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 *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 significatively outperforms any of them in isolation. Moreover, we have built a word-based translation model based on Word-Net 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 *tokens* instead of *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 *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 *PoS*, *lemma*, and *chunk label*.

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

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

2 System Description

The *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 (Cwc) chunk of words and chunk labels (Cwpc) chunk of words, PoS and chunk labels. By chunk label we refer to the IOB label associated to every word inside a chunk, e.g. 'I_{B-NP} declare$_{B-VP}$ resumed$_{I-VP}$ the$_{B-NP}$ session$_{I-NP}$ of$_{B-PP}$ the$_{B-NP}$ European$_{I-NP}$ Parliament$_{I-NP}$.$_O$'). We build chunk tokens by explicitly connecting words in the same chunk, e.g. '$(I)_{NP}$ (declare_resumed)$_{VP}$ (the_session)$_{NP}$ (of)$_{PP}$ (the_European_Parliament)$_{NP}$'. See examples of some of these data views in Table 1.

Then, running *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 *GIZA++*. See details in subsection 2.2. *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 *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 *SVMTool*[1] has been used for PoS-tagging (Giménez and Màrquez, 2004). The *Freeling*[2] package (Carreras et al., 2004) has been used for lemmatizing. Finally, the *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:

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:

> **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.

> **noMRG** Passing translation models directly to the 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.

GPHEX Global phrase extraction. To build a single phrased-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 *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

---

WPC	It$_{[PRP:B-NP]}$ would$_{[MD:B-VP]}$ appear$_{[VB:I-VP]}$ that$_{[IN:B-SBAR]}$ a$_{[DT:B-NP]}$ speech$_{[NN:I-NP]}$ made$_{[VBN:B-VP]}$ at$_{[IN:B-PP]}$ the$_{[DT:B-NP]}$ weekend$_{[NN:I-NP]}$ by$_{[IN:B-PP]}$ Mr$_{[NNP:B-NP]}$ Fischler$_{[NNP:I-NP]}$ indicates$_{[VBZ:B-VP]}$ a$_{[DT:B-NP]}$ change$_{[NN:I-NP]}$ of$_{[IN:B-PP]}$ his$_{[PRP\$:B-NP]}$ position$_{[NN:I-NP]}$ ·$_{[.:O]}$ Fischler$_{[VMN:B-VP]}$ pronunció$_{[VMI:B-VP]}$ un$_{[DI:B-NP]}$ discurso$_{[NC:I-NP]}$ este$_{[DD:B-NP]}$ fin$_{[NC:I-NP]}$ de$_{[SP:B-PP]}$ semana$_{[NC:B-NP]}$ en$_{[SP:B-PP]}$ el$_{[DA:B-SBAR]}$ que$_{[PR0:I-SBAR]}$ parecía$_{[VMI:B-VP]}$ haber$_{[VAN:I-VP]}$ cambiado$_{[VMP:I-VP]}$ de$_{[SP:B-PP]}$ actitud$_{[NC:B-NP]}$ ·$_{[Fp:O]}$
Cwpc	(It$_{[PRP:B-NP]}$) (would$_{[MD:B-VP]]}$ appear$_{[VB:I-VP]}$) (that$_{[IN:B-SBAR]}$) (a$_{[DT:B-NP]}$ speech$_{[NN:I-NP]}$) (made$_{[VBN:B-VP]}$) (at$_{[IN:B-PP]}$) (the$_{[DT:B-NP]}$ weekend$_{[NN:I-NP]}$) (by$_{[IN:B-PP]}$) (Mr$_{[NNP:B-NP]}$ Fischler$_{[NNP:I-NP]}$) (indicates$_{[VBZ:B-VP]}$) (a$_{[DT:B-NP]}$ change$_{[NN:I-NP]}$) (of$_{[IN:B-PP]}$) (his$_{[PRP\$:B-NP]}$ position$_{[NN:I-NP]}$) (·$_{[.:O]}$) (Fischler$_{[VMN:B-VP]}$) (pronunció$_{[VMI:B-VP]}$) (un$_{[DI:B-NP]}$ discurso$_{[NC:I-NP]}$) (este$_{[DD:B-NP]}$ fin$_{[NC:I-NP]}$) (de$_{[SP:B-PP]}$) (semana$_{[NC:B-NP]}$) (en$_{[SP:B-PP]}$) (el$_{[DA:B-SBAR]}$ que$_{[PR0:I-SBAR]}$) (parecía$_{[VMI:B-VP]}$ haber$_{[VAN:I-VP]}$ cambiado$_{[VMP:I-VP]}$) (de$_{[SP:B-PP]}$) (actitud$_{[NC:B-NP]}$) (·$_{[Fp:O]}$)

