The Robustness of Domain Lexico-Taxonomy: Expanding Domain Lexicon with CiLin

Chu-Ren Huang
Institute of Linguistics, Academia Sinica, Taipei
churen@sinica.edu.tw

Xiang-Bing Li
Institute of Information Science, Academia Sinica, Taipei
dreamer@hp.iis.sinica.edu.tw

Jia-Fei Hong
Institute of Linguistics, Academia Sinica, Taipei
jiafei@gate.sinica.edu.tw

Abstract.

This paper deals with the robust expansion of Domain Lexico-Taxonomy (DLT). DLT is a domain taxonomy enriched with domain lexica. DLT was proposed as an infrastructure for crossing domain barriers (Huang et al. 2004). The DLT proposal is based on the observation that domain lexica contain entries that are also part of a general lexicon. Hence, when entries of a general lexicon are marked with their associated domain attributes, this information can have two important applications. First, the DLT will serve as seeds for domain lexica. Second, the DLT offers the most reliable evidence for deciding the domain of a new text since these lexical clues belong to the general lexicon and do occur reliably in all texts. Hence general lexicon lemmas are extracted to populate domain lexica, which are situated in domain taxonomy. Based on this previous work, we show in this paper that the original DLT can be further expanded when a new language resource is introduced. We applied CiLin, a Chinese thesaurus, and added more than 1000 new entries for DLT and show with evaluation that the DLT approach is robust since the size and number of domain lexica increased effectively.

1. Introduction

Domain-based language processing has an inherent research dilemma when the construction of domain lexicons is involved. The standard approach of building domain lexicon from domain corpora requires a very high threshold of existing domain resources and knowledge. Since only well-documented domains can provide enough quality corpora, it is likely these fields already have good manually constructed domain lexica. Hence this approach is can only deal with domains where only marginal benefit can be achieved, while it cannot deal with domains where it can make most contribution since there is not enough resources to work with.

It was observed that the type of domain language processing that has the widest application and best potentials are cross-domain and multi-domain in nature. For instance, a typical web-search is a search for specific domain information from the www as an archive of mixed and heterogeneous domains. The contribution will be immediate and salient to be able to acquire resources and information for a new domain that is not well documented yet.

A new approach towards domain language processing by constructing an infrastructure for multi-domain language processing called the Domain Lexico-Taxonomy (DLT) was proposed in Huang et al. (2004). In the DLT approach, domain lexica are semi-automatically acquired to populate domain taxonomy. This lexically populated domain taxonomy serves two purposes: as the basis of stylo-statistical prediction of the domain of a new text, and as the core seed of complete domain lexica. For the first purpose, the DLT approach relies crucially on the ability to effectively identify words that are good indicators of specific domains. For the second purpose, the DLT needs to be robust enough to allow incremental expansion when new content resources are integrated. In this study, we integrate CiLin, a Chinese thesaurus, to show that the DLT architecture is indeed robust.

2. Related Work

Typical studies on domain lexica focuses on assigning texts to specific classes, hence they use a limited taxonomy augmented with a small set of features (e.g. Avancini et al. 2003, Sebastiani 2002, and Yand and Pederson 1997). However, specialized lemmas cannot be useful in multi-domain processing. To achieve domain versatility in processing, it is necessary to identify lemmas with wider distributions and yet is associated with particular domain(s). We follow the DLT architecture (Huang et al. 2004), which was shown to be effective in predicting the domain of documents extracted from the web. We aim to elaborate that framework by proposing a domain lexica can be incrementally expanded with knowledge from a new resource.

3. Domain Taxonomy

A domain taxonomy containing 549 nodes was manually constructed. The main sources of domain classification are from Chinese Library Classification system, Encyclopedia Britannica and the Global View English-Chinese dictionary. Two important criteria were chosen: that the taxonomy is bilingual and that it is maintained locally. First, the bilingual taxonomy is essential for future cross-lingual processing but also allows us to access relevant resources in both languages. Second, since our emphasis was not on the correctness of a dogmatic taxonomy but on the flexibility that allows monotonic extensions, it is essential to be able to monitor any changes in the taxonomy.

