The LDV-COMBO system for 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 LDV-COMBO system presented at the Shared Task. Our approach explores the possibility of working with alignments at different levels of abstraction using different degrees of linguistic analysis from the lexical to the shallow syntactic level. Translation models are built on top of combinations of these alignments. We present results for the Spanish-to-English and English-to-Spanish tasks. We show that liniguistic information may be helpful, specially when the target language has a rich morphology.

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. In the last years, many efforts have been devoted to this matter (Yamada and Knight, 2001; Gildea, 2003).

Following our previous work (Giménez and Màrquez, 2005), we use shallow syntactic information to generate more precise alignments. Far from full syntactic complexity, we suggest going back to the simpler alignment methods first described by IBM (1993). Our approach exploits the possibility of working with alignments at two different levels of granularity, lexical (words) and shallow parsing (chunks). Apart from redefining the scope of the alignment unit, we may use different linguistic data views. We enrich tokens with features further than lexical such as *part-of-speech (PoS)*, *lemma*, and *chunk IOB label*.

For instance, suppose the case illustrated in Figure 1 where the lexical item *'plays'* is seen acting as a verb and as a noun. Considering these two words, with the same lexical realization, as a single token adds noise to the word alignment process. Representing this information, by means of linguistic data views, as *'plays*$_{VBZ}$*'* and *'plays*$_{NNS}$*'* would allow us to distinguish between the two cases. 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*$_{VBG}$ *football*$_{NN}$*'*. Of course, there is a natural trade-off between the use of linguistic data views and data sparsity. Fortunately, we hava data enough so that statistical parameter estimation remains reliable.

The approach which is closest to ours is that by Schafer and Yarowsky (2003) who suggested a combination of models based on shallow syntactic analysis (part-of-speech tagging and phrase chunking). They followed a backoff strategy in the application of their models. Decoding was based on Finite State Automata. Although no significant improvement in MT quality was reported, results were promising taking into account the short time spent in the development of the linguistic tools utilized.

Our system is further described in Section 2. Results are reported in Section 3. Conclusions and further work are briefly outlined in Section 4.

Proceedings of the Workshop on Statistical Machine Translation, pages 166–169,
New York City, June 2006. ©2006 Association for Computational Linguistics

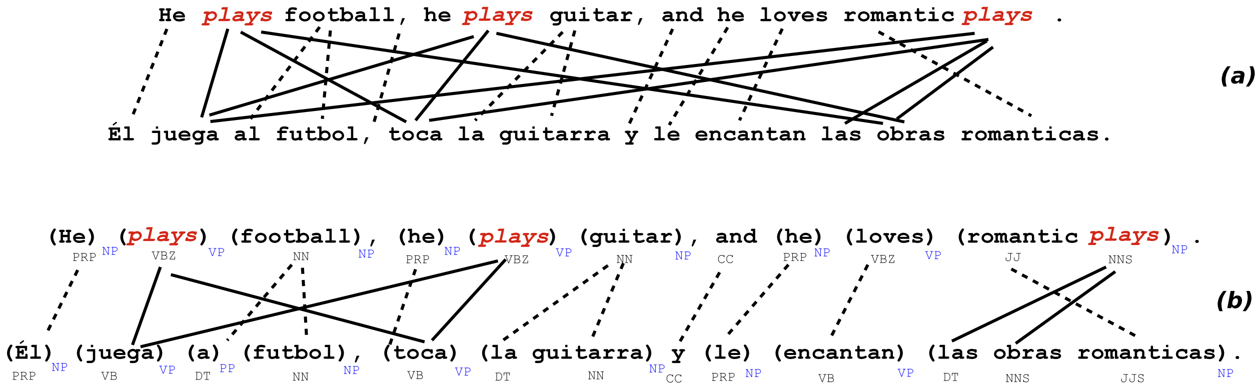
He plays football, he plays guitar, and he loves romantic plays .
Él juega al futbol, toca la guitarra y le encantan las obras romanticas.
(a)
(He) PRP NP (plays) VBZ VP (football) NN NP , (he) PRP NP (plays) VBZ VP (guitar) NN NP , and CC (he) PRP NP (loves) VBZ VP (romantic JJ plays NNS) NP .
(Él) PRP NP (juega) VB VP (a) DT PP (futbol) NN NP , (toca) VB VP (la DT guitarra NN) NP y CC (le) PRP NP (encantan) VB VP (las DT obras NNS romanticas JJS) NP .
(b)

Figure 1: A case of word alignment possibilities on top of lexical units (a) and linguistic data views (b).

