11 Conclusions

We propose the entropy-guided structure learning framework that extends the general SL framework by integrating an automatic feature generation approach – the entropy-guided feature generation method that is based on the conditional entropy of local decision variables given input basic features.

We compare EFG with two important alternative feature generation methods, namely manual template generation and polynomial kernel functions. Our empirical results on dependency parsing show that EFG outperforms the best known template set on the Portuguese CoNLL-2006 dataset. We compare EFG with polynomial kernel methods on two text chunking datasets: the Portuguese Bosque dataset and the English CoNLL-2000 dataset. EFG outperforms both. Furthermore, our method presents additional advantages over these two alternative methods. It is faster than kernel methods and avoid the overfitting issue. Compared to manual feature templates, the fact that EFG bypasses domain experts is highly valuable.

We evaluate ESL on nine datasets involving five natural language processing tasks and four different languages. ESL presents state-of-the-art comparable performances on all evaluated datasets. Moreover, it outperforms the previous best performing systems on six datasets, namely the Mac-Morpho dataset for Portuguese part-of-speech tagging, the Bosque dataset for Portuguese text chunking, the GloboQuotes dataset for Portuguese quotation extraction, the CoNLL-2012 Shared Task datasets for Arabic, Chinese, and multilingual coreference resolution. Additionally, on the Portuguese dependency parsing task, we demonstrate the power of ESL by automatically including two basic features in our model, lifting the final performance by around 2.4 points.

We propose a novel modeling for coreference resolution based on latent structure learning. The ESL systems based on this modeling achieve the best results for Arabic and English coreference resolution. Moreover, the ESL coreference systems – for Arabic, Chinese and English – achieve the very first place on the renowned CoNLL-2012 Shared Task. The proposed modeling is highly language-independent, allowing us to apply it on three very different languages with no more than minor adaptations, which are mainly necessary due to lack of features for some languages.

Koo et al. (2010) and McDonald and Pereira (2006) show significant improvements on the performance of DP systems by extending the first-order model used in this work to include second- and third-order features. As future work, we plan to apply these modelings to dependency parsing and coreference resolution.

For coreference resolution, some authors (Lee et al., 2011) report that features based on partial clusters bring substantial improvements on performance. We also plan to extend our latent modeling in order to include such type of features.

Text chunking and named entity recognition have been recurrently recast as sequence labeling problems. Nevertheless, these tasks require sentence segmentation and, additionally, segment classification. In this way, we can apply our modeling based on weighted interval scheduling to such tasks, just as we have done for quotation extraction. By using a sequence segmentation modeling, we are able to use more meaningful features for these tasks and, consequently, improve performance.