools and a paper on translating Chengyu into English. Section 3 lists the project dependencies. Section 4 is installation instructions which are also in the README.md file, Section 5 is about data processing, including preprocessing the corpora, translation difficulties and other lexicographic decisions made. It also deals with the search mechanism and database with technical details. Then, in section 6, we present the user interface. Section 7 recorded the division of labor and section 8 is our discussion of this tool and vision for future improvements.

2 Related Works

The following are existing tools that share similar functions with our web application:

- Chengyu Dictionary provided by *Chinese-Tools.com*
  Chengyu Dictionary is an online dictionary that provide pinyin, explanations in English (only for some entries) and detailed explanations in Chinese for 30000 Chengyu. The information provided in Chinese includes meaning, context(origin), example, synonyms, antonyms and even grammatical information. A user may search by pinyin or Chinese character input, but not English. The results are displayed in a list of links which lead to the entries' own page, where the above information are presented.

  Available at: https://www.chinese-tools.com/chinese/chengyu/dictionary

- lists of Chengyu that are introduced on websites, for example *List of 148 well-known Chengyu*, available at: https://www.saporedicina.com/english/list-chengyu/
  There are a lot of similar resources like this on the Internet. They are mainly summaries of frequently encountered Chengyu by Chinese learners, so they differ

individually. Upon further inspection, they are not listed in the most-frequent-to-least-frequent manner, but overall these are very common Chengyu. Information provided for each Chengyu includes Chinese characters, pinyin and meaning in English.

There is also a paper on translation strategies: *A Study of Idiom Translation Strategies between English and Chinese*
This paper discussed the different strategies that are commonly used when translating idiomatic phrases, Chengyu in particular, into English. The four main strategies are: literal translation, free translation, abridged translation and borrowing translation. It also outlined the principles to live by when translating idioms.

3 Project Dependencies

A list of project dependencies used in this project (managed by Maven):

- Google Web Toolkit and Mojo's Maven Plugin for GWT (gwt-maven-plugin, version 2.8.1)
- mysql-connector-java(version 8.0.13)
- GwtBootstrap3 and GwtBootstrap3-extras (gwtbootstrap3 and gwtbootstrap3-extras, version 0.9.4) for button group, button dropdown and multiple select.

4 How to install

To install our application, please follow the listed instructions:

1. First, you need an environment where GWT can be deployed. At this moment, we are using Eclipse plus a GWT plugin where we retrieved it from Eclipse marketplace.
2. Then, MySQL is needed in order to import our data.
3. MySQL Workbench is optional but it helped comparing to MySQL command line tools.
4. To import our data, you need to go to data folder and run data_structure.sql file where our data structures and relations were stored.
5. Then, you should first import chengyu data from chengyu_data.json file and tags data from tags.csv file, since there are foreign key dependencies.
6. Then, you can import chengyu tags relations from chengyu_tag.csv file.

7. Following that, you need to go to DictionaryServiceImpl.java
   (located in chengyu.dict\src\main\java\com\colewe\ws1819\server)
   where database connection information was hard-coded in. Variables DB_URL, USER and PASS need to be modified in order to make our program connect to the database.
8. Please run the application through Eclipse.
9. Open the page http://127.0.0.1:8888/semesterproject.html via a browser. We have tested and passed with Safari, Google Chrome and Firefox.
10. Enjoy it!

5 Method

5.1 Pre-processing Steps

Currently, there was no ready-made data resources we could find. Therefore we found a Chinese Chengyu dictionary online resource, calculate all Chengyus' frequencies against the OpenSubtitles corpus as Dr.Johannes Dellert recommended, translated and tagged about 150 entries.

5.1.1 Chengyu Data

Available at: https://github.com/by-syk/chinese-idiom-db
These are Chengyu collected from the web, with a total count of 12976 entries. Each entry contains a Chengyu, its Pinyin. Other information includes explanation, origin and example sentences all in Chinese. However, some Chengyu are without origin and/or example sentences.

