Mining Parenthetical Translations for Polish-English Lexica

Filip Graliński

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

Abstract. Documents written in languages other than English sometimes include parenthetical English translations, usually for technical and scientific terminology. Techniques had been developed for extracting such translations (as well as transliterations) from large Chinese text corpora. This paper presents methods for mining parenthetical translation in Polish texts. The main difference between translation mining in Chinese and Polish is that the latter is based on the Latin alphabet and it is more difficult to identify English translations in Polish texts. On the other hand, some parenthetically translated terms are preceded with the abbreviation "ang." (=English), a kind of an "anchor", allowing for querying a Web search engine for such translations.

1 Introduction

Bilingual lexica are of paramount importance because of their applications in such natural language processing domains as (both statistical and rule-based) machine translation, computer-assisted translation or cross-language information retrieval. With the rapid growth of the Internet, a natural question arises: how to extract bilingual lexicon entries from the huge volume of Web data, not only Web pages, but also PDF documents or files in Microsoft Word format.

One line of research is to collect bilingual sentence-level Web corpora, e.g. by exploiting pairs of Web pages that are mutual translations [1], and to automatically acquire lexical data from them [2]. Sometimes comparable rather than strictly parallel corpora (e.g. Wikipedia) are used [3].

Interestingly, some bilingual lexical data can be extracted from (purely or mostly) monolingual corpora. One method is to combine frequency information and cognate analysis [4]. Another technique exploits short bilingual snippets repeated in a similar manner in a mostly monolingual Web page [5]. Finally, some bilingual lexicon entries can be extracted from semi-structured Web data sources such as bilingual keyword listings [6].

[0] A. Gelbukh (Ed.): CICLing 2010, LNCS 6008, pp. 464–472, 2010.
Springer-Verlag Berlin Heidelberg 2010

In this paper, experiments on mining bilingual data from *parenthetical translations* put in (mostly) monolingual Polish Web texts are reported. The idea comes from the observation that Polish authors sometimes annotate words, terms, book or film titles with their translations in English. The following example illustrates this phenomenon:

> Stosować się będzie bowiem ona do działalności nie tylko operatora i dysponenta sieci telekomunikacyjnej (ang. network providers) oraz dostawcy dostępu do Internetu (ang. access providers), ale również dostawców usług w sieciach (ang. Internet Service Providers).[1]
>
> [For it will be applicable to the operations of not only an operator and owner of the telecommunications network (Eng. network providers) as well as of a provider of the access to the Internet (Eng. access providers), but also of providers of services in networks (Eng. Internet Service Providers).]

(A literal translation of the sentence is given in square brackets, the terms for which parenthetical translations were specified in the original texts are underlined and translated word-by-word here.) The three parenthetical translations were given in round brackets and were preceded by the word *ang.*, which is an abbreviated form of the adjective *angielski* (= *English*).

Even though parenthetical translations are typical for academic papers, PhD and master's theses and other types of formal texts, they can be occasionally encountered in virtually any kind of Web texts. Their frequency is rather low but the sheer size of the Web makes their number large (even for medium-sized languages such as Polish). They are valuable needles which come in thousands in the huge haystack of the Web. What makes them interesting is the very reason they are used: they are new and/or technical terms usually with no standard Polish translation, absent from conventional dictionaries.

The idea to mine parenthetical translation is not new: techniques for supervised [7], semi-supervised [8] and unsupervised [9] lexicon mining were proposed for English parenthetical translations in Chinese texts. No experiments, however, have been reported for languages other than Chinese, and in particular for languages the writing system of which is based on the Latin alphabet. It should be noted that the case of Polish is different, to some extent, from the Chinese language as far as English parenthetical expressions are concerned. First, as the same Latin alphabet is used in Polish and English,[2] it would be more difficult (for computers as well as for humans) to identify English insertions if only parentheses were to be used, therefore some additional clues (like the abbreviation *ang.*) are usually applied. Second, English parenthetical expressions rarely refer to Polish transliterations.[3] Third, the volume of English parenthetical translit-

[1] `http://www.piit.org.pl/piit2/index.jsp?place=Lead07&news_cat_id=51&news_id=1422&layout=2&page=text`

[2] Except that 9 characters with diacritics (*ą*, *ć*, *ę*, *ł*, *ń*, *ó*, *ś*, *ź*, *ż*) are used in Polish.

