Two-Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora

Rogelio Nazar[1], Leo Wanner[2] and Jorge Vivaldi[1]

[1] Institut Universitari de Lingüística Aplicada, Pompeu Fabra University
Pl. de la Mercè 10-12. Barcelona.

[2] Fundació Barcelona Media, Pompeu Fabra University
Ocata 1. Barcelona.

{rogelio.nazar; leo.wanner; jorge.vivaldi}upf.edu

Abstract. This paper presents a language independent methodology for automatically extracting bilingual lexicon entries from the web without the need of resources like parallel or comparable corpora, POS tagging, nor an initial bilingual lexicon. It is suitable for specialized domains where bilingual lexicon entries are scarce. The input for the process is a corpus in the source language to use as example of real usage of the units we need to translate. It is a two-step flow process because first we extract single-word units from the source language and then the multi-word units where the initial single units are instantiated. For each of the multi-word units, we see if they appear in texts from the web in the target language. The unit of the target language that appears more frequently across the sets of multi-word units is usually the correct translation of the initial single-word source language entry.

Keywords: Bilingual Lexicon Extraction, Specialized Terminology, Machine Translation, Corpus Linguistics, Knowledge-poor methods, statistical methods.

1 Introduction

Strategies that involve the use of parallel corpora were among the first attempts to extract bilingual lexicons, using measures of statistical association to study the co-occurrence of pairs of entries in the aligned sentences ([1]; [2]; among others). This methodology has yielded accurate results. However, the shortcoming is that parallel corpora are not easy to compile, particularly in the case of specialized domains.

There have also been a number of attempts to extract bilingual lexicons without the need of parallel corpora, but using bilingual dictionaries as seed words. In this line of research there are two main trends. The first one is represented by authors such as [3]; [4]; [5]; [6] and [7]. Briefly, most of these approaches involve a similarity metric between a word in the source language and a candidate for translation in the target language. The rationale behind this strategy is that both the source language word and its equivalent are supposed to share the same profile of co-occurrence, in the same

manner that synonyms do ([8]; [9]; [10]). The process is then to study the units that co-occur significantly with the input word, and then try to translate as many as possible co-occurrents with the help of the initial bilingual lexicon. Once this information is gathered, the next step is to select as a candidate for translation the unit of the target language that co-occurs more often (in some corpus) with the greatest number of the translations obtained with the bilingual lexicon.

The other trend in the literature ([11]; [12]) is more related to the present proposal. In order to obtain documents where equivalents co-occur, [11] exploited search engines using pairs of equivalent terms in Japanese and English obtained from a bilingual dictionary as queries. This yields bilingual glossaries as well as other partially parallel texts among the downloaded collection. In the case of [12], they mine English-to-Chinese bilingual translations and transliterations from monolingual Chinese Web pages. Their idea is to extract equivalent pairs searching for a pattern of an English expression enclosed by parenthesis in a Chinese document. All possible sequences of words before the English words are considered possible translation candidates. In this way they collect a large number of translation candidates keeping only the most probable ones. The ranking of the equivalent pairs depends on different features but mainly on a machine learning method trained with an initial bilingual lexicon. With this they build a character bigram language model which yields a transliteration probability from English to Chinese.

In this paper we present a different approach that has encouraging results even when we do not use any of the resources that other authors need, such as comparable corpora, lemmatization, POS tagging or initial bilingual lexicons. The only input needed is Internet access and a corpus of the studied domain in the source language where the words we need to translate occur, with an extension of at least 40,000 tokens. Hereafter, this corpus will be called DSCSL, which stands for Domain Specific Corpus in the Source Language. The purpose of this knowledge-poor approach is to determine to what extent we can have a quality result with the minimum resources and the maximum amount of generalization possible. Not using this type of resources means that our conclusions can be extrapolated to other languages and domains. In future work we will explore hybrid methods, that is, also taking into account knowledge of the domain and the language, although at the moment we are casting the problem of bilingual terminology acquisition purely as a mathematical problem, in the line of previous work ([13]).

The rest of the paper is organized as follows: the next section gives a basic outline of the algorithm; section 3 explains some support actions that improve accuracy and section 4 shows some evaluation figures for the results. In section 5 we discuss conclusions and in section 6 a few promising lines of future work.

