Domain Lexico-Taxonomy: An Approach Towards Multi-domain Language Processing

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 domain barrier issues in language processing. Our work centers on Domain Lexico-Taxonomy (DLT), a domain taxonomy enhanced by domain lexicons. We propose DLT as an infrastructure for crossing domain barriers. By using DLT with WordNet and Domain Taxonomy, we can get 15160 Chinese lemmas in 463 domains. We estimate the accuracy of five domain's lemmas, and get 89.74% in average. Look from all lemmas, each lemma is assigned 1.38 domains in average. By gathering the web pages which Google query returned as testing data, we make some experiments to confirm effectiveness of domain lexicons. Finally, we analyze the result to prove the usefulness of domain lexicons obtained by the DLT approach.

Key word: multi-domain language processing, domain lexicon, domain taxonomy, WordNet.

1 Introduction

Domain-based language processing is one of the most active and productive directions in NLP. Its task-oriented nature allows the research to be focused and productive. And the focus on domain knowledge often facilitates restricted language and controlled vocabulary approaches. However, there is 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. To start with, it requires good quality domain corpora. Since only well-documented domains can provide enough quality corpora, it is likely these fields already have good manually constructed domain lexicons. 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.

We observe that the type of domain language processing that has the widest application and best potentials are cross-domain and multi-domain in nature. NLP in a specific domain does not differ from general NLP except for the restriction on the type of resources (corpora, lexica, etc.) The real challenge and rewards lies in multi-domain processing. 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. Lastly, high value addition can be achieved if domain information can be extracted from a resource classified to belong to a different domain.

In this paper, we propose a new approach towards domain language processing by constructing an infrastructure for multi-domain language processing. A domain taxonomy is constructed, and domain lexicons are semi-automatically acquired to populate the taxonomy. This lexically populated domain taxonomy, called Domain Lexico-Taxonomy, will provide the core information for identifying and processing of multiple domains information

2 Related Work

Unlike previous work, we aim to populate each the domain lexicon attached to each node in the taxonomy with lemmas from a general lexicon. Previous studies targets on assigning texts to specific categories, hence they use a limited taxonomy augmented with a small set of features (e.g. Yand and Pedersen (1997), Sebastiani (2002) and Avancini, et al. (2003)). However, although specialized lemmas are very useful in dealing with a single specific domain, they cannot be useful in multi-domain processing. The rationale is straightforward: since a specialized lemma has very restricted distribution, it is not likely to occur regularly in a multi-domain corpus. To achieve domain versatility in processing, it is necessary to identify lemmas with wider distributions and yet is associated with particular domain(s).

3 Resources Used

3.1 Domain Taxonomy

A domain taxonomy containing 549 domains is 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 be bilingual and that it be 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/). The Chinese version can also be found in Huang (2004).

3.2 WordNet and Sinica BOW

WordNet, an electronic lexical database, is considered to be one of the most important resources available to researchers in computational linguistics, text analysis, and many related areas (Miller et al., 1993; Fellaum, 1998). Its design is inspired by current psycholinguistic and computational theories of human lexical memory. 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).

There are several versions of WordNet, with WordNet 2.0 being the most recent one. The differences between these versions include the quantity of synsets and their definition. The version of WordNet that we use in this research is version 1.6, since this is the version most widely used by computational linguists. There are nearly 100,000 synsets in this version.

We mentioned earlier that we adopted a bilingual domain taxonomy to increase the versatility of our domain processing. Similarly, we use a bilingual wordnet as our lexical knowledgebase to achieve bilingual support to our study at the lexico-conceptual level. Each 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 resulted bilingual wordnet is further linked to the SUMO ontology to form the Academia Sinica Bilingual Ontological Wordnet (Sinica BOW, Huang and Chang, 2004). We use the semantic relations in bilingual resource to expand and predict domain classification when it cannot be judged directly from a lexical lemma.

4 Domain Lexico-Taxonomy

Recall that we define our Domain Lexico-Taxonomy (DLT) as a domain taxonomy populated with lexical entries. In other words, each domain taxonomy node will become a small domain lexicon. And the lemmas populating these lexica will be generally used lemmas as defined by the general lexical knowledgbase WordNet. In other words, we hope that cues from these domain classified lexical items will help us to identify a domain without using domain specific resources. There are hence two test for our approaches: whether the DLT can be efficiently built, and whether the DLT can successfully predict domain of a unknown text.

