Domain Specific Word Extraction from Hierarchical Web Documents: A First Step Toward Building Lexicon Trees from Web Corpora

Jing-Shin Chang
Department of Computer Science & Information Engineering
National Chi-Nan University
1, University Road, Puli, Nantou, Taiwan 545, ROC.
jshin@csie.ncnu.edu.tw

Abstract

Domain specific words and ontological information among words are important resources for general natural language applications. This paper proposes a statistical model for finding domain specific words (DSW's) in particular domains, and thus building the association among them. When applying this model to the hierarchical structure of the web directories node-by-node, the document tree can potentially be converted into a large semantically annotated lexicon tree. Some preliminary results show that the current approach is better than a conventional TF-IDF approach for measuring domain specificity. An average precision of 65.4% and an average recall of 36.3% are observed if the top-10% candidates are extracted as domain-specific words.

1 Domain Specific Words and Lexicon Trees as Important NLP Resources

Domain specific words (DSW's) are important "anchoring words" for natural language processing applications that involve word sense disambiguation (WSD). It is appreciated that multi-sense words appearing in the same document tend to be tagged with the same word sense if they belong to the same common domain in the semantic hierarchy (Yarowsky, 1995). The existence of some DSW's in a document will therefore be a strong evidence of a specific sense for words within the document. For instance, the existence of "basketball" in a document would strongly suggest the "sport" sense of the word "活塞" ("Pistons"), rather than its "mechanics" sense. It is also a personal belief that DSW-based sense disambiguation, document classification and many similar applications would be easier than sense-based models since sense-tagged documents are rare while domain-aware training documents are abundant on the Web. DSW identification is therefore an important issue.

On the other hand, the semantics hierarchy among words (especially among sets of domain specific words) as well as the membership of domain specific words are also important resources for general natural language processing applications, since the hierarchy will provide semantic links and ontological information (such as "is-A" and "part-of" relationships) for words, and, domain specific words belonging to the same domain may have the "synonym" or "antonym" relationships. A hierarchical lexicon tree (or a network, in general) (Fellbaum, 1998; Jurafsky and Martin, 2000), indicative of sets of highly associated domain specific words and their hierarchy, is therefore invaluable for NLP applications.

Manually constructing such a lexicon hierarchy and acquiring the associated words for each node in the hierarchy, however, is most likely unaffordable both in terms of time and cost. In addition, new words (or new usages of words) are dynamically produced day by day. For instance, the Chinese word "活塞" (pistons) is more frequently used as the "sport" or "basketball" sense (referring to the "Detroit

Pistons") in Chinese web pages rather than the "mechanics" or "automobile" sense. It is therefore desirable to find an automatic and inexpensive way to construct the whole hierarchy.

Since the hierarchical web pages provide semantic tag information (explicitly from the HTML/XML tags or implicitly from the directory names) and useful semantic links, it is desirable that the lexicon construction process could be conducted using the web corpora. Actually, the directory hierarchy of the Web can be regarded as a kind of classification tree for web documents, which assigns an implicit hidden tag (represented by the directory name) to each document and hence the embedded domain specific words. Converting such a hierarchy into a lexicon tree is therefore feasible, provided that we can remove non-specific terms from the associated document sets.

For instance, the domain-specific words for documents under the "sport" hierarchy are likely to be tagged with a "sport" tag. These tags, in turn, can be used in various word sense disambiguation (WSD) tasks and other hot applications like anti-spamming mail filters. Such rich annotation provides a useful knowledge source for mining various semantic links among words.

We therefore will explore a non-conventional view for constructing a lexicon tree from the web hierarchy, where domain-specific word identification turns out to be a key issue and the first step toward such a construction process. An inter-domain entropy (IDE) measure will be proposed for this purpose.

2 Conventional Clustering View for Constructing Lexicon Trees

One conventional way to construct the lexicon hierarchy from web corpora is to collect the terms in all web documents and measure the degree of word association between word pairs using some well-known association metrics (Church and Hanks, 1989; Smadja et al., 1996) as the distance measure. Terms of high association are then clustered bottom-up using some clustering techniques to build the hierarchy. The clustered hierarchy is then submitted to lexicographers to assign a semantic label to each sub-cluster. The cost will be reduced in this way, but could still be unaffordable. Besides, it still depends on the lexicographers to assign appropriate semantic tags to the list of highly associated words.

