IOP Conf. Series: Materials Science and Engineering **322** (2018) 052054 doi:10.1088/1757-899X/322/5/052054

Natural-Annotation-based Unsupervised Construction of Korean-Chinese Domain Dictionary

Wuying Liu[1,*] and Lin Wang[2]

[1] Laboratory of Language Engineering and Computing,Guangdong University of Foreign Studies,Guangzhou 510420, Guangdong, China
[2]Xianda College of Economics and Humanities, Shanghai International Studies University,Shanghai 200083, China

*Corresponding author e-mail: wyliu@gdufs.edu.cn
lwang@xdsisu.edu.cn

Abstract. The large-scale bilingual parallel resource is significant to statistical learning and deep learning in natural language processing. This paper addresses the automatic construction issue of the Korean-Chinese domain dictionary, and presents a novel unsupervised construction method based on the natural annotation in the raw corpus. We firstly extract all Korean-Chinese word pairs from Korean texts according to natural annotations, secondly transform the traditional Chinese characters into the simplified ones, and finally distill out a bilingual domain dictionary after retrieving the simplified Chinese words in an extra Chinese domain dictionary. The experimental results show that our method can automatically build multiple Korean-Chinese domain dictionaries efficiently.

1. Introduction

The fast-paced development of deep learning and statistical learning makes the large-scale language resource is particularly important in natural language processing [[1]]. On one hand, the explosion of language big data has increased the complexity of multilingual information processing [[2]]. On the other hand, it will bring a new opportunity to build language resources automatically [[3]].

Currently, the single-language resource [[4]] and the bilingual parallel resource [[5]] are two kinds of widely used resources. Comparing with the construction of the bilingual parallel resource, that of the single-language resource is relatively straightforward [[6]]. How to build the bilingual parallel resource efficiently from language big data has become a significant scientific problem [[7]].

The bilingual domain dictionary is one of the basic parallel lexical resources, which is important to machine translation [[8]] and literary translation. In this paper, we investigate the automatic construction issue [[9]] of the Korean-Chinese domain dictionary. Previous investigations have found that there were a lot of annotations written by traditional Chinese characters in parentheses after a Hangeul word or phrase [[10]]. Motivated by this, we propose an unsupervised construction architecture based on the natural annotation in a large-scale Korean text corpus.

2. Unsupervised Construction Architecture

Figure 1 shows our unsupervised construction architecture, which mainly includes four processing units. The Preprocessor Unit receives many Korean text documents in their chronological sequence

from a Korean text corpus, and splits each text document into several candidate segments which must include Hangeul letters, Chinese characters, and parentheses. The Word Pair Extractor Unit receives the candidate Korean text segments, and separates each Chinese word from its relevant Hangeul word to form a Korean-Chinese word pair. All the word pairs will be collected into a Korean-Chinese dictionary. If our task were just to build a Korean-Chinese dictionary from a Korean text corpus, the final Korean-Chinese dictionary would be a perfect ending. Actually, the Korean-Chinese domain dictionary is our ultimate target indeed. Because the Chinese word of each word pair is mostly composed by traditional Chinese characters, the Chinese Simplifier Unit has to transform the traditional Chinese word into the simplified one. Ultimately, the Distiller Unit will distill out a Korean-Chinese domain dictionary after retrieving the simplified Chinese words in an extra Chinese domain dictionary. Moreover, if you have multiple Chinese dictionaries from different domains, you will obtain multiple Korean-Chinese domain dictionaries.

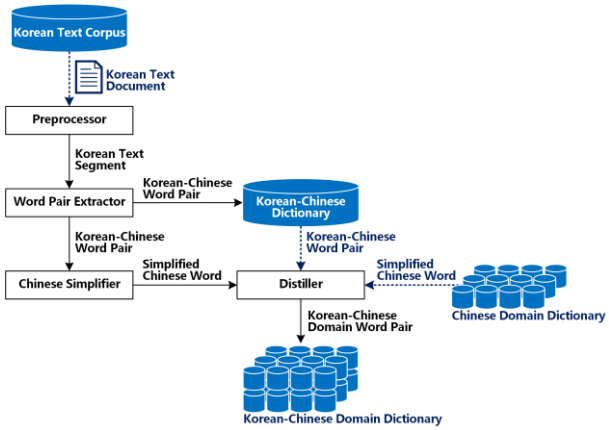

Figure 1. Unsupervised Construction Architecture.

