LIHLA: A lexical aligner based on language-independent heuristics

Helena de Medeiros Caseli[1,2] , Maria das Graças Volpe Nunes[1] , Mikel L. Forcada[2]

[1]NILC – Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, CP 668P, 13560-970 São Carlos-SP, Brazil
`http://www.nilc.icmc.usp.br`

[2]Grup Transducens – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
`http://transducens.dlsi.ua.es`

`{helename, gracan}@icmc.usp.br, mlf@dlsi.ua.es`

Abstract. *Alignment of words and multiword units plays an important role in many natural language processing applications, such as example-based machine translation, transfer rule learning for machine translation, bilingual lexicography, word sense disambiguation, etc. In this paper we describe LIHLA, a lexical aligner which uses bilingual probabilistic lexicons generated by a freely available set of tools (NATools) and language-independent heuristics to find links between single words and multiword units in Brazilian Portuguese, Spanish and English parallel texts. The method has achieved a precision of 92.48% and 84.35% and a recall of 88.32% and 76.39% on Brazilian Portuguese–Spanish and Brazilian Portuguese–English parallel texts, respectively.*

1. Introduction

Alignment of words and multiword units plays an important role in many natural language processing (NLP) applications, such as example-based machine translation (EBMT) [Somers 1999] and statistical machine translation (SMT) [Ayan et al. 2004, Och and Ney 2000], transfer rule learning [Carl 2001, Menezes and Richardson 2001], bilingual lexicography [Gómez Guinovart and Sacau Fontenla 2004], and word sense disambiguation [Gale et al. 1992], among others.

Aligning two (or more) texts means finding correspondences (translation equivalences) between segments (paragraphs, sentences, words, etc.) of the source text and segments of its translation (the target text). In this paper the focus is on lexical alignment, that is, alignment between single words and multiword units in Brazilian Portuguese (`pt`), Spanish (`es`) and English (`en`) parallel texts.

In the last years, several lexical alignment systems have been proposed in the literature achieving precision and recall values between 71–84% and 61–81%, respectively, for several different language pairs. Among all of them, statistical systems are considered to be the state of the art (e.g., [Hiemstra 1998] and [Och and Ney 2000]). Although these systems provide quite satisfactory results they can not deal properly with syntactic differences between languages, such as non-consecutive phrasal information, long-range dependencies [Ayan et al. 2004] and alignments involving multiword units. These problems are very frequent in lexical alignment and unfortunately also very difficult to handle.

Following the same idea of many recently proposed approaches on lexical alignment (e.g., [Wu and Wang 2004] and [Ayan et al. 2004]), the method described in this paper, LIHLA (Language-Independent Heuristics Lexical Aligner), tries to solve some of these problems by using statistical alignments between single words (defined in bilingual probabilistic lexicons) as a starting point, and by applying language-independent heuristics to them, aiming at finding the best alignments between words or multiword units.

Although the most frequent alignment category is $1 : 1$ (in which one source word is translated exactly as one target word), other categories such as omissions ($1 : 0$ or $0 : 1$) or those involving multiword units ($n : m$, with n and/or $m \geq 1$) are also possible. An example of alignment involving a multiword unit is the $1 : 2$ alignment between `pt` word *dos* and `es` multiword unit *de los*.

This paper is organized as follows: section 2 presents an overview of bilingual lexicon generation and section 3 explains how LIHLA works. Section 4 describes some experiments carried out with LIHLA and their results. Finally, in section 5, some concluding remarks are presented.

2. Bilingual lexicon generation

As its first step, LIHLA uses alignments between single words defined in two statistical bilingual lexicons (source–target and target–source) generated using NATools.[1] To generate these bilingual lexicons, the parallel texts have to be sentence-aligned. In the experiments described here, sentence alignment was carrried out using a version of Translation Corpus Aligner (TCA) [Hofland 1996], but any other sentence alignment method proposed in the literature could be similarly used, such as the well-known method of Gale and Church (1991).[2]

The automatically sentence-aligned texts were not post-processed for correction of misalignments because we believe that a few misaligned sentences will not significantly degrade the translation probabilities of all words in the corpus considering the way NATools generates the bilingual lexicons. Furthermore, it is important to say that the sentence alignment produced in this step is not used by LIHLA (version 1.0) since it takes raw texts in spite of aligned ones as its input.

