Acquiring bilingual lexica from keyword listings

Filip Graliński*, Krzysztof Jassem*, Roman Kurc†

*Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Umultowska 87, 61-614 Poznań, Poland
{filipg, jassem}@amu.edu.pl

†Wrocław University of Technology
Faculty of Computer Science and Management
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
roman.kurc@pwr.wroc.pl

Abstract

In this paper, a new method for acquiring bilingual dictionaries from on-line text corpora is presented. The method merges rule-based techniques for obtaining dictionaries from structuralised data, such as paper dictionaries (in electronic form) or on-line glossaries, with methods used by aligning tools, such as GIZA. The basic idea is to search for anchor words such as *abstract* or *keywords* followed by their equivalents in another language. Text fragments that follow anchor words are likely to supply new entries for bilingual lexica.

1. Introduction

The origin of the idea to use large text corpora for finding translations of words was formed in Weaver's famous memorandum (Weaver, 1949), where the author predicted that meaning (and proper translation of a word) may be determined on the analysis of its context in a text – the longer the context, the lesser unambiguity. However, his ideas had not been fully put into practice until the breakthrough work of Brown et al (Brown et al., 1990). Since then the idea of Statistical Machine Translation (SMT) has gradually overcome former ideas on automatic translation. The main assumption of SMT is that given a bilingual corpus of source texts and their translations (target texts) we may presume the most likely translation of a word (phrase) to be the word (phrase) that occurs most frequently in the similar position in the target text. This idea is often used in the creation of bilingual dictionaries, in particular domain-specific ones, when the spectrum of texts under consideration is limited to a certain topic. In order to obtain a corpus-based bilingual dictionary, three phases are executed:

i) collection of bilingual corpora,

ii) paragraph and sentence alignment,

iii) word alignment — to obtain mappings between words, i.e. a raw bilingual dictionary.

Each step poses specific issues.

i) Resnik (Resnik, 1998) as well as Xiaoyi and Liberman (Xiaoyi and Liberman, 1999) suggested algorithms for finding parallel texts on the Internet. The papers report very good results for pairs of languages such as English and German (the latter reports 97% recall and 99% precision). However, the same experiments carried out for a language pair less frequently represented on the Internet (Rosińska, 2007).

ii) There exist tools for sentence alignment such as described in (Moore, 2002), for which the precision over 95% was reported. The up-to-date open source implementation of the Moore aligner has been described in (Lipski, 2007). However, all algorithms give significantly worse results when applied to texts that are not well structured (which is often the case with texts acquired from the Internet).

iii) The most widely used tool for word alignment is GIZA and its successors (Och and Ney, 2003). The main problem is that GIZA, as purely statistical tool, requires a corpus sufficiently large to yield a reliable alignment. SMT is rather robust to word-alignment noise (Lopez and Resnik, 2006), whereas lexicon acquisition is more demanding.

The combination of three imprecise tools for each of the three steps makes the SMT approach not fully satisfactory if used for the creation of bilingual dictionaries. This is why other corpus-based methods are searched for.

The usage of monolingual domain-specific resources and the Internet was described in (Nazar et al., 2008). For a given word they look for multiword terms including the word that occur most frequently in the collected monolingual corpus. Next, they search Internet target texts for the same multiword terms – it is assumed that target language texts may have the source language insertions. The translation of the given word is the most frequent meaningful word occurring in the target texts that include the multiword terms. The idea would not give high recall for general-domain words but it works well for domain-specific texts and specialised terms which often are cited in texts together with their translations.

A method for acquiring bilingual dictionaries solely from monolingual web pages was presented in (Cao et al., 2007). The authors of that paper observed that "many Chinese terms (e.g., named entities that are not stored in a conventional dictionary) are accompanied by their English

translations in the Chinese web pages".

The method presented in this paper is similar to that of Cao: the algorithm works on a single document rather than on a pair of documents. The document may be only partly bilingual, for example a scientific paper might be written in Polish with only the abstract and the list of keywords given in Polish and English. In order to find the bilingual fragments we introduce a new factor for searching: the anchor. For the sake of this paper we assume the following definition:

Definition 1 ***Anchor*** *is a pair of words or multiword terms that indicates two positions in a text: beginning of the text which should be rich in source terms; beginning of the target text fragment which should contain their translations.*

For instance, the following anchors could be used for acquiring a Polish-English dictionary: {*słowa kluczowe*, *keywords*} {*streszczenie*, *summary*}, {*streszczenie*, *abstract*}, {*abstrakt*, *abstract*}, {*składniki*, *ingredients*}[1]. In what follows, we present the idea implemented for the first anchor ({*słowa kluczowe*, *keywords*}). The goal is to extract all terms and their translations that are defined as **keywords** in a text (usually a **scientific paper**).

