LIHLA: Shared task system description

Helena M. Caseli, Maria G. V. Nunes
NILC – ICMC – Univ. São Paulo
CP 668P, 13560-970 São Carlos–SP, Brazil
{helename,gracan}@icmc.usp.br

Mikel L. Forcada
Transducens – DLSI – Univ. d'Alacant
E-03071 Alacant, Spain
mlf@dlsi.ua.es

Abstract

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 sentence-aligned parallel texts. The method has achieved an alignment error rate of 22.72% and 44.49% on English–Inuktitut and Romanian–English parallel sentences, 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). 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) starts from statistical alignments between single words (defined in bilingual lexicons) and applies 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.

This paper is organized as follows: section 2 explains how LIHLA works; section 3 describes some experiments carried out with LIHLA together with their results and, in section 4, some concluding remarks are presented.

2 How LIHLA works

As the first step, LIHLA uses alignments between single words defined in two bilingual lexicons (source–target and target–source) generated from sentence-aligned parallel texts using NATools.[1]

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

[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/.

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 111–114,
Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

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.

The construction of the bilingual lexicons is an independent prior step for the alignment performed by LIHLA and the same bilingual lexicons can be used several times to align parallel sentences.

So, using the two bilingual lexicons generated by NATools and some language-independent heuristics, LIHLA tries to find the best alignment between source and target tokens (words, numbers, special characters, etc.) in a pair of parallel sentences. For each source token s_j in source sentence S, LIHLA will look for the best token t_i in the target parallel sentence T applying these heuristics in sequence:

1. **Exact match**
   LIHLA creates a $1:1$ alignment between s_j and t_i if they are identical. This heuristic stays for exact matches, for instance, between proper names and numbers.

2. **Best candidate according to the bilingual lexicon**
   LIHLA looks for possible translations of s_j in the source–target bilingual lexicon (B_S) and makes an intersection between them and the words in T. In this intersection, if no candidate word identical to those in B_S is found, then LIHLA tries to look for cognates for those words using the longest common subsequence ratio (LCSR).[2] 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.
   Then, LIHLA selects the best target candidate word t_i for s_j —the best candidate word according to B_S among those in a position which is favorably situated in relation to s_j— and looks for multiword units involving s_j and t_i —those words that occur immediately before and/or after s_j (for source multiword units) or t_i (for target multiword units) and are not possible translations for other words in T and S, respectively. According to the multiword units that have (or not) been found, a $1:1$, $1:n$, $m:1$ or $m:n$ alignment is established. An omission alignment for s_j $(1:0)$ can also be established if no target candidate word t_i that satisfies this heuristic is available.

3. **Cognates**
   If no possible translation for s_j is found in the bilingual lexicon and the target sentence (T) at the same time, LIHLA uses the LCSR to look for cognates for s_j in T and sets a $1:1$ alignment between s_j and its best cognate or a $1:0$ alignment if there is no cognate available.

These heuristics are applied while alignments can still be produced and a maximum number of iterations is not reached (see section 3 for the number of iterations performed in the experiments described in this paper). Furthermore, at the first iteration, all words with a frequency higher than a set threshold are ignored to avoid erroneous alignments since all subsequent alignments are based on the previous ones.

In its last step (which is optional and has not been performed in the experiments described in this paper), LIHLA aligns the remaining unaligned source and target tokens between two pairs of already aligned tokens establishing several $1:1$ alignments when there are the same number of source and target tokens, or just one alignment involving all source and target tokens if they exist in different quantities. 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 $n:n$ one.

3 Experiments

In this section we present the experiments carried out with LIHLA for the "Shared task on word alignment" in the Workshop on Building and Using Parallel Texts during ACL2005. Systems participating in this shared task were provided with training data (consisting of sentence-aligned parallel texts) for three pairs of languages: English–Inuktitut,

[2] 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 Portuguese word *alinhamento* and 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*.

Romanian–English and English–Hindi. Furthermore, the systems would choose to participate in one or both subtasks of “limited resources” (where systems were allowed to use only the resources provided) and ”unlimited resources” (where systems were allowed to use any resources in addition to those provided). The system described in this paper, LIHLA, participated in the subtask of limited resources aligning English–Inuktitut and Romanian–English test sets.

The training sets —composed of 338,343 English–Inuktitut aligned sentences (omission cases were excluded from the whole set of 340,526 pairs) and 48,478 Romanian–English aligned ones— were used to build the bilingual lexicons. Then, without changing any default parameter (threshold for LCSR, maximum number of iterations, etc.), LIHLA aligned the 75 English–Inuktitut and the 203 Romanian–English parallel sentences on test sets. The whole alignment process (bilingual lexicon generation and alignment itself) did not take more than 17 minutes for English–Inuktitut (3 iterations per sentence, on average) and 7 minutes for Romanian–English (4 iterations per sentence, on average).

