Automatic Detection of Multilingual Dictionaries on the Web

Gintarė Grigonytė[♠] and **Timothy Baldwin**[♡]

♠ Department of Linguistics, Stockholm University
♡ Department of Computing and Information Systems, The University of Melbourne
gintare@ling.su.se tb@ldwin.net

Abstract

This paper presents an approach to query construction to detect multilingual dictionaries for predetermined language combinations on the web, based on the identification of terms which are likely to occur in bilingual dictionaries but not in general web documents. We use eight target languages for our case study, and train our method on pre-identified multilingual dictionaries and the Wikipedia dump for each of our languages.

1 Motivation

Translation dictionaries and other multilingual lexical resources are valuable in a myriad of contexts, from language preservation (Thieberger and Berez, 2012) to language learning (Laufer and Hadar, 1997), cross-language information retrieval (Nie, 2010) and machine translation (Munteanu and Marcu, 2005; Soderland et al., 2009). While there are syndicated efforts to produce multilingual dictionaries for different pairings of the world's languages such as `freedict.org`, more commonly, multilingual dictionaries are developed in isolation for a specific set of languages, with ad hoc formatting, great variability in lexical coverage, and no central indexing of the content or existence of that dictionary (Baldwin et al., 2010). Projects such as `panlex.org` aspire to aggregate these dictionaries into a single lexical database, but are hampered by the need to identify individual multilingual dictionaries, especially for language pairs where there is a sparsity of data from existing dictionaries (Baldwin et al., 2010; Kamholz and Pool, to appear). This paper is an attempt to automate the detection of multilingual dictionaries on the web, through query construction for an arbitrary language pair. Note that for the method to work, we require that the dictionary occurs in "list form", that is it takes the form of a single document (or at least, a significant number of dictionary entries on a single page), and is not split across multiple small-scale sub-documents.

2 Related Work

This research seeks to identify documents of a particular type on the web, namely multilingual dictionaries. Related work broadly falls into four categories: (1) mining of parallel corpora; (2) automatic construction of bilingual dictionaries/thesauri; (3) automatic detection of multilingual documents; and (4) classification of document genre.

Parallel corpus construction is the task of automatically detecting document sets that contain the same content in different languages, commonly based on a combination of site-structural and content-based features (Chen and Nie, 2000; Resnik and Smith, 2003). Such methods could potentially identify parallel word lists from which to construct a bilingual dictionary, although more realistically, bilingual dictionaries exist as single documents and are not well suited to this style of analysis.

Methods have also been proposed to automatically construct bilingual dictionaries or thesauri, e.g. based on crosslingual glossing in predictable patterns such as a technical term being immediately proceeded by that term in a lingua franca source language such as English (Nagata et al., 2001; Yu and Tsujii, 2009). Alternatively, comparable or parallel corpora can be used to extract bilingual dictionaries based on crosslingual distributional similarity (Melamed, 1996; Fung, 1998). While the precision of these methods is generally relatively high, the recall is often very low, as there is a strong bias towards novel technical terms being glossed but more conventional terms not.

Also relevant to this work is research on lan-

Grigonyte, Gintare and Timothy Baldwin (to appear) Automatic Detection of Multilingual Dictionaries on the Web, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA.

guage identification, and specifically the detection of multilingual documents (Prager, 1999; Yamaguchi and Tanaka-Ishii, 2012; Lui et al., 2014). Here, multi-label document classification methods have been adapted to identify what mix of languages is present in a given document, which could be used as a pre-filter to locate documents containing a given mixture of languages, although there is, of course, no guarantee that a multilingual document is a dictionary.

Finally, document genre classification is relevant in that it is theoretically possible to develop a document categorisation method which classifies documents as multilingual dictionaries or not, with the obvious downside that it would need to be applied exhaustively to all documents on the web. The general assumption in genre classification is that the type of a document should be judged not by its content but rather by its form. A variety of document genre methods have been proposed, generally based on a mixture of structural and content-based features (Matsuda and Fukushima, 1999; Finn et al., 2002; zu Eissen and Stein, 2005).