Our method is intermediate between algorithms for extraction of terms from structuralised materials (like on-line lexica, glossaries) with algorithms for extraction of terms from raw, free texts. Note that keywords are usually listed in their basic form[2], which makes the extraction more reliable compared to that executed on raw texts. Furthermore, a phrase listed as a keyword is more likely to make sense as a lexicon entry (than a phrase extracted from a GIZA alignment). Another advantage of using terms defined as keywords as dictionary entries is the relatively high reliability of translation. Keywords are usually used in scientific or reliable business articles.

The techniques described in this paper show how the vast amount of text resources on the Internet makes it possible to acquire a useful amount of data (here: bilingual dictionaries) by using "tricky" methods.

The paper is organised as follows: We start with some brief remarks on how the raw corpus was collected (Section 2.). In Section 3., algorithms for extraction of bilingual dictionaries on the basis of anchors are presented. The evaluation by means of comparison to GIZA is given in Section 4. We end with conclusions in Section 5.

2. Collecting the corpus

The great advantage of using anchors is that they can be used not only for the extraction of desired text fragments in a given document, but also for seeking out the document itself on the Internet. In other words, documents with text fragments in question might represent a tiny percentage of the whole Web, however, if they are amassed from the Internet, a textual material of considerable value can be obtained.

In order to collect the corpus of papers with bilingual keyword listings, we started with manually querying the Google search engine with queries like `"słowa kluczowe" keywords abstract summary`[3]. A tentative list of websites was obtained in this manner. Each of these websites was crawled by our in-house web robot. We obtained 20945 documents, mostly PDF files, containing the anchors. Not all of these documents contained a bilingual keyword listing (more on this in Section 4.2.), however, some of them contained more than one keyword listing.

3. Procedure for the Keywords Extraction

The procedure for the keywords extraction consists of two phases: **scan** of the document in search for keyword lists and **alignment** of keywords in the lists found.

The scan is executed by means of regular expressions with anchors. The alignment applies various heuristic techniques.

The pseudo-code representation of the whole procedure is as follows:

```
foreach document in documents do
a.    choose anchor_schema;
b.    while (kwdsSRClist, kwdsTRGlist) =
      anchor_schema.scan_and_split(document) do
c.        if correct(kwdsSRClist, kwdsTRGlist) then
d.            aligned_pairs = align(kwdsSRClist,
              kwdsTRGlist);
          end
      end
end
```

A general run of the procedure will be shown by Example 1.

Example 1 *Piotrowo 3, 60-965 Poznań, tel. (61)6652065*

Słowa kluczowe: wózek inwalidzki aktywny, wybór wózka inwalidzkiego, konfiguracja wózka inwalidzkiego Streszczenie Wózek inwalidzki z napędem ręcznym w dużym ... możliwości socjalizacyjne osoby niepełnosprawnej ruchowo. 1 Chosen problems of manual wheelchair selection and configuration

Keywords: wheelchair for active use, wheelchair selection, wheelchair configuration Abstract A manual wheelchair highly determines possibility of adaptation of a person with permanent dysfunction of locomotion for daily living. Selecting type of the wheelchair, geometrical sizes, adjustments regulations and choosing additional equipments determine regular choosing and selection of the wheelchair to individual user The paper characterizes wheelchairs for activ[4]

[1] *Słowa kluczowe*, *streszczenie*, *abstrakt*, *składniki* are simply Polish equivalents of respectively: *keywords*, *summary*, *abstract*, *ingredients*.

[2] With the exception of, for example, names of chemicals, which are sometimes given in plural form in keyword listings. However, even in this case, the keywords are in the nominative, which is an important advantage considering the rich declension system of the Polish language.

[3] We used less frequently used synonyms of *słowa kluczowe* (e.g. *hasła kluczowe*, *hasła indeksowe*) as well.

[4] A fragment of a paper by Maciej Sydor and Marek Zabłocki taken from http://www.au.poznan.pl/sdwtd/sydor/sydor_zablocki_wybrane_probl_2006.pdf.

Re a.: The chosen anchor schema for the passage is: *Słowa kluczowe:* SRC-KEYWORD DELIMITER SRC-KEYWORD DELIMITER ... OTHER TEXT *Keywords:* TRG-KEYWORD DELIMITER TRG-KEYWORD DELIMITER ...