5.2 Frequency of Chengyu

In order to allow Chinese Chengyu learner to learn the most commonly used Chengyu, we count each Chengyu against their occurrence in the OpenSubtitles corpus, which contains subtitles of films. Although Chengyu usage is limited in this corpus, we managed to obtain a general idea of how they are used. To count the occurrence of each Chengyu, we created a separate Python program located in the data folder named chengyu_frequency.py.

5.3 Data Format

The original data we retrieved was in Comma-separated Value (CSV) formatting. Since we wrote a Python program to count frequency and Python support JSON, and since MySQL also support JSON regarding importing data, we used JSON as our data format (see data.json in the data folder). As for tags and chengyu-tag relations, we kept the format as CSV format (tags.csv and chengyu_tag.csv in the data folder) and for data structures, we remain in SQL format.

5.4 Linguistic Problems and Lexicographic Decisions

When we started out designing this Chengyu Dictionary, we had in mind a profile of our target user. We want to provide a tool for an intermediate/advanced Chinese language learner (who is also an English speaker) so that he/she can get a better grasp at Chengyu. At the same time, we think it is also a tool for any Chinese speaker who wants to refresh their Chengyu knowledge. Therefore the target language we used is both English and Chinese, but mainly focused on English (as can be seen from the user interface) because English speakers are our main intended users.
Then we thought about what they need from an online dictionary for idiomatic expressions. First of all, the expressions themselves as well as the pronunciation information (pinyin), then the meaning. But how do we present a meaning of an idiomatic expression without going into details of cultural background, conventional symbolism and even moral codes? This brings about the most difficult part of the construction of this dictionary, translation. Wang, Lanchun and Wang, Shuo (2013) presented in their *A Study of Idiom Translation Strategies between English and Chinese* 4 translational approaches: literal translation, free translation, abridged translation and borrowing translation. Literal translation, which translate the expression as it is laid out in the source language, has the merits of being true to the source language, not much information is lost, but at the same time one has to question how much of it can be perceived if not lost. Should the intended user make guessworks and use his/her full imagination without any aid to try to interpret why such metaphor is made in a Chengyu? On the other hand, if we were to discard the literal part completely, and give only the practical figurative meaning of a Chengyu to a learner, how is it different from a regular noun? It would become a longer sequence of characters that just encode this particular meaning, which be hard for learners to use them without sounding awkwardly. Therefore we decided to include both a literal translation and a figurative translation. However, we wanted to be as literal as we could so instead of just a literal translation, we did an almost character-by-character translation, in the hope that the intended user (intermediate/advanced), knowing the characters presumably, now understand what meaning it takes in this particular Chengyu, then with the figurative meaning also provided, hopefully he/she can make an educated guess about the connection in-between. We did not use the abridged translation strategy as we were trying to do character-by-character literal translation. We did, however, use borrowing translations if there are equivalent expressions in En-

glish that we know and used them in the figurative translation. We believe this can make better connections from Chinese to English, for the learners to be able to use them in the correct context. Chinese explanations are added to accompany the English translations, they are also good for Chinese users.
We also included origin (if there were an origin) of a Chengyu, and if possible an example usage of it. The frequency of the respective Chengyu in the OpenSubtitles corpus is shown at the bottom, as an attempt of indicating whether it is rarely- or commonly used in Chinese. However, in this version, it shouldn't be taken very seriously.
This is because written language and spoken language differ, and film language is more so. Film subtitles might subject to the limitation of space and time, people may use them in a somewhat unnatural way. Therefore our choice of the OpenSubtitles corpus might have affected the frequency of Chengyu observed, depending on the genres of the film and the language used in the film. One improvement to be made could be to incorporate different kinds of corpora. But then another problem may emerge as the literature used might be old and language might be archaic. In this case, since OpenSubtitles contains contemporary language used in film subtitles, this corpus could be a good choice, other corpora of the same time period may be used for improvement of calculating frequency.
Since we didn't want to translate all of the 12976 entries, we had to base our choice of Chengyu on this calculated frequency, although this is definitely better than choosing randomly or based on only two people's judgement, it is still not satisfying and should be improved.
The results are displayed in alphabetical order, as they would regularly do in a paper dictionary. The information is ordered in a way that the English part comes first. Then the Chinese explanation, origin and example. And last is the frequency (its place should be promoted for the improved version).