There are several disadvantages with this approach. Firstly, the hierarchical relationship among the web documents, and hence the embedded DSW's, is lost during the document collection process, since the words are collected without considering where they come from in the document hierarchy. The loss of such hierarchical information implies that the clustered one will not match human perception quite well. Secondly, the word association metric and the clustering criteria used by the clustering algorithm are not directly related to human perception. Therefore, the lexicographers may not be able to adjust the clustered hierarchy comfortably. Thirdly, most clustering algorithms merge terms in a binary way; this may not match human perception as well. As far as the computation cost is concerned, computation of word association based on *pairwise* word association metrics will be time consuming.

Actually, such an approach may not be the only option today, thanks to the large number of web documents, which are natively arranged in a *hierarchical* manner.

3 Lexicon Tree Construction as Domain Specific Word Detection from Web Hierarchy

Since the web documents virtually form an extremely huge document classification tree, we propose here a simple approach to convert it into a lexicon tree, and assign implicit semantic tags to the domain specific words in the web documents automatically.

This simple approach is inspired by the fact that most text materials (webpages) in websites are already classified in a hierarchical manner; the hierarchical directory structures implicitly suggest that the domain specific terms in the text

materials of a particular subdirectory are closely related to a common subject, which is identified by the name of the subdirectory.

If we can detect domain specific words within each document, and remove words that are non-specific, and tag the DSW's thus acquired with the directory name (or any appropriate tag), then we virtually get a hierarchical lexicon tree. In such a tree, each node is semantically linked by the original web document hierarchy, and each node has a set of domain specific words associated with it.

For instance, a subdirectory entitled 'entertainment' is likely to have a large number of web pages containing domain specific terms like 'singer', 'pop songs', 'rock & roll', 'Ah-Mei' (nickname of a pop song singer), 'album', and so on. Since these words are highly associated with the 'entertainment' domain, we will be able to collect the domain specific words of the 'entertainment' domain from such a directory.

In the extraction process, the directory names can be regarded as *implicit sense labels* or *implicit semantic tags* (which may be different from linguistically motivated semantic tags), and the action to put the web pages into properly named directories can be regarded as an *implicit tagging* process by the webmasters. And, the hierarchical directory itself provides information on the hierarchy of the semantic tags.

From a well-organized web site, we will then be able to acquire an implicitly tagged corpus from that site. Thanks to the webmasters, whose daily work include the implicit tagging of the corpora in their websites, there is almost no cost to extract DSW's from such web corpora. This idea actually extends equally well for other Internet resources, such as news groups and BBS articles, that are associated with hierarchical group names. Extending the idea to well organized book chapters, encyclopedia and things like that would not be surprised too.

The advantages of such a construction process, by removing non-specific terms, are many folds. First, the original hierarchical structure reflects human perception on document (and term) classification. Therefore, the need for adjustment may be rare, and the lexicographers may be more comfortable to adjust the hierarchy even if necessary. Second, the directory names may have higher correlation with linguistically motivated sense tags than those assigned by a clustering algorithm, since the web hierarchy was created by a human tagger (i.e., the webmaster). As far as the computation cost is concerned, pairwise word association computation is now replaced by the computation of "domain specificity" of words against domains. The reduction is significant, from O(|W|x|W|) to O(|W|x|D|), where |W| and |D| represent the vocabulary size and number of domains, respectively.

4 Domain Specific Word Extraction as the Key Technology: An Inter-Domain Entropy Approach

Since the terms (words or compound words) in the documents include general terms as well as domain-specific terms, the only problem then is an effective model to exclude those domain-independent terms from the implicit tagging process. The degree of domain independency can be measured with the inter-domain entropy (IDE) as will be defined in the following **DSW (Domain-Specific Word) Extraction Algorithm**. Intuitively, a term that distributes evenly in all domains is likely to be independent of any domain. We therefore weight such terms less probable as DSW's. The method can be summarized in the following algorithm:

Domain-Specific Word Extraction & Lexicon Tree Construction Algorithm:

Step1 (Data Collection): Acquire a large collection of web documents using a web spider while preserving the directory hierarchy of the documents. Strip unused markup tags from the web pages.

Step2 (Word Segmentation or Chunking): Identify word (or compound word) boundaries in the documents by applying a word segmentation process, such as

(Chiang et al., 1992; Lin et al., 1993), to Chinese-like documents (where word boundaries are not explicit) or applying a compound word chunking algorithms to English-like documents (where word boundaries are clear) in order to identify interested word entities.

