Auto-extracting Paraphrases of Letter-word Phrases in Live Texts[1]

Zezhi ZHENG

Dep. of Chinese Language & Literature, Xiamen University, Xiamen 361005
Zhengzz@xmu.edu.cn

Abstract: In this paper we will discuss the Auto-extraction of paraphrases of letter-word phrases in live Chinese texts. The paper discusses the modes of conventional dictionaries firstly, and then gives the principles of paraphrase of letter-word phrases; with an analysis of the examples of letter-word phrases paraphrases secondly, and then gives their formalized denotations and presents an auto-recognizing algorithm for bilingual synonymous letter-word phrases; lastly, based on the labeled result of our auto-labeling software of letter-word phrase, uses the vector space distance to extract the paraphrase of letter-word phrases in live Chinese texts.

Key words: Phrases; Letter-word phrase; Auto-extract

1 introduction

Traditional dictionaries have the veracious meaning expression, and rational, normative language, and better systematic semantic systems. The paraphrases in dictionaries are generalized from the collection of example sentences which are gathered from the fields that seemed most feasible. Therefore compiling a dictionary always cost a great deal manpower, material resources and time. The renewed period of traditional dictionaries has not adapted to the increasing and changing of new words and new meanings, so using computers to help the compile of dictionaries is brought forward.

About the extraction for the definition, [zhang Yan, 2003] instructed a definition extraction means for Chinese terms, which based on Chinese syntactic parsing. They segmented and tagged the corpora of computer domain with part-of-speech firstly, then used two parsers to gain structures and phrases of sentences. They summarized the structures characteristics of term definitions and automatically extracted the patterns of definitions. Finally they gave an algorithm to define a new term according to their knowledge database. Their algorithm to define a term needs concept semantic knowledge and other knowledge, and segmented and tagged texts. In 2004, XU Yong et al presented an experiment of web based term definition retrieval system. For a given term, the system used the algorithm based on term definition patterns with indicated words to extract its definition. The two means of definition extraction are all for given terms, and have available to some extent.

We think that definitions are for terms, the paraphrases are suitable for letter-word phrases (hereinafter we will use "LWP" to denote "letter-word phrase") in live Chinese texts in general field. Therefore in this paper we'll discuss the principles of paraphrases of LWPs firstly, and then generalize their formalized patterns based on the collection of paraphrases of LWPs. Secondly, we'll analyse the especial paraphrases, bilingual synonymous LWPs, and present an algorithm to obtain this kind paraphrases. Thirdly, we'll label the LWPs using our auto-labeling software, and then put forward an algorithm to extract the paraphrases of LWPs.

2 The principles of our paraphrase of letter-word phrases

LWPs, especially those prevailed in general media, many of them are terms or proper noun, and most readers have no acquaintance with them, so they want to know the meaning of them. With their

[1] Sustentation fund: science fund (code: 60475022), and XIAMEN University start-up fund.

increasing quickly, we have to use machines to help extracting their paraphrases. But what is the extent? What are the principles?

With analyzing different requirements of dictionaries and investigating paraphrases of LWPs in live texts, we would conclude our interpretation criterion.

From the book by FU Huiqing, and the paper by LI Xiyin, we can conclude that traditional dictionaries use about four forms to interpret a word.

a) Genus name and differentiae addition. This pattern is used usually to define a term. The differentiae would be properties, reasons, degrees, functions, effects, or extents etc..
b) Synonymous words interpretation. Interpreting a word with its synonymous words.
c) Translation with bilingual synonymous words.
d) Depiction, explanation. Depicting or explaining a word with its phenomenon, characters, relations, parable etc.

[ZHENG Shupu, 2005] Russian Scelba considered that cyclopedias and common dictionaries are different in defining a word. Common dictionaries provide an interpretation to help readers to understand the meaning of a word, not a complete interpretation of the word. And long before ,Russian term scholar Liefu'ermaciji presented that terms in textbooks or in news papers are secondhand, which are derived from scientific terms, their interpretation demands are different from scientific term system, for they just relate to some keywords, and don't relate to whole term system.

