Domain Specific Corpora from the Web[1]

Avinesh PVS, Diana McCarthy, Dominic Glennon & Jan Pomikálek

Keywords: *domain corpus*, *DANTE*, *WebBootCat*.

Abstract

Language usage is dependent on domain and, as a consequence, domain specific corpora are extremely useful for language learning and lexicography. It is possible to label heterogeneous data for domain either manually or automatically using human knowledge or machine learning. State-of-the-art text classification uses supervised techniques whereby a system learns from previously annotated data. This works well when such data is available in sufficient quantities for supervised machine learning, though often that is not the case depending on the domain and language required. Moreover, this approach assumes that the heterogeneous data in the available corpus covers the required domains. In this paper we present the results of an approach using WebBootCat to retrieve data from the web in eight specific domains. A key component of this work was the use of the DANTE database for generating seed words for initial web data retrieval. To tailor the corpus to the nuances of the domain categorisation that we required, we used some of our own corpus data already annotated with subject codes (domain codes) to help refine the seed words used at the start of the iterative web retrieval process. Human effort was needed to refine a whitelist of words for each domain to reduce the chance of irrelevant data due to ambiguous terms in the seeds and extracted keywords used for subsequent retrieval. The domain corpora retrieved are loaded in the Sketch Engine. The word sketches and sketch difference functionality help reveal appropriate domain specific behaviour of words in the respective corpora.

1. Introduction

Language usage is dependent on domain (Hanks, 2000) and domain specific corpora are consequently extremely useful for language learning and lexicography (Barrière, 2009; Drouin, 2004). It is possible to label heterogeneous data for domain either manually (Atkins et al., 2010) or automatically (for a survey see (Sebastiani, 2002)) using human knowledge or machine learning. State-of-the-art text classification uses supervised techniques whereby a system learns from previously annotated data. This works well when such data is available in sufficient quantities for supervised machine learning, though often that is not the case depending on the domain and language required. Moreover, this approach assumes that the heterogeneous data in the available corpus covers the required domains. In this paper we present the results of an alternative approach proposed by Baroni et al. (2006a) for creating domain specific corpora using the WebBootCat tool (Baroni et al., 2006b). Our work uses this technology to retrieve data from the web ineight specific domains that we require data for, using seed words generated from a lexical resource known as DANTE (Atkins et al., 2010). To tailor the corpus to the nuances of the domain categorisation that we required, we used some of our own corpus data, already annotated with subject codes (domain codes) to help refine the seed word list.

The paper is structured as follows. First we give details of the WebBootCat method and describe how it can be used for building domain specific corpora. Next we describe DANTE which has lexical entries with domain mark up which we exploit in this work. In section 4 we provide more details of the method we employed and in section 5 we give details of the specific corpora acquired in this project and some example domain specific lexicographic analysis we can now perform with the aid of the corpus query tool Sketch Engine (Kilgarriff et al., 2004).

2. WebBootCat

WebBootCat is a tool for producing corpora by retrieving documents from the web using either a set of seed words or a collection of urls. In this paper we use the seed word method which follows a bootstrapping approach first proposed by Baroni and Bernardini (2004). This is a two step iterative approach. In the first step a small seedlist of words[2] is used to collect the first version of a corpus from the Web by sending queries made up of these words to a search engine API. In the second step automatic term extraction is used to identify the words for the queries in the next iteration. We followed a refinement of this approach using WebBootCat as implemented within Sketch Engine with the following steps:

1. Gather a list of domain-specific 'seed words' as initial query words.
2. Repeat the following steps until the corpus is sufficiently large:
    a. Randomly select a small set ('tuple') of the query words to create a search query.
    b. Send this search query to a search engine API, which returns a list of 'search hits'.
    c. Apply a blacklist of problematic urls for removal and retain the remaining list.
    d. Filter the list of urls according to:
        1. size of the document
        2. body text extraction: extract 'cleaned text' from URLs using Justext[3] (Pomikálek, 2011): this removes recurring material like navigation bars, advertisements, links etc...
        3. ratio of stopwords in the document
        4. relevancy: ratio of 'whitelisted' words in the document. A whitelist of words is used to reduce ambiguity of the words in the queries.
    e. Remove near-duplicates.
    f. Semi-automatically extract keywords from the corpus so far created using the statistic described by Kilgarriff (2009). These will be used for step 2a in the next iteration.[4]
3. Tokenise, lemmatise and part-of-speech tag the data.
4. Finally load the corpus into the Sketch Engine corpus query tool.