There are four layers in the constructed domain taxonomy. Fourteen (14) domains are in the upper layer, including Humanities, Social Science, Formal Science, Natural Science, Medical Science, Engineering Science, Agriculture and Industry, Fine Arts, Recreation, Proper Name, Genre/Strata, Etymology, Country Name, Country People. The Second layer has 147 domains. The third layer has 279 domains. Lastly the fourth layer has only 109 domains since not all branches need to be expanded at this level. In sum, there are 549 possible domain tags when the hierarchy is ignored. The domain taxonomy is available online at the Sinica BOW website (http://BOW.sinica.edu.tw/, Huang and Chang 2004).

4. Detection of Domain Lexicon in DLT

The challenge in integrating heterogeneous language resources for domain information is that conceptual classification varies from one resource to another and hence cannot be directly harvested. We propose to utilize the inheritance relations of these resources, instead of their hierarchy. In other words, lexical (and hence conceptual) identity is established first, following by expanding this matching with logical inheritance but without branching out on the conceptual hierarchy.

DLT establish the correspondences between the taxonomic nodes of domains and the linguistic resources of sub-lexica. Note that a lexical knowledgebase, in a Wordnet fashion, also contains hierarchical relations. The domain taxonomy can be enriched by taking the hierarchical information internal to the lexica. If these resources directly encodes the 'is-a' relation by hyponymy, we assume that both the node (lexicons) and their hyponym node (lexicons) belong to that domain. Using the simple supposition, we can observe the domain knowledge with various resources, and strengthens the domain lexica for domain taxonomy. The process of populating DLT is shown in Fig. 1.

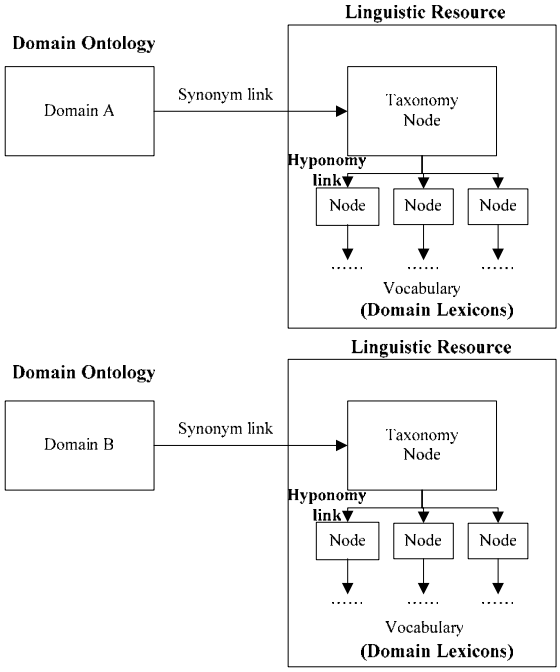

Figure. 1. Populating DLT from Linguistic Resource

5. Experiment

5.1. The Original Study with Bilingual WordNet

The original DLT work was based on bilingual Wordnet (Huang et al. 2004). This is because of the Wordnet lexical knowledgebase is highly enriched with lexical semantic relation information. In addition, the bilingual Wordnet adds an unparallel dimension of knowledge coverage. The bilingual Wordnet used is Sinica BOW (The Academia Sinica Bilingual Ontological WordNet, Huang and Chang (2004)). Sinica BOW is bilingual lexical knowledgebase connecting WordNet and SUMO and mapping both between English and Chinese. The study reported in Huang et al. (2004) also contains a small domain identification experiment to show the application of DLT.