The LWPs in the paper were from general media, most of them were terms or proper nouns. According to the argumentation above, we think that paraphrases of LWPs are not the cyclopaedic definitions, they are generally common interpretations. In fact, for most readers it is better to give an interpretation like "a standard of information communication.", "a sort of computer virus." than a scientific definition. Thereinafter, our paraphrases of LWPs are common dictionary interpretations.

3 The paraphrases patterns of letter-word phrases in live texts

Our paraphrases extracting process is based on no segmented and tagged texts, we have to find out LWPs firstly, and then retrieval the paraphrase sentences.

The paraphrase examples are a byproduct of our collating software, videlicet, during our collating we can enregister the encountered interpretations by the record function of the software. So we get hold of 400 paraphrases of LWPs. Some of these paraphrases are repeated, but they were from different texts, and with different contents, with different points of view. They are beneficial for us to find patterns of paraphrase.

Through investigation, we find that there were two types paraphrases of LWPs, one has steady forms, the other has no form. The former paraphrases with mark words. Such as:

FAD2 是一种将饱和脂肪酸转化成不饱和脂肪酸的酶，对哺乳动物生长具有重要作用，在菠菜等植物中存在。但是哺乳动物，包括人类体内不含这种酶，必须从食物中摄取。

可怕的电磁武器家族中还有"杀人雷达"，简称 RF/MO 武器。它是功率特强的射频、超高频或微波武器，其发射功率从几亿瓦到几十万亿瓦不等。根据所发射电磁波不同的频率、调制方式和功率强度，RF/MO 武器可以在人体的不同组织或器官产生不同的效应。人体最易遭受电磁武器攻击的组织和器官是大脑、脖子、胸部和生殖腺。受到攻击时的症状是身心疲惫、记忆紊乱、皮肤生病、眼睛出血、白内障、角膜和视网膜损伤，甚至罹患癌症。

The "LWP是一种……的……"，"LWP，它是/是……"are mark words. The latter interprets a LWP using its contexts without mark words, and readers may see the meaning from the contexts, such as:

据《日刊工业新闻》报道，硼同位素有 B10 和 B11，二者的存在比例为 20∶80。B10 能吸收中子，用中子束照射浓缩硼，会产生阿尔法射线。它的能量足以杀死癌细胞，但又不伤害正常组织，因而不会产生副作用。

除来自英国、加拿大、德国、法国、意大利、日本、美国和俄罗斯的 8 国领导人外，非洲的南非、尼日利亚、埃及、塞内加尔、安哥拉 5 国的领导人也将参加会议。这在 G8 首脑会议历史上还是第一次，凸显本次会议的非洲议题。联合国秘书长安南也将参加会议。

This sort LWP may have deep structure, but now we haven't do farther analysis. This paper we only study these LWP's paraphrases with steady forms.

From our invesitigation, we find that just three types of common dictionarie paraphrases are shown in the paraphrases of LWP, such as genus name and differentiae addition, translation with bilingual synonymous words and depiction, explanation.

The paraphrases of LWP, which could be formalized, can be divided into six groups, such as:

a) the form with "是", like "LWP 是一种……", "LWP 是……的缩写|简称，是一种……", etc.
b) the form with "即", like "……LWP 即|，即……"
c) the form with "名为", like "研制出|推出的|生产的|发现的名为 LWP……的……"
d) the form with "称为", like "被称为|简称为|称为 LWP ……的……"
e) translation with bilingual synonymous words, like "中国石油集团物探局（BGP）,世界贸易组织（WTO）"
f) dash paraphrases, like "……——LWP……".

We don't differentiate between simple interpretations and complete definitions in extracting process. We obtain interpretation sentences from LWP whereabouts to next sentence, for most interpretations of LWP are covered to this extent.

We find that the two types interpretation, translation with bilingual synonymous words (we call them bracket interpretations) and "是" structure interpretation, take about 90 percent by our investigation, so this time we just extract the two types interpretation.