2 Basic Algorithm

English is a widely used language in scientific and technical domains, therefore it is not surprising to find terms or fragments of text in English in specialized literature even when it is written in other languages. Abstracts and keywords in English and in the language of the document are commonly included in scientific papers; titles in the

bibliographical references may include terms in English relevant to the topic of the document and even authors often include the English version of the terminology they introduce in their native languages. As a consequence, many specialized terms that are currently found on the web are statistically associated to their equivalents in different languages. Thus, we can obtain equivalent terms in different languages using the DSCSL as input and Internet access. The process is as follows:

1. Take as initial units single words from the DSCSL. This is a set *V*.
2. Extract multi-word units from the DSCSL where the single words appear. Then for each element V_i we have another set $V_i = \{V_{i,1}, V_{i,2}, V_{i,3} \ldots V_{i,n}\}$, where every $V_{i,j}$ is a unit that has V_i as a component (including its only component). For instance, if V_i is *light*, then $V_{i,1}$ is the same element, *light*, $V_{i,2}$ is *incident light*; $V_{i,3}$ is *transmitted light*; $V_{i,4}$ is *light beams*, and so on.
3. For each multi-word download *n* documents in the target language and sort their vocabulary by decreasing frequency order, as shown in tables 1, 2 and 3.
4. For each source language single word, see which is the single word that has occurred more times in the multi-word alignments. If V_i is *light* and the target language is Spanish, then the most recurrent element is *luz*.
5. Return to the multi-word alignments and select the candidate that shares an associated pair of words. For instance, in table 1, link *incident light* to *luz incidente* because they share the associated pair *luz-light.*

Table 1. Expressions appearing with *incident light* in Spanish documents.

Rank	Term	Frequency
1)	sekonic	76
2)	**luz incidente**	44
3)	incident light	22

Table 2. Expressions appearing with *transmitted light* in Spanish documents.

Rank	Term	Frequency
1)	microscope	158
2)	illumination	59
3)	scales	47
4)	transmitted light	39
5)	**luz transmitida**	32

Table 3. Expressions appearing with *light beams* in Spanish documents.

Rank	Term	Frequency
1)	photoshop	231
2)	optics	70
3)	photoshop comentarios	58
4)	light beams	41
...	...	...
13)	**haces de luz**	5

3 Some Support Strategies

There are several possibilities to improve the performance of the algorithm explained above that do not need external knowledge sources like bilingual dictionaries. In this section we explain those that are already implemented at the time of writing. Other strategies, that also seem interesting but have not yet been tested, are included in section 6, future work.

3.1 Including a Reference Corpus of General Language

We use reference corpora of general language of the source and the target language as a model of the expected frequency of a word in a text. Such reference corpora can be automatically acquired from the web using random queries of high or mid-frequency words. Two million tokens of text include approximately 20,000 types with a frequency greater than 5. We can use this model to eliminate false candidates that are frequent but non informative, like *would* or *has been*, because they are very frequent in the source language reference corpus. In addition, we can expect that both the source unit and the correct equivalent in the target language have a similar frequency in their respective reference corpora. Using the model of the language we can infer that *este trabajo* is not a good candidate for the translation of *alkyl group*, because *este trabajo* is more frequent in the Spanish reference corpus than *alkyl group*. In contrast, the correct candidate, which is *grupo alquilo*, has the same zero frequency in the reference corpus as *alkyl group*.

3.2 Using a Measure of Dispersion

We do not expect the correct translation to be only the most frequent among the downloaded collection, but also the most dispersed. If the downloaded collection has more than five documents, we can safely remove all units that appear in only one or two documents, and then the vocabulary size and computational cost will be significantly reduced. A simple measure of dispersion for candidates can be tf.df being tf the term frequency in the collection and df the number of documents where it occurs.

3.3 Using a Similarity Measure

It is usually observed, specially in scientific or technical domains, that term equivalents in different languages are cognates. Thus we can use some similarity measure to detect morphological resemblance between the candidates and the unit we are trying to translate. The unit *reflected energy* in our corpus generates a set of equivalent candidates. Among these we find the Spanish term *energía reflejada*. It is possible to automatically detect the relation between those two using a vector similarity measure. First we transform both units to vectors X and Y that have sequences of two characters as components. We compute a Dice similarity

coefficient, which is defined as (1). The correct translation is not always the cognate, however the result of this similarity measure will add some points to the final score of a candidate.

$$\mathrm{sim}(X,Y)=\frac{2|X\cap Y|}{|X|+|Y|} \quad (1)$$

3.4 Length Ratio

Equivalent terms should have a relatively similar length. Therefore, assuming that lr(*t*) is the length in characters of term *t*, we can define a ratio ln(*i*,*j*) for a term *i* and a translation candidate *j* as (2). Since we are not interested in an exact match in length (equivalents rarely have exactly the same length) we will take into account this variable only if it is less than a threshold of .7. Otherwise, it has a value of 1.

