VisualLIHLA: the visual online tool for lexical alignment

Helena de M. Caseli
Federal University of São Carlos
Computer Science Department – NILC
Rod. Washington Luiz, km 235
13565-905 – São Carlos, SP – Brazil
helenacaseli@dc.ufscar.br

Felipe T. Gomes, Thiago A. S. Pardo,
Maria das Graças V. Nunes
University of São Paulo – ICMC – NILC
CP 668P – 13560-970
São Carlos, SP – Brazil
felipe.gomes@gmail.com,
{taspardo,gracan}@icmc.usp.br

ABSTRACT
This paper presents a freely available online lexical alignment tool based on the LIHLA lexical aligner. LIHLA aligns tokens, words and multiword units based on language-independent heuristics (cognates, position, etc.) and automatically built language-dependent resources (bilingual dictionaries). VisualLIHLA allows the online usage, visualization and download of the lexical alignments produced by LIHLA with 84–92% of precision and 76–91% of recall.

Categories and Subject Descriptors
H.5 [**Information Interfaces and Presentation**]: User Interfaces—*Natural Language*

General Terms
Design, Algorithms

Keywords
Lexical alignment, machine translation, Brazilian Portuguese

1. INTRODUCTION
Lexical alignment is an useful previous step for many natural language processing (NLP) applications, such as machine translation [1, 8], bilingual lexicography [6], and word sense disambiguation [4].

The alignment of two (or more) texts is the process of finding translation equivalences between segments (paragraphs, sentences, words, etc.) of one text (the source text) and segments of its translation (the target text). In this paper we are concerned with the lexical alignment, that is, the alignment between tokens (single words, characters, etc.) and sequences of tokens (e.g., multiword units).

Among the several automatic lexical aligners proposed in the literature, the statistical systems are considered to be the state of the art (e.g., [7] and [8]). Although they provide good results, the statistical aligners are not able to deal properly with the syntactic divergences between languages which bring about some frequent problems such as the non-consecutive phrasal information, long-range dependencies [1] and alignments involving multiword units.

In an attempt to handle some of these problems, many approaches were proposed recently (e.g., [10], [1] and [2, 3]). In this paper we describe a freely available visual online tool developed based on a hybrid lexical aligner: LIHLA [2, 3]. LIHLA tries to find the best alignments between tokens, words and multiword units based on language-independent heuristics and statistical alignments between single words defined in automatically built bilingual probabilistic dictionaries.

This paper is organized as follows: section 2 explains how the LIHLA works. Section 3 describes the visual online tool developed based on LIHLA and section 4 finishes this paper with some concluding remarks and proposals for future work.

2. THE LIHLA LEXICAL ALIGNER
The lexical alignment performed by LIHLA is briefly described in this section, for a more detailed description see [2, 3].

To perform the lexical alignment, LIHLA uses the statistical alignments between single words defined in two bilingual dictionaries (source–target and target–source). These dictionaries are automatically built from sentence-aligned parallel texts using NATools.[1] Given two sentence-aligned corpus files, NATools counts the co-occurrences of words in all aligned sentence pairs and builds a sparse matrix of word-to-word probabilities using an iterative expectation-maximization algorithm. Then, the two probabilistic bilingual dictionaries are composed by the elements with the highest probability values in the matrix [9].

Besides the word-by-word correspondences found in the bilingual dictionaries, LIHLA also uses four language-independent heuristics, in the following order:

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

1. **exact matches** – LIHLA prioritizes a target token which is identical to the source token being aligned. This heuristic is very useful, for example, in the alignment of proper names, numbers and special characters.

2. **cognates** – LIHLA tries to look for cognates for a source word using the longest common subsequence ratio (LCSR). The LCSR of two words is the length of their longest common subsequence by the length of the longest word. For example, the LCSR of the Portuguese word *alinhamento* and the Spanish 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*.

3. **best translation** – LIHLA chooses the best translation for a source word following one of two criteria (an input parameter): the target word with the highest probability according to the *bilingual dictionary* or the one at the best *position* regarding the source word position. By default, LIHLA uses the position criterion.

4. **multiword units** – After finding the best translation for a given source word based on the previous heuristics, LIHLA looks for multiword units for them. To find a source multiword unit LIHLA looks for the words occurring in the neighbourhood of the source word (in the source sentence) that: (1) are possible translations of the target word and (2) are not a possible translation of any other target word in the neighbourhood. A similar process is carried out to find a target multiword unit.