5.1.1 Description of WordNet and Sinica BOW

WordNet is inspired by current psycholinguistic and computational theories of human lexical memory (Fellbaum (1998), Miller et al. (1993)). English nouns, verbs, adjectives, and adverbs are organized into synonym sets, each representing one underlying lexicalized concept. Different semantic relations link the synonym sets (synsets). The version of WordNet that Sinica BOW implemented is version 1.6, with nearly 100,000 synsets.

In Sinica BOW, ach English synset was given up to 3 most appropriate Chinese translation equivalents. And in cases where the translation pairs are not synonyms, their semantic relations are marked (Huang et al. 2003). The bilingual WordNet is further linked to the SUMO ontology. We use the semantic relations in bilingual resource to expand and predict domain classification when it cannot be judged directly from a lexical lemma.

5.1.2 Experiment and Result with WordNet

463 of the 549 nodes in the domain taxonomy were successfully mapped to a WordNet synset through an identical lemma. 452 or 463 mappings were manually confirmed to be correct, a precision score of over 97%. These domains were expanded to cover a total of 11,918 synsets corresponding to 15,160 Chinese lemmas. Note that both English and Chinese correspondences are used since our resources (WordNet and domain taxonomy) are both bilingual.

Due mostly to hyponymy expansion, each lemma is mapped to 1.38 domains in average. While each lemma is assigned to no more than 8 domains, with the majority (6,464) assigned to only one. These mapped lemmas populate a set of domain lexica. The number of entries in these domain lexica ranges from 1 to 3762. The average size of these domain lexica is 32.8 lemmas. Only 41 domains lexical contain 33 or more lemmas. Since we cannot know the effective of the lexicon of a domain a priori, we take those whose size are above average as the effective domain lexica.

These domain lexica and their sizes are shown in Table 1.

5.1.3 Evaluation: precision of domain lexica

It is impossible to formally evaluate the recall rate of this domain lexica study since we do not know the total number of entries to be recalled. However, it is possible to evaluate the precision rate of the constructed domain lexica. First, the precision of all recalled lemmas is tested. Among the mapped lemmas, 8696 (out of 15,160) lemmas are assigned to multiple domains, while 6,464 are assigned to single domain. The single domain mappings were spot-checked to be correct. On the other hand, the precision of all 8,696 multi-domain lemmas are carefully evaluated. Among these lemmas, only 4.81% (418) proves to be wrong; and an overwhelming majority of 95.19% turns out to be correct (8278).

Second, a more meaningful test is to evaluate how well the domain lexica are defined. Five effective domain lexica with over 100 entries were randomly chosen for evaluation: Insect (515 entries), Natural Science (262 entries), Sports (180 entries), Dance (124 entries) and Religious Music (48 entries).

The manually checked precision of these domain lexica is listed below the Table 2:

Domain	Domain	Domain	Domain
Vertebrates 脊椎動物 3676	Food 食品 2968	Bird 鳥類 1059	Fish 魚類 729
Language 語言 699	Recreation 休閒娛樂 548	Insect 昆蟲 515	Natural Science 自然科學 262
Country 國家 250	contest 競賽 207	music 音樂 192	Indian 印地安 188
Sports 運動 180	commerce 商業 144	Business 生意 144	Dance 舞蹈 124
Heraldic design 紋章設計 120	Medical Science 醫療科學 85	Medicine 醫學 76	Pathological medicine 病理醫學 76
Clinical medicine 臨床醫學 76	Mathematics 數學 69	Humanities 人文學科 64	Social Science 社會科學 62
physics 物理學 56	Religion 宗教 52	Religious Music 宗教音樂 48	Plastic art 造形藝術 45
Pure mathematics 純數學 44	Anthropology 人類學 42	Earth science 地球科學 39	drawing 素描 39
Norse Mythology 北歐神話 39	Philosophy 哲學 37	Telecommunication 電信通訊 35	theater 戲劇 34
Fine Arts 藝術 33			

Table 1. Domain lexica containing 33 or more lemmas

Domain Label	# of entries	Precision (%)
Insect	515	99.03
Natural Science	262	69.85
Sports	180	86.11
Dance	124	100.00
Religious Music	48	93.75

Table 2. Size and Precision of selected domain lexica

Table 2 shows an overall precision of over 95%, while no other lexica has precision lower than 86%, natural science is lowest at just below 70%. This is because "Natural Science" is a higher level domain and hence open to more noises in the detection process. This study clearly showed that the WordNet helped to effectively build core domain lexica.

