Supplementary Material

A Details of Deterministic Annealing

In practice, deterministic annealing (§6.3) is implemented in a way that dynamically increases the number of clusters k (Friedman et al., 2001), leading to a hierarchical clustering. First, we initialize with one cluster, and all the word tokens are mapped to that cluster with probability 1. Second, for each cluster i, duplicate the cluster C_i to form C_{ia}, C_{ib}, and divide the probabilities associated with C_i approximately evenly (with perturbation) between the two clusters, i.e., set $p(c_{ia}|x) = \frac{1}{2}p(c_i|x) + \epsilon_x$ and $p(c_{ib}|x) = \frac{1}{2}p(c_i|x) - \epsilon_x$. Third, update $\beta \leftarrow \beta/\alpha$, and run optimization until convergence. Fourth, for each former cluster i, if C_{ia} and C_{ib} have not differentiated from each other, remerge them by setting $p(c_i|x) = p(c_{ia}|x)+p(c_{ib}|x)$. (Optimization will have pulled them together again for higher β values and pushed them apart for lower β values.) Our heuristic is to re-merge them if for all word tokens x, $|p(c_{ia}|x) - p(c_{ib}|x)| \leq 0.01$. Finally, loop back to the second step, unless the β value has fallen below a given threshold β_{min} or we have reached a desired maximum number of clusters.

B Additional Tradeoff Curves

Figure 5 supplements the tradeoff curves in Figure 2 by plotting the relationship between $\mathrm{I}(T_i; X \mid \hat{X}_i)$ vs. $\mathrm{I}(Y; T)$, and $\mathrm{I}(T_i; X \mid \hat{X}_i)$ vs. LAS. Moving leftward on the graphs, each T_i contains less *contextual* information about word i (because γ in equation (2) is larger) as well as less information overall about word i (because we always set $\gamma = \beta$, so β is larger as well). The graphs show that the tag sequence T then becomes less informative about the parse Y.

C Additional t-SNE plots

Recall that Figure 3 (in §6.2) was a row of t-SNE visualizations of the continuous *token* embeddings $p_\theta(t_i \mid x_i)$ under no compression, moderate compression, and too much compression. Figure 6 gives another row visualizing the continuous *type* embeddings $s_\xi(t_i \mid \hat{x}_i)$ in the same way. In both cases, the “moderate compression” condition shows $\beta = 0.01$.

Figure 6 also shows rows for the *discrete* type and token embeddings. In both cases, the “moderate compression” condition shows $\beta = 0.001$.

In the continuous case, each point given to t-SNE is the mean of a Gaussian-distributed stochastic embedding, so it is in $\mathbb{R}^d$. In the discrete case, each point given to t-SNE is a vector of k tag probabilities, so it is in $\mathbb{R}^k$ and more specifically in the $(k-1)$-dimensional simplex. The t-SNE visualizer plots these points in 2 dimensions.

The message of all these graphs is that the tokens or types with the same gold part of speech (shown as having the same color) are most nicely grouped together in the moderate compression condition.

D Syntactic Feature Classification

Figure 7 shows results for the **Syntactic Features** paragraph in §6.2, by showing the prediction accuracy of subcategorization frame, tense, and number from t_i as a function of the level of compression. We used an SVM classifier with a radial basis function kernel.

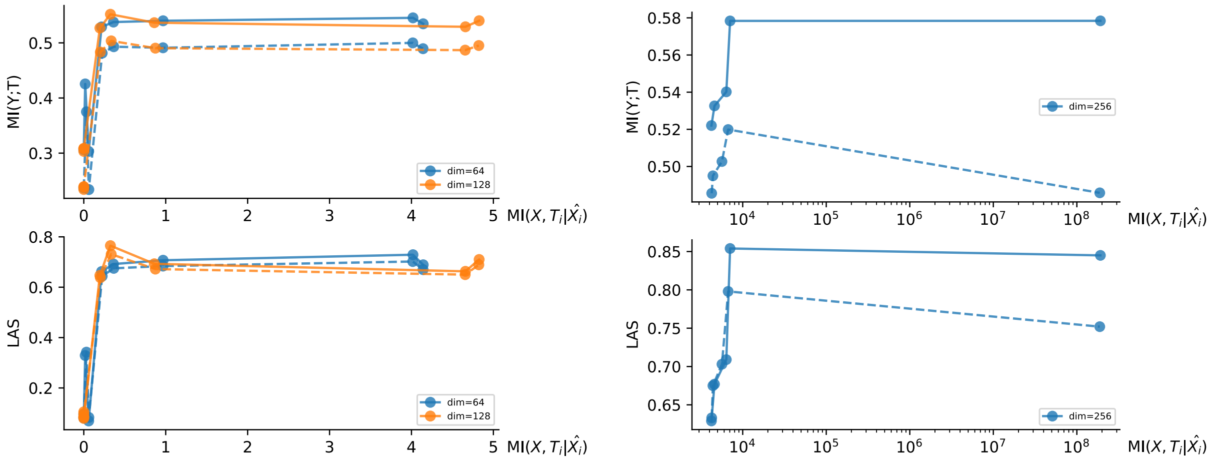
All results are on the English UD data with the usual training/test split. To train and test the classifiers, we used the gold UD annotations to identify the nouns and verbs and their correct syntactic features.