While all of these lines of research are relevant to this work, as far as we are aware, there has not been work which has proposed a direct method for identifying pre-existing multilingual dictionaries in document collections.

3 Methodology

Our method is based on a query formulation approach, and querying against a pre-existing index of a document collection (e.g. the web) via an information retrieval system.

The first intuition underlying our approach is that certain words are a priori more "language-discriminating" than others, and should be preferred in query construction (e.g. *sushi* occurs as a [transliterated] word in a wide variety of languages, whereas *anti-discriminatory* is found predominantly in English documents). As such, we prefer search terms w_i with a higher value for $\max_l P(l|w_i)$, where l is the language of interest.

The second intuition is that the lexical coverage of dictionaries varies considerably, especially with multilingual lexicons, which are often compiled by a single developer or small community of developers, with little systematicity in what is including or not included in the dictionary. As such, if we are to follow a query construction approach to lexicon discovery, we need to be able to predict the likelihood of a given word w_i being included in an arbitrarily-selected dictionary D_l incorporating language l (i.e. $P(w_i|D_l)$). Factors which impact on this include the lexical prior of the word in the language (e.g. $P(\textit{paper}|\text{en}) > P(\textit{papyrus}|\text{en})$), whether they are lemmas or not (noting that multilingual dictionaries tend not to contain inflected word forms), and their word class (e.g. multilingual dictionaries tend to contain more nouns and verbs than function words).

The third intuition is that certain word *combinations* are more selective of multilingual dictionaries than others, i.e. if certain words are found together (e.g. *cruiser*, *gospel* and *noodle*), the containing document is highly likely to be a dictionary of some description rather than a "conventional" document.

Below, we describe our methodology for query construction based on these elements in greater detail. The only assumption on the method is that we have access to a selection of dictionaries D (mono- or multilingual) and a corpus of conventional (non-dictionary) documents C, and knowledge of the language(s) contained in each dictionary and document.

Given a set of dictionaries D_l for a language l and the complement set $D_{\bar{l}} = D \setminus D_l$, we first construct the lexicon L_l for that language as follows:

$$L_l = \left\{ w_i | w_i \in D_l \cap w_i \notin D_{\bar{l}} \right\} \quad (1)$$

This creates a language-discriminating lexicon for each language, satisfying the first criterion.

Lexical resources differ in size, scope and coverage. For instance, a well-developed, mature multilingual dictionary may contain over 100,000 multilingual lexical records, while a specialised 5-way multilingual domain dictionary may contain as few as 100 multilingual lexical records. In line with our second criterion, we want to select words which have a higher likelihood of occurrence in a multilingual dictionary involving that language. To this end, we calculate the weight $\text{sdict}(w_{i,l})$ for each word $w_{i,l} \in L_l$:

$$\text{sdict}(w_{i,l}) = \sum_{d \in D_l} \begin{cases} \frac{|L_l| - |d|}{|L_l|} & \text{if } w_{i,l} \in d \\ -\frac{|d|}{|L_l|} & \text{otherwise} \end{cases} \quad (2)$$

where $|d|$ is the size of dictionary d in terms of the number of lexemes it contains.

The final step is to weight words by their typicality in a given language, as calculated by their

likelihood of occurrence in a random document in that language. This is estimated by the proportion of Wikipedia documents in that language which contain the word in question:

$$\text{Score}(w_{i,l}) = \frac{df(w_{i,l})}{N_l}\text{sdict}(w_{i,l}) \quad (3)$$

where $df(w_{i,l})$ is the count of Wikipedia documents of language l which contain w_i, and N_l is the total number of Wikipedia documents in language l.

In all experiments in this paper, we assume that we have access to at least one multilingual dictionary containing each of our target languages, but in absence of such a dictionary, sdict$(w_{i,l})$ could be set to 1 for all words $w_{i,l}$ in the language.

The result of this term weighing is a ranked list of words for each language. The next step is to identify combinations of words that are likely to be found in multilingual dictionaries and not standard documents for a given language, in accordance with our third criterion.