[3] With some minor exceptions, like Russian names, which are traditionally transliterated in Polish in a different way than in English, e.g. *Maja Plisiecka (ang. Maya Plisetskaya)*.

erations seems to be much smaller in Polish than in Chinese, which makes some of the quantitative methods unfeasible.

The paper is organised as follows: Sect. 2 is a discussion of the conventions for parenthetical transliterations used by Polish authors. Section 3 presents methods for gathering Web corpora and Sect. 4 – the methods for extracting parenthetical translations. The results come in Sect. 5 and remarks concerning future work and conclusions are provided in Sect. 6.

2 Parenthetical Translation Conventions

There are two main conventions for specifying parenthetical English translations in Polish texts:

A. $p_1\ p_2 \ldots p_m$ `(ang.` $e_1\ e_2 \ldots e_n$ `)` – the English translation is given in parentheses and is preceded by the abbreviation *ang.* (= *English*), see the example given in the Introduction;
B. $p_1\ p_2 \ldots p_m$ ($e_1\ e_2 \ldots e_n$) – the English translation is given in parentheses, in italics.

(Here, p_1, p_2,..., p_m denote Polish words, whereas e_1, e_2,..., e_n – English words.) Some variations can, however, be observed. Some of them are:

- $p_1\ p_2 \ldots p_m$ `(ang.` $e_1\ e_2 \ldots e_n$ `)` – (A) and (B) combined,
- $p_1\ p_2 \ldots p_m$ `(z ang.` $e_1\ e_2 \ldots e_n$ `)` – *z ang.* = *from English*,
- „$p_1\ p_2 \ldots p_m$” `(ang. “`$e_1\ e_2 \ldots e_n$`” )` – additional quotes are used,
- $p_1\ p_2 \ldots p_m$ `(ang.:` $e_1\ e_2 \ldots e_n$ `)` – colon used after the abbreviation *ang.*,
- $p_1\ p_2 \ldots p_m$ `[ang.` $e_1\ e_2 \ldots e_n$ `]` – non-round brackets are used.

Sometimes Polish or English synonyms or glosses are given within the parenthesis, e.g.: *Rozwój zrównoważony (ekorozwój, ang. sustainable development)*[4] or *uwalnianie leku z jego postaci farmaceutycznej*[5] *(ang. liberation, drug release)*. Acronyms are described in parenthetical expressions even more often, e.g.: *rdzeniowej atrofii mięśni (ang. spinal muscular atrophy, SMA)*. Quite frequently, instead of a Polish term, only the acronym is given and the parenthetical expressions is just the English term for which it stands, see the following example:

> W metodach wektorowych wykorzystuje się między innymi algorytmy FDTD (ang. Finite Difference Time Domain) i FMM (ang. Fourier Modal Method).[6]
>
> [In vector methods FDTD (Eng. Finite Difference Time Domain) and FMM (ang. Fourier Modal Method) algorithms are used among others.]

[4] *Rozwój zrównoważony* = lit. *stable development*, *ekorozwój* = lit. *eco-development*
[5] = lit. *release of the drug from its pharmaceutical form*
[6] `http://pl.wikipedia.org/wiki/Transformacja_genetyczna`

Such abbreviations along with their full forms might be of interest (e.g. for acronym lexicons), I decided, however, to filter them out (see Sect. 4.2) as, strictly speaking, they are not translations.

This paper focuses on convention (A) and its variations, i.e. only parenthetical expressions with the abbreviation *ang.* are considered. The reason is that visual formatting markup is usually discarded while generating text corpora from Web pages,[7] which makes recognising the convention (B) more difficult. Also, as we will see in the next section, the abbreviation *ang.* makes it possible to seek out texts with parenthetical translations on the Internet.