The Korean text corpus and the Chinese domain dictionary in the unsupervised construction architecture are both single-language resources, which can be easily obtained from the Internet in the era of big data. While the Korean-Chinese dictionary and the Korean-Chinese domain dictionaries in the architecture are all bilingual parallel lexical resources, which can be automatically generated without any manual intervention.

3. Unsupervised Construction Algorithm

Within the unsupervised construction architecture, we design a detailed unsupervised construction algorithm, which is shown in Figure 2. In the algorithm, there are also four functions corresponding to the above four processing units.

Normally, a Korean text mainly uses the words consisting of Hangul letters, and occasionally between these Hangul words, there are some annotations written by Chinese characters. The most common format is "␣HangulWord(ChineseWord)␣", which can be regarded as a natural annotation. We can use a regular expression in the algorithm (line 12 in Figure 2) to match the natural annotation. But the matching is so strict to rule out many non-standard Korean-Chinese co-occurrence representations, which is an unavoidable sacrifice to the unsupervised learning. In order to save these candidate representations, a supervised-based or semi-supervised-based manual intervention will be essential.

```
1.// Unsupervised Construction Algorithm
2.Input:  String[] ktc; // Korean Text Corpus
3.        String[] cdd; // Chinese Domain Dictionary
4.Output:<String, String>[] kcdd; // Korean-Chinese Domain Dictionary
5.Function String[]: preprocessing(String[] ktc)
6.String[] kts;
7.For Integer i ← 1 To ktc.size Do
8.        String ktd ← ktc[i];
9.        String[] segs ← ktd.split(" ");
10.       For Integer j ← 1 To segs.size Do
11.       String seg ← segs[j];
12.       If
          (seg.contains(HangeulLetter&ChineseCharacter&Parentheses))
13.       Then kts.add(seg);
14.       End If
15.       End For
16.End For
17.Return kts.
18.Function <String, String>[]: wordpairextracting(String[] segs)
19.<String, String>[] kcd;
20.For Integer i ← 1 To segs.size Do
21.       String seg ← segs[i];
22.       String kw ← seg.substring(0, index("("));
23.       String cw ← seg.substring(index("(")+1, index(")"));
24.       kcd[i] ← <kw, cw>;
25.End For
26.Return kcd.
27.Function String: chinesesimplifing(String cw)
28.For Integer i ← 1 To cw.charlen Do
29.       cw.char(i) ← traditionaltosimplified(cw.char(i));
30.End For
31.Return cw.
32.Function <String, String>[]: distilling(<String, String>[] kcd; String[]
   cdd)
33.<String, String>[] kcdd;
34.For Integer i ← 1 To kcd.size Do
35.       String cw ← chinesesimplifing(kcd[i].chineseword);
36.       If (cdd.contains(cw))
37.       Then kcdd.add(kcd[i]);
38.       End If
39.End For
40.Return kcdd.
```

Figure 2. Unsupervised Construction Algorithm.

During the process of the algorithm implementation, the space overhead is mainly used to store dictionaries, which is negligible for the current memory capacity. Though the main time overhead is proportional to the size of the Korean text corpus. The stream processing method, only one-pass scanning for each text document, will make it time-efficient. The space-time complexity of the unsupervised construction algorithm is acceptable in practical applications.

4. Experiment

In order to validate the effectiveness of our unsupervised construction architecture and algorithm, we firstly prepare a large-scale Korean text corpus and several Chinese domain dictionaries. Supported by

the ***Wiki Dump*** tool, we download the Korean Wikipedia dataset (**kowiki-20170101-pages-articles.xml.bz2**) on the Internet. From the dataset, the ***Wikipedia Extractor*** software extracts 368646 Korean plain-text documents, which will be treated as the Korean text corpus. Furthermore, we already have 21 Chinese dictionaries from the following 21 domains in Table 2. We will use them as the Chinese domain dictionaries. Secondly, we implement the unsupervised construction algorithm, and run it in the above corpus and domain dictionaries. Finally, we analyze the experimental results and suggest some discussions.