5.5 Search Mechanism

We split our search function into two parts. The initial search and a tag filter.

In our initial search, we first determine which mode the user choose and pass target and mode variables to our search logic class. In this class, we already made SQL statement templates and inserted variables into those statement and request database to search.

After database returning back, we retrieved the data and parsed the data into an arraylist of Entries. Then, we used the pre-made function to filter the results and sent them to the front.

Intuitively, we wanted our filter function takes place in the front-end to be written by JavaScript as the industrial standards usually do. However, we found GWT having problems with this method where user interface was designed in Java and written JavaScript was actually very hard to step in. Thus, we moved our tag filter function back to Java.

5.6 Database

We are using MySQL which is the world's most commonly used and free relational database and it is easy to use and maintain.

To design our data tables, we follow first, second and third normalization where we eliminated all repeating records, partial dependency and transfer dependency. Thus, we have Chengyu table and Tags table and one more associate table which stores the relationships of Chengyu and Tags where we have IDs for both Chengyu and Tags as primary keys and foreign keys.

Figure 1 shows the relationships of our tables.

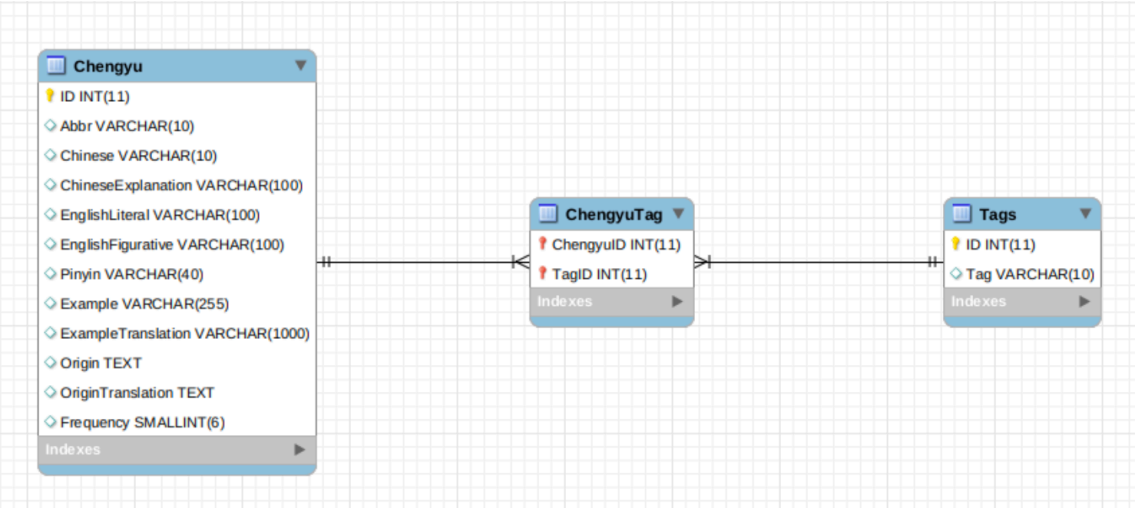

Figure 1: The relationships of our tables

6 User Interface

6.1 Three Search Modes

To search a Chengyu, you can directly search it by typing in its Chinese characters. If you have forgotten the Chengyu and remember only a part of it, you can search the Chengyu by typing in one of the characters contained in the Chengyu. However, you'll have to go through the results to find your desired Chengyu. In addition you may also search in pinyin.

As for English mode, sometimes you may want to search a Chengyu which has a specific meaning. In this case you can type in its English meaning. For example if you want

to search for Chengyu containing the meaning of *mistake* in it, you may select English mode, type in *mistake* and search.

