Combining Linguistic Data Views for Phrase-based SMT

Jesús Giménez and Lluís Màrquez
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

Abstract

We describe the Spanish-to-English \textit{LDV-COMBO} system for the Shared Task 2: “Exploiting Parallel Texts for Statistical Machine Translation” of the ACL-2005 Workshop on “Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond”. Our approach explores the possibility of working with alignments at different levels of abstraction, using different degrees of linguistic annotation. Several phrase-based translation models are built out from these alignments. Their combination significantly outperforms any of them in isolation. Moreover, we have built a word-based translation model based on WordNet which is used for unknown words.

1 Introduction

The main motivation behind our work is to introduce linguistic information, other than lexical units, to the process of building word and phrase alignments. Many other authors have tried to do so. See (Och and Ney, 2000), (Yamada and Knight, 2001), (Koehn and Knight, 2002), (Koehn et al., 2003), (Schafer and Yarowsky, 2003) and (Gildea, 2003).

Far from full syntactic complexity, we suggest to go back to the simpler alignment methods first described by (Brown et al., 1993). Our approach exploits the possibility of working with alignments at two different levels of granularity, lexical (words) and shallow parsing (chunks). In order to avoid confusion so forth we will talk about \textit{tokens} instead of \textit{words} as the minimal alignment unit.

Apart from redefining the scope of the alignment unit, we may use different degrees of linguistic annotation. We introduce the general concept of \textit{data view}, which is defined as any possible representation of the information contained in a bitext. We enrich data view tokens with features further than lexical such as \textit{PoS}, \textit{lemma}, and \textit{chunk label}.

As an example of the applicability of data views, suppose the case of the word ‘\textit{plays}’ being seen in the training data acting as a verb. Representing this information as ‘\textit{plays}_{VBZ}’ would allow us to distinguish it from its homograph ‘\textit{plays}_{NNS}’ for ‘\textit{plays}’ as a noun. Ideally, one would wish to have still deeper information, moving through syntax onto semantics, such as \textit{word senses}. Therefore, it would be possible to distinguish for instance between two realizations of ‘\textit{plays}’ with different meanings: ‘\textit{he}\_{PRP} \textit{plays}_{VBG} \textit{guitar}_{NN}’ and ‘\textit{he}\_{PRP} \textit{plays}_{VBG} \textit{basketball}_{NN}’.

Of course, there is a natural trade-off between the use of data views and data sparsity. Fortunately, we have data enough so that statistical parameter estimation remains reliable.

2 System Description

The \textit{LDV-COMBO} system follows the SMT architecture suggested by the workshop organizers.

First, training data are linguistically annotated for the two languages involved (See subsection 2.1). 10 different data views have been built. Notice that it is not necessary that the two parallel counterparts of a bitext share the same data view, as
long as they share the same granularity. However, in all our experiments we have annotated both sides with the same linguistic information. See token descriptions: (W) word, (WL) word and lemma, (WP) word and PoS, (WC) word and chunk label, (WPC) word, PoS and chunk label, (Cw) chunk of words, (Cwl) chunk of words and lemmas, (Cwp) chunk of words and PoS, (Cwpc) chunk of words and chunk labels. By chunk label we refer to the IOB label associated to every word inside a chunk, e.g., ‘\text{I}_\text{NP} \text{declare}_\text{VP} \text{resumed}_\text{VP} \text{the}_\text{NP} \text{session}_\text{NP} \text{of}_\text{PP} \text{the}_\text{NP} \text{European}_\text{NP} \text{Parliament}_\text{NP} \text{.O}’). We build chunk tokens by explicitly connecting words in the same chunk, e.g., ‘\text{(I)}_\text{NP} \text{declare}_\text{VP} \text{resumed}_\text{VP} \text{(the}_\text{PP} \text{of}_\text{PP} \text{the}_\text{NP} \text{European}_\text{NP} \text{Parliament}_\text{NP} \text{)}_\text{NP} \text{.O}’’. See examples of some of these data views in Table 1.

Then, running \textit{GIZA++}, we obtain token alignments for each of the data views. Combined phrase-based translation models are built on top of the Viterbi alignments output by \textit{GIZA++}. See details in subsection 2.2. \textit{Combo-models} must be then post-processed in order to remove the additional linguistic annotation and split chunks back into words, so they fit the format required by \textit{Pharaoh}.