So, given two sentence-aligned corpus files, the NATools word aligner —based on the Twenty-One system [Hiemstra 1998]— counts the co-occurrences of words in all aligned sentence pairs and builds a sparse matrix of word-to-word probabilities (Model A) using an iterative expectation-maximization algorithm (5 iterations by default). Finally, the elements with higher values in the matrix are chosen to compose two probabilistic bilingual lexicons (source–target and target–source) [Simões and Almeida 2003]. For each word in the corpus, each bilingual lexicon gives: the number of occurrences of that word in the corpus (its absolute frequency) and its most likely translations together with their probabilities.

Figure 1 shows an entry in the `pt–es` bilingual lexicon to the `pt` word *dos*. In this example, the best translation is *los* and the second one is *de*. It is due to the fact that the

[1] NATools is a set of tools developed to work with parallel corpora, which is freely available in `http://natura.di.uminho.pt/natura/natura/`.

[2] For more information on sentence alignment methods see PESA (Portuguese-English Sentence Alignment) project home-page: `http://www.nilc.icmc.usp.br/projects/PESA.html`.

pt word *dos* can be translated into Spanish as several combinations of other prepositions plus the definite article *los* or just the article. Also, the probability of omission of its translation (indicated by \(null\)) is specified which is higher than the probability of its translation as *dos* or *la*.

```
"dos" => {
          count => 2196,
          trans => {
                          "los" => 0.74646669626236,
                          "de" => 0.178675398230553,
                          "\(null\)" => 0.0156443770974874,
                          "dos" => 0.0111551126465201,
                          "la" => 0.00522150145843625,
          },
},
```

Figure 1. Possible translations for pt word *dos* in the pt–es bilingual lexicon

3. How LIHLA works

In spite of the fact that sentence alignment has been performed previously for lexicon generation (see section 2), the texts are presented to LIHLA (version 1.0) without any sentence alignment marks. So, given a pair of raw parallel texts and two bilingual lexicons (those generated by NATools in the previous step), LIHLA tries to find the best alignment between source and target tokens (words, numbers, special characters, etc.) following the algorithm on Figure 2. As its output, LIHLA produces a set A of alignments $(\alpha : \beta)$ where α is a sequence of one or more source tokens (separated by '+'), and β is a similar sequence of target tokens.

For each source token s_j, LIHLA establishes a correspondence window (the area where the correspondences will be looked for) and takes those source and target tokens inside this window with the same type (word or special character[3]) as s_j as possible translations of each other. Those source and target tokens are stored in source (C_S) and target (C_T) candidate sets, respectively (line 4). The length of the correspondence window is calculated as the average sentence length on both texts and the window is centered regarding the positions of s_j and the previously aligned tokens around it (if possible).

Then, if s_j is a special character LIHLA looks for a special character in C_T —priorizing identical ones— by means of the `align_char` function (line 6) which returns an alignment $(s_j : \beta)$ where β can be a target special character or the `null` word (indicating an omission alignment). Otherwise, if s_j is a word, language-independent heuristics are applied to the words in C_S and the words in C_T (lines 8 to 31) aiming at finding the best possible lexical alignments between s_j (and maybe other words in C_S) and one or more words in C_T.

First of all, LIHLA priorizes a target word which is identical to s_j, to find exact matches, for instance, between proper names and numbers. If this word is found then a

[3]Each token in a source/target sentence is classified as a word if it contains at least one alphanumeric character or as a special character otherwise.

algorithm LIHLA

Input:

a source text $S = \{s_1, s_2, ..., s_x\}$ with x words
a target text $T = \{t_1, t_2, ..., t_y\}$ with y words
a source–target bilingual lexicon B_S
a target–source bilingual lexicon B_T

Output:

a set A of alignments between words in S and T

Pseudo code:

$A \leftarrow \emptyset$
while alignments can still be produced **and** not maximum number of iterations **do**
  for $j \leftarrow 1$ **to** x
    set_correspondence_window(s_j, C_T, C_S)
    if s_j is a special character
    then $A \leftarrow A \cup \{$align_char$(s_j, C_T)\}$
    else
      if $(\exists\ t_i \in C_T \mid t_i = s_j)$ **then** $A \leftarrow A \cup \{(s_j : t_i)\}$ [1 : 1 alignment]
      else
        $C'_T \leftarrow C_T \cap$ look_for_translation(s_j, B_S)
        if $C'_T \neq \emptyset$ **then**
          $t_i \leftarrow$ best_position(s_j, C'_T)
          $M_T \leftarrow$ look_for_multiword(t_i, C_T, C_S)
          if $(|M_T| > 1)$ **then** $A \leftarrow A \cup \{(s_j : M_T)\}$ [1 : n alignment]
          else
            $C'_S \leftarrow C_S \cap$ look_for_translation(t_i, B_T)
            if $s_j \in C'_S$ **then**
              $M_S \leftarrow$ look_for_multiword(s_j, C_S, C_T)
              if $(|M_T| > 1)$ **then** $A \leftarrow A \cup \{(M_T : t_i)\}$ [n : 1 alignment]
              else $A \leftarrow A \cup \{(s_j : t_i)\}$ [1 : 1 alignment]
            end_then
          end_else
        end_then
        else
          $C'_T \leftarrow$ look_for_cognate(s_j, C_T)
          if $C'_T \neq \emptyset$ **then**
            $t_i \leftarrow$ best_cognate(C'_T)
            $A \leftarrow A \cup \{(s_j : t_i)\}$ [1 : 1 alignment]
          end_then
        end_else
      end_else
    end_else
  end_for
end_while
for-each $(s_j : t_i) \in A$ **and** $(s_{j+k} : t_{i+l}) \in A$ **with** $k, l > 1$ **do**
  if $(k = l)$ **then for** $z \leftarrow 1$ **to** $(k-1)$ $A \leftarrow A \cup \{(s_{j+z} : t_{i+z})\}$ [1 : 1 alignment]
  else
    $M_S \leftarrow s_{j+1} + ... + s_{j+k-1}$
    $M_T \leftarrow t_{i+1} + ... + t_{i+l-1}$
    $A \leftarrow A \cup \{(M_S : M_T)\}$ [n : m alignment]
  end_else
end_for-each
return A

Figure 2. Lexical alignment algorithm of LIHLA (version 1.0)

$1 : 1$ alignment is established (line 8); otherwise, LIHLA looks for possible translations in the source–target bilingual lexicon (B_S) and makes an intersection between them and the words in C_T (line 10).

In this intersection, if no candidate word identical to those in B_S is found in C_T then, for each word in B_S, LIHLA tries to look for cognates for this word in C_T using the longest common subsequence ratio (LCSR).[4] The cognates which have been found are added to C'_T and the search follows with the next word in B_S until all words in B_S and/or C_T have already been processed. By doing this, LIHLA can deal with small changes in possible translations such as different forms of the same verb, changes in gender and/or number of nouns, adjectives, and so on. Furthermore, if a \(`null`\) is found in B_S it is added to C'_T to allow an omission alignment to be set.

In the next step, LIHLA selects the best target candidate word, that is, that in C'_T which is at the best position in relation to s_j (line 12), and tries to find a multiword unit involving it (line 13). A multiword unit, in this case, is composed of words in C_T that occur immediately before and/or after the best target word (t_i) and are not possible translations of other words in C_S. If a multiword unit is found then a $1 : n$ alignment is established (line 14); otherwise LIHLA will try to confirm the alignment verifying if s_j is a possible translation for t_i (lines 16 and 17). If the translation is possible in both sides, then a multiword unit involving s_j is also looked for (line 18) and, if it is found, a $n : 1$ alignment is established (line 19), otherwise a $1 : 1$ alignment is set (line 20).

LIHLA can also deal with target words that do not occur in the source–target bilingual lexicon B_S and the set of target candidate words C_T at the same time by looking for cognate words for s_j in C_T using the LCSR and setting a $1 : 1$ alignment between s_j and its best cognate (lines 25 to 29). It is important to say that the steps 3 to 33 are repeated while alignments can still be produced and a maximum number of iterations (10 by default) is not reached (in the experiments described in this paper LIHLA has performed on average 4 iterations for each pair of parallel texts). Furthermore, at the first iteration, all frequent words[5] are ignored to avoid erroneous alignments since all subsequent alignments are based on the previous ones.

In the last step (lines 35 to 42), LIHLA aligns the remaining unaligned source and target tokens between two pairs of already aligned ones (in A) establishing several $1 : 1$ alignments when there are the same number of source and target tokens (line 36), or just one alignment involving all source and target tokens if they exist in different quantities (lines 38 to 40). The decision of creating n $1 : 1$ alignments in spite of just one $n : n$ alignment when there is the same number of source and target tokens is due to the fact that a $1 : 1$ alignment is more likely to be found than a $n : n$.