Step3 (Acquiring Normalized Term Frequencies for all Words in Various Domains): For each subdirectory d_j, find the number of occurrences n_{ij} of each term w_i in all the documents, and derive the normalized term frequency $f_{ij} = n_{ij} / N_j$ by normalizing n_{ij} with the total document size, $N_j \equiv \sum_i n_{ij}$, in that directory. The directory is then associated with a set of $< w_i, d_j, f_{ij} >$ tuples, where w_i is the i-th words of the complete word list for all documents, d_j is the j-th directory name (refer to as the domain hereafter), and $f_{ij} = n_{ij} / N_j$ is the normalized relative frequency of occurrence of w_i in domain d_j.

Step4 (Identifying Domain-Independent Terms): Domain-independent terms are identified as those terms which distributed evenly in all domains. That is, terms with large **Inter-Domain Entropy** (IDE) defined as follows:

$$H_i \equiv H(w_i) \equiv -\sum_j P_{ij} \log P_{ij}$$

$$P_{ij} \equiv \frac{f_{ij}}{\sum_j f_{ij}}$$

Terms whose IDE's are above a threshold are likely to be removed from the lexicon tree since such terms are unlikely to be associated with any particular domain. Terms with a low IDE, on the other hand, may be retained in a few domains with high normalized frequencies.

To appreciate the fact that a high frequency term may be more important in a domain, the IDE is further weighted by the term frequency in the particular domain when deciding whether a term should be removed. Currently, the weighting method is the same as the conventional TF-IDF method (Baeza-Yates and Ribeiro-Neto, 1998; Jurafsky and Martin, 2000) for information retrieval. In brief, a word with entropy *Hi* can be think of as a term that spreads in *2**Hi* domains on average. The equivalent number of domains a term could be found then can be equated to $Nd_i = 2^{H_i}$. The term weight for w_i in the *j-th* domain can then be estimated as:

$$W_{ij} = n_{ij} \times \log_2 \left(\frac{N}{Nd_i} \right)$$

where N is the total number of domains. Unlike the conventional TF-IDF method, however, the expected number of domains that a term could be found is estimated by considering its probabilistic distribution, instead of simple counting.

Step5 (Output): Sort the words in each domain by decreasing weights, W_{ij}, and output the top-k% candidates as the domain specific words of the domain. The percentage (k) can be determined empirically, or based on other criteria, such as their classification performance in a DSW-based text classifier (Chang, 2005). The directory tree now represents a hierarchical classification of the domain specific terms for different domains.

Since the document tree may not be really perfect, we have the option to adjust the hierarchy or the sets of words associated with each node, after eliminating domain-independent terms from the directory tree. The terms can be further clustered into highly associated word lists, with other association metrics. On the other hand, we can further move terms that are less specific to the current domain upward toward the root. This action will associate such terms with a slightly more general domain. All these issues will be left as our future works.

However, the current method is independent of the source web hierarchy. Given a web organized as an encyclopedium of biology, the current method is likely to find out the living species associated with each node of the underlying taxonomy automatically. With more and more well organized web sites of various kinds of knowledge online, the problems with imperfect web hierarchy will hopefully become a less important issue.

5 Evaluation

To see how the above algorithm could be useful as a basis for building a large lexicon tree from web pages, some preliminary results will be examined in this section.

A large collection of Chinese web pages was collected from a local news site. The size of the HTML web pages amounts to about 200M bytes in 138 subdomains (including the most specific domains at the leaf nodes and their ancestor domains). About 16,000 unique words are identified after word segmentation is applied to the text parts.

It was observed, from some small sample domains, that only around 10% of the words in each subdomain are deemed domain specific. (The percentages, however, may vary from domain to domain.) The large vocabulary size and the small percentage of DSW's suggest that the domain specific word identification task is not an easy one.

Table 1 shows a list of highly associated domain-specific words of low inter-domain entropies and their domain names. (Literal English translation for each term is enclosed in the parenthesis.) They are sampled from 4 out of 138 subdomains. The domain names virtually act as the semantic tags for such word lists. The tags, being extracted from manually created directory, well reflect the senses of the words in each subdomains.

Table 1 shows that many domain-specific words can really be extracted with the proposed approach in their respect domains. For instance, the word "pitcher" ("投手") is specifically used in the "baseball" domain. The domain specific words and their domain tags are well associated.

As a result of such association, low inter-domain entropy words in the same domain are also highly correlated. For instance, the term "部長" for calling a *Japanese* baseball team "manager" is specifically used with "日本職棒" (Japanese professional baseball team), instead of a *Chinese* team, where "manager" is called differently.

In addition, new usages of words, such as "活塞 (Pistons)" with the "basketball" sense, could also be identified by the current approach. Furthermore, it was also observed that many irrelevant words (such as those words in the webmasters' advertisement) are rejected as the DSW candidates automatically since they have very high inter-domain entropies.