Moreover, we have used the Multilingual Central Repository (MCR), a multilingual lexical-semantic database (Atserias et al., 2004), to build a word-based translation model. We back-off to this model in the case of unknown words, with the goal of improving system recall. See subsection 2.3.

2.1 Data Representation

In order to achieve robustness the same tools have been used to linguistically annotate both languages. The \textit{SVMTool} has been used for PoS-tagging (Giménez and Márquez, 2004). The \textit{Freeling} package (Carreras et al., 2004) has been used for lemmatizing. Finally, the \textit{Phreco} software by (Carreras et al., 2005) has been used for shallow parsing.

No additional tokenization or pre-processing steps other than case lowering have been performed. Special treatment of named entities, dates, numbers, currency, etc., should be considered so as to further enhance the system.

2.2 Building Combined Translation Models

Because data views capture different, possibly complementary, aspects of the translation process it seems reasonable to combine them. We consider two different ways of building such combo-models:

\textbf{LPHEX} Local phrase extraction. To build a separate phrase-based translation model for each data view alignment, and then combine them. There are two ways of combining translation models:

\textbf{MRG} Merging translation models. We work on a weighted linear interpolation of models. These weights may be tuned, although a uniform weight selection yields good results. Additionally, phrase-pairs may be filtered out by setting a score threshold.

\textbf{noMRG} Passing translation models directly to the \textit{Pharaoh} decoder. However, we encountered many problems with phrase-pairs that were not seen in all single models. This obliged us to apply arbitrary smoothing values to score these pairs.

\textbf{GPHEX} Global phrase extraction. To build a single phrase-based translation model from the union of alignments from several data views.

In its turn, any MRG operation performed on a combo-model results again in a valid combo-model.

In any case, phrase extraction\(^3\) is performed as depicted by (Och, 2002).

2.3 Using the MCR

Outer knowledge may be supplied to the \textit{Pharaoh} decoder by annotating the input with alternative translation options via XML-markup. We enrich every unknown word by looking up every possible translation for all of its senses in the MCR. These are scored by relative frequency according to the number of senses that lexicalized in the same manner. Let \(w_f, p_f\) be the source word and PoS, and \(w_e\) be the target word, we define a function

\[^{3}\text{We always work with the union of alignments, no heuristic refinement, and phrases up to 5 tokens. Phrase pairs appearing only once have been discarded. Scoring is performed by relative frequency. No smoothing is applied.}\]
It would appear that a speech made by Mr. Fischler positions of the development set.

Table 3: MT Results without tuning, for some data view combinations on the development set.

<table>
<thead>
<tr>
<th>System</th>
<th>NIST</th>
<th>BLEU</th>
<th>GTM-1</th>
<th>GTM-2</th>
<th>5C-MRG</th>
<th>5W-MRG</th>
<th>5W-GPHEX</th>
<th>5C-GPHEX</th>
<th>10-MRG</th>
</tr>
</thead>
</table>

Table 1: An example of 2 rich data views: (WPC) word, PoS and IOB chunk label (Cwpc) chunk of word, PoS and chunk label.

\[
\text{score}(w_f, p_f | w_e) = \frac{\text{Scount}(w_f, p_f, w_e)}{\sum_{(w_f, p_f)} \text{Scount}(w_f, p_f, w_e)}
\]

Better results would be expected working with word sense disambiguated text. We are not at this point yet. A first approach could be to work with the most frequent sense heuristic.

3 Experimental Results

3.1 Data and Evaluation Metrics

We have used the data sets and language model provided by the organization. No extra training or development data were used in our experiments.

We evaluate results with 3 different metrics: GTM F1-measure (\(e = 1, 2\)), BLEU score (\(n = 4\)) as provided by organizers, and NIST score (\(n = 5\)).

3.2 Experimenting with Data Views

Table 2 presents MT results for the 10 elementary data views devised in Section 2. Default parameters are used for \(\lambda_m\), \(\lambda_i\), and \(\lambda_w\). No tuning has been performed. As expected, word-based views obtain significatively higher results than chunk-based. All data views at the same level of granularity obtain comparable results.