3 Corpora

3.1 Pre-existing Corpora

I started with the available Polish corpora (i.e. not collected with parenthetical translations in mind), namely: a general Web corpus of over 2.8 million web pages and PDF files collected from the Polish Internet, a dump of the Polish Wikipedia and a collection of Polish academic papers and abstracts (see Table 1). The frequency of the abbreviation *ang.* token turned out to be much higher in Wikipedia and scientific texts than in general Web texts. The total number of occurrences of *ang.* was 41,714. As this is the upper bound of the number of parenthetical translations with *ang.* (*ang.* can be used for other purposes than parenthetical translations, see the next subsection), the results were somewhat unsatisfactory. This is why the decision was made to actively seek parenthetical translations on the Internet.

Table 1. Corpora initially used (#*ang.* – number of *ang.* tokens in the text).

Corpus	**Bytes**	**Tokens**	**#*ang.* (/ 1M tokens)**
Web corpus	15.6GB	2.0G	24487 (12.2)
Wikipedia dump	498MB	61.5M	11300 (183.7)
Corpus of academic papers	425MB	56.2M	5927 (105.4)
Total	16.5GB	2.1G	41714 (19.7)

3.2 Dedicated Corpus

The interesting thing about the abbreviation *ang.* is that it can be used not only for the extraction of desired parenthetical expressions in a given document, but also for seeking out the document itself on the Internet, i.e. a query with

[7] The problem is even more complicated for PDF files.

ang. can be constructed to locate document with parenthetical translations using Web search engines.

One obstacle is that the periods (full stops) are usually discarded by search engines and *ang.* would be probably normalised to `ang`, the same goes for such words as *ang*, *Ang* and *ANG*. Fortunately, the tokens normalised to `ang` are not frequently used for purposes other than parenthetical translations, in particular *ang* is not a valid Polish word. Some of the cases which nevertheless should be taken into account are:

- *ang.* in *j. ang.* or *jęz. ang.*, a short form for *język angielski* (*the English language*),
- *ang.-pol.* (or *pol.-ang.*), a short form for *angielsko-polski* (*English-(to-)Polish*), which is usually tokenized and normalised by search engines into two strings: `ang` and `pol`,
- *Ang* as the first name of the film director Ang Lee.

In order to avoid on-line dictionaries and Web sites for Polish students of the English language (where *ang.* is often used for purposes other than parenthetical translations), three additional words were added as "negative" terms in the constructed query: *słownik* (*dictionary*), *język* (*language/tongue*), *angielski* (*English*). Hence, the final query was as follows:

```
ang -"j ang" -jęz -pol -lee -słownik -język -angielski
```

This query (and its variations) was entered into the Google and Bing search engines (with the language option set to Polish). The websites with the largest number of hits were additionally crawled by an in-house web robot. A list of 91,872 URLs was obtained in this manner. 69,493 files were successfully downloaded and converted[8] into plain text. The characteristics of the corpus are given in Table 2.

It should be noted that no dedicated corpora were gathered in the experiments concerning Chinese-English parenthetical translations ([7], [8], [9]).

Table 2. Dedicated corpus.

Bytes	**Tokens**	**#*ang.* (/ 1M tokens)**
1.36GB	177M	141227 (798.9)

[8] Some PDFs could not be converted into text by the tool available (`pdftotext`).

4 Translation Extraction

4.1 Preprocessing

Snippets containing *ang.* were first extracted using hand-crafted regular expressions. The limit for the number of tokens to the left and to the right of *ang.* was set to 7. Some words (mostly Polish conjunctions) were then "blacklisted" and discarded from the beginning of a snippet. Anomalous snippets, e.g. with no letters in pre-parenthesis nor in in-parenthesis fragments, were discarded as well.

If the fragment of a snippet suspected of being an English parenthetical translation contained characters with Polish diacritics, the snippet as a whole was discarded.