We take the domain "Dance" as an example to explain the process. First, we map "Dance" to the Wordnet synset—"dance", and we look for the hyponym synsets. Table3 will be shown the expanding lexica of one of hyponym synsets. These lexical entries are associated with domain "Dance" and populate the domain lexicon.

Level	synset
1	social_dancing
2	folk_dancing, folk_dance
3	country-dance, country_dancing,
4	square_dance, square_dancing
5	quadrille

Table 3. The expanding hyponym synsets of "dance"

5.2 For CiLin

5.2.1 Description of CiLin

CiLin, a short name for Tongyici CiLin, is a Chinese thesaurus published in 1984 (Mei et al. 1984). The terms in CiLin are organized in a conceptual hierarchy, with near-synonym terms forming a set. There are five levels in the taxonomy structure of CiLin. The CiLin terms between Level1 to Level4 are taxonomy categories. Level1 is the upper class, and it includes 12 categories, like as people, object, time and space, abstract etc. Level2 has 106 categories. Level3 has 3,948 categories. Level4 has 4,014 categories. There are 64,157 terms in Level5 since all branches need to be expanded at this level. These terms are classified to 12,193 sets by the meaning. The average number of terms in each set is 5.34. Fig. 2 shows the structure of CiLin.

Figure. 2. The structure of CiLin

5.2.2. Experiment and Result with CiLin

First, we map the 549 domains to CiLin's taxonomy. Unlike the previous study, only Chinese terms were available on CiLin. The result is given in Table 4.

	# of entries	# of domains	entries/domains
Leve1 1	146	1	146
Level 2	1,587	3	529
Level 4	1,222	32	38.19

Table 4. Number of expanding entries and mapping domains

Manual checking showed that mappings to Level 1 and Level 2 are both imprecise and small in number. Hence we take Level 4 as the lexical anchor for enriching domain lexica. 1,222 lexical items are expanded from 32 domains, and these domain lexica and their sizes are shown in Table 5.

Domain	Domain
Insect(昆蟲) -- 146	Sewing(縫紉) -- 25
Country(國家) -- 128	Movie(電影) -- 25
Theater(戲劇) -- 116	Game(遊戲) -- 25
Painting(繪畫) -- 88	Photography(攝影) -- 21
Capital(資本) -- 54	Payment(支付) -- 20
Cookery(烹飪) -- 52	Printing(印刷) -- 20
Dance(舞蹈) -- 52	Literature(文學) -- 18
Law(法律) -- 50	Investment(投資) -- 14
Education(教育) -- 47	Swimming(游泳) -- 12
Martial_art(武術) -- 45	Broadcasting(廣播) -- 11
Religion(宗教) -- 39	Ranching_and_animal_husbandry(畜牧) -- 10
Architecture(建築) -- 38	Textile_industry 紡織 10
Carving(雕刻) -- 37	Boating(划船) -- 8
Language(語言) -- 37	Trade(貿易) -- 7

Table 5. Domain lexica

When all mappings are evaluated, 873(71.44%) of them are correct, and 349 (28.56%) incorrect. Five effective domain lexica are evaluated, as shown below in Table 6:

Domain Label	# of entries	Precision (%)
Insect	146	58.9
Country	128	55.47
Theater	116	80.17
Painting	88	80.68
Dance	52	80.77