4.1 Populating DLT using WordNet lexical knowledge

The current study maps WordNet synsets to domain taxonomy. Note that we have English-Chinese bilingual pairs both for our WordNet synsets and our domain taxonomy. Two types of relations between WordNet and Domain Taxonomy are explored: Identity, and Hyponymy. That is, we tried to link the lemmas to domain taxonomy when the domain has an identical WordNet lemma or when the domain is a hyponym of a WordNet lemma. Post-mapping checking reveals that hypernymy link yields low accuracy and will not be discussed here.

463 of the 549 domain labels have a directly corresponding WordNet synset through an identical lemma. The mapping relation is over 97% correct.

4.2 Expansion with hyponymy

Since WordNet directly encodes the 'is-a' relation by hyponymy, we assume that both the synset members and their hyponym synsets belong to that domain. The process of populating DLT is shown in Figure 1. Thus, the 463 domains are 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.

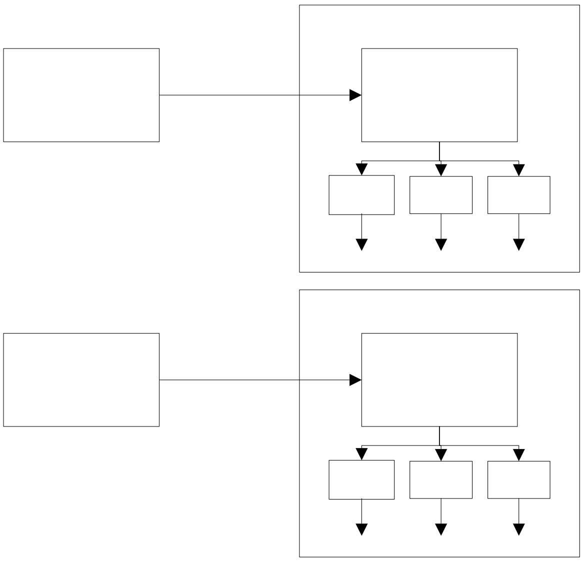

Figure 1. Populating DLT from WordNet

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.

# of domain each lemma is assigned	# of lemmas
1	6464
2	2417
3	529
4	173
5	16
6	7
7	6

Table 1. Multi-domain lemmas

With regard to the size of resultant domain lexicon, the number of entries ranges from 1 to 3762. The average size of these domain lexica is 32.8 lemmas. Due to the a great number of lexica with small number of entries assigned, only 41 domains lexical contain more than 32.8 lemmas. These domain lexica and their sizes are shown in Table 2.

Domain	Domain
Vertebrates (脊椎動物) -- 3676	Mathematics (數學) – 69
food(食品) -- 2968	Humanities (人文學科) – 64
Bird(鳥類) -- 1059	Social Science (社會科學) – 62
Fish(魚類) -- 729	physics(物理學) -- 56
language(語言) -- 699	Biology(生物學) -- 56
Recreation (休閒娛樂) -- 548	Distribution(物流) -- 54
Insect(昆蟲) -- 515	computing(計算) -- 54
Natural Science (自然科學) -- 262	Genre(語體) -- 54
Country(國家) -- 250	Religion(宗教) -- 52
contest(競賽) -- 207	Religious Music (宗教音樂) – 48
music(音樂) -- 192	Plastic art (造形藝術) – 45
Indian(印地安) -- 188	Pure mathematics (純數學) -- 44
Sports(運動) -- 180	Anthropology (人類學) -- 42
commerce(商業) -- 144	Earth science (地球科學) -- 39
business(生意) -- 144	drawing(素描) -- 39
Dance(舞蹈) -- 124	Norse Mythology (北歐神話) -- 39
Heraldric design (紋章設計) – 120	philosophy(哲學) -- 37
Medical Science (醫療科學) – 85	Telecommunication (電信通訊) -- 35
medicine(醫學) -- 76	theather(戲劇) -- 34
Pathological medicine (病理醫學) -- 76	Fine Arts(藝術) -- 33
Clinical medicine (臨床醫學) -- 76	

Table 2. Domain lexica containing more than 32.8 lemmas

4.3 Evaluation: precision of domain lexica

Since the effective size of a domain lexicon is highly dependent upon the nature of the size and task involved, we will not attempt to pre-assign a threshold number for an effective domain lexicon. However, we can offer some data from our experiment to indicate the scale of our study. Of the domain lexica where entries are successfully assigned, 88 lexica have 10 or more entries while only 17 lexica have 100 or more entries. The above number also underlines the fact that we cannot formally evaluate the recall rate of this 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. Note that among the mapped lemmas, 8696 (out of 15,160) lemmas are assigned to multiple domains, while 6,464 are assigned to single domain. We can assume that the uniquely assigned domains to be highly reliable. Hence we look at the precision of all 8,696 multi-domain lemmas first. Among these lemmas, only 4.81% (418) proves to be wrong; and an overwhelming majority of 95.19% turns out to be correct (8278). In other words, we showed that our bootstrapping from bilingual WordNet yields reliable data.