The methodology is described more fully in section 4 below.

3. DANTE

DANTE[5] (Atkins et al., 2010) is a lexical database produced by a team of lexicographers scrutinising a 1.7 billion word corpus of English produced as a starting point for the New English Irish Dictionary. The database comprises approximately 92,000 entries for words and phrases with information and examples on every variety of lexical information that the lexicographers have deemed potentially relevant for a thorough and accurate description of English. One type of lexical information contained in DANTE that we exploit for this project is subject field (domain). There are 156 domains in this taxonomy and these have been used to mark the word senses within DANTE. We use eight of the domains that we specifically require corpora for: Medical, Finance, Law, Cooking, Food, Employment, Commerce and IT.

4. Methology

4.1. Seed Words

For each domain, we used monosemous words from DANTE within that domain as seeds. Using unambiguous words helped reduce the impact of ambiguity. To supplement these seeds, and to tailor the corpus to the specific nuances of our required domain classification we used some data within our own corpus (the Cambridge International Corpus: CIC) where the documents have been marked with subject codes that related to those eight domains we were interested in. We determined which subject codes related to the eight domains by a manual mapping and then used the primary subject code listed with each CIC document. We exploited this data by using it to extract seeds for each target corpus as follows. For any target corpus, for example Medical, we obtained a training counterpart from the CIC for the purpose of extracting seed words. The counterpart (referred to here as CICdom) is used to generate a normalised word list which is then compared with that for a reference corpus: the BNC.
The keywords for the CICdom corpus are calculated as follows and used as putative seeds for step 2a of the WebBootCat algorithm described above in section 2.

For both the CICdom corpus and reference corpus (BNC):

1. make a word frequency list
2. Normalise the list to frequency per-million
3. add 100 to each normalised frequency following (Kilgarriff, 2009)
4. for each word (w) calculate:
   Score (w) =
   (freq-per-million (w, CICdom) + 100) / (freq-per-million (w, BNC) +100)
5. re-rank the word list according to this score

These seeds were augmented with the DANTE seeds and then filtered and extended by manual inspection

4.2. *Details of the Data Collection*

For each domain, the software automatically constructs queries for the search engine by putting a tuple (size 3) of the seeds together.[6] When filtering we excluded files smaller than 5KB to increase the chance of connected text, and we removed files greater than 2MB to avoid any particular files dominating the composition of the corpus, which also tend to contain unconnected text. Connected text usually contains high proportions of function words Baroni (2005). We set a threshold of at least 36 function words and a ratio of function words of 0.25.

Usually texts have multiple instances on the Web. Most common types of duplicates and near-duplicates include mirrored websites, many presentation styles, similar or identical articles at various sources. Duplication artificially inflates the frequencies of some words in a

corpus, which may bias any analysis. We used Onion[7] for deduplication. This software is slightly adapted from that described in Pomikálek (2011).

Due to the inherent ambiguity of words, some extracted seeds from our own corpora, or in the bootstrapping process may be ambiguous and make it more likely that documents irrelevant to the domain are retrieved. To reduce the impact of this we produced a list of whitelist words from manual inspection of our seeds and introspection. The whitelisted words were selected as being relatively unambiguous and generally applicable to that domain e.g. *employee* for employment. We set thresholds on the total number of whitelist words in a document (unique types and tokens) and the ratio of whitelisted words: (no. of whitelisted words / total words in a document). The thresholds were set empirically for each domain.

5. Results from the corpus

Table 1. Domain Corpus Statistics.