3.1 Apriori-based query generation

We perform query construction for each language based on frequent item set mining, using the Apriori algorithm (Agrawal et al., 1993). For a given combination of languages (e.g. English and Swaheli), queries are then formed simply by combining monolingual queries for the component languages.

The basic approach is to use a modified support formulation within the Apriori algorithm to prefer word combinations that do not cooccur in regular documents. Based on the assumption that querying a (pre-indexed) document collection is relatively simple, we generate a range of queries of decreasing length and increasing likelihood of term co-occurrence in standard documents, and query until a non-empty set of results is returned.

The modified support formulation is as follows:

$$\text{cscore}(w_1, ..., w_n) = \begin{cases} 0 & \text{if } \exists d, w_i, w_j : \text{co}_d(w_i, w_j) \\ \prod_i Score(w_i) & \text{otherwise} \end{cases}$$

where $\text{co}_d(w_i, w_j)$ is a Boolean function which evaluates to true iff w_i and w_j co-occur in document d. That is, we reject any combinations of words which are found to co-occur in Wikipedia documents for that language. Note that the actual calculation of this co-occurrence can be performed efficiently, as: (a) for a given iteration of Apriori, it only needs to be performed between the new word that we are adding to the query ("item set" in the terminology of Apriori) and each of the other words in a non-zero support itemset from the previous iteration of the algorithm (which are guaranteed to not co-occur with each other); and (b) the determination of whether two terms collocate can be performed efficiently using an inverted index of Wikipedia for that language.

en - natives unenjoyable

de - andeuten tau anwuchs fügung

fr - collègue étouffée hybride

es - encendedor juntarse tensión

it - ardenne gradevole calcolare mancia

ar - الجيج بي أقواس الحربية عناصر الشرم ال

zh - 球员 胡同 粒子

ja - 冷房 メモリ 巡洋艦 福音 丼

Figure 1: Examples of learned queries for different languages

In our experiments, we apply the Apriori algorithm exhaustively for a given language with a support threshold of 0.5, and return the resultant item sets in ranked order of combined score for the component words.

A random selection of queries learned for each of the 8 languages targeted in this research is presented in Figure 1.

4 Experimental methodology

We evaluate our proposed methodology in two ways:

1. against a synthetic dataset, whereby we injected bilingual dictionaries into a collection of web documents, and evaluated the ability of the method to return multilingual dictionaries for individual languages; in this, we naively assume that all web documents in the background collection are not multilingual dictionaries, and as such, the results are potentially an underestimate of the true retrieval effectiveness.

2. against the open web via the Google search API for a given combination of languages, and hand evaluation of the returned documents

Lang	Wikipedia articles (M)	Dictionaries	Queries learned	Avg. query length
en	3.1	26	2546	3.2
zh	0.3	0	5034	3.6
es	0.5	2	356	2.9
ja	0.6	0	1532	3.3
de	1.0	13	634	2.7
fr	0.9	5	4126	3.0
it	0.6	4	1955	3.0
ar	0.1	2	9004	3.2

Table 1: Details of the training data and queries learned for each language

Note that the first evaluation with the synthetic dataset is based on *monolingual* dictionary retrieval effectiveness because we have very few (and often no) multilingual dictionaries for a given pairing of our target languages. For a given language, we are thus evaluating the ability of our method to retrieve multilingual dictionaries containing that language (and other indeterminate languages).

For both the synthetic dataset and open web experiments, we evaluate our method based on mean average precision (MAP), that is the mean of the average precision scores for each query which returns a non-empty result set.

To train our method, we use 52 bilingual Freedict (Freedict, 2011) dictionaries and Wikipedia[1] documents for each of our target languages. As there are no bilingual dictionaries in Freedict for Chinese and Japanese, the training of Score values is based on the Wikipedia documents only. Morphological segmentation for these two languages was carried out using MeCab (MeCab, 2011) and the Stanford Word Segmenter (Tseng et al., 2005), respectively. See Table 1 for details of the number of Wikipedia articles and dictionaries for each language.