To summarise, we designed these three search modes where you can search by Chinese, pinyin or English. The switches are displayed as a button group just above the input box, shown in Figure 2.

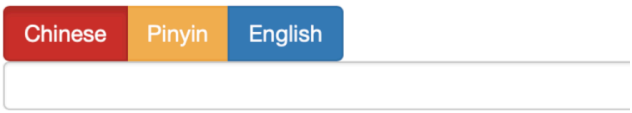

Figure 2: Three Search Modes

6.2 Tags

One interesting thing about Chengyu is that they love to use metaphors of animals, human body parts and numbers. Therefore we have character tags like *numbers* and *color* indicating these metaphors are incorporated in the Chengyu.

Another interesting thing is that there are certain patterns to form a Chengyu, such as ” 一心一意”, ” 三天两头”, ” 七上八下” where the first and third characters are all numbers. And ” 一生一世” is in the form of AXAY, where the first and third character are the same, and is also a number. Therefore this Chengyu has two tags: AXAY and numbers. Tags like *AXAY*, *AABB* are form tags.

Attitude is also important to Chengyu, a same meaning can be separated to different Chengyu regarding its attitude. For example, ” 臭味相投” and ” 志同道合” both mean similar people go together, but the former is used for bad persons such as gangsters and the later is used for normal people such as business partners. To further illustrate this point, let’s continue with our *mistake* example, if you performed the search mentioned above, you’ll get three Chengyu: ” 将功赎罪”, ” 万无一失” and ” 铸成大错”. You may expect all of them to mean something negative as they are related to mistakes. In fact, the first two are quite the opposite. The first one makes up a mistake and the second indicates no mistakes are made, therefore they each have a positive tag. Tags like *positive*, *negative* are the sense tags.

That is why, we designed this tag filter to help people narrow-down all kinds of Chengyu. To perform a tag search, you may use any of the three mode and input some character/pinyin/English word (this is necessary for the search) and select the tags you want. Multiple selection is enabled. However, notice that we have only annotated around 150 Chengyu so there won’t be many results if you select more than 2 tags at the same time. If you want to cancel you selection of a specific tag, just click it again and it will be unselected. You can either perform a tag search right from the beginning or narrow-down

you search result by selecting tags later. Just press the search button again after your selection. The advanced search button works only as a label.

Table 1 shows a table of Chengyu-tag examples and Figure 3 and 4 shows how it looks like in our application.

Chengyu	唇齿相依	虎头蛇尾	一心一意	七上八下	臭味相投	志同道合
Tags	Organs	Animals, Negative	Organs, Numbers, AXAY	Numbers	Negative	Positive

Table 1: Chengyu-Tags Examples

6.3 Results Display

Search results are displayed in the lower section of our page, shown as Figure 5, where each Chengyu takes a block and each column of information in the database takes a row in the block. Tags are attached as badges to the title.

6.4 Download

You may want to download your current search results as a XML format for future reference. This is performed by the Download button. It works exactly like Search button, but the results would not be displayed on the page. Instead, an XML file will be downloaded with the search results.

7 Division of Labor

We worked as a team for the majority of the project parts. However, we do have specialties regarding what we are good at and what we are not. Therefore, we make a list of all sections with percentages of what we have done.

- Data collection and Processing(Xuefeng 30% and Jingwen 70%)
- Database designs (Xuefeng 70% and Jingwen 30%)
- Front-end/client side coding (Xuefeng 50% and Jingwen 50%)
- Back-end/server side coding (Xuefeng 70% and Jingwen 30%)
- Project Report (Xuefeng 30% and Jingwen 70%)