Domain	**# of seeds**	**# of URLS**	**# of URLs (final)**	**#Total Tokens**
Medicine	172	40670	10281	35 M
Law	54	42339	10624	34 M
IT	48	42606	7328	30M
Commerce	37	35233	1321	17M
Cook	36	45989	10531	28M
Employment	28	27893	371	13M
Food	29	44488	6326	22M
Finance	24	43442	11928	32M

Table 1. contains statistics for each domain corpus, including the number of seed words used at the start of the WebBootCat algorithm, the number of documents (URLs) before and after the filtering, cleaning and deduplicating process, and the final number of tokens (words and punctuation). The data has been tokenised, lemmatised and part-of-speech tagged with treetagger (Schmid, 1994)[8] and then loaded into the Sketch Engine and processed with a sketch grammar to reveal, as well as new keywords and terminology, the key grammatical collocations in the corpora.

In figure 1. for example, we show a small part of the word sketch for the verb *induce* in the medical corpus. The corpora and functionality enable the lexicographer to quickly determine key collocates and usages of words within the domain. Due to new functionality in Sketch Engine, it is also possible to explore differences in word sketches across domains, for example we can use the sketch difference function to see the difference in meaning of the noun *stock* in the Cook corpus compared to the Finance corpus and determine that *stock* tends to be a subject of *simmer* in the former and subject of *outperform* in the latter. Figure 2 highlights further examples of the behaviour of the noun *stock* in the Cook and Finance corpora with respect to attached prepositional phrases with i*n* and *to*. Salient collocates for the Finance corpus are highlighted in red while those for the Cook corpus are shown in green. The exact shade depends on the salience of the collocate where salience is the log dice calculated as described by Rychlý (2008).

object	22482	5.3	subject	7508	3.0	modifier	3052	1.2
apoptosis	334	8.7	stimulation	51	5.95	experimentally	108	9.67
ovulation	140	7.46	radiation	81	5.77	chemically	88	9.01
coma	148	7.34	injection	60	4.97	fraudulently	53	8.34
activation	185	7.21	stress	91	4.87	medically	77	8.24
maturation	127	6.86	acid	57	4.52	artificially	81	8.21
abortion	271	6.78	alcohol	41	3.72	also	277	1.97
remission	87	6.63	drug	183	3.54	often	54	1.95
vomiting	95	6.51	treatment	72	2.37	then	68	1.66
anesthesia	104	6.38	agent	50	2.35	even	42	1.1
gvbd	52	6.2	factor	53	2.35	only	62	0.97

Figure 1. A Small portion of a word sketch of the verb induce in the medical domain.

pp_in-i	119	749	-1.8	0.9
portfolio	0	61	0.0	4.0
anticipation	0	4	0.0	2.8
index	0	22	0.0	2.6
sector	0	30	0.0	1.5
corporation	0	9	0.0	1.4
quantity	0	4	0.0	1.0
trade	0	16	0.0	0.8
general	0	7	0.0	0.6
fund	0	29	0.0	0.4
account	0	19	0.0	0.4
bowl	4	0	1.0	0.0
pot	8	0	1.6	0.0
pan	6	0	2.4	0.0
fridge	6	0	2.9	0.0
saucepan	12	0	4.7	0.0

pp_to-i	84	171	-1.6	-0.6
public	0	12	0.0	0.8
investor	0	13	0.0	0.2
rice	5	0	1.1	0.0
boil	24	0	5.8	0.0

Figure 2. An example of a small portion of a sketch difference for the noun stock when comparing the cook and finance domain corpora. Salient collocates for the finance corpus are highlighted in red while those for the cook corpus are shown in green.

6. Conclusion

In this paper we presented work to acquire domain specific corpora for lexicographic purposes using seeds and the WebBootCat tool to iteratively retrieve documents from the web. The seed words were obtained from DANTE, supplemented with some automatically extracted keywords from corpora tagged with related subject codes. Human effort was needed to refine the whitelisted words for each domain to reduce the chance of irrelevant data. The domain corpora retrieved are loaded in the Sketch Engine and the word sketches

and sketch difference functionality help reveal appropriate domain specific behaviour of words in the respective corpora.

Some domains are less prevalent than others and have less distinctive keywords, for example employment has less distinctive keywords and some, such as *occupation*, are ambiguous and prevalent in web data in other meanings (i.e. military occupation). Some domains such as Cook and Food, and Finance and Commerce domains are semantically close and have considerable overlap of keywords. For future work it would be useful to explore further ways of generating and validating the seed and whitelist words to increase the relevance of the retrieved data for the given domain while not compromising on the variety of data within that domain.