Table 1 presents some examples of `pt`–`es` lexical alignments produced by LIHLA, together with their categories and the steps of the algorithm in which they were established. The first example illustrates the case in which LIHLA did not find any target word for the given source word during the alignment process (`null` is a special word

[4]The LCSR of two words is computed by dividing the length of their longest common subsequence by the length of the longer word. For example, the LCSR of `pt` word *alinhamento* and `es` word *alineamiento* is $\frac{10}{12} \simeq 0.83$ as their longest common subsequence is *a-l-i-n-a-m-e-n-t-o*.

[5]LIHLA considers as source/target frequent words those whose absolute frequencies together give 30% of the total frequency of all words in the lexicon for that language.

used in omission alignments); and the last one is an example of an alignment generated based on two previously aligned pairs: (*tão:tan*) and (*que:que*).

Table 1. Examples of `pt–es` lexical alignments generated by LIHLA

Category	Alignment	Algorithm step
$1:0$	(*por:*`null`)	none
$1:1$	(*vida:vida*)	5
$1:1$	(*atmosfera:atmósfera*)	25
$1:1$	(*apelo:llamado*)	17
$1:2$	(*dos:de+los*)	11
$2:1$	(*só+que:pero*)	16
$3:2$	(*a+respeito+do:referente+al*)	36
$1:1$ $1:1$ $1:1$	(*tão:tan*) (*bons:halagüeños*) (*que:que*)	17 32 5

4. Evaluation and results

For testing and evaluation purposes, we used a `pt–es` parallel corpus (`CorpusFAPESP`) composed of 1,292 articles (646 in `pt` and 646 in `es`) from the online version of the Brazilian scientific magazine *Pesquisa FAPESP*.[6] The `pt–es` `CorpusFAPESP` has 908,656 tokens (431,169 in `pt` and 477,487 in `es`).

A manual reference alignment has been built with 20 pairs of parallel texts (3%) randomly selected from the whole set. The 31,471 tokens (14,756 in `pt` and 16,719 in `es`) in the reference corpus were manually aligned by two bilingual annotators following the guidelines established in [Caseli et al. 2005][7] and the observed inter-annotator agreement rate of 95% indicates that the annotations are reasonably reliable. As expected, most of the alignments on the `pt–es` reference corpus as annotated by the human annotators are $1:1$ (83.85%), but other categories such as omissions (6.60%) or those involving multiword units (9.55%) can also be found.

Alignments in the reference corpus were used to automatically evaluate those produced by LIHLA using the well-known precision, recall and alignment error rate (AER) metrics. Let R be the set of reference alignments and A the set of alignments proposed by the method; $|A \cap' R|$ stands for the number of source and target tokens found in reference (R) and proposed (A) alignments at the same time, splitting the tokens in reference alignment between more than one proposed alignment if needed. Precision, recall and AER (the complement of the F-measure, a combination of precision and recall metrics) are shown below. In these experiments, AER was calculated considering all alignments as sure links[8] —as in [Wu and Wang 2004]— and not as possible and sure links— as done in [Och and Ney 2000].

[6]The *Pesquisa FAPESP* magazine is available at `http://revistapesquisa.fapesp.br` with parallel texts written in Brazilian Portuguese (original), English (version) and Spanish (version).

[7]The guidelines defined in [Caseli et al. 2005] are based on those defined for ARCADE [Véronis and Langlais 2000] and Blinker [Melamed 1998] projects.

[8]A sure link is an unambiguous alignment while a possible link is an alignment that might or might not be established since there is not a straight correspondence between source and target tokens.

$$\text{Precision} = \frac{|A \cap' R|}{|A|} \qquad \text{Recall} = \frac{|A \cap' R|}{|R|} \qquad \text{AER} = 1 - 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 2. Evaluation of LIHLA per alignment category on `pt–es` parallel texts

Category	Precision (%)	Recall (%)	AER (%)
1 : 1	80.21	86.59	16.72
1 : 1 – omissions	89.86	88.40	10.87
omissions	30.49	64.18	58.66
multiword units	74.21	68.72	28.64
all	83.23	86.68	15.08
all – omissions	**92.48**	**88.32**	**9.65**

Table 3. Evaluation of `pt–es` alignments not considering omission cases

Method (Train×Test)	Precision (%)	Recall (%)	AER (%)
LIHLA 646×20	92.48	88.32	9.65
NATools 646×20	85.99	72.64	21.25
LIHLA 626×20	91.37	86.61	11.07
NATools 626×20	85.06	65.76	25.82
LIHLA 20×20	87.95	81.82	15.22
NATools 20×20	79.87	57.66	33.03

Table 2 shows the metric values per alignment category in `pt–es` parallel texts and, as can be noticed, the worst AER is on omission category (58.66%) and the AER for all categories except omissions (all – omissions) is 9.65%. From this table it is also possible to notice the promissing results of LIHLA in the alignment involving multiword units: 74.21% of precision and 68.72% of recall.