In Table 3 MT results for different data view combinations are showed. Merged model weights are set equiprobable, and no phrase-pair score filtering is performed. We refer to the W model as our baseline. In this view, only words are used. The 5W-MRG and 5W-GPHEX models use a combination of the 5 word-based data views, as in MRG and GPHEX, respectively. The 5C-MRG and 5C-GPHEX system use a combination of the 5 chunk based data views, as in MRG and GPHEX, respectively. The 10-MRG system uses all 10 data views combined as in MRG. The 10-GPHEX/MRG system uses the 5 word based views combined as in GPHEX, the 5 chunk based views combined as in GPHEX, and then a combination of these two combo-models as in MRG.

<table>
<thead>
<tr>
<th>System</th>
<th>NIST</th>
<th>BLEU</th>
<th>GTM-1</th>
<th>GTM-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
<tr>
<td>3W-MRG</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
<tr>
<td>3W-GPHEX</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
<tr>
<td>5C-MRG</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
<tr>
<td>5C-GPHEX</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
<tr>
<td>10-MRG</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
<tr>
<td>10-GPHEX/MRG</td>
<td>6.6207</td>
<td>6.6403</td>
<td>7.1576</td>
<td>7.1496</td>
</tr>
</tbody>
</table>
It can be seen that results improve by combining several data views. Furthermore, global phrase extraction (GPHEX) seems to work much finer than local phrase extraction (LPHEX).

Table 4 shows MT results after optimizing $\lambda_{tm}$, $\lambda_{tm}$, $\lambda_w$, and the weights for the MRG operation, by means of the Downhill Simplex Method in Multidimensions (William H. Press and Flannery, 2002). Observe that tuning the system improves the performance considerably. The $\lambda_w$ parameter is particularly sensitive to tuning.

Even though the performance of chunk-based models is poor, the best results are obtained by combining the two levels of abstraction, thus proving that syntactically motivated phrases may help. 10-MRG and 10-GPHEX models achieve a similar performance. The 10-MRG-bestyN system corresponds to the 10-MRG model using WordNet. The 10-MRG-subyN system is this same system at the time of submission. Results using WordNet, taking into account that the number of unknown words in the development set was very small, are very promising.

<table>
<thead>
<tr>
<th>data view</th>
<th>GTM-1</th>
<th>GTM-2</th>
<th>BLEU</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0.6174</td>
<td>0.2583</td>
<td>28.13</td>
<td>7.1540</td>
</tr>
<tr>
<td>5W-MRG</td>
<td>0.6206</td>
<td>0.2605</td>
<td>28.50</td>
<td>7.2076</td>
</tr>
<tr>
<td>5W-GPHEX</td>
<td>0.6207</td>
<td>0.2603</td>
<td>28.38</td>
<td>7.1992</td>
</tr>
<tr>
<td>5C-MRG</td>
<td>0.5882</td>
<td>0.2426</td>
<td>25.06</td>
<td>6.6773</td>
</tr>
<tr>
<td>5C-GPHEX</td>
<td>0.5816</td>
<td>0.2387</td>
<td>24.40</td>
<td>6.5595</td>
</tr>
<tr>
<td>10-MRG</td>
<td>0.6218</td>
<td>0.2625</td>
<td>28.88</td>
<td>7.2491</td>
</tr>
<tr>
<td>10-GPHEX/MRG</td>
<td>0.6229</td>
<td>0.2622</td>
<td>28.82</td>
<td>7.2414</td>
</tr>
<tr>
<td>10-MRGwN</td>
<td>0.6228</td>
<td>0.2625</td>
<td>28.90</td>
<td>7.2583</td>
</tr>
<tr>
<td>10-MRG-subyN</td>
<td>0.6228</td>
<td>0.2622</td>
<td>28.79</td>
<td>7.2528</td>
</tr>
</tbody>
</table>

Table 4: MT Results for some data view combinations after tuning on the development set.

4 Conclusions

We have showed that it is possible to obtain better phrase-based translation models by utilizing alignments built on top of different linguistic data views. These models can be robustly combined, significantly outperforming all of their components in isolation. We leave for further work the experimentation of new data views such as word senses and semantic roles, as well as their natural porting and evolution from the alignment step to phrase extraction and decoding.

Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology (ALIADO TIC2002-04447-C02). Authors are thankful to Patrik Lambert for providing us with the implementation of the Simplex Method used for tuning.

References