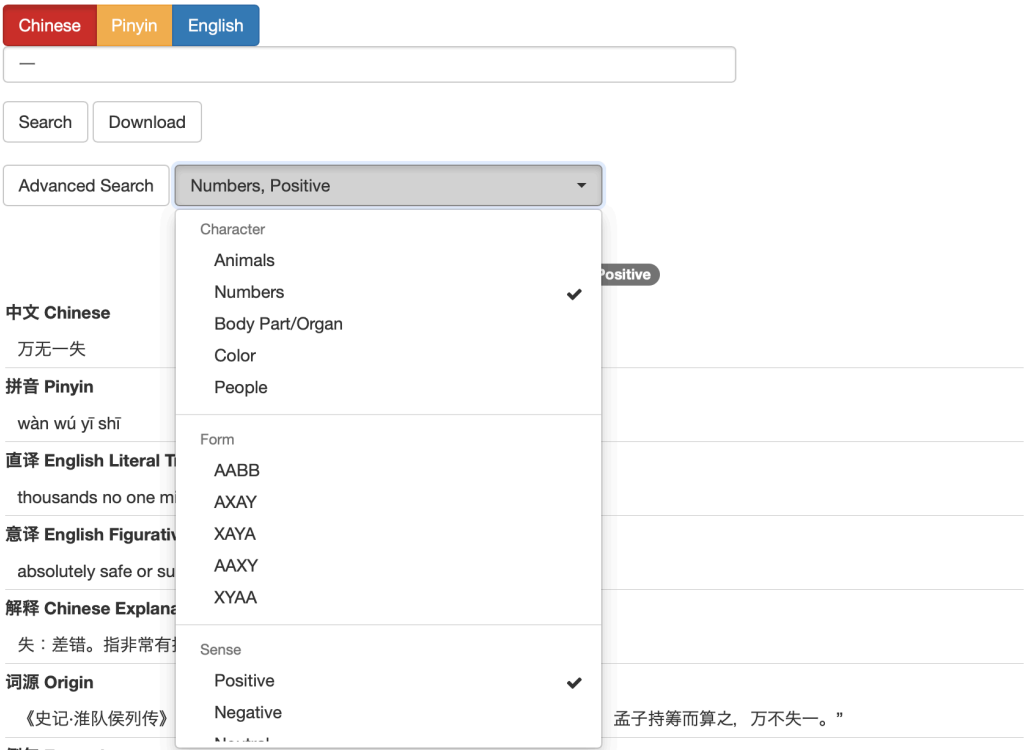

Figure 3: Tag Filter

Figure 4: Tags are displayed as badges attached to result title

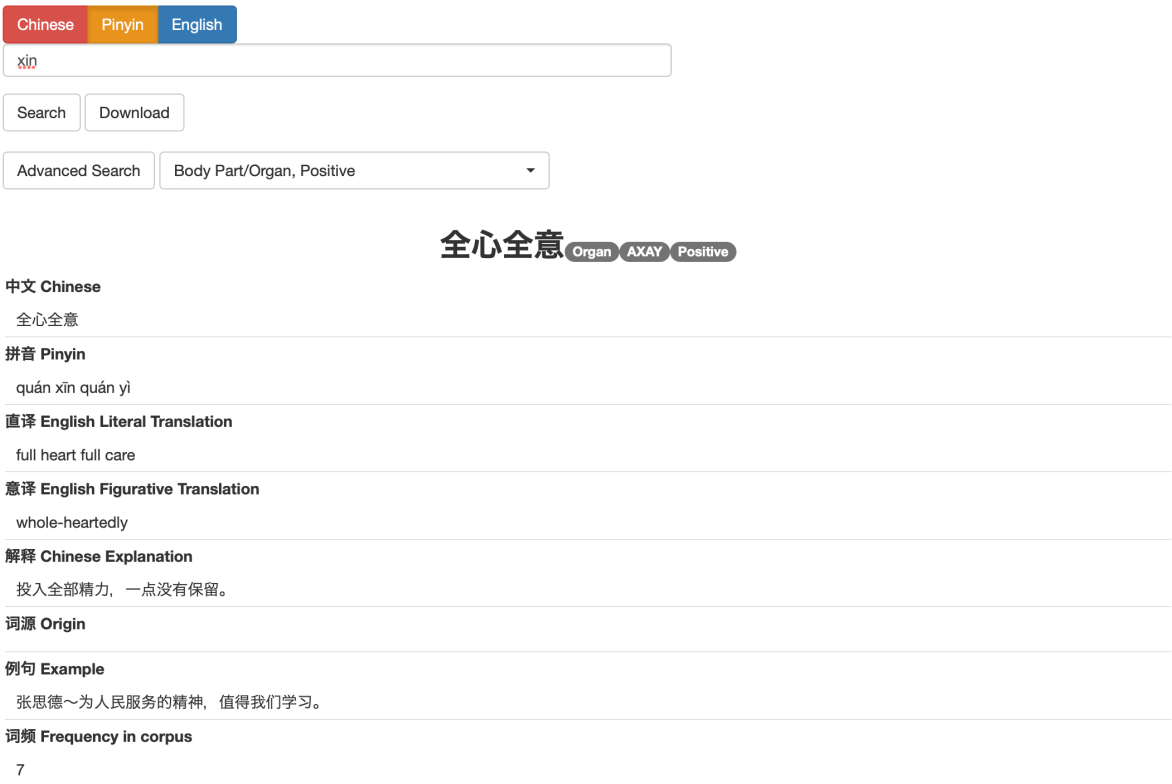

Figure 5: Search result for pinyin search mode, input ”xin”, select tags ”Body Part/Organ” and ”Positive”

8 Discussion and Future Work

8.1 Linguistic Discussion

As mentioned in Section 5.4, we can improve the estimation of how frequent a Chengyu occurs and determine its frequency rank by using more diverse corpora.
Also, given more time, the English translation of the origin of a Chengyu could be added. This again is a daunting task because the origins are quite often written in ancient Chinese. Even a Chinese speaker might need to consult a Traditional Chinese Dictionary to make sure he/she understands it correctly and he/she must also have the ability to translate it into English.
Language-wise, we have also noticed that some of the frequently-used Chengyu are not strictly speaking a Chengyu at all. Because it may just be an fixed repetition of a 2-character word or a fixed abbreviation for something said or even a fixed noun-modifier phrase. It lacks the sense of a traditional Chengyu and is more of a common usage. But because it is already fixed into a 4-character, or even more character, combination, it has been granted the Chengyu status. This is quite an interesting phenomenon as traditionally Chengyu has to come from some story that happened in the past, or a very famous saying that is unbreakable, nowadays people seem to condense daily language and squeeze them into the 4-character box to make it a Chengyu. For example the expression " 步调一致" can be broken into 2 parts, " 步调" and "一致" , the former means 'the pace of walking' and the latter means "to be the same", which make the whole