Notes

[1] This work was partially supported by an EU ICT-2009.2.2: Language-based Interaction-2001-34460 project: PRESEMT (Pattern REcognition-based Statistically Enhanced MT).
[2] A seedlist is a list of words which are anticipated to be salient in the domain.
[3] http://code.google.com/p/justext/
[4] https://trac.sketchengine.co.uk/wiki/SimpleMaths. We extract the keywords automatically and then manually select from the top 150.
[5] See http://www.webDANTE.net/ where you can find details of the project as well as data samples and a search facility.
[6] We used a tuple size of 4 for Employment to increase the chance of obtaining relevant data since it was less easy to find distinctive seeds for this domain.
[7] http://code.google.com/p/onion/
[8] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

References

Atkins, S., M. Rundell and A. Kilgarriff. 2010. 'Database of ANalysed Texts of English (DANTE).' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy / Afuk, 549–556.

Baroni, M. 2009. 'Distributions in text.' In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: An international handbook.* Berlin: Walter de Gruyter, 803–821.

Baroni, M. and S. Bernardini. 2004. 'Bootcat: Bootstrapping corpora and terms from the web.' In M. T. Lino et al. (eds.), *Fourth International Conference on Language Resources and Evaluation, held in memory of Antonio Zampolli: proceedings*. Paris: ELRA, 1313–1316.

Baroni, M., A. Kilgarriff, J. Pomikalek and P. Rychly. 2006a. 'Webbootcat: Instant domain-specific corpora to support human translators.' In *EAMT-2006 11th Annual Conference of the European Association for Machine Translation June 19 & 20, 2006 Oslo University (Norway)*, 247–252.

Baroni, M., A. Kilgarriff, J. Pomikálek and P. Rychlý. 2006b. 'Webbootcat: a web tool for instant corpora.' In E. Corino, C. Marello and C. Onesti (eds.), *Atti del XII Congresso Internazionale di Lessicografia : Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso, 123–131.

Barrière, C. 2009. 'Finding domain specific collocations and concordances on the web.' In I. Ilisei, V. Pekar and S. Bernardini (eds.), *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography,*

and Language Learning, Borovets, Bulgaria, September. Stroudsburg, PA: Association for Computational Linguistics, 1–8.

Drouin, P. 2004. 'Detection of domain specific terminology using corpora comparison.' In M. T. Lino et al. (eds.), *Fourth International Conference on Language Resources and Evaluation, held in memory of Antonio Zampolli: proceedings*. Paris: ELRA, 79–82.

Hanks, P. 2000. 'Contributions of lexicography and corpus linguistics to a theory of language performance.' In U. Heid et al. (eds.), *Proceedings of the 9th Euralex International Congress, EURALEX 2000, Stuttgart, Germany, August 8th - 12th, 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 3–13.

Kilgarriff, A. 2009. 'Simple maths for keywords.' In M. Mahlberg, V.González-Díaz and C. Smith (eds.). *Proceedings of the Corpus Linguistics Conference, CL2009, University of Liverpool, UK , 20-23 July 2009*. Liverpool: University of Liverpool.

Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell. 2004. 'The sketch engine.' In G. Williams and S.Vessier (eds.), *Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004*. Lorient: Université de Bretagne-Sud, 105–116. Reprinted in Patrick Hanks (ed.) 2007. Lexicology: Critical concepts in Linguistics. London: Routledge.

Pomikálek. J. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora.* PhD Thesis, Masaryk University.

Rychlý, P. 2008. 'A lexicographer-friendly association score.' In P. Sojka and A. Horák (eds.), *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*, Brno: Masaryk University, 6–9.

Schmid. H. 1994. 'Probabilistic Part-of-Speech Tagging Using Decision Trees.' In D. Jones (ed.), *Proceedings of the Conference : International Conference on New Methods in Language Processing: (NeMLaP), September 14-16 1994, The University of Manchester Institute of Science and Technology Manchester United Kingdom*. Manchester: Centre for Computational Linguistics, 44–49.

Sebastiani. F. 2002. 'Machine learning in automated text categorization.' *ACM Computing Surveys* 34(1): 1–47.