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

$Scount(w_f, p_f, w_e)$ which counts the number of senses for (w_f, p_f) which can lexicalize as w_e. A translation pair is scored as:

$$score(w_f, p_f | w_e) = \frac{Scount(w_f, p_f, w_e)}{\sum_{(w_f, p_f)} Scount(w_f, p_f, w_e)} \quad (1)$$

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 F$_1$-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 λ_{tm}, λ_{lm}, and λ_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

data view	GTM-1	GTM-2	BLEU	NIST
W	0.6108	0.2609	25.92	7.1576
WL	0.6110	0.2601	25.77	7.1496
WP	0.6096	0.2600	25.74	7.1415
WC	0.6124	0.2600	25.98	7.1852
WPC	0.6107	0.2587	25.79	7.1595
Cw	0.5749	0.2384	22.73	6.6149
Cwl	0.5756	0.2385	22.73	6.6204
Cwp	0.5771	0.2395	23.06	6.6403
Cwc	0.5759	0.2390	22.86	6.6207
Cwpc	0.5744	0.2379	22.77	6.5949

Table 2: MT Results for the 10 elementary data views on the development set.

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.

data view	GTM-1	GTM-2	BLEU	NIST
W	0.6108	0.2609	25.92	7.1576
5W-MRG	0.6134	0.2631	26.25	7.2122
5W-GPHEX	0.6172	0.2615	26.95	7.2823
5C-MRG	0.5786	0.2407	23.18	6.6754
5C-GPHEX	0.5739	0.2368	22.80	6.5714
10-MRG	0.6130	0.2624	26.24	7.2196
10-GPHEX/MRG	0.6142	0.2600	26.58	7.2542

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

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 λ_{tm}, λ_{lm}, λ_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 λ_w parameter is particularly sensitive to tuning.

Even though the performance of chunk-based models is poor, the best results are obtained by combinining 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-best$_{WN}$* system corresponds to the 10-MRG model using WordNet. The *10-MRG-sub$_{WN}$* system is this same system at the time of submission. Results using WordNet, taking into account that the number of unknown[4] words in the development set was very small, are very promising.

data view	GTM-1	GTM-2	BLEU	NIST
W	0.6174	0.2583	28.13	7.1540
5W-MRG	0.6206	0.2605	28.50	7.2076
5W-GPHEX	0.6207	0.2603	28.38	7.1992
5C-MRG	0.5882	0.2426	25.06	6.6773
5C-GPHEX	0.5816	0.2387	24.40	6.5595
10-MRG	0.6218	0.2623	28.88	7.2491
10-GPHEX/MRG	0.6229	0.2622	28.82	7.2414
10-MRG$_{WN}$	0.6228	0.2625	28.90	7.2583
10-MRG-sub$_{WN}$	0.6228	0.2622	28.79	7.2528

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.

---

[4]Translation for 349 unknown words was found in the MCR.

References

Jordi Atserias, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *Proceedings of GWC*, Brno, Czech Republic, January. ISBN 80-210-3302-9.

Peter E Brown, Stephen A. Della Pietra, Robert L. Mercer, and Vincent J. Della Pietra. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th LREC*.

Xavier Carreras, Lluís Màrquez, and Jorge Castro. 2005. Filtering-ranking perceptron learning for partial parsing. *Machine Learning*, 59:1–31.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of ACL*.

Jesús Giménez and Lluís Màrquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *Proceedings of 4th LREC*.

Philipp Koehn and Kevin Knight. 2002. Chunkmt: Statistical machine translation with richer linguistic knowledge. Draft.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL*.

Franz Josef Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Germany.

Charles Schafer and David Yarowsky. 2003. Statistical machine translation using coercive two-level syntactic transduction. In *Proceedings of EMNLP*.

William T. Vetterling William H. Press, Saul A. Teukolsky and Brian P. Flannery. 2002. *Numerical Recipes in C++: the Art of Scientific Computing*. Cambridge University Press.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL*.