This paper presents a novel approach for inducing unsupervised part-of-speech (POS) taggers for languages that lack labeled training data but have translated text in a resource-rich language. The approach uses graph-based label propagation for cross-lingual knowledge transfer and incorporates the projected labels as features in an unsupervised model. The authors evaluate their approach on eight European languages and show significant improvements over state-of-the-art baselines. The proposed method is applicable to a wide array of resource-poor languages and does not assume any knowledge about the target language.

The authors propose a method for building part-of-speech (POS) taggers for foreign languages using an English POS tagger and parallel text between the two languages. The approach involves constructing a bilingual similarity graph from the parallel corpus, where the vertices on the foreign language side correspond to trigram types and the vertices on the English side correspond to individual word types. The graph is built using similarity functions based on co-occurrence and unsupervised word alignment statistics. The authors initialize the graph by tagging the English side of the parallel text using a supervised model, and then propagate the POS labels from the English side to the foreign side using label propagation. The POS distributions over the foreign trigram types are then used as features to learn a better unsupervised POS tagger. The authors evaluate their approach on eight European languages and show that it provides significant improvements in POS tagging accuracy compared to previous methods. The final average POS tagging accuracy of 83.4% compares favorably to the average accuracy of 73.0% for a monolingual unsupervised model and bridges the gap to fully supervised POS tagging performance (96.6%).

In this section, the authors discuss the graph vertices used in their bilingual setup. They explain that the information flow in their graph is asymmetric, from English to the foreign language, so they use different types of vertices for each language. The foreign language vertices correspond to foreign trigram types, while the English vertices correspond to word types. The authors also mention that since all English vertices are labeled, they do not need to disambiguate them by embedding them in trigrams. Additionally, they only connect the English vertices to the foreign language vertices, not to each other. The graph vertices are extracted from a parallel corpus and an unlabeled monolingual foreign corpus. The authors use two different similarity functions to define the edge weights among the foreign vertices and between vertices from different languages. They briefly review their monolingual similarity function, which is the same as the one used by Subramanya et al. (2010), and describe how they compute the similarity between trigram types based on co-occurrence statistics of different feature concepts. They also define a bilingual similarity function based on high-confidence word alignments between the English and foreign language sentences in the parallel corpus.

. In this stage, the tag distributions from the periphery are propagated to the rest of the graph using the similarity metric between the foreign language vertices. The label distribution for each foreign language vertex is updated based on the label distributions of its neighbors. This two-stage label propagation process allows us to generate soft POS labels for all the vertices in the graph, including those that are not directly connected to any English vertices. The resulting label distributions can be used for various downstream tasks, such as POS tagging or syntactic parsing, in the foreign language.

The objective function in the graph is optimized to minimize the cost function C(q). The objective function is defined as: C(q) = Î£(uiâˆˆVf\Vl) [Î£(f,ujâˆˆN(ui)) wij / |qi - qj|^2 + Î½ Î£(uiâˆˆVf\Vl) f / |qi - U|^2] The optimization is subject to the following constraints: - Î£y qi(y) = 1 for all ui - qi(y) >= 0 for all ui and y - qi = ri for all ui âˆˆ Vl The label distributions qi (i=1,...,|Vf|) represent the probabilities of different labels for the foreign language vertices. The hyperparameters Âµ and Î½ are used to tune the importance of the different terms in the objective function. The first term in the objective function is a graph smoothness regularizer that penalizes neighboring vertices with different label distributions. The second term is a regularizer that encourages the label distributions to be uniform. To solve this convex objective function, an iterative update method is used. The update equation for qi(y) is given by: qi(y) = ri(y) if ui âˆˆ Vlf Î³i(y) / Îºi otherwise where Î³i(y) and Îºi are defined as: Î³i(y) = Î£(ujâˆˆN(ui)) wij q(m-1)j(y) + Î½ U(y) Îºi = Î½ + Î£ (ujâˆˆN(ui)) wij This update is performed for 10 iterations. After running label propagation, tag probabilities for foreign word types are computed by marginalizing the POS tag distributions of foreign trigrams over the left and right context words. A set of possible tags is then extracted by eliminating labels with probabilities below a threshold value. The POS induction model is based on a feature-based HMM. The emission distribution in the HMM is replaced with a log-linear model that incorporates various features of the observation. The model uses a locally normalized log-linear model to calculate the probability of an observation given a state. Overall, the objective of the optimization is to find the label distributions that minimize the cost function, taking into account the graph structure and the desired uniformity of label distributions.

The passage describes a model used for POS tagging in natural language processing. The model checks features of a word, such as whether it contains digits or hyphens, whether the first letter is uppercase, and suffix features up to length 3. These features are conjoined with a state variable. The model is trained by optimizing an objective function that involves marginalizing all possible state configurations for a sentence. The L-BFGS algorithm is used for optimization. The model outperformed other methods in English POS tagging by 12%. The model is then adapted to incorporate a constraint feature that uses information from a smoothed graph and prunes hidden states inconsistent with a thresholded vector. The feature is equivalent to using a tagging dictionary extracted from the graph. The model is tested on monolingual treebanks and large amounts of parallel text in eight Indo-European

languages. The universal POS tagset is used, consisting of 12 coarse-grained tags. The same hyperparameters are used for all language pairs.

The authors used a tagset consisting of 12 categories, including nouns, verbs, adjectives, adverbs, pronouns, determiners, prepositions, conjunctions, interjections, punctuation marks, and a catch-all category for abbreviations or foreign words. They also used a mapping from the language-specific POS tags in the treebank to universal POS tags. They evaluated their approach using three baselines (EM-HMM, Feature-HMM, Projection) and two oracles (TB Dictionary, Supervised). They set the number of latent HMM states for each language to the number of fine tags in the language's treebank and used the same hyperparameters for all languages.

The table shows the part-of-speech tagging accuracies for various baselines and oracles, as well as our approach, for eight different languages. The "Avg" column represents the macro-average across the eight languages. The baselines include EM-HMM, Feature-HMM, and Projection models. Our approach includes two variations: "No LP" and "With LP", which refer to our model without and with label propagation, respectively. The oracles include TB Dictionary and Supervised models. The results show that the vanilla HMM trained with EM performs the worst. The feature-HMM model works better for all languages. The Projection baseline improves upon the monolingual baselines but falls short of our "No LP" model. The "No LP" model performs better for six out of eight languages and gives an average improvement of 8.3% over the unsupervised feature-HMM model. Our full model, "With LP", outperforms all other models, including the baselines and the "No LP" setting, for all languages except German. It performs 10.4% better than the feature-HMM baseline and 4.6% better than direct projection on average. The increase in vocabulary sizes for all languages confirms that our full model has better vocabulary coverage and allows the extraction of a larger set of constraint features. An example from the Italian test set is shown in Figure 2, where our best model achieves the most accurate tagging compared to the other models.