E Plot & Table for Stem Prediction

Figures 8–9 supplement the **Stem** paragraph in §6.2. Figure 8 plots the error rate of reconstructing English stems as a function of the level of compression. Figure 9 shows the reconstruction error rate for the other 8 languages.

F Additional Table of Parsing Performance

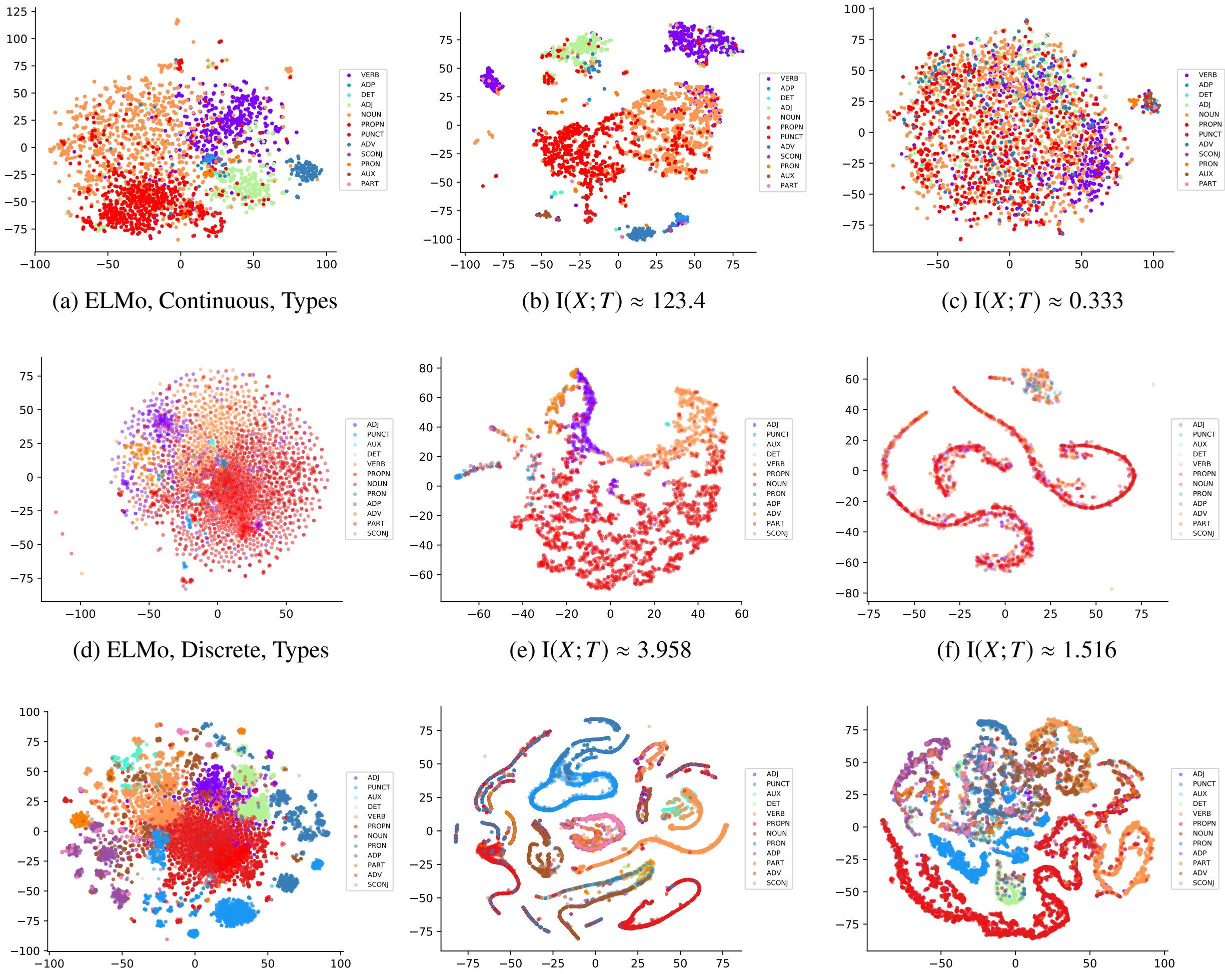
Table 3 is an extended version of Table 2 in §7. It includes parsing performance (measured by LAS and UAS) using ELMo layer 0, 1, and 2.