2 System Description

The LDV-COMBO system follows the SMT architecture suggested by the workshop organizers. We use the *Pharaoh* beam-search decoder (Koehn, 2004).

First, training data are linguistically annotated. 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 (Carreras et al., 2005) has been used for shallow parsing. In this paper we focus on data views at the word level. 6 different data views have been built: (W) word, (L) lemma, (WP) word and PoS, (WC) word and chunk IOB label, (WPC) word, PoS and chunk IOB label, (LC) lemma and chunk IOB label.

Then, running *GIZA++* (Och and Ney, 2003), 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++*. Phrase extraction is performed following the phrase-extract algorithm depicted by Och (2002). We do not apply any heuristic refinement. We work with phrases up to 5 tokens. Phrase pairs appearing only once have been discarded. Scoring is performed by relative frequency. No smoothing is applied.

In this paper we focus on the global phrase extraction (GPHEX) method described by Giménez and Màrquez (2005). We build a single translation model from the union of alignments from the 6 data views described above. This model must match the input format. For instance, if the input is annotated with word and PoS (WP), so must be the translation model. Therefore either the input must be enriched with linguistic annotation or translation models must be post-processed in order to remove the additional linguistic annotation. We did not observe significant differences in either alternative. Therefore, we simply adapted translations models to work under the assumption of unannotated inputs (W).

3 Experimental Work

3.1 Setting

We have used only the data sets and language model provided by the organization. For evaluation we have selected a set of 8 metric variants corresponding to seven different families: BLEU ($n = 4$) (Papineni et al., 2001), NIST ($n = 5$) (Lin and Hovy, 2002), GTM F$_1$-measure ($e = 1, 2$) (Melamed et al., 2003), 1-WER (Nießen et al., 2000), 1-PER (Leusch et al., 2003), ROUGE (ROUGE-S*) (Lin and Och, 2004) and METEOR[3] (Banerjee and Lavie, 2005). Optimization of the decoding parameters (λ_{tm}, λ_{lm}, λ_w) is performed by means of the *Downhill Simplex Method in Multidimensions* (William H. Press and Flannery, 2002) over the BLEU metric.

[1] The SVMTool may be freely downloaded at http://www.lsi.upc.es/~nlp/SVMTool/.

[2] Freeling Suite of Language Analyzers may be downloaded at http://www.lsi.upc.es/~nlp/freeling/

[3] For Spanish-to-English we applied all available modules: exact + stemming + WordNet stemming + WordNet synonymy lookup. However, for English-to-Spanish we were forced to use the exact module alone.

Spanish-to-English

System	1-PER	1-WER	BLEU-4	GTM-1	GTM-2	METEOR	NIST-5	ROUGE-S*
Baseline	**0.5514**	**0.3741**	0.2709	0.6159	0.2579	0.5836	**7.2958**	0.3643
LDV-COMBO	0.5478	0.3657	0.2708	**0.6202**	0.2585	**0.5928**	7.2433	**0.3671**

English-to-Spanish

System	1-PER	1-WER	BLEU-4	GTM-1	GTM-2	METEOR	NIST-5	ROUGE-S*
Baseline	0.5158	**0.3776**	0.2272	0.5673	0.2418	0.4954	6.6835	0.3028
LDV-COMBO	**0.5382**	0.3560	**0.2611**	**0.5910**	**0.2462**	**0.5400**	**7.1054**	**0.3240**

Table 1: MT results comparing the LDV-COMBO system to a baseline system, for the test set both on the Spanish-to-English and English-to-Spanish tasks.