Re b.: Both keyword blocks are cleaned and split into keyword lists. In this phase the source block *Słowa kluczowe: wózek inwalidzki aktywny, wybór wózka inwalidzkiego, konfiguracja wózka inwalidzkiego Streszczenie Wózek* is converted into:

kwdsSRClist = [*wózek inwalidzki aktywny*, *wybór wózka inwalidzkiego*, *konfiguracja wózka inwalidzkiego*]

Let us notice that the last two words of the block are removed in this step. Accordingly *Keywords: wheelchair for active use, wheelchair selection, wheelchair configuration Abstract A manual* is converted into:

kwdsTRGlist = [*wheelchair for active use*, *wheelchair selection*, *wheelchair configuration*]

Re c.: At this point the structures of keyword lists are recognised. The goal is to delete ill-balanced list pairs, i.e. pairs of lists in which the number of elements differ significantly. This step will qualify the exemplary pair of lists as both of the lists have an equal number of elements (3).

Re d.: The last step is the alignment of keywords. As the result of this step, lists of aligned pairs are obtained:

aligned_pairs = [
(*wózek inwalidzki aktywny*, *wheelchair for active use*),
(*wybór wózka inwalidzkiego*, *wheelchair selection*),
(*konfiguracja wózka inwalidzkiego*, *wheelchair configuration*)]

Sections 3.1. and 3.2. focus on technical details of scan and alignment respectively.

3.1. Scan and split

Two issues have to be considered here:

1. how to locate keywords;
2. how to extract anything but keywords, skipping unnecessary text before and after keyword blocks.

We solve problem (1) by using regular expressions. Let us discuss two general patterns:

Pattern 1 – Two lists are placed one after the other as in Example 1:

SRC-ANCHOR SRC-KEYWORD DELIMITER SRC-KEYWORD DELIMITER ... some text ... TRG-ANCHOR TRG-KEYWORD DELIMITER TRG-KEYWORD DELIMITER

Pattern 2 – Each keyword is followed by its translation (e.g. written in brackets):

SRC-ANCHOR TRG-ANCHOR SRC-KEYWORD TRG-KEYWORD DELIMITER SRC-KEYWORD TRG-KEYWORD DELIMITER ...

Pattern 2, extracts keywords form text given as Example 2.

Example 2 *Słowa kluczowe (key words): kamica nerkowa (renal stone disease), szczawian wapnia (calcium oxalate),* ...

The second issue (cleaning) is handled by two heuristic techniques: First, we set a maximum length of the keyword to 5, i.e. if a fragment between keyword separators (usually commas) is composed of more than 5 words then it is treated as a beginning of the non-keyword block. Second, we use regular expressions to delete unwanted strings from keywords, e.g. strings in braces or brackets or certain strings at the end of the keyword lists such as those starting with *abstract* or *summary*. In Example 3 a list of keywords is presented and the strings that are removed during cleaning are highlighted.

Example 3 *Keywords: wheelchair for active use, wheelchair selection (since 1990), wheelchair configuration* ***Abstract A manual***

Once the keywords are found and cleaned, they are split into lists.

3.2. Alignment

In order to align elements of keyword listings we consider the following cases, reflecting the possible ways in which authors of scientific papers list and translate keywords:

1. At least one list is sorted (we assume that the author of a given paper wrote the keywords in alphabetical order on purpose[5] and therefore there may be no correspondence between the order in which source language keywords were given and the order in which their translations were listed).
2. Both lists are unsorted **and** the lists are of different length.
3. Both lists are unsorted **and** the lists are of the same length (presumably because the author of a given papers listed the translations of keywords in the same order as the original keywords).

Each case is treated by in a different way, which is illustrated by the following pseudo-code:

```
1. align(kwdsSRClist,kwdsTRGlist):aligned_pairs;
   begin
      if kwdsSRClist or kwdsTRGlist is sorted
      ascending then
         align_similar(kwdsSRClist, kwdsTRGlist,
         true)
      else
         if length(kwdsSRClist) ≠
         length(kwdsTRGlist) then
2.          align_similar(kwdsSRClist,
            kwdsTRGlist, false)
         else
3.          simple_align(kwdsSRClist,
            kwdsTRGlist);
         end
      end
   end
```

[5] Obviously, in some cases keywords occur in alphabetical order just by accident.