Table 6. Size and Precision of selected domain lexica

Compared with the work reported in (Huang et al. 2004), both the number of lemma (1,222 vs. 15,160) and precision (71.44% vs. nearly 95%) are lower. This result is expected since CiLin has a simple taxonomy without the rich lexical information of a Wordnet. The crucial fact shown, however, is that DLT can be incrementally enhanced with the new mappings. Of the 873 correct domain lexica entries, 79.5% (694) are new entries that were not identified previously. Even more impressive is the effectiveness of increase in lexica sizes for applicable domains, as shown below in Table 7.

domain	WN/old	CiLin/new	increase	domain	WN/old	CiLin/new	increase
戲劇	34	80	0.7018	文學	12	15	0.5556
昆蟲	515	65	0.1121	印刷	22	15	0.4054
繪畫	17	61	0.7821	縫紉	28	14	0.3333
國家	250	44	0.1497	音樂	192	12	0.0588
舞蹈	124	34	0.2152	支付	20	11	0.3548
武術	7	33	0.8250	攝影	23	10	0.3030
資本	26	33	0.5593	游泳	15	9	0.3750
烹飪	14	32	0.6957	投資	5	8	0.6154
建築	2	29	0.9355	紡織	9	7	0.4375
雕刻	2	27	0.9310	廣播	2	7	0.7778
法律	22	24	0.5217	畜牧	0	5	1.0000
教育	26	23	0.4694	遊戲	16	5	0.2381
語言	699	22	0.0305	划船	27	4	0.1290
宗教	52	21	0.2877	貿易	4	4	0.5000
計算	55	19	0.2568	料理	6	2	0.2500
電影	21	17	0.4474	鍛造	1	2	0.6667

Table 7. Increase in Domain Lexicon Size after CiLin Integration

Table 7 shows that, even though adding CiLin only helped 32 domain lexica, 14 of them have their lexicon size more than doubled. One of them, ranching and animal husbandry is a new domain lexicon where no mapping was possible with WordNet. In other words, adding the CiLin resource substantially enhanced effective domain coverage of DLT.

6. Conclusion

In this paper, test the robustness of the DLT architecture. We show both the coverage and the sizes of the domain lexica on DLT can be effectively expanded by integrating a new language resource. The robustness is convincing given that the coverage and quality of the new resource is actually not as good as the original reference resources. In other words, we showed the open architecture of DLT facilitates integration of new domain information without imposing any high threshold on the format and quality of new resources. We also verify partial results of previous work since 205 lemma mappings were repeated. For future work, we plan to continue to populate DLT, as well as to explore other possibilities for putting DLT to actual applications.

References

Chu-Ren Huang, Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. Languages and Linguistics. 4.3. (2003)509-532

Chu-Ren, Huang and Ru-Yng Chang. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO". Presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May (2004)

Chu-Ren Huang, Xiang-Bing Li, Jia-Fei Hong. "Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing", Asian Symposium on Natural Language Processing to Overcome Language Barriers, The First International Joint Conference on Natural Language Processing (IJCNLP-04). Sanya City, Hainan Island, China. 22-24 March (2004)

F. Sebastiani., "Machine learning in automated text categorization". ACM Computing Surveys, 34(1) (2002)1-47

Fellbaum C.. WordNet: An Electronic Lexical Database. Cambridge: MIT Press (1998)

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. "Introduction to WordNet: An On-line Lexical Database," In Proceedings of the fifteenth International Joint Conference on

Artificial Intelligence. Chambéry, France. 28 August- 3 September (1993)

Henri Avancini, Alberto Lavelli, Bernardo Magnini, Fabrizio Sebastiani, Roberto Zanoli. Expanding Domain-Specific Lexicons by Term Categorization. Proceedings of the 2003 ACM symposium on Applied computing. Melbourne, Florida, USA. 9-12 March (2003)

Jia-ju Mei, Yi-Ming Zheng, Yun- Qi Gao and Hung-Xiang Yin. TongYiCi CiLin. Shanghai: the COMMERCIAL Press (1984)

Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, Proceedings of ICML-97, 14th International Conference on Machine Learning, pages 412 420. San Francisco: Morgan Kaufmann (1997)