4.1 Rusult and Discussion about Korean-Chinese Dictionary

After running of our unsupervised construction algorithm, we have extracted 178,354 Korean-Chinese word pairs to form a Korean-Chinese dictionary. The effectiveness of our method depends on the natural annotations written by traditional Chinese characters in raw Korean Wikipedia dataset. These traditional Chinese characters can be transformed into simplified Chinese ones straightforwardly, and therefore they can be regarded as a natural bridge across the two languages of Korean and Chinese.

Table 1. Partial Examples of Korean Homonym.

Korean	**Chinese**
이정(21)	李霆/理程/以定/夷靖/李祯/李贞/移定/而丁/李桢/里正/李锃/而净/李整/利贞/李靖/李颋/李渟/二程/李婷/李挺/李精
전주(19)	前柱/伝奏/传奏/前奏/转注/全周/田畴/笺注/铨注/田主/笺奏/专注/前主/殿主/転住/田柱/全州/澶州/电柱
조정(19)	赵挺/祖丁/祖珽/赵政/赵祯/赵侹/助丁/赵靖/赵鼎/曹鼎/曹正/汉朝/调整/曹整/朝廷/赵珽/赵晟/调停/赵亭
유지(19)	佑司/类智/庾智/有志/幽志/刘挚/酉地/油纸/谕旨/乳脂/维持/柳枝/遗址/遗旨/刘志/留止/遗志/有旨/油脂
유수(19)	有水/有数/洧水/柳洙/柳树/柳宿/柳谭/潍水/柳绥/留数/刘修/濡须/类数/刘秀/流水/幽囚/渝水/刘寿/留守
전기(18)	战气/战记/前期/传奇/全寄/転机/传记/田忌/田既/电气/田琦/全期/电机/战机/伝记/伝奇/全纪/转机
유정(18)	刘綎/裕靖/刘桢/有情/榴亭/刘政/维贞/惟政/刘梃/刘靖/庾靖/惟静/柳町/维祯/油井/有顶/柳汀/留正
영주(18)	令洲/瀛州/灵州/领主/宁州/永州/永畴/荣州/迎州/颍州/郢州/霊珠/营州/永住/霊州/英州/英主/瀛洲
아키라(17)	佐藤昭/藤浪鉴/今谷明/小野哲/鸟山明/鱼住昭/浅田彰/平山谛/大冈玲/黑泽明/铃木哲/黒泽明/武藤章/有吉明/太田亮/关口明/藤枝晃
정주(17)	庭州/亭主/贞州/静州/定主/定州/定住/程朱/汀州/祯州/郢州/鼎州/郑胄/净住/郑注/汀洲/郑州
정원(17)	侦员/政院/定远/丁元/定员/丁瑗/靖远/私园/贞元/静园/正员/庭园/庭院/正元/丁原/净源/净远
기수(17)	淇水/技手/沂水/汽水/基洙/旗手/奇数/耆叟/机首/气数/骑手/器数/箕宿/既遂/气随/奇兽/基数
이유(17)	蜼蝓/李愈/耳犹/李需/李瑜/李瑠/李濡/理由/夷维/李由/李裕/李儒/二酉/李攸/李遗/李有/李栐
이상(17)	异想/理想/异常/李相/李瑞/李鸑/异相/李像/二相/履祥/李翔/履常/李晌/以上/异像/李箱/李祥
유성(17)	惟性/游星/流星/柳城/油性/有性/儒城/惟圣/有声/刘晟/幼成/刘圣/有诚/刘成/由盛/柳惺/有成

Table 1. Cont.