In the first and the second case (`align_similar` with different values of the third argument) the first step of the alignment is the search for translation pair candidates. All elements of the first list are compared with all elements of the second list. A pair {SRC-KEYWORD, TRG-KEYWORD} is a translation candidate if one of the two conditions is met:

- the keywords comprise of words that are equivalent according to an external bilingual lexicon (a very large, albeit noisy, lexicon of over 2.5 million translation pairs was used; words are lemmatised before the equivalence check);
- the keywords are similar as far as their surface form is concerned, which is determined among others by the Levenshtein distance.

Example 4 shows an exemplary linkage between Polish and English keywords.

Example 4 *[Błoto pochromowe, chrom [1], chromian sodu [2], ekstrakcja [3], filtracja [4], kalcynacja [5], ługowanie, metody otrzymywania chromianu sodu, obróbka surowców chromonośnych, recyrkulacja [6], ruda chromowa [7]];*

[Calcination [5], chromic mud, chromic ore [7], chromium [1], extraction [3], filtration [4], methods of sodium chromate production, processing of chromium materials, recirculation [6], separation, sodium chromate [2]];

Links from [1] to [7] were detected by lexicon look-up. For example, translation pair *chrom* = *chrome* (link [1]) was simply found in the lexicon. Obtaining [7] required some more processing. Namely, the Polish word *ruda* is homonymous (it is either a feminine form of adjective *rudy* or the basic form of noun *ruda*) and the following translations can be found for the two interpretations of the word:

1. (adj. *rudy*) *redhead*; *rusty*; *red*; *reddish-brown*; *foxy*; *ginger*;*red-haired*; *rufous*, *ginger-haired*, *russet*; *ruddy*; *brownish-red*; *bay*
2. (noun *ruda*) *ore*; *mineral ore*

The Polish word *chromowy* (the lemma for the feminine form *chromowa*) have a few equivalents as well: *chromic*, *chromium*, *chrome*, *chromie*. Similar type of information may be found in the English-to-Polish direction, where the word *ore* have a few Polish equivalents, such as: *rudowy*, *rudny*, *kruszec*, *kopalina*, *ruda metalu*, *ruda*. In order to find the link between *ruda chromowa* and *chromic ore* all combinations of equivalents have to be processed.

Once the translation candidate pairs are found, the alignment algorithm tries to determine other equivalents. If one of the keyword lists is sorted alphabetically, it is assumed that the order of elements gives no further clue for the alignment: thus only translation candidate pairs (found in the external lexicon) are aligned, leaving remaining elements unaligned (this is the case in Example 4). However, if exactly **one** pair of keywords remains unaligned, the algorithm assumes the keywords to be equivalent and aligns them as well.

When keyword lists have different lengths and both are unsorted, the procedure `align_similar` aligns sublists of keywords. The algorithm looks for translation candidate pairs. If any two candidate pairs are separated by the same number of keywords, the algorithms aligns all the keywords between them, as shown in Example 5.

Example 5 *[przetwarzanie rozproszone [1], układy programowalne, FPGA [2], kompresja obrazu [3], sieć w układzie [4], sieć NoC [5], AVC [6], VC-1 [7]];*

[scattered processing [1], FPGA [2], video compression [3], Network on Chip [4], NoC network [5], AVC [6], VC-1 [7]]

Let us assume that [1], [2], [3], [6] and [7] are candidate pairs (i.e. obtained by lexicon look-up). There is a redundant keyword between the first and the third pair. It is removed. [3] and [6] are translation candidates and keywords between them ([4] and [5]) form continuous sequences of the same length. Therefore we assume that they are parallel translations and we join them together.

If the lists are of the same length and both of them are unsorted the algorithm assumes that they are parallel translations of each other. All elements of the list are paired according to their positions on the lists. This case produces the highest amount of new keywords. The precision of this lexically unsupervised case is surprisingly high (see Section 4.).

4. Evaluation

4.1. Alignment with GIZA as a baseline

GIZA (Och and Ney, 2003) is a well-known tool for word alignment. We decided to use GIZA (IBM Model 4) on the same material on which our algorithm had worked, for comparison. Lists of keywords were organised as a parallel corpus. Since GIZA uses statistical mechanisms, we decided to add the entries of the external lexicon that were used in the procedure `align_similar` (see Section 3.2.) and/or abstracts that follow keywords to the parallel corpus, thus supplying GIZA with more data. (A parallel corpus of Polish and English abstracts/summaries is an interesting by-product of our experiment. We managed to extract 1656 abstract/summary pairs.)

An in-house program for extracting Polish/English translation pairs from GIZA output files was used.