$$\ln(ij)=\frac{\operatorname{argmin}(\mathrm{lr}(i),\mathrm{lr}(j))}{\operatorname{argmax}(\mathrm{lr}(i),\mathrm{lr}(j))} \quad (2)$$

3.5 Statistical Noise Reduction

Another problem that we can observe is the presence of a repetitive noise that is domain specific. Candidates such as *Buenos Aires*, *Facultad de Ciencias*, *Universidad Nacional*, *Departamento de Ciencias*, etc., appear frequently as candidates for translation. We can reduce this noise statistically using a distributional criterion. These units have an exaggerated dispersion among the sets, thus we can reduce their weight accordingly to make up for their high frequency.

$$w'(t)=\frac{w(t)}{d(t)} \quad (3)$$

In (3), if *w(t)* is the weight that *t* had as a translation candidate for some term, *d(t)* would be the number of times *t* has been proposed as a candidate. With a threshold *h*, the size of the initial single-word sample over 7, we reduce the effect of *d(t)* as in (4).

$$d'(t)=\begin{cases}1 & \text{if } d(t)\le h\\ \dfrac{d(t)}{h} & \text{if } d(t)>h\end{cases} \quad (4)$$

3.6 Analyzing the Language of the Context

A simple strategy is to study the contexts of occurrence of a translation candidate. If the contexts are in the target language, then the probability of being a correct translation is higher, except if it is enclosed by parenthesis or marked with some typographical convention such as italics, which may indicate that the sign is extraneous to the language of the text. We can confidently eliminate a candidate if the ratio of Spanish words vs. English words in the contexts where the candidate occurs is below a certain threshold. If *s(t)* is the total number of Spanish words found in the contexts of term *t* and *e(t)* the total number of English words found in the same context, then we can define the condition (5).

$$w(t)=0 \quad if \frac{s(t)}{e(t)+1}<.7 \tag{5}$$

3.7 Final Weighting of Candidates

As stated earlier, the final weighting technique is divided into two-steps. The first step is to find, for each single word unit, the equivalent candidates for the multi-word units where the single unit appears. We have defined a collection of measures of weighting, such as the frequency of the term in the collection, fr(*t*), that we will express as (6).

$$\mathrm{fr}(t)=\log(\mathrm{fr}(t)+1) \tag{6}$$

The frequency of a word in a reference target language corpus, Sfr(*t*), is also expressed in log scale, as Efr(*t*), the frequency of that term in the English reference corpus. Naturally, we will prefer as a Spanish translation a unit that has a greater frequency on the Spanish corpus than in the English one. We can define a binary value *m(t)* with value of 1 if this is the case. The frequency on the reference corpus of a term *i* and its translation candidate *j*, as explained in section 3.1, also gives us pr(*i*, *j*), defined as (7).

$$\mathrm{pr}(ij)=\frac{\mathrm{argmin}(\mathrm{Sfr}(i),\mathrm{Efr}(j))}{\mathrm{argmax}(\mathrm{Sfr}(i),\mathrm{Efr}(j))} \tag{7}$$

Other variables we defined are df(*t*), in section 3.2; sim(*t*), the similarity metric of section 3.3. and two more binary variables, *y(t)* and *n(t)*. The first one has value 1 if the term *t* has as an internal component (not at the beginning nor at the end) a very frequent Spanish word, such as “de” in *medio de almacenamiento*, while *n(t)* will punish with value 1 a candidate with a very frequent English word inside. For every

$V_{i,j}$, as a multi-word instance of English word V_i , the first weighting of a candidate $V_{i,j,k}$ is then defined as (8).

$$w\left(V_{ijk}\right)=\mathrm{fr}(k).\frac{\mathrm{Sfr}(k)}{\mathrm{Efr}(k)}.\frac{(y(k)+1)}{(n(k)+1)}.\mathrm{df}(k).\mathrm{sim}\left(V_{ij},k\right).\mathrm{pr}\left(V_{ij},k\right) \quad (8)$$