4.2 Filtering out Acronyms

As I mentioned in Sect. 2, sometimes there is no Polish equivalent to the English term in parenthesis – with only the English acronym provided. Such abbreviations were filtered out by comparing the letters of an acronym with the initial letters of the term words and taking into account some standard acronym conventions (such as using *2* instead of *to*). A random sample of ten filtered out snippets is listed in Table 3.

Table 3. A sample of snippets filtered out.

Snippet
EURIBOR (ang. Euro Interbank Offered Rate)
ON (ang. over night)
odróżniającą go od większości MTA (ang. mail transport agent)
stopa depozytów jednodniowych rozpoczynających się dziś SW (ang. spot week)
jest protokołem wykorzystywanym do przeglądania WWW (ang. World Wide Web)
Systemy MES (ang. Manufacturing Execution System)
PIN (ang. Personal Identification Number)
SCORM i AICC. LMS (ang. Learning Management System)
LIBOR (ang. London Interbank Offered Rate)
Forex (ang. Foreign Exchange)

The number of candidate translation pairs after preprocessing and filtering out was 82,434 (all the corpora mentioned in Sect. 3 were used).

4.3 Word Alignment

In order to extract a parenthetical translation the *first* word of the Polish equivalent of the parenthetical translation ought to be determined. Following [9] I used

a word alignment algorithm for determining the left boundary: the first pre-parenthesis word aligned with an in-parenthesis word is assumed to be the left boundary of the Polish equivalent of the English in-parenthesis translation. However, as the collection of snippets was much smaller than that obtained in [9] an external Polish-English lexicon had to be consulted. The lexicon contained 474,265 translation pairs (both single words and multi-word units), it was based on heterogeneous acquisition techniques and data sources. The translation pairs obtained from parenthetical expressions are planned to be yet another source of lexical data for this still growing lexicon.

Competitive Linking [10], a simple yet effective [11] algorithm, was used for word alignment. The algorithm is a kind of greedy, best-first search: a pair of words can be linked on condition that none of the two words were previously aligned to any other words. Potential word associations are sorted by some score.

As it was mentioned already, an external lexicon was used as the source of scores for pairs of Polish and English words. The scores had been calculated based on the number and the quality of sources confirming the given translation pair. Contrary to [9], consecutive sequences of words are not allowed to be linked independently to one word on the other side, however, lexicon multi-word units are taken into account during linking, so many-to-many links are allowed for words being part of multi-word units.

For word pairs not listed in the external lexicon, cognate analysis was introduced as an additional source of scores:

1. The Polish and English words are normalised to abstract from most frequent differences in Polish and English spelling: $ks \rightarrow x$, $ph \rightarrow f$, $sz \rightarrow sh$, $k \rightarrow c$, $w \rightarrow v$, $y \rightarrow i$.
2. The longest common prefix for the Polish and English word (after normalisation) is determined. If it is longer than 4, the words can be aligned, the longer is the common prefix, the higher is the score.

5 Results

A sample of 600 snippets with the abbreviation *ang.* was randomly selected for evaluation.[9] The sample was manually inspected and 333 (55,5%) correct translation pairs were identified and marked up. The automatic translation extraction procedure described in Sect. 4 was then applied to the sample. The results are presented in Table 4. The **baseline** is simply taking the same number of Polish words as on the English side. **One-word backup** is used when no lexicon/cognate alignments were found: if the parenthetical expression is just one English word, take the last pre-parenthesis word as its Polish translation.

It should be noted that if the external lexicon is used for alignment (**lexicon** method) and if a single link for the whole parenthetical English expression (one word or a multi-word unit) can be found in the lexicon then the correct translation pair (i.e. attested in the lexicon) will be extracted. Translation extraction

[9] The sample was not used during the development.

could be viewed more as confirmation rather than as discovery in such a case. Therefore, the **"fair" lexicon** method was introduced for comparison. "Fair lexicon" means that links for the whole parenthetical expression are not used during alignment.