Thus, using the two bilingual dictionaries built by NATools and the heuristics described before, LIHLA tries to find the best lexical alignment in a pair of parallel sentences by means of an iterative process. In the experiments described in [2] and [3], LIHLA has achieved 84–92% of precision and 76–91% of recall for Portuguese–English and Portuguese–Spanish parallel texts, respectively.[2]

3. THE VISUALLIHLA TOOL

The VisualLIHLA[3] is a freely available tool developed to allow the online usage, visualization and download of the lexical alignments produced by LIHLA. As shown in Figure 1, to align a pair of texts in VisualLIHLA, the user has to: (1) enter the source and target texts, respectively, in the left and right text boxes; (2) choose the source and target languages and (3) press the "Align" button. The language selection is very important since this information determines which dictionaries LIHLA will use to align the source and target texts.[4]

To improve the alignment performance, the input parallel texts must be sentence aligned, that is, the source sentences and their translations (the target sentences) must occur in parallel lines. For example, the target sentence(s) at the first line in the right text box must be the translation of the source sentence(s) at the first line in the left text box.

During the alignment, a progress window displays the alignment types being produced: omissions (1 : 0, 0 : 1), one-to-one (1 : 1) and many-to-many (1 : 2, 2 : 1, 2 : 2, etc.). Once finished the lexical alignment, the progress window closes and the user is able to view the resulting alignments along with some statistics. To visualize the alignments, it is necessary just pass the mouse over the desired source or target token. By doing this, the selected token and the corresponding (aligned) token(s) in the other text will be highlighted in a different background color as shown in Figure 2 for the 2 : 1 alignment between the two Portuguese words *em conjunto* and the English word *jointly*.

In addition to the visualization of the lexical alignments in the parallel texts, it is possible to see how many alignments of each type were produced and also to select some of them by clicking on the corresponding colored box, as shown in Figure 3. Finally, the user can download the input and aligned (output) texts by clicking on "Save this alignment".

4. CONCLUDING REMARKS AND FUTURE WORK

In this paper we have presented the VisualLIHLA, a freely available online tool for lexical alignment. VisualLIHLA uses the LIHLA [2, 3] lexical aligner to find the best alignments between tokens, words and multiword units.

As future work, we intend to extend the VisualLIHLA to allow the manual edition of the automatic lexical alignments aiming at correcting misalignments or improving the aligner's performance. We also want to adopt the XML[5] output format and integrate the VisualLIHLA with the VisualTCA [5] —the online visualization tool for sentence alignment.[6]

5. ACKNOWLEDGMENTS

We thank the financial support of the Brazilian agencies FAPESP, CAPES and CNPq.

6. REFERENCES

[1] N. F. Ayan, B. J. Dorr, and N. Habash. Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In *Proceedings of AMTA 2004*, pages 17–26, 2004.

[2] H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. *Procesamiento del Lenguaje Natural*, (35):237–244, 2005.

[3] H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. LIHLA: A lexical aligner based on language-independent heuristics. In *Proceedings of ENIA 2005*, pages 641–650, São Leopoldo, RS, Brazil, 2005.

[2] Precision is the number of correct alignments divided by the number of proposed alignments while recall is the number of correct alignments divided by the number of alignments in the reference.

[3] `http://www.nilc.icmc.usp.br/nilc/tools/visuallihla/lihla.htm`.

[4] There are 4 language pairs allowed in the current version of VisualLIHLA: Portuguese-Spanish, Portuguese-English, Spanish-Portuguese and English-Portuguese.

[5] `www.w3.org/XML`.

[6] `http://www.nilc.icmc.usp.br/nilc/tools/pagina-visualtca/visualtca/tca.htm`.