Once the score for each alignment $V_{i,j,k}$ is defined, we will calculate the best candidate for V_i. Some of the variables are the same, only changing the parameters: pr(*i*,*k*) or sim(*i*,*k*), remembering that V_i is the initial single word in the target language and *k* the term of the weighted alignment $V_{i,j,k}$. Weights are normalized before the final score is defined, ranging then from 0 to 1. The final weight of the alignment of elements V_i and *k* is shown in (9). The only variable that is new here is st(V_i, *k*), which is the number of subsets of V_i (V_{ij}) where *k* is present. In the example given in section 2, *k* would be how recurrent is the element *luz* among the translation sets of the phrases with the element *light.*

$$w\left(V_i,k\right)=w\left(V_{ijk}\right)+\mathrm{st}\left(V_i,k\right)+m(k)+\mathrm{sim}(i,k)+\mathrm{pr}(i,k) \quad (9)$$

Certainty over multi-word alignment can be recalculated now on the basis of the results of (9). If two single-word units are highly associated, the multi-word units where they appear will be associated too, as explained at the end of section 2. Thus, if $w(V_i, k)$ is above a certain threshold, for instance, if *dispositivo* and *device* are strongly associated, then, from all the candidates available for *safety device* the algorithm will select *dispositivo de seguridad* because it contains a member of an associated pair. If two multi-word equivalent candidates *i* and *j* share an associated pair, then the final certainty score sc(*i*,*j*) is defined as (10). The new element here is sw(*i*,*j*) that is the number of associated pairs in common.

$$\mathrm{cs}(i,j)=\mathrm{sw}(i,j)+\mathrm{sim}(i,j)+\mathrm{pr}(i,j)+\ln(i,j)-d(j) \quad (10)$$

4 Evaluation

As a preliminary and small scale evaluation, we show an experiment translating from English to Spanish[1], 76 randomly selected single-words from the DSCSL. For each

1 A complete evaluation for the claim of language independence would be to translate, for example, from French to Spanish using English as intermediate step. It is not yet clear then if problems would be magnified by the iteration or if, on the contrary, the triangle could be used as an extra source of information and greater certainty.

single-word, the computer program written for this evaluation selected up to 10 different multi-word units (of a maximum extension of five words). For each multi-word unit, the program attempted to download up to 100 documents in Spanish where the unit occurs. After downloading thousands of documents from the web and selecting the most frequent, informative, dispersed and similar units (as explained in sections 2 and 3), the system outputs tables of candidates ordered by their final score as equivalents. Table 4 shows the results we obtained.

Table 4. Accuracy by position in the rank of candidates on a random sample of 76 English head-words.

1#	2#	3#	5#	10#	15#
39%	44%	50%	51%	57%	59%

Table 5 shows some of the alignments. It must be borne in mind that these equivalences are valid only in the domain we are studying. For instance, an equivalence between *emitting* and *emisor* may seem strange because it is a gerund and therefore it should be translated as *emitiendo*. But in this context it is correct because we have terms like *light emitting diode* that are translated as *diodo emisor de luz*.

Table 5. A few examples of the obtained single-word alignments.

English Term	Spanish Equivalent
acetate	acetato
apparatus	aparatos
beam	rayo / haz
clock	reloj
curing	curado
cutting	corte
deflection	deflexión
device	dispositivo

In respect to the alignment of multi-word units, we need to minimize the compounding effect of errors made during the single-word alignment running again in the multi-word alignment step. Therefore, we rank the multi-word pairs by the degree of certainty explained in subsection 3.7-(10). In this case, the program has to select only one translation for each multi-word unit. The trial is considered a success if the aligned pair is correct, such as *carbon material* and *materiales de carbono* or *position sensor* and *sensor de posición*. All other cases, including partial matches such as *optical disc drive* and *disco óptico* were considered failures. From a sample of 150 multi-word alignments, only a small subset has a minimum degree of certainty. However, certainty and precision are linked, as shown in Figure 1. The vertical axis indicates the cumulative precision while the results are ranked in the horizontal axis according to certainty. We can see that, among the first 40 positions on the ranking, precision is above 50%. From that point, the curve rapidly decreases and gets steady from position 100 at around 25% precision. These figures are consistent with the small proportion of terms in a random sample of *n*-grams.