대사(17)	大祀/才谈/大赦/大使/大志/台辞/台词/大社/大事/带沙/大蛇/大史/大师/代谢/大舍/大士/大寺
조사(16)	吊辞/吊词/措辞/祖师/照射/造士/朝士/助辞/诏使/赵奢/赵师/组士/早死/助事/操丝/助词
전사(16)	转写/典事/传舍/传写/前司/戦史/典祀/前事/前史/战使/讨死/战史/田舍/填词/战士/战死
유기(16)	刘畿/留记/柳器/鎏器/有机/诱起/幼期/遗弃/刘锜/游妓/流沂/唯气/刘基/幼起/刘琦/唯几
사상(16)	写象/史上/狮象/死伤/私商/舍象/舍上/事象/泗上/四象/谢尚/思想/事相/四相/士常/丝状
사도(16)	思道/沙岛/使徒/司徒/佐渡/使者/沙道/士道/茶道/使道/司道/思悼/蛇岛/砂岛/邪道/私道
조성(15)	调性/赵城/赵碹/赵成/造声/鸟城/造成/赵声/照星/组成/洮城/曹诚/赵晟/赵岘/曹性
쓰보네(15)	大进局/矢岛局/飧庭局/大姥局/丹后局/大贰局/小督局/朝仓局/今参局/京极局/阿波局/东殿局/小弁局/西鄉局/春日局
정사(15)	精查/情事/正使/郑泗/丁巳/亭舍/情死/上使/正史/精舍/正师/政史/正思/政事/呈辞
장수(15)	蒋修/张铢/张秀/将帅/张宿/漳水/长修/长水/樟寿/张繡/张数/张绣/张修/长寿/良将
이원(15)	尼院/李元/而元/李瑗/梨园/吏员/李愿/移园/李源/而远/李黿/李园/李原/李圆/利原
이수(15)	李修/伊水/李寿/耳叟/离水/而寿/李琇/李铢/泥水/李燧/理水/履修/螭首/异数/里数
소사(15)	少使/少志/烧死/少史/所事/少师/小事/小祀/昭思/小史/小社/小使/小舍/所司/素沙
사성(15)	嗣圣/斯珹/赐姓/四圣/写成/四姓/司成/蛇城/司星/思诚/莎城/司城/四声/四星/师圣
사전(15)	辞典/史伝/祀典/史传/用语/辞书/私战/私田/谢笺/社殿/事前/私戦/寺田/赐田/寺传
사고(15)	蛇蛊/师考/四苦/四库/佐护/师古/史藁/思考/士高/史高/社告/私稿/史库/事故/死苦
수성(15)	守成/水城/寿星/秀星/输城/随城/受姓/隋城/遂成/寿成/水星/修城/修省/穗城/寿城

We also calculate and analyze the phenomenon of Korean homonym in the Korean-Chinese dictionary. The experimental result shows that there are 17,408 Korean words with more than two meanings mapping to different Chinese words. Table 1 shows the partial examples of Korean homonym with more than 15 meanings. We can at least find that each pair of [电气, 战机], [瀛州, 永州], [定远, 靖远], and [刘基, 刘琦] will have a same Hangul word. But unfortunately, the meanings of the two Chinese words in each pair are completely different. So, many Korean words only represented by Hangeul letters without any Chinese annotation will cause serious semantic confusions.

4.2 Rusult and Discussion about Korean-Chinese Domain Dictionary

Supported by our 21 Chinese domain dictionaries, the distilling function of the unsupervised construction algorithm has selected out 21 Korean-Chinese domain dictionaries from the Korean-Chinese dictionary with 178,354 words. Table 2 shows the detailed number of word pairs in each Korean-Chinese domain dictionary. The experimental result shows that the Chinese word is not only a bridge to connect Korean and Chinese, but also a bridge to connect single-language classified dictionary and bilingual classified one.

Table 2. Number of Word Pairs.

Domain	**Number**	**Domain**	**Number**	**Domain**	**Number**
Geography(地理)	1139	Economy&Trade(经贸)	192	Astronomy(天文)	199
PlaceName(地名)	1806	Military(军事)	613	Physics(物理)	789
Law(法律)	451	History(历史)	3461	Medicine(医药)	1315
Buddhism(佛学)	3052	Agronomy(农学)	370	Art(艺术)	1735
Sinology(国学)	1401	PersonName(人名)	2189	Film&Television(影视)	2515
Mechanics&Electronics(机电)	788	Biology&Chemistry(生化)	884	Philosophy(哲学)	1319
Architecture(建筑)	592	Mathematics(数学)	653	TraditionalChineseMedicine(中医)	526

In order to further elucidate the effect of Korean-Chinese domain dictionary construction, we choose the Korean-Chinese Economy&Trade dictionary for a specific analysis. Although the total number (192) of word pairs in the dictionary is not too much, the auto-generated domain dictionary is splendid enough to be used as a seed dictionary to snowball an incremental construction.