(a) Discrete Version – ELMo layer-1

(b) Continuous Version – ELMo layer-1

Figure 5: . Tradeoff curves for $\text{I}(T_i; X \mid \hat{X}_i)$ vs. $\text{I}(Y;T)$ and $\text{I}(T_i; X \mid \hat{X}_i)$ vs. LAS, complementary to Figure 2.



(a) ELMo, Continuous, Types

(b) $\text{I}(X;T) \approx 123.4$

(c) $\text{I}(X;T) \approx 0.333$

(d) ELMo, Discrete, Types

(e) $\text{I}(X;T) \approx 3.958$

(f) $\text{I}(X;T) \approx 1.516$

(g) ELMo, Discrete, Tokens

(h) $\text{I}(X;T) \approx 4.755$

(i) $\text{I}(X;T) \approx 1.475$

Figure 6: t-SNE visualization of our continuous tags ($d = 256$) and our distributions over discrete tags ($k = 128$), supplementing Figure 3. Each marker in the figure represents a word token, colored by its gold POS tag. For each row, the series of figures (from left to right) shows a transition from no compression to moderate compression and to too-much compression. The first row (a-c) shows the continuous type embeddings; the second row (d-f) shows the discrete type embeddings; the third row (g-i) shows the discrete token embeddings.

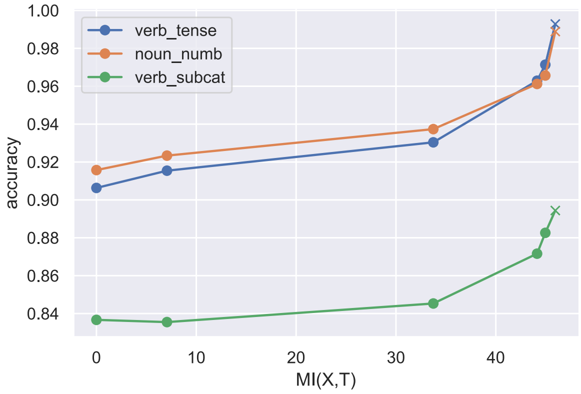


Figure 7: The accuracy of predicting the subcategorization frame of verbs (transitive/intransitive), number of nouns (plural/singular), and tense of verbs (past/present/future), as we change the level of compression of ELMo layer-1 (see the **Subcategorization frame** paragraph in §6.2). As we move from right to left and squeeze irrelevant information out of the tags, they retain these three syntactic distinctions quite well.

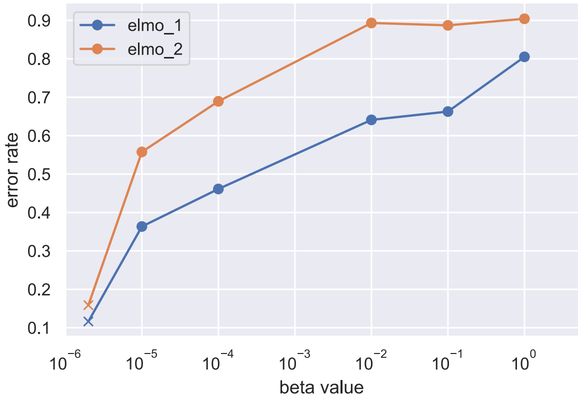


Figure 8: The error rate of reconstructing the stem of a word from the specialized continuous tags. The legend indicates whether we are compressing the ELMo layer-1 or ELMo layer-2 (see the **Stem** paragraph of §6.2).

Compression	layer	Arabic	Spanish	French	Hindi	Italian	Portuguese	Russian	Chinese
Slight	1	26.7%	24.0%	25.5%	20.5%	29.9%	32.5%	38.8%	26.7%
Moderate	1	89.5%	79.8%	66.7%	94.6%	94.7%	93.7%	94.0%	89.6%
Slight	2	34.9%	34.9%	34.9%	34.9%	34.9%	34.9%	34.9%	34.9%
Moderate	2	94.3%	94.3%	94.3%	94.3%	94.3%	94.3%	94.3%	94.3%

Figure 9: Error rate in reconstructing the stem of a word from the compressed version of the ELMo layer-1 and layer-2 embedding). Slight compression refers to $\beta = 0.0001$, and moderate compression refers to $\beta = 0.01$.