expression just mean 'walk(noun) at the same pace(modifier)'. This is more of an invention that is consolidated through political slogans. But then should they not be considered as Chengyu? The difference seems to be, we used to condense 4 words together to indicate something different or something more general, but now the newcomers seem like fixed combination of words that happen to have 4 characters, which one can definitely find as many as one want because Chinese words now are mostly 2-character words.

8.2 Technical Problems

One of the most discouraging thing is that GWT would not re-compile a file after we modified them, specially for our CSS file and Entry file. Currently, we could not find out why this happens and how to fix it. Our temporary solution to this was to delete all files and re-clone and re-import them.

Another problem is that GWT seems not to work well with front-end framework such as BootStrap. Currently, we require a BootStrap plug-in for GWT which was already out-of-date and has limited supports for BootStrap components. This is why we want to do more front-end design but we were not able to do so.

Thus, if we had another chance, we would probably move our project to another frame-

work, such as Spring for Java or Laravel for PHP. Those frameworks are not only supporting all web application technologies but also have a large community group to discuss and maintain them.

9 References

By_syk, 2017, 成语数据 Chinese Idiom data. From: https://github.com/by-syk/chinese-idiom-db

P. Lison and J. Tiedemann, 2016, OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)

Wang, Lanchun and Wang, Shuo, 2013, A Study of Idiom Translation Strategies between English and Chinese. Theory and Practice in Language Studies. 3. 10.4304/tpls.3.9.1691-1697.