Second, a more meaningful test is to look at individual lexicon to see how well the domain lexica are defined. We randomly chose five effective domain lexica for evaluation. To prevent potential data sparseness problems, we chose among those lexica with over 100 entries: 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 3:

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 3. Size and Precision of selected domain lexica

Given the overall precision of over 95%, Table 3 shows that the lowest precision, of all five manually checked lexica is 69.85% for the domain lexicon of “Natural Science”. There are two main reasons to account for the lower precision. First, “science” is a generic term easily confused with other domain concepts. For example, "Medical science", "Formal science" and so on all contribute to wrongly assigned entries. Second, "Natural Science" is a first layer domain term, hence it is very general and more easily overlap with related terms, such as “Medicine”. On the other hand, the fact that the other domain lexica have a much higher precision range, mostly in the nineties, is indeed very strong support that the bilingual wordnet based approach to DLT is promising.

5 Applying DLT to overcome the domain knowledge barrier

We pointed out in the last section that since we started out with an empty slate, there is no real baseline for the construction of DLT. We also showed that the current approach yields high precision. In this section, further support of our approach will be given by a small but successful application to domain knowledge processing. In this study we choose to use Google as the baseline since it is the default tool for obtaining domain knowledge on the web.

The domain barrier that we proposed to overcome is the lack of specific domain knowledge. Hence, we simulate this by using only the domain name but no other terms in Google search. We search for three domain terms in both English and Chinese: Insect 昆蟲: 515 entries, Dance 舞蹈: 124 entries, and Religious Music 宗教音樂: 48 entries. Please note that we chose three very different domains with different domain lexicon sizes. The first 30 web pages returned by Google are taken as the test data, a total of 180 web pages.

Our application of DLT to obtain domain knowledge is straightforward. A web page describing a given domain should contain occurrences of the domain lexicon. And a web with more prominent presence of the domain lexicon is the more relevant page to that domain. Since a Google query does not directly access the web content, it is well-known that human users must do further screening. We try to see if DLT can be used to successfully evaluate the web content. According the assumption, we set the below formula to calculate the relevance scores between one webpage and a specific domain.

$$Score(P,D) = \alpha \times tf + (1-\alpha) \times wf$$

P : Webpage of Google Query returned
D : A specific domain
tf: The term (i.e. lemma) frequency of Domain lexical items in P
wf: The word (i.e. token) frequency of Domain lexical items in P

For each webpage P, we calculate its relevance score in the domain D. According the scores, we can rearrange the relevant rank of the web pages.

For evaluation, we ask 10 users to judge these web pages for their domain relevance. A 5-point scale is used. And the average score of the 10 scorers are taken as our target. We take the scores of above 3 as the threshold for being highly relevant for the target domain.

5.1 Accuracy and recall

Since we take Google search result as the baseline and take the N web pages presented on the first page returned by Google will be the baseline result. For Google, N is usually default to 10. Hence our accuracy evaluation is to compare the precision (defined as receiving a grade of higher than 3 from the human grader) of our top 10 web pages. This will then be compared with the baseline of 10 web pages returned by Google.

Webpage	**Google top 10 (%)**	**Our top 10 (%)**
Insect (Traditional Chinese)	70	70
Insect(English)	60	80
Dance (Traditional Chinese)	30	60
Dance(English)	30	70
Religious Music (Traditional Chinese)	50	70
Religious Music (English)	60	60

Table 4. The accuracy compression of Google top 10 and our top 10 results in each domain

The above table shows that the top 10 query result based on DLT is equal or better than Google result. In average, our accuracy rate is 18.3% better than Google. And the difference can be as large as 30%. In addition, it is also observed that even though the variation is large for Google top 10 (40%), the variation is relatively small for DLT (20%).

We next compare our recall with the Recall from Google. Again, all web pages rated greater than 3 are consider a correct return. And for the DLT based recall, we use the hreshold score of 1 to eliminate the occurrences of a domain lexical item by accident. The result is given in Table 5.