The evaluation was run with respect to precision, recall, F-measure, and alignment error rate (AER) considering sure and probable alignments but not NULL ones (Mihalcea and Pedersen, 2003). Tables 1 and 2 present metric values for English–Inuktitut and Romanian–English alignments, respectively, as provided by the organization of the shared task.

Metric	Sure	Probable
Precision	46.55%	79.53%
Recall	73.72%	18.71%
F-measure	57.07%	30.30%
AER	22.72%	

Table 1: LIHLA results for English–Inuktitut

Metric	Sure	Probable
Precision	57.68%	57.68%
Recall	53.51%	53.51%
F-measure	55.51%	55.51%
AER	44.49%	

Table 2: LIHLA results for Romanian–English

The results obtained in these experiments were not so good as those achieved by LIHLA on the language pairs for which it was developed, that is, 92.48% of precision and 88.32% of recall on Portuguese–Spanish parallel texts and 84.35% of precision and 76.39% of recall on Portuguese–English ones.[3]

The poor performance in the English–Inuktikut task may be partly due to the fact that Inuktikut is a polysynthetic language, that is, one in which, unlike in English, words are formed by long strings of concatenated morphemes. This makes it difficult for NATools to build reasonable dictionaries and lead to a predominance of $n : 1$ alignments, which are harder to determine —this fact can be confirmed by the better precision of LIHLA when probable alignments were considered (see table 1). The performance in the English–Romanian task, not very far from the English–Portuguese task used to tune up the parameters of the algorithm, is harder to explain without further analysis.

The difference in precision and recall between the two language pairs is due to the fact that on the English–Inuktitut reference corpus in addition to sure alignments the probable ones were also annotated while in Romanian–English only sure alignments are found. This indicates that evaluating alignment systems is not a simple task since their performance depends not only on the language pairs and the quality of parallel corpora (constant criteria in this shared task) but also the way the reference corpus is built.

So, at this moment, it would be unfair to blame the worse performance of LIHLA on its alignment methodology since it has been applied to the new language pairs without changing any of its default parameters. Maybe a simple optimization of parameters for each pair of languages could bring better results and also the impact of size and quality of training and reference corpora used in these experiments should be investigated. Then, the only conclusion that can be taken at this moment is that LIHLA, with its heuristics and/or default parameters, can not be indistinctly applied to any pair of languages.

Despite of its performance, LIHLA has some

[3] For more details of these experiments see (Caseli et al., accepted paper).

advantages when compared to other lexical alignment methods found in the literature, such as: 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 to handle texts (as in Gómez Guinovart and Sacau Fontenla (2004)). Furthermore, the whole alignment process (bilingual lexical generation and alignment itself) has proved to be very fast as mentioned previously.

4 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.

In the experiments carried out at the "Shared task on word alignment" which took place at the Workshop on Building and Using Parallel Texts during ACL2005, LIHLA has been evaluated on English–Inuktitut and Romanian–English parallel texts achieving an AER of 22.72% and 44.49%, respectively.

As future work, we aim at investigating the impact of using additional linguistic information (such as part-of-speech tags) on LIHLA's performance. Also, as a long-term goal, LIHLA will be part of a system implemented to learn transfer rules 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.

References

Necip F. Ayan, Bonnie J. Dorr, and Nizar Habash. 2004. Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In R. E. Frederking and K. B. Taylor, editors, *Proceedings of the 6th Conference of the AMTA (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 17–26. Springer-Verlag Berlin Heidelberg.

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

Helena M. Caseli, Maria das Graças V. Nunes, and Mikel L. Forcada. (accepted paper). LIHLA: A lexical aligner based on language-independent heuristics. In *Proceedings of the V Encontro Nacional de Inteligência Artificial (ENIA05)*, São Leopoldo, RS, Brazil.

William A. Gale, Kenneth W. Church, and David Yarowsky. 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, June.

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

Djoerd Hiemstra. 1998. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In Peter Arno Coppen, Hans van Halteren, and Lisanne Teunissen, editors, *Proceedings of the 8th CLIN meeting*, pages 41–58.

Arul Menezes and Stephen D. Richardson. 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.

Rada Mihalcea and Ted Pedersen. 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, May–June.

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

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

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

Hua Wu and Haifeng Wang. 2004. Improving domain-specific word alignment with a general bilingual corpus. In R. E. Frederking and K. B. Taylor, 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.