Moreover, some interesting conclusions can be drawn from the above dictionary. We firstly find out two groups of synonymous expressions [가능성, 있음직] and [삼각형, 세모꼴]. The former means POSSIBILITY (可能性), and the latter means TRIANGLE (三角形). Secondly, we find out three groups of variant characters [国務院, 國務院], [点数, 點數], and [周期性, 週期性]. There is no doubt that the variant characters are understandable and acceptable by Korean readers. Finally, there is a slightest blemish about the pair of <중항, 中行> within the economy and trade domain. Because there are two meanings about "中行". One is an ancient surname and is pronounced as "zhōng xíng". The other is an abbreviation for BANK OF CHINA (中国银行, 중국은행), and is pronounced as "zhōng háng". Although there is no any mistake about <중항, 中行> and <중행, 中行>, the <중행, 中行> is more appropriate in the Economy&Trade dictionary standing with <공행, 工行>, <교행, 交行>, and other banks.

5. Conclusion

This paper presents an unsupervised architecture and a relevant algorithm for Korean-Chinese domain dictionary construction. The experimental results show that our method can automatically construct bilingual domain dictionaries efficiently from language big data, and whose effectiveness depends on the natural annotations in the raw corpus.

Further research will concern the influence of manual interventions. It is undeniable that manual work is indispensable for language experts to obtain a more precision bilingual dictionary. Therefore, the supervised learning and semi-supervised learning will be more expected for optimal bilingual domain dictionary construction. We will also transfer above research productions to other suitable oriental languages like Japanese, Vietnamese, and so on.

Acknowledgements
The research is supported by the Key Project of State Language Commission of China (No.ZDI135-26) and the Featured Innovation Project of Guangdong Province (No.2015KTSCX035).

References
[1] Yang Yuqing. A Review of Term Semantic Hierarchy Induction for Domain-specific Chinese Text Information Processing. International Journal of Knowledge and Language Processing, 6(4):50-60, 2015.
[2] Liu Xiaodong, Duh Kevin, Matsumoto Yuji. Multilingual Topic Models for Bilingual Dictionary Extraction. ACM Transactions on Asian and Low-resource Language Information Processing, 14(3):11, 2015.
[3] Fu Ruiji, Qin Bing, Liu Ting. Generating Chinese Named Entity Data from Parallel Corpora. Frontiers of Computer Science, 8(4):629-641, 2014.
[4] Liang Hu and Xuri Tang. Multiword Expression Extraction Based on Word Relativity. International Journal of Knowledge and Language Processing, 5(1):27-40, 2014.
[5] Kim Jae-Hoon, Kwon Hong-Seok, Seo Hyeong-Won. Evaluating a Pivot-based Approach for Bilingual Lexicon Extraction. Computational Intelligence and Neuroscience, 2015:434153, 2015.
[6] Ge Xu and Chu-Ren Huang. Extracting Chinese Polarity Shifting Patterns from Massive Text Corpora. Lingua Sinica, 2(1):5, 2016.
[7] Xiang Lu, Zhou Yu, Zong Chengqing. An Efficient Framework to Extract Parallel Units from Comparable Data. Communications in Computer and Information Science, 400:151-163, 2013.
[8] Pal Santanu, Pakray Partha, Gelbukh Alexander, van Genabith Josef. Mining Parallel Resources for Machine Translation from Comparable Corpora. Lecture Notes in Computer Science, 9041:534-544, 2015.
[9] Seo Hyeong-Won and Kim Jae-Hoon. Analyzing Errors in Bilingual Multi-word Lexicons Automatically Constructed through a Pivot Language. Journal of the Korean Society of Marine Engineering, 39(2):172-178, 2015.
[10] 권홍석, 서형원, 김재훈. Enhancing Performance of Bilingual Lexicon Extraction through Refinement of Pivot-Context Vectors (중간언어 문맥벡터의 정제를 통한 이중언어 사전 구축의 성능개선). Journal of KIISE: Software and Applications, 41(7):492-500, 2014.