Webpage of	Recall rate (%)
Insect (Traditional Chinese)	69.23
Insect(English)	86.67
Dance (Traditional Chinese)	54.55
Dance(English)	46.15
Religious Music (Traditional Chinese)	88.89
Religious Music (English)	54.55

Table 5. The recall rate of the DLT method

Our result varies greatly from just under 50% to nearly 90%. A couple of observations are important. First, since the recall of the Religious domain is better than the Dance domain, it shows that the absolute lexicon size is not crucial in domain processing. We have suggested that the domain lexicon size required may vary form domain to domain. And we showed that a small lexicon size (48 for Religious Music) does not hurt our approach. Second, the low recall of the Dance domain needs some explanation. We found that the web pages in question are mostly visual files, which is reasonable given the topic. However, the DLT approach relies crucially on textual content. Hence the result is not as good.

5.2 Ranking discrepancy

The last test we will show is the measurement of ranking discrepancy. This is an attempt to see if the information extracted matches the expectation of human users. This is a very important attribute to measure the success of an information extraction system but not fully explored in the field. In this preliminary attempt, we compare the top 10 ranked return from our approach and that of Google with human ranking results.

It is obvious that we are very far away from actually mimicking human ranking results yet. Hence the only feature we measure now is ranking discrepancy. In this measure, we compare the relation of the nth ranked return and the n+1th ranked return in any given extraction system. If the ranking matches that of human (i.e. the nth ranked item is also higher ranked by human), than there is no discrepancy. And the pair receives a sore of +1. On the other hand, if there is a mismatch with human ranking, then the mismatch is given a score of –1. We take the Insect domain based on Chinese lexicon for example. For our no. 1, and no. 2, scores are both 3.5, therefore no discrepancy. For our no. 2 (3.5) > no. 3 (3.125), there is again no discrepancy. However, for our no. 3 (3.125) < no. 4 (3.375), there is discrepancy

Hence for the top 10 return, the range of possible scores are from +9 to –9. And we can say that a positive score indicates that the ranking is generally in line with human expectation, while a negative score indicates that the ranking does not reflect human expectation.

Webpage of	Google top 10	Our top 10
Insect (Traditional Chinese)	+3	+5
Insect(English)	+5	-1
Dance (Traditional Chinese)	+1	+3
Dance(English)	-3	+1
Religious Music (Traditional Chinese)	-1	-1
Religious Music (English)	-1	+3

Table 6. The ranking discrepancy of our top 10 and Google top 10

The above data shows that the DLT approach does in general improve the Google results and yields better ranking. Four ranking results are better, with only one worse than the original Google ranking. It can also be observed that four rankings received positive scores and can be considered to be close to human ranking, while the other two (-1) are only marginally unlike human ranking.

6 Conclusion and future work

DLT is our proposed first step towards an infrastructure for multi-domain language processing. In the paper, there are 15,160

Chinese lemmas that linked and distributed in 463 domain lexico-taxonomy nodes.

We first show that there is a high precision of the current assignment and hence support the claim that DLT can be effectively built. Since we were able to assign lexical items to 463 (of 579) domains using only one general lexical knowledgebase, we expect that the merging of other lexical and textual resources will help greatly to populate the DLT. In addition, in our test application, we also showed that it is not necessary to build large-scale lexicon for domain information extraction. We showed that when a domain is small, a small-scale lexicon (48 entries for Religious Music) is often sufficient in identifying domain information.

We next showed, bootrapping with Google search result, that DLT can be used to improve domain information processing. While applying a simply DLT based equation, we were able to improve on the initial Google search results in terms of accuracy, recall, as well as similarity to human ranking. Although the small experiment does not yet have real application value, it does prove that DLT contains the right domain knowledge to improve domain processing. In other words, it is shown the DLT approach is indeed promising as a tool to overcome the barrier of lack of domain knowledge.

In the future, we will continue to explore ways to populate LDT, such as integrating other resources or feedbacks from cross- and multi-domain processing work.

References

Avancini, F. Henri, Alberto Lavelli, Bernardo Magnini, Fabrizio Sebastiani, Roberto Zanoli. 2003. Expanding Domain-Specific Lexicons by Term Categorization. Proceedings of the 2003 ACM symposium on Applied computing.

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

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

Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. *Languages and Linguistics*. 4.3.509-532.

Miller G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. 1993. "Introduction to WordNet: An On-line Lexical Database," In Proceedings of the fifteenth International Joint Conference on Artificial Intelligence.

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

Yand Y. and J. O. Pedersen. 1997. 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.