Table 4. Results for the sample (E – number of extracted translations, C – number of correct translations)

Method	E	C	Prec.	Recall	F-score
baseline	368	169	0.459	0.508	0.482
cognates	84	40	0.476	0.120	0.192
lexicon	204	147	0.721	0.441	0.547
lexicon + cognates	216	154	0.713	0.462	0.561
lexicon + cognates + one-word backup	318	175	0.550	0.526	0.538
"fair" lexicon + cognates	168	114	0.679	0.342	0.455
"fair" lexicon + cognates + one-word backup	315	170	0.540	0.511	0.525

Finally, the translation extraction procedure was applied to the full corpus of 82,434 snippets. 46,728 unique translation pairs were extracted using the lexicon+cognates method. A sample of extracted translations is listed in Table 5.

Table 5. A sample of extracted translations. Extracted translations are underlined

Correct?	Snippet
yes	serwery domeny głównej (ang. root servers)
yes	będące integracją infrastruktury hurtowni danych (ang. Data Warehouse)
yes	szafa stelażowa (ang.: rack)
yes	Marynarka Wojenna Stanów Zjednoczonych (ang. United States Navy,
too long	Przerwa ta (ang. Intermission)
too long	Nicolas Dauphas z Uniwersytetu w Chicago (ang. University of Chicago)
too short	może oznaczać wystrój "bojowy" (ang. war color)
too short	posiadanie szczególnych przymiotów moralnych (ang. moral insight

6 Conclusions and Future Work

The number of extracted parenthetical translations reported here is much smaller than obtained for Chinese texts ([8], [9]), even if to take into account that the Polish corpus was smaller. The main reason is that the frequency of parenthetical English translation in Polish is simply much lower than in Chinese. There is

nevertheless some room for improvement: part-of-speech could be taken into account, machine learning techniques could be used for filtering out incorrect translation pairs, parenthetical translations without the abbreviation *ang.*[10] could be identified (e.g. using methods with which semantics relations are extracted [12]).

Even though the results presented in this paper are less encouraging than those reported for Chinese, the parenthetical expressions can be used as a supplementary source of Polish-English lexical data (for other examples of such sources see [6]).

The methods proposed in this paper could probably be adopted for other European languages provided that the frequency of the expression analogical to *ang.* is high enough.

Acknowledgements

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

References

1. Resnik, P., Smith, N.A.: The web as a parallel corpus. Comput. Linguist. **29**(3) (2003) 349–380
2. Melamed, I.D.: Automatic discovery of non-compositional compounds in parallel data. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing. (1997)
3. Shao, L., Ng, H.T.: Mining new word translations from comparable corpora. In: COLING '04: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2004) 618
4. Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D.: Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL-08: HLT, Columbus, Ohio, Association for Computational Linguistics (June 2008) 771–779
5. Jiang, L., Yang, S., Zhou, M., Liu, X., Zhu, Q.: Mining bilingual data from the web with adaptively learnt patterns. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, Association for Computational Linguistics (August 2009) 870–878
6. Graliński, F., Jassem, K., Kurc, R.: Acquiring bilingual lexica from keyword listings. In Vetulani, Z., ed.: Proceedings of 4th Language & Technology Conference, Poznań, Wydawnictwo Poznańskie Sp. z o.o. (2009) 326–330
7. Cao, G., Gao, J., Nie, J.Y.: A system to mine large-scale bilingual dictionaries from monolingual web pages. In: MT Summit XI. (2007) 57–64
8. Wu, X., Okazaki, N., Tsujii, J.: Semi-supervised lexicon mining from parenthetical expressions in monolingual web pages. In: NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2009) 424–432

[10] It should be noted, however, that parenthetical expressions with *ang.* constitute a substantial part (if not the majority) of all the parenthetical translation.

9. Lin, D., Zhao, S., Van Durme, B., Paşca, M.: Mining parenthetical translations from the web by word alignment. In: Proceedings of ACL-08: HLT, Columbus, Ohio, Association for Computational Linguistics (June 2008) 994–1002
10. Melamed, I.D.: Models of translational equivalence among words. Comput. Linguist. **26**(2) (2000) 221–249
11. Tiedemann, J.: Word to word alignment strategies. In: COLING '04: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2004) 212
12. Pantel, P., Pennacchiotti, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2006) 113–120