4 Interpretations of translation with bilingual synonymous words

In Chinese texts bilingual LWPs can be classified as follows:

A：complete Chinese character+(complete English corresponding words)
B：complete Chinese character+(English abbreviation)
C：English abbreviation LWP+(complete English)
D：English abbreviation LWP +(complete Chinese corresponding words)
E：bilingual LWP+(complete English or English abbreviation)
F：complete English LWP+(English abbreviation)
G：complete English LWP +(Chinese corresponding words)
H：bilingual LWP +(Chinese or bilingual words)

For instance:

A：全球环境基金（Global Environment Facility，简称GEF）
　　无线局域网（WLAN，Wireless Local Area Network）
B：全球环境基金（GEF）
　　国际标准化组织（ISO）
C：TD—SCDMA（Time Divided Syn－chronization CDMA）
　　GEF（Global Environment Facility）
D：DOI（数字对象标识符）技术
　　COD（生化耗氧量）
　　"M—office"（商务干线）
E：黑客QQ（Hack.OicqHack）
　　X射线扫描断层检查仪（CT）
F：National Council for Social Security Fund，PRC（缩写：NaCSSeF）
G：EVALUATION BASED ON RESEARCH（基于研究的评估）
　　EXCHANGE INFORMATION（交流信息）
H：奥运艺术Swatch（斯沃琪）
　　ERP软件（enterprise resource planning，企业资源管理）
　　北京CBD（商务中心区）

About this type interpretation, we designed a module to process.

In order to extract this type interpretation, we must consider three didymous punctuations, such as "《补贴与反补贴措施协定》（"《SCM 协定》"）", it contains 书名号， marks and brackets. We design a code system firstly, for codes are simple and contain more information than texts, and then an algorithm was put forward based the code system.

a) The code system

this system memorizes the state about three didymous punctuations used in a bilingual LWP with six codes. Thereinto, the former twain codes denote brackets, the middle twain codes denote quotation marks, the latter twain codes denote 书名号. For every twain codes, the first one denotes whether one didymous punctuation appeared in a LWP, it has three meaning, such as "1 " means the left punctuation appeared, "2" means the right punctuation appeared, and "3" the left or the right punctuation don't appeared at first position of the LWP. The second one denotes position of one punctuation, it has three forms and four states, "1" the left punctuation appeared at first position of the LWP, "2" means the right punctuation appeared at first position of the LWP, "0" means one punctuation don't appeared or had appeared in couples. For example, "113131" can denotes a left bracket appeared at left first position, and a left quotation marks and a left "《" appeared at other positions, such as "（"《SCM". "103030" mean three didymous punctuations appeared in couples, such as "（"GATT1994"）".

b) The process flow

The first step, we need to scan a letter string in one text, and transfer the function iskuohao（）to encode the letter string with our code system. The second step, according to the codes, judge whether any punctuations had not appeared in couples, and if there are such punctuations, then transfer the function iskq（）to scan two sides of the letter string to obtain the other one of these punctuations. The operation order is brackets, then quotation marks, and then "《,》". This step generally circulate two times. The third step, examine the string obtained from above step, if there are any quotation marks, then examine if there are any irregular punctuations within the quotation marks, if it is true, then take out the quotation marks part. If all punctuations in the string are in couples, then transfer the module, which obtains the collocation Chinese characters, and then record the bilingual LWP.

We have 712 items bilingual LWPs from the People's Daily in 2002.

Here, we will chose "是" structure to illustrate our extracting experiment, the others we'll discuss in future work.