Table 3 presents the results of an experiment carried out in order to compare LIHLA with another alignment system. In this experiment, the 20 pairs of parallel texts used for testing were also aligned using the bilingual lexicons generated by NATools but, now, considering only the best candidates (those with the highest probabilities) in both sides (source–target and target–source) and performing only $1:1$ alignments. This experiment was repeated using three pairs of bilingual lexicons which were generated from the whole `pt–es CorpusFAPESP` (646 pairs of parallel texts), the set without the 20 pairs used for testing (a kind of training set composed of 626 pairs of parallel texts) and only the test set (20 pairs of parallel texts).

As it can be noticed from Table 3, LIHLA has improved the results of the method based only on the best $1:1$ alignments in all "Train×Test" sets. Improvements are more than 6% in precision and more than 15% in recall. Furthermore, when LIHLA was ran with small lexicons (generated only from the test set, that is, just 20 pairs of texts) the precision and recall were above 87% and 81% respectively, showing that the method is robust and can achieve a satisfactory performance even with a small amount of data.

A small-scale experiment was also carried out with `pt–en CorpusFAPESP` aiming at verifying how LIHLA would perform on a different pair of languages. The `pt–en CorpusFAPESP` has also 646 pairs of parallel texts and 893,141 tokens (431,169 in `pt` and 461,972 in `en`). The `pt-en` reference corpus has 10 pairs of parallel texts and

15,900 tokens (7,631 in `pt` and 8,269 in `en`) and was aligned following the same guidelines mentioned previously in this section. Table 4 presents metric values (per alignment category) on `pt`–`en` lexical alignments produced by LIHLA while Table 5 shows the results of the experiments carried out with 3 "Train×Test" sets for comparing with those generated based only on the best candidates as pointed out by NATools.

As expected, the metric values for `pt`–`en` were lower than those for `pt`–`es` (a decrease of 8% in precision and 11% in recall) due to the larger difference between the languages involved. However, LIHLA has achieved interesting results 84.35% of precision and 76.39% of recall considering all categories except omissions (all – omissions) and 64.25% of precision and 60.29% of recall on alignments involving multiword units. Once again LIHLA has improved in more than 4% precision and in more than 14% recall the performance of an alignment produced based only on the best candidates as pointed out by NATools.

Table 4. Evaluation of LIHLA per alignment category on pt–en parallel texts

Category	Precision (%)	Recall (%)	AER (%)
1 : 1	63.46	75.02	31.25
1 : 1 – omissions	77.67	76.63	22.85
omissions	20.64	58.36	69.51
multiword units	64.25	60.29	37.79
all	70.32	75.01	27.41
all – omissions	**84.35**	**76.39**	**19.83**

Table 5. Evaluation of pt–en alignments not considering omission cases

Method (Train×Test)	Precision (%)	Recall (%)	AER (%)
LIHLA 646×10	84.35	76.39	19.83
NATools 646×10	78.30	62.14	30.71
LIHLA 636×10	83.40	74.53	21.29
NATools 636×10	78.50	56.98	33.97
LIHLA 10×10	67.02	56.34	38.78
NATools 10×10	61.33	36.38	54.33

In the literature, several lexical alignment evaluations point at the statistical alignment systems based on IBM and HMM models as the state of the art in this field. Some results of these evaluations using the same metrics showed previously and only sure links (to allow a comparison with our results), are presented below. As shown by the results of the shared task described in [Mihalcea and Pedersen 2003], considering the subtask of limited resources (in which the systems were allowed to use only the resources provided) the best systems on Romanian–English and English–French were those based on IBM Model 4 (82.65% best precision and 62.44% best recall) and IBM Model 2 (72.54% best precision and 80.61% best recall), respectively.