Below, we detail the construction of the synthetic dataset.

4.1 Synthetic dataset

The synthetic dataset was constructed using a subset of ClueWeb09 (ClueWeb09, 2009) as the background web document collection. The original ClueWeb09 dataset consists of around 1 billion web pages in ten languages that were collected in January and February 2009. The relative proportions of documents in the different languages in the original dataset are as detailed in Table 2.

We randomly downsampled ClueWeb09 to 10 million documents for the 8 languages targeted in this research (the original 10 ClueWeb09 languages minus Korean and Portuguese). We then sourced a random set of 246 multilingual dictionaries that were used in the construction of `panlex.org`, and injected them into the document collection. Each of these dictionaries contains at least one of our 8 target languages, with the second language potentially being outside the 8. A total of 49 languages are contained in the dictionaries.

Language	Proportion
en (English)	48.41%
zh (Chinese)	17.05%
es (Spanish)	7.62%
ja (Japanese)	6.47%
de (German)	4.89%
fr (French)	4.79%
ko (Korean)	3.61%
it (Italian)	2.8%
pt (Portuguese)	2.62%
ar (Arabic)	1.74%

Table 2: Language proportions in ClueWeb09.

We indexed the synthetic dataset using Indri (Indri, 2009).

5 Results

First, we present results over the synthetic dataset in Table 3. As our baseline, we simply query for the language name and the term *dictionary* in the local language (e.g. *English dictionary*, for English) in the given language.

For languages that had bilingual dictionaries for training, the best results were obtained for Spanish, German, Italian and Arabic. Encouragingly, the results for languages with only Wikipedia documents (and no dictionaries) were largely comparable to those for languages with dictionaries, with Japanese achieving a MAP score comparable to the best results for languages with dictionary training data. The comparably low result for

[1] Based on 2009 dumps.

Lang	Dicts	MAP	Baseline
en	92	0.77	0.00
zh	7	0.75	0.00
es	34	0.98	0.04
ja	5	0.94	0.00
de	75	0.97	0.08
fr	34	0.84	0.03
it	8	0.95	0.01
ar	3	0.92	0.00
AVERAGE:	32.2	0.88	0.04

Table 3: Dictionary retrieval results over the synthetic dataset ("Dicts" = the number of dictionaries in the document collection for that language.

English is potentially affected by its prevalence both in the bilingual dictionaries in training (restricting the effective vocabulary size due to our L_l filtering), and in the document collection. Recall also that our MAP scores are an underestimate of the true results, and some of the ClueWeb09 documents returned for our queries are potentially relevant documents (i.e. multilingual dictionaries including the language of interest). For all languages, the baseline results were below 0.1, and substantially lower than the results for our method.

Looking next to the open web, we present in Table 4 results based on querying the Google search API with the 1000 longest queries for English paired with each of the other 7 target languages. Most queries returned no results; indeed, for the en-ar language pair, only 49/1000 queries returned documents. The results in Table 4 are based on manual evaluation of all documents returned for the first 50 queries, and determination of whether they were multilingual dictionaries containing the indicated languages.

The baseline results are substantially higher than those for the synthetic dataset, almost certainly a direct result of the greater sophistication and optimisation of the Google search engine (including query log analysis, and link and anchor text analysis). Despite this, the results for our method are lower than those over the synthetic dataset, we suspect largely as a result of the style of queries we issue being so far removed from standard Google query patterns. Having said this, MAP scores of 0.32–0.92 suggest that the method is highly usable (i.e. at any given cutoff in the document ranking, an average of at least one in three documents is a genuine multilingual dictionary), and any non-dictionary documents returned by the method could easily be pruned by a lexicographer.

Lang	Dicts	MAP	Baseline
zh	16	0.55	0.19
es	17	0.92	0.13
ja	13	0.32	0.04
de	34	0.77	0.09
fr	36	0.77	0.08
it	23	0.69	0.11
ar	8	0.39	0.17
AVERAGE:	21.0	0.63	0.12

Table 4: Dictionary retrieval results over the open web for dictionaries containing English and each of the indicated languages ("Dicts" = the number of unique multilingual dictionaries retrieved for that language).