5 the paraphrases patterns of "是" structure

We conclude the paraphrases patterns from 400 examples of LWP paraphrases, such as:

a) "LWP 是……的……缩写"

b) "LWP 是……的……缩写，……，是……"

c) "LWP 是……的……缩写，……，意即……"

Such as:

CMM是软件"能力成熟度模型"的缩写，是一种用于评价软件承包能力并帮助其改善软件质量的方法，是评估软件能力与成熟度的一套标准，侧重于软件开发过程的管理及工程能力的提高与评估。

d) LWP 是……的……简称

e) LWP 是……的……简称，……，是……

f) LWP 是……的……简称，……，意即……

Such as:

QFII是Qualified Foreign Institutional Investors（合格的境外机构投资者）的简称，QFII机制是指外国专业投资机构到境内投资的资格认定制度。作为一种过渡性制度安排，QFII制度是在资本项目尚未完全开放的国家和地区，实现有序、稳妥开放证券市场的特

殊通道。包括韩国、台湾、印度和巴西等市场的经验表明，在货币未自由兑换时，QFII不失为一种通过资本市场稳健引进外资的方式。

g) LWP 是……

h) LWP 是……的……

Such as:肯尼·G是当代广受欢迎的演奏家和音乐家，曾获得格莱美奖、美国音乐奖等奖项。他的《回家》早为中国听众所熟悉。

i) LWP，是……的……

第二阶段削减战略武器条约（START Ⅱ），是1993年1月由叶利钦和老布什两位总统签署的。条约规定10年内各自削减2/3的核弹头，并将可携带多弹头的陆基洲际弹道导弹全部销毁。

j) LWP 是指……

QFII机制是指外国专业投资机构到境内投资的资格认定制度。

k) LWP 是指……的……

GDP是指一个国家（或地区）在一定时期内所有常住单位生产经营活动的全部最终成果。GDP是按国土原则核算的生产经营的最终成果。比方说，外资企业在中国境内创造的增加值就应该计算在GDP中。

l) LWP，是指……的……

m)LWP，指的是……

而人们通常所说的3G，指的是下一代多媒体移动通信系统，它具有更宽的带宽，更高的频率和传输速率，可以方便地进行图像甚至活动画面的传输。

n) LWP 是一种……

o) LWP，是一种……的……

FAD2是一种将饱和脂肪酸转化成不饱和脂肪酸的酶，对哺乳动物生长具有重要作用，在菠菜等植物中存在。

p) LWP，简称 LWP，是……

q) 简称 LWP，是一种……

r) ……简称 LWP，它是……

小灵通又称无线市话（Personal Phone System），简称PHS，是一种个人无线接入系统。它采用微蜂窝技术，通过微蜂窝基站实现无线覆盖，将用户端（即无线市话手机）以无线的方式接入本地电话网，使传统意义上的固定电话不再固定在某个位置，可在无线网络覆盖范围内移动使用，随时随地接听、拨打本地和国内、国际电话。

s) ……是 LWP……就是……

t) ……是 LWP……

u) ……就是……LWP……

"……" are the other characters in paraphrases sentences excluding the label words, when we regard punctuations and the other characters as coordinate, the above patterns could be reduce to fifteen patterns. We find that "是指" and "是……"arranged in pairs or groups with "简称", "缩写" always a paraphrase sentence.

6 The auto-extracting algorithm

When we could recognize a LWP, then we could extract a paraphrase sentence, so we make use of our auto-labeling software [see also ZHENG Zezhi, 2005] to tag the LWP firstly, and then extract LWP paraphrases.

The experiment steps

1) Tag the texts choused with auto-labeling software;
2) Express the sentences and patterns with vectors;
3) Extract LWPs and their paraphrases;
4) Estimate the results from auto-extracting manually and conclude erratum rules;
5) Go back 2), continue training.

The auto-extracting algorithm of LWP paraphrases

1) Express the paraphrase patterns with the mark vector patterns, $\vec{T_i}$=(w, b_1, b_2, b_3, b_4),（i=1, 2, 3,…,15）, thereinto, w is a LWP, b_j (j=1, 2, 3 , 4) are mark words, b_1 can't be a null, b_2, b_3, b_4 could be a null. Found a vector space with the fifteen mark vector patterns;
2) Extract the sentences with a LWP or more;
3) Express these sentences extracted at 2) step with mark vector patterns, $\vec{S}$ =(LWP, b_1', b_2', b_3', b_4');
4) Calculate the distance between the two vectors: $\vec{T_i}$ and $\vec{S}$, $D_{\vec{T_i}\vec{S}} = (\sum_{i=1}^{4}(b_i - b_i')^2)^{1/2}$, when $D_{\vec{T_i}\vec{S}}$=0, then the current sentence is paraphrase sentence.
5) Take out the paraphrase of current LWP. Scan back forward to the head of current sentence, and scan along to next sentence end, and then take the two sentences as the paraphrase.