					UAS					
Models	Layer	Arabic	Hindi	English	French	Spanish	Portuguese	Russian	Chinese	Italian
Iden	0	0.817	0.914	0.793	0.836	0.851	0.844	0.859	0.775	0.904
Iden	1	0.821	**0.915**	0.868	0.833	0.852	0.842	0.860	0.771	0.903
Iden	2	0.820	0.914	0.843	0.833	0.856	0.841	0.859	0.773	0.901
PCA	0	0.814	0.912	0.787	0.814	0.847	0.857	0.831	0.773	0.897
PCA	1	0.815	0.912	0.865	0.807	0.846	0.855	0.828	0.759	0.899
PCA	2	0.814	0.915	0.832	0.808	0.846	0.858	0.829	0.766	0.902
MLP	0	0.830	0.918	0.742	0.856	0.829	0.869	0.852	0.797	0.910
MLP	1	0.831	0.923	0.823	0.870	0.832	0.867	0.852	0.800	0.908
MLP	2	0.833	0.918	0.787	0.859	0.813	0.871	0.849	0.790	**0.914**
VIBc	0	0.852	**0.915**	0.866	**0.879**	**0.881**	0.871	0.862	0.800	0.831
VIBc	1	**0.860**	**0.913**	**0.871**	**0.877**	**0.880**	**0.877**	**0.865**	**0.814**	**0.913**
VIBc	2	0.851	0.894	**0.880**	0.876	0.879	**0.877**	0.843	0.768	0.878
POS	-	0.722	0.819	0.762	0.800	0.802	**0.808**	0.739	0.570	**0.843**
VIBd	0	**0.783**	0.823	0.784	**0.821**	**0.821**	0.793	**0.777**	0.671	**0.855**
VIBd	1	**0.784**	**0.862**	**0.825**	**0.822**	**0.822**	**0.805**	**0.776**	**0.691**	**0.857**
VIBd	2	0.754	**0.861**	0.816	**0.822**	0.812	0.790	0.768	0.672	0.849

					LAS					
Models	layer	Arabic	Hindi	English	French	Spanish	Portuguese	Russian	Chinese	Italian
Iden	0	0.747	**0.867**	0.745	0.789	0.806	0.812	0.788	0.713	**0.864**
Iden	1	0.751	**0.870**	0.824	0.784	0.808	0.813	0.783	0.709	**0.863**
Iden	2	0.743	0.867	0.798	0.782	0.811	0.813	0.787	0.713	0.861
PCA	0	0.746	0.864	0.742	0.758	0.804	0.811	0.781	0.706	0.856
PCA	1	0.743	**0.866**	0.823	0.749	0.802	0.808	0.777	0.697	0.857
PCA	2	0.744	**0.870**	0.787	0.750	0.801	0.811	0.780	0.700	**0.865**
MLP	0	0.754	**0.869**	0.801	0.814	0.772	0.817	0.798	0.739	**0.871**
MLP	1	0.759	**0.871**	0.839	0.816	**0.835**	0.821	0.800	0.734	**0.867**
MLP	2	0.760	**0.871**	0.834	0.814	0.755	0.822	0.797	0.726	0.869
VIBc	0	**0.778**	0.865	0.822	0.822	**0.839**	0.827	0.807	0.739	0.862
VIBc	1	**0.779**	**0.866**	**0.851**	**0.828**	**0.837**	**0.836**	**0.814**	**0.754**	**0.867**
VIBc	2	0.777	0.838	0.840	**0.826**	0.840	0.829	0.786	0.710	0.818
POS	-	0.652	0.713	0.712	0.718	**0.739**	**0.743**	**0.662**	0.510	0.779
VIBd	0	**0.671**	0.702	0.721	**0.723**	**0.724**	0.710	0.648	0.544	**0.780**
VIBd	1	**0.672**	**0.736**	**0.742**	**0.723**	**0.725**	0.710	**0.651**	**0.591**	**0.781**
VIBd	2	0.643	**0.735**	**0.741**	0.721	0.719	0.698	0.646	0.566	0.763

Table 3: Parsing accuracy of 9 languages (LAS and UAS); Table 2 is a subset of this table. Black rows use continuous tags; gray rows use discrete tags (which does worse). The "layer" column indicates the ELMo layer we use. In each column, the best score for each color is boldfaced, along with all results of that color that are not significantly worse (paired permutation test, $p < 0.05$).