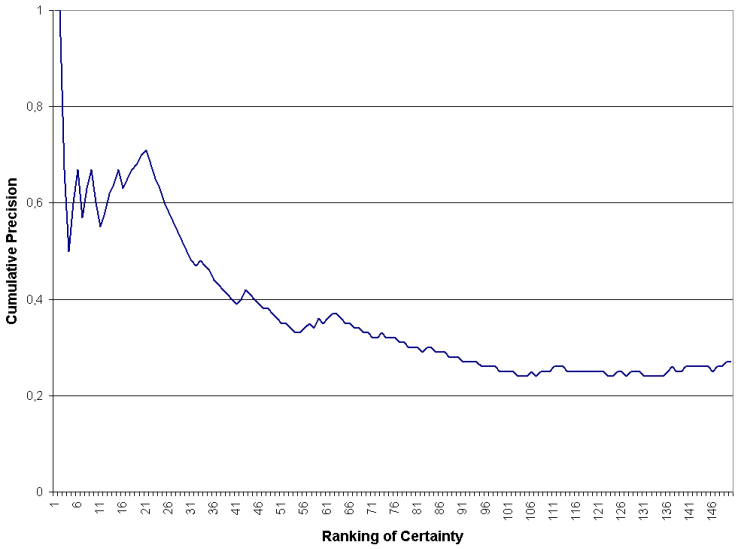

Fig. 1. Precision vs. Certainty in multi-word alignment using only the top candidate. Most of the correct trials are in the first positions of the ranking.

6 Conclusions

We have presented a method for the extraction of a bilingual lexicon requires practically no external resources except a corpus with the units to translate and Internet access. It is an interesting methodology from an engineering, terminographic or lexicographic point of view. However, it is also an attractive subject of research from a purely theoretical perspective, since it states a fact about macroscopic and structural regularities of language that are visible only now, when the massive amount of data from the web offer us the possibility to extract valid conclusions out of a great number of apparently chaotic individual behaviors, the decisions made by each author in every language and discipline. From this unorganized social behavior, a remarkable regularity emerges, that is the statistical association of equivalent terms in different languages.

7 Future Work

We are extending this work in different directions. Most new ideas will be included as support and refinement strategies, like those described in section 2. The most important pending work now is replication of this experiment with bigger data sets and with different domains and languages. The second most important thing will be to

use a terminology extraction system for the selection of the units to translate. This would have undoubtedly yielded better results than with the simple random sampling we used, and its replacement will not affect the general architecture of this system. Another line is to try a hybrid method. Using different degrees of knowledge of the language and/or the domain in question may improve the quality of the results. There is yet another strategy that is conceptually simple but computationally costly. One of the possible ways to eliminate false candidates would be to iterate the process in the opposite direction. That means, repeating the process with each of the equivalent candidates this time as input to find their translation in what was originally the source language. The correct translation will have the original term among the equivalent candidates in the original source language.

Acknowledgments
This paper was possible thanks to fundings from the project RICOTERM3 lead by Dr. Mercè Lorente, which is in turn funded by the Ministry of Education and Science of the Government of Spain (HUM2007-65966-C02-01/FILO). We would like to thank the anonymous reviewers for their comments and to Edmund Maklouf for proofreading.

References

1. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roosin, P.: A Statistical Approach to Machine Translation. Computational Linguistics, 16, pp. 79--85 (1990).
2. Gale, W., Church, K.: Identifying word correspondences in parallel texts. Proceedings of the DARPA SNL Workshop (1991).
3. Fung. P.: Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. Proceedings of the Third Workshop on Very Large Corpora, pp. 173--183 (1995).
4. Fung, P.: A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. Proceedings of the AMTA Conference, pp. 1--16 (1998).
5. Fung, P., McKeown, K.: Finding Terminology Translations From Non-Parallel Corpora. The 5th Annual WVLC, pp. 192--202, Hong Kong (1997).
6. Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. Proceedings of 37th ACL Annual Meeting, pp. 5190--526 (1999).
7. Tanaka, T., Matsuo, Y.: Extraction of Translation Equivalents from Non-Parallel Corpora. Proceedings of the 8th TMI Conference, pp. 109--119 (1999).
8. Harris, Z.: Distributional Structure. In: Katz, J.J. The Phylosophy of Linguistics, pp 26--47. Oxford University Press, New York (1954/1985).
9. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Norwell, MA (1994).
10. Schütze, H., Pedersen, J.: A Co-occurrence-based Thesaurus and Two Applications to Information Retrieval. Information Processing and Management. 33(3), p.307--318 (1997).
11. Nagata, M., Saito, T.; Suzuki, K.: Using the Web as a Bilingual Dictionary. Proceedings of ACL DD-MT Workshop (2001).
12. Guihong C., Jianfeng G., Jian-Yun N.: A System to Mine Large-Scale Bilingual Dictionaries from Monolingual Web Pages. Proceedings of MT Summit XI (2007).
13. Nazar, R.: Bilingual Terminology Acquisition from Unrelated Corpora. Proceedings of the XIII EURALEX Congress, Barcelona (2008).