One can also find interesting lexical relations (Fellbaum, 1998) among the domain tags and domain specific words, form Table 1, such as:

Hypernym/Hyponym: athlete (運動) *vs.* baseball game (棒球賽); car (汽車) *vs.* small car (小型車).

Has-Member/Member-Of: baseball team (球團) *vs.* manager (部長), pitcher (投手).

Has-Part/Part-Of: car (汽車) *vs.* engine cover (引擎蓋), tank (水箱), safety system (安全系統), trunk (行李箱).

Antonym: shot (投籃) *vs.* defense (防守).

Such lexical relations are, in general, interesting to lexical database builders. Furthermore, for data driven applications, such fine details are unlikely to be listed in a general purpose lexical database. Extracting DSW's with the inter-domain entropy (IDE) metric is therefore well founded.

Baseball	Broadcast-TV	Basketball	Car
日本職棒 (Japanese professional baseball)	有線電視 (cable TV)	一分 (one minute)	千西西 (Kilo-c.c.)
棒球賽 (baseball games)	東風 (the Dong Fong TV Station)	三秒 (three seconds)	小型車 (small car)
熱身 (warm up)	開工 (start to work)	女子組 (girl's teams)	中古 (used car)
運動 (athlete)	節目中 (on air)	包夾 (fold; clip)	引擎蓋 (engine cover)
場次 (time table)	廣電處 (radio-tv office)	外線	水箱 (tank)
價碼 (cost)	收視	犯規 (foul)	加裝
球團 (baseball team)	和信 (Ho-Hsin TV Station)	投籃 (shot)	市場買氣 (market atmosphere)
部長 (manager)	新聞局 (government information office)	男子組 (male team)	目的地 (destination)
練球 (practicing)	開獎	防守 (defense)	交車 (car delivery)
興農 (Hsin-Lung team)	頻道 (channel)	冠軍戰(championship)	同級 (of the same grade)
球場(course; diamond)	電視 (TV)	後衛 (fullback)	合作開發 (co-development)
投手 (pitcher)	電影(movie)	活塞 (Pistons team)	安全系統 (safety system)
球季 (season)	熱門 (hot)	國男 (national male team)	行李 (luggage)
賽程 (schedule)	影視 (video)	華勒(Wallace)	行李廂 (trunk)
太陽 (the Sun team)	娛樂 (entertainment)	費城 (Philadelphia)	西西 (c.c.)

Table 1. Sampled domain specific words with low entropies.

In order to have a quantitative evaluation, we have inspected a few domains of small sizes (each containing about 300 unique words or less) for a preliminary estimation. The top-10% candidates with lowest inter-domain entropy, weighted by their term frequencies in their respect domains, are evaluated. (The 10% threshold is selected arbitrarily.) Table 2 shows the results in terms of precision (P), recall (R) and F-measure (F). The column with the '#Words' label shows the numbers of unique words used in the 5 domains.

Since it is difficult sometimes to have a consistent judgment on "domain specificity", the estimation could vary drastically on other domains by other persons. For this reason, the degree of domain specificity is ranked from 0 (irrelevant) to 5 (absolutely specific to the domain) points. Therefore, when computing the precision and recall measures, a completely "correct" answer should have a grading point '5'. Fortunately, most terms are assigned the grading point 5, with a few less certain cases assigned '3' or '4'.

Domain	#Words	R	P	F
Baseball	149	29.7	68.0	41.3
Basketball	277	26.2	60.7	36.6
Broadcast -TV	161	47.6	50.0	48.8
Education	255	40.0	81.5	53.7
Health -care	263	38.2	66.9	48.6
(Average)		36.3	65.4	45.8

Table 2. Performance of the Top-10% DSW candidate lists in 5 sample domains.

Table 2 shows that, by only gathering the first 10% of the word lists, we can identify about 36% of the embedded domain specific words, and the precision is as high as 65%. Therefore, we can identify significant amount of DSW's about every 1.5 entries from the top-10% list of low entropy words.

Since the TF-IDF (term frequency-inverse document frequency) approach (Baeza-Yates and Ribeiro-Neto, 1998; Jurafsky and Martin 2000) is widely used in information retrieval applications for indexing important terms within documents, it can also be applied to identify domain specific words in various domains. To make a comparison, the TF-IDF term weighting method is also applied to the same corpus. The "baseball" domain is then inspected for their differences. It turns out that the top-10% candidate lists of both methods show the *same* performance. However, the IDE measure appears to reach the highest precision faster than the TF-IDF approach. Furthermore, the IDE measure has a better top-20% performance than that of the TF-IDF approach as listed in Table 3.