English Reference: *consider* germany , where some leaders [...]
Spanish Reference: **pensemos** en alemania , donde algunos dirigentes [...]

English-to-Spanish	**Baseline**	*estiman* que alemania , donde algunos dirigentes [...]
	LDV-COMBO	**pensemos** en alemania , donde algunos dirigentes [...]

Table 2: A case of error analysis.

3.2 Results

Table 1 presents MT results for the test set both for the Spanish-to-English and English-to-Spanish tasks. The variant of the LDV-COMBO system described in Section 2 is compared to a baseline variant based only on lexical items. In the case of Spanish-to-English performance varies from metric to metric. Therefore, an open issue is which metric should be trusted. In any case, the differences are minor. However, in the case of English-to-Spanish all metrics but '1-WER' agree to indicate that the LDV-COMBO system significantly outperforms the baseline. We suspect this may be due to the richer morphology of Spanish. In order to test this hypothesis we performed an error analysys at the sentence level based on the GTM F-measure. We found many cases where the LDV-COMBO system outperforms the baseline system by choosing a more accurate translation. For instance, in Table 2 we may see a fragment of the case of sentence 2176 in the test set. A better translation for "consider" is provided, "pensemos", which corresponds to the right verb and verbal form (instead of "estiman"). By inspecting translation models we confirmed the better adjustment of probabilities.

Interestingly, LDV-COMBO translation models are between 30% and 40% smaller than the models based on lexical items alone. The reason is that we are working with the union of alignments from different data views, thus adding more constraints into the phrase extraction step. Fewer phrase pairs are extracted, and as a consequence we are also effectively eliminating noise from translation models.

4 Conclusions and Further Work

Many researchers remain sceptical about the usefulness of linguistic information in SMT, because, except in a couple of cases (Charniak et al., 2003; Collins et al., 2005), little success has been reported. In this work we have shown that liniguistic information may be helpful, specially when the target language has a rich morphology (e.g. Spanish).

Moreover, it has often been argued that linguistic information does not yield significant improvements in MT quality, because (i) linguistic processors introduce many errors and (ii) the BLEU score is not specially sensitive to the grammaticality of MT output. We have minimized the impact of the first argument by using highly accurate tools for both languages. In order to solve the second problem more sophisticated metrics are required. Current MT evaluation metrics fail to capture many aspects of MT quality that characterize human translations with respect to those produced by MT systems. We are devoting most of our efforts to the deployment of a new MT evaluation framework which allows to combine several similarity metrics into a single measure of quality (Giménez and Amigó, 2006).

We also leave for further work the experimentation of new data views such as word senses and semantic roles, as well as their natural porting 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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

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 Márquez, and Jorge Castro. 2005. Filtering-Ranking Perceptron Learning for Partial Parsing. *Machine Learning*, 59:1–31.

Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based Language Models for Machine Translation. In *Proceedings of MT SUMMIT IX*.

Michael Collins, Philipp Koehn, and Ivona Kucerová. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of ACL*.

Daniel Gildea. 2003. Loosely Tree-Based Alignment for Machine Translation. In *Proceedings of ACL*.

Jesús Giménez and Enrique Amigó. 2006. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th LREC*.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of 4th LREC*.

Jesús Giménez and Lluís Màrquez. 2005. Combining Linguistic Data Views for Phrase-based SMT. In *Proceedings of the Workshop on Building and Using Parallel Texts, ACL*.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA*.

G. Leusch, N. Ueffing, and H. Ney. 2003. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In *Proceedings of MT Summit IX*.

Chin-Yew Lin and E.H. Hovy. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, National Institute of Standards and Technology.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of ACL*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT/NAACL*.

S. Nießen, F.J. Och, G. Leusch, and H. Ney. 2000. Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, rc22176. Technical report, IBM T.J. Watson Research Center.

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*.