7 the experiment result

Our experiment is based on 500 text files, which are from the corpus of the "People's Daily" in 2002 and contains LWPs. From the 500 text files, we got 59 paraphrase pieces by manual work, in which 22 pieces are of "是"structure paraphrases, 18 pieces are with mark words, others are without marker phrases and not of "是"structure paraphrases.

With the algorithm, we do an experiment with "是"structure paraphrases. From the experiment result, we got 29 pieces are of "是"structure paraphrases, in which 19 pieces are Useful results and 10 pieces are not., so the usefulness rate is 19/29=65.5%, and recall rate is 19/22=86.4%.

The formula of usefulness rate is: $p = \frac{cordef(d)}{extdef(d)}$

The cordef(w) means the number of useful paraphrase of extracted results. The extdef(w) means the number of extracted results.

The formula of recall rate is: $r = \frac{cordef(d)}{orgdef(d)}$

The cordef(w) means the number of useful paraphrase of extracted results. The orgdef(w) means the number of useful paraphrase of "是"structure paraphrases from 500 text files.

An example of useful results:

<zm>P38 通路</zm>是少有的几个在进化上非常保守的生物信息传导机制。此通路的激活与失活和休克、关节炎、动脉粥样硬化等急慢性免疫疾病有重大关系。

An example of no-paraphrase results:

<zm>CDMA</zm>是个好技术，关键还要看联通经营得怎么样，这是不少业内人士的看法。

In fact, no-paraphrase results are these paraphrases which can't provide enough information to interpret or to make readers understand the LWP's meaning.

As a result, we find that these patterns always gave useful results, such as:

ELWP是……的……**缩写/**简称;
ELWP是……的……缩写/简称，……，是……;
ELWP是……的……缩写/简称，……，意即……;
简称ELWP，是一种……

And the no-paraphrase results were from these patterns, such as:

ELWP是……　　ELWP是……的……

If we provide category names, like "系统|标准|组织|企业|公司……", i.e. add category names in the "是" structure patterns, the number of useful results will be increased, for this measure would get rid of some no-paraphrase results. Also if we add no-paraphrase definitive, like "良机|机遇……", some no-paraphrase results would be eliminated.

8 Farther work

By now, our training set only had correct instances, so our farther work is to tidy up the no-paraphrase results from our experiment to form our exceptional rules, and collect category names and no-paraphrase definitive, in order to increase the precision rate of our algorithm.

We have finished extracting "是" structure paraphrases, but is not all-sided, we must try other no-"是" paraphrases extracting. And we should study the relation between our patterns and the normative patterns of dictionaries, ascertain the power coefficient, and then rank the same LWP's paraphrases.

Reference

1.ZHENG Zezhi, ZHANGPu, et al, The Research on Lettered-word Extraction in Chinese Texts, Chinese Information Transaction, 2005.1, vol. 19.

2.LI Xiyin, Ten Relation Groups in Dictionaries, Dictionary Study, 2005.1, SHANGHAI dictionary publishing company.

3.FU Huaiqing, Analysis and depiction of acceptations, BEIJING, Chinese publishing company, 1996.1

4.ZHANG Yan, ZONG Chengqing, et al, Structure Analysis and Extraction for the Definitions of Chinese terms, Chinese Information Transaction, 2003. 6,vol. 17.

5.XU Yong, XUN Endong, et al, A Web Term Definition Extracting System, Chinese Information Transaction, 2004. 4, vol. 18.

6.ZHENG Shupu, the Compendium of Russian Terminological Thesaurus theory, Dictionary Study, 2005.1, SHANGHAI dictionary publishing company.