Among the 7 language pairs, en-es, en-de, en-fr and en-it achieved the highest MAP scores. In terms of unique lexical resources found with 50 queries, the most successful language pairs were en-fr, en-de and en-it.

6 Conclusions

We have described initial results for a method designed to automatically detect multilingual dictionaries on the web, and attained highly credible results over both a synthetic dataset and an experiment over the open web using a web search engine.

In future work, we hope to explore the ability of the method to detect domain-specific dictionaries (e.g. training over domain-specific dictionaries from other language pairs), and low-density languages where there are few dictionaries and Wikipedia articles to train the method on.

Acknowledgements

We wish to thank the anonymous reviewers for their valuable comments, and the Panlex developers for assistance with the dictionaries and experimental design. This research was supported by funding from the Group of Eight and the Australian Research Council.

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207–216.

Timothy Baldwin, Jonathan Pool, and Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Demo Volume*, pages 37–40, Beijing, China.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language IR. In *Proceedings of Recherche d'Informations Assistee par Ordinateur 2000 (RIAO'2000)*, pages 62–77, College de France, France.

ClueWeb09. 2009. The ClueWeb09 dataset. `http://lemurproject.org/clueweb09/`.

Aidan Finn, Nicholas Kushmerick, and Barry Smyth. 2002. Genre classification and domain transfer for information filtering. In *Proceedings of the 24th European Conference on Information Retrieval (ECIR 2002)*, pages 353–362, Glasgow, UK.

Freedict. 2011. Freedict dictionaries. `http://www.freedict.com`.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of Association for Machine Translation in the Americas (AMTA 1998): Machine Translation and the Information Soup*, pages 1–17, Langhorne, USA.

Indri. 2009. Indri search engine. `http://www.lemurproject.org/indri/`.

David Kamholz and Jonathan Pool. to appear. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Batia Laufer and Linor Hadar. 1997. Assessing the effectiveness of monolingual, bilingual, and "bilingualised" dictionaries in the comprehension and production of new words. *The Modern Language Journal*, 81(2):189–196.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2(Feb):27–40.

Katsushi Matsuda and Toshikazu Fukushima. 1999. Task-oriented world wide web retrieval by document type classification. In *Proceedings of the 1999 ACM Conference on Information and Knowledge Management (CIKM 1999)*, pages 109–113, Kansas City, USA.

MeCab. 2011. `http://mecab.googlecode.com`.

I. Dan Melamed. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA 1996)*, Montreal, Canada.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the web as a bilingual dictionary. In *Proceedings of the ACL 2001 Workshop on Data-driven Methods in Machine Translation*, pages 1–8, Toulouse, France.

Jian-Yun Nie. 2010. *Cross-language information retrieval*. Morgan and Claypool Publishers, San Rafael, USA.

John M. Prager. 1999. Linguini: language identification for multilingual documents. In *Proceedings the 32nd Annual Hawaii International Conference on Systems Sciences (HICSS-32)*, Maui, USA.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Stephen Soderland, Christopher Lim, Mausam, Bo Qin, Oren Etzioni, and Jonathan Pool. 2009. Lemmatic machine translation. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, Ottawa, Canada.

Nicholas Thieberger and Andrea L. Berez. 2012. Linguistic data management. In Nicholas Thieberger, editor, *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press, Oxford, UK.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 171.

Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. 2012. Text segmentation by language using minimum description length. In *Proceedings the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 969–978, Jeju Island, Korea.

Kun Yu and Junichi Tsujii. 2009. Bilingual dictionary extraction from Wikipedia. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, pages 379–386, Ottawa, Canada.

Sven Meyer zu Eissen and Benno Stein. 2005. Genre classification of web pages. In *Proceedings of the 27th Annual German Conference in AI (KI 2005)*, pages 256–269, Ulm, Germany.