Model	R	P	F
IDE	48.3	55.3	51.6
TF-IDF	44.8	51.3	47.8

Table 3. Comparison of the top-20% candidate list performance between IDE and TFIDF -based approaches.

Although it is not sure whether the superiority of the IDE approach will retain when examining larger corpora, it does have its advantages in indicating the "degree of specificity". In particular, the degree of domain specificity of a term is estimated by considering the cross-domain *probability distribution* of the term in the current IDE-based approach. Instead, the TF-IDF approaches only count the number of domains a term was found as a measure of randomness. The IDE approach is therefore gaining a little bit performance than a TF-IDF model.

The results partially confirm our expectation to build a large semantically annotated lexicon tree from the web pages using the implicit tags associated with the web directory hierarchy and removing non-specific words from web documents.

6 Error Sources and Future Works

In spite of some encouraging results, we also observed some adverse effects in using the single inter-domain entropy metric to identify domain specific words. For instance, some non-specific words may also have low entropy simply because they appear in only one domain (IDE=0). Since such words cannot be distinguished from real "domain-specific" words, there should be other knowledge sources to reduce false alarms of this kind (known as the Type II errors.)

On the other hand, some multiple sense words may have too many senses such that they are considered non-specific in each domain (although the sense is unique in each respect domain). This is a typical Type I error we have observed.

As a result, further refinement of the purely statistical model is required to improve the precision of the current approach. Currently, we prefer a co-training approach inspired by the works in (Chang and Su, 1997; Chang, 1997), which is capable of augmenting a single IDE-like metric with other information sources.

We have also assumed that the directories of all web sites are well organized in the sense that the domain labels (directory names) are appropriate representatives of the documents under the directories. This assumption is not always satisfied since it depends on the site owners' views on the documents. There are good

chances that the hierarchies differ from site to site. Therefore, we may need some measures of site similarity, and approaches to unify the different hierarchies and naming policies as well. The answers to such problems are not yet clear. However, we believe that the hierarchy of the directories (even though not well named) had substantially reduces the cost for lexicogrophers who want to build a large semantically annotated lexicon tree. And the whole process will become more and more automatic as we refine the above model against more and more data.

7 Concluding Remarks

The major contribution of the proposed model is to extract highly associated sets of domain-specific words, and keeping their hierarchical links with other sets of domain specific words at low cost. These sets of highly associated domain specific words can thus be used directly for sense disambiguation and similar applications. The proposed model takes advantages of the rich web text resources and the implicit semantic tags implied in the directory hierarchy for web documents. Therefore, the requirement for manual tagging is negligible. The extracted lists of DSW's are not only useful for word sense disambiguation but also useful as a basis for constructing lexicon databases with rich semantic links. So far, an average precision of 65.4% and an average recall of 36.3% are observed if the top-10% candidates are extracted as domain-specific words. And it outperforms the TF-IDF method for term weighting in the current task.

Acknowledgements

Part of the work was supported by the National Science Council (NSC), ROC, under the contract NSC 90-2213-E-260-015-.

References

Christiane Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, NJ, USA.

David Yarowsky. 1995. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196, MIT, MA, USA.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, 22(1):1-38.

Jing-Shin Chang and Keh-Yih Su. 1997. "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", *International Journal of Computational Linguistics and Chinese Language Processing (CLCLP)*, 2(2):97-148.

Jing-Shin Chang. 1997. *Automatic Lexicon Acquisition and Precision-Recall Maximization for Untagged Text Corpora*, Ph.D. dissertation, Department of Electrical Engineering, National Tsing-Hua University, Hsinchu, Taiwan, R.O.C.

Jing-Shin Chang. 2005. "Web Document Classification into a Hierarchical Document Tree Based on Domain Specific Words", submitted.

Ken Church and Patrick Hanks. 1989. "Word Association Norms, Mutual Information, and Lexicography," *Proc. 27th Annual Meeting of the ACL*, pp. 76-83, University of British Columbia, Vancouver, British Columbia, Canada.

Ming-Yu Lin, Tung-Hui Chiang and Keh-Yih Su. 1993. "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," *Proceedings of ROCLING VI*, pp. 119-142.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval,* Addison Wesley, New York.

Tung-Hui Chiang, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su. 1992. "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING-V*, pp. 123-146, Taipei, Taiwan, R.O.C.