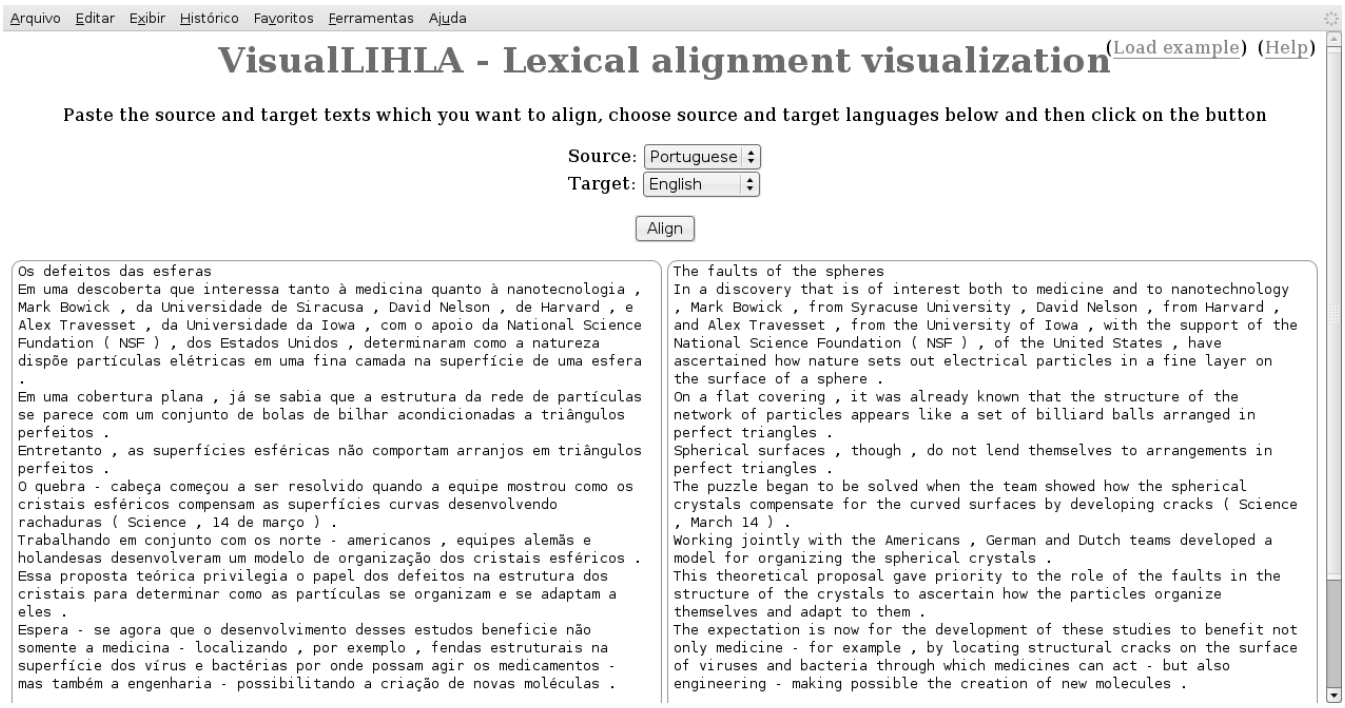

Figure 1: Screenshot of the VisualLIHLA alignment tool

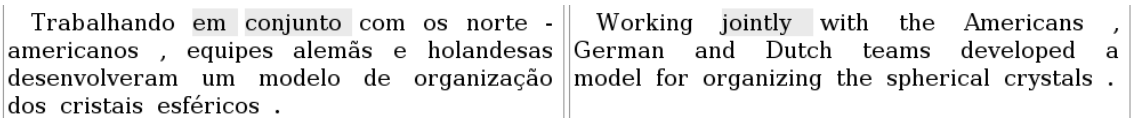

Figure 2: Highlighted alignment between *em conjunto* and *jointly*

Figure 3: Lexical alignment result

[4] W. A. Gale, K. W. Church, and D. Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of TMI 1992*, pages 101–112, Montreal, Canada, June 1992.

[5] F. T. Gomes, T. A. S. Pardo, and H. M. Caseli. VisualTCA: Uma Ferramenta Visual On-line para Alinhamento Sentencial de Textos Paralelos. In *Proceedings of TIL 2007*, pages 1729–1732, Rio de Janeiro-RJ, Brazil, July, 5-6 2007.

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

[7] D. Hiemstra. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proceedings of the 8th CLIN meeting*, pages 41–58, 1998.

[8] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of ACL 2000*, pages 440–447, Hong Kong, China, October 2000.

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

[10] H. Wu and H. Wang. Improving domain-specific word alignment with a general bilingual corpus. In *Proceedings of AMTA 2004*, pages 262–271, 2004.