4.2. Recall from documents

The recall from documents was designed to prove a high quality of the first part of the algorithm: the scan. The results of the scan are the base for an alignment and thus for the comparison between our method of alignment and GIZA. In order to estimate the recall from documents we used documents that were rejected during the scan. Out of 20945 documents, 8513 were rejected. We identified which of them were rejected by mistake and the reason for

	Anchors	GIZA [abstracts]	GIZA [lexicon]	GIZA [abstracts + lexicon]
# of translation pairs	17825	15065	25730	23866
# of confirmed pairs	4621	1423	2374	1935
% of confirmed pairs	0.259	0.094	0.092	0.081
precision (estimated)	**0.97**	-	**0.52**	-
# of correct translation pairs (estimated)	**17290**	-	**13380**	-

Table 1: Results of various methods of extraction of keyword pairs.

the mistake. A sample of 385 rejected documents was inspected. In this way, we estimated the document-level recall (the ratio of the number of successfully scanned documents to the number of all documents containing parallel Polish/English keyword listings) to be 0.87%. Such an outcome is not surprising, since the texts were searched on the Internet using the elements of the anchors. Most of the rejections resulted from a malformed encoding, garbled text (e.g. the result of unsuccessful PDF-to-text conversion), papers containing only keywords in one language, etc. Some mistakes came from the fact that we had missed some anchor patterns.

4.3. Precision

Our anchor algorithm found 17825 different translation pairs (see Table 1), 14142 of which were obtained with the `simple_align` procedure (i.e. without a lexicon lookup, see Section 3.2.). In order to estimate the precision, we checked the number of translation pairs confirmed in the lexicon (the same lexicon as used in the `align_similar` procedure) – 4621 pairs were found in the lexicon. Then the sample of 390 unconfirmed translation pairs was manually checked, only 18 translation pairs were marked as incorrect. Consequently, the estimation of the overall precision of our method is 0.97%.

25730 translation pairs were obtained with GIZA run on keyword listings and appropriate lexicon entries (interestingly, adding summaries/abstracts make the number of obtained translation pairs smaller). It is more than the number of pairs obtained with anchors, but the estimated precision (the same estimation procedure was used as for our method) is much lower (0.52%). It is quite obvious that the corpus (even with abstracts) processed by GIZA was too small to give reasonable results for statistical methods.

5. Conclusions

This paper shows a new technique of extracting translation pairs from Web texts. The method relies on the existence of special words (called anchors) that indicate fragment of texts likely to contain words together with their translations. The described method has been used for extracting keyword pairs from scientific papers. It turns out that such tailored, "tricky" techniques significantly overcome standard statistical methods for specific data.

Obviously, extracting keyword translations from scientific papers is not enough to create a complete lexicon, it may, however, make a valuable contribution to (or confirmation of) lexical material obtained with other techniques.

6. Acknowledgment

The paper is based on research funded by the Polish Ministry of Science and Higher Education (Grant No 003/R/T00/2008/05).

7. References

Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Cao, G., J. Gao, and J.-Y. Nie, 2007. A system to mine large-scale bilingual dictionaries from monolingual web pages. *Machine Translation*, Summit XI:57–64.

Lipski, Jarosław, 2007. *Urównoleglanie tekstów dwujęzycznych na poziomie zdania*. Master's thesis, Adam Mickiewicz University in Poznań.

Lopez, Adam and Philip Resnik, 2006. Word-based alignment, phrase-based translation: What's the link? *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*:90–99.

Moore, Robert C., 2002. Fast and accurate sentence alignment of bilingual corpora. In *In Stephen D*. Springer-Verlag.

Nazar, R., L. Wanner, and J. Vivald, 2008. Two step flow in bilingual lexicon extraction from unrelated corpora. In *In Proceedings of the EAMT(European Association for Machine Translation) 2008 Conference (Hamburg, Germany, 22-23 September 2008)*.

Och, Franz Josef and Hermann Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Resnik, Philip, 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *In Third Conference of the Association for Machine Translation in the Americas*. Springer.

Rosińska, M., 2007. Collecting Polish-German parallel corpora in the Internet. *Proceedings of the International Multiconference on Computer Science and Information Technology*, Volume 2, XXIII Autumn Meeting of Polish Information Processing Society.

Weaver, Warren, 1949. Translation. In *Mimeographed*. MIT Press.

Xiaoyi, M. and M. Liberman, 1999. BITS. a method for bilingual text search over the Web. *Machine Translation*, Summit VII.