In [Ayan et al. 2004], on Spanish–English alignments, GIZA++ achieved a precision of 72.32% and a recall of 72.28% versus 73.01% and 73.36%, respectively, achieved by the system proposed in this paper. In [Wu and Wang 2004], on English–Chinese alignments, GIZA++ achieved a precision of 71.40%, a recall of 69.42% and an AER of

29.61% while the method proposed by the authors has achieved 83.63%, 76.73% and 19.97%, respectively. In that paper, the authors also evaluated the multiword alignments (56.65% of precision, 40.83% of recall and 52.54% of AER) with values worse than ours (see Tables 2 and 4).

Therefore, with respect to the values reported in the refered papers for several language pairs and considering only sure links (our case) it can be concluded that precision lies between 71% and 84% and recall is between 61% and 81%.[9]

5. Concluding remarks

This paper has presented a lexical alignment method, LIHLA, which aligns words and multiword units based on initial statistical word-to-word correspondences and language-independent heuristics. LIHLA has been evaluated on `pt–es` and `pt–en` parallel texts and has achieved, respectively: 92.48% and 84.35% of precision, 88.32% and 76.39% of recall and 9.65% and 19.83% of AER. These values are in accordance with those values reported in the literature for other language pairs, that is, 71–84% of precision and 61–81% of recall, being even above them in the `pt–es` parallel texts.

Furthermore, LIHLA has some advantages when compared to other lexical alignment methods: it does not need to be trained for a new pair of languages (as in [Och and Ney 2000]) and neither does it require pre-processing steps (apart from tokenization) to handle texts (as in [Gómez Guinovart and Sacau Fontenla 2004]) or a large parallel corpus since it has achieved interesting results even with a very small amount of data. LIHLA also can deal quite well with multiword units as pointed out by the 74.21% and 64.25% of precision and 68.72% and 60.29% of recall on `pt–es` and `pt–en` parallel texts, respectively.

Finally, the best contribution of LIHLA is that it is based on language-independent heuristics and, therefore, it can be applied to a new pair of languages without any modification (as has been done with `pt–es` and `pt–en`). As future work, we aim at evaluating LIHLA on different corpora from other genres and languages, investigating better ways to deal with multiword units and also the impact of using sentence-aligned parallel texts and/or additional linguistic information (such as part-of-speech tags) as its input. As a long-term goal, LIHLA will be part of a system to learn transfer rules to machine translation from sequences of aligned words.

Acknowledgments

We thank FAPESP, CAPES, CNPq and the Spanish Ministry of Science & Technology (Project TIC2003-08681-C02-01) for financial support. We also thank Miriam A. G. Scalco for her help in the manual lexical alignment process.

References

Ayan, N. F., Dorr, B. J., and Habash, N. (2004). Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the AMTA (AMTA-2004), LNAI*, number 3265, pages 17–26.

[9]Since not all methods refered in this paper have shown AER values in their evaluations, we believe that it is better not to arrive at conclusions about the lowest and the highest values for this metric.

Carl, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of CoNLL-2001*, pages 145–151, Toulouse, France.

Caseli, H. M., Scalco, M. A. G., and Nunes, M. G. V. (2005). Manual para anotação de alinhamentos lexicais. Série de Relatórios do ICMC 256 (NILC-TR-05-09), NILC, http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-05-09.pdf.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992)*, pages 101–112, Montreal, Canada.

Gómez Guinovart, X. and Sacau Fontenla, E. (2004). Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural*, 33:133–140.

Hiemstra, D. (1998). Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In Coppen, P. A., van Halteren, H., and Teunissen, L., editors, *Proceedings of the 8th CLIN meeting*, pages 41–58.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. In Hockey, S., Ide, N., and Perissinotto, G., editors, *Research in Humanities Computing*, pages 165–178, Oxford. Oxford University Press.

Melamed, I. D. (1998). Annotation style guide for the Blinker project, version 1.0.4. Technical Report 98-06, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

Menezes, A. and Richardson, S. D. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the ACL (ACL-2001)*, pages 39–46, Toulouse, France.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL (ACL-2000)*, pages 440–447, Hong Kong, China.

Simões, A. M. and Almeida, J. J. (2003). NATools – A statistical word aligner workbench. *Processamiento del Lenguaje Natural*, 31:217–224.

Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

Véronis, J. and Langlais, P. (2000). Evaluation of parallel text alignment systems: The ARCADE project. In Véronis, J., editor, *Parallel text processing: Alignment and use of translation corpora*, pages 369–388. Kluwer Academic Publishers.

Wu, H. and Wang, H. (2004). Improving domain-specific word alignment with a general bilingual corpus. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the AMTA (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 262–271. Springer-Verlag Berlin Heidelberg.