Unsupervised Disambiguation of Syncretism in Inflected Lexicons

Ryan Cotterell, Christo Kirov, Sebastian J. Mielke, Jason Eisner

Department of Computer Science, Johns Hopkins University, Baltimore, USA

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

A lexicon contains syncretism

Wort	Wort	N;SG;NOM
Wort	Wortes	N;SG;GEN
Wort	Wort	N;SG;ACC
Wort	Worte	N;SG;DAT
Wort	Wörter	N;PL;NOM
Wort	Wörter	N;PL;GEN
Wort	Wörter	N;PL;ACC
Wort	Wörtern	N;PL;DAT
Herr	Herr	N;SG;NOM
Herr	Herrn	N;SG;GEN
Herr	Herrn	N;SG;ACC
Herr	Herrn	N;SG;DAT
Herr	Herren	N;PL;NOM
Herr	Herren	N;PL;GEN
Herr	Herren	N;PL;ACC
Herr	Herren	N;PL;DAT
…	…	…

UniMorph (Kirov et al., 2018)

How frequent are any of these forms in natural text?

We could just count "Wortes" in unlabeled data, but what about "Wort"?

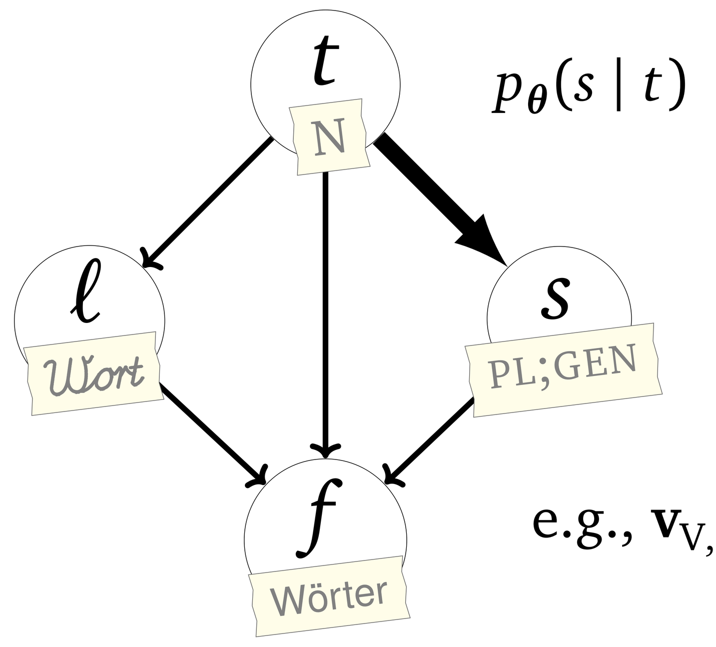
Wort (N)	SG	PL
NOM	Wort	Wörter
GEN	Wortes	Wörter
ACC	Wort	Wörter
DAT	Worte	Wörtern

Herr (N)	SG	PL
NOM	Herr	Herren
GEN	Herrn	Herren
ACC	Herrn	Herren
DAT	Herrn	Herren

Note: Syncretism commonly refers to intra-paradigmatic ambiguity of forms, but we also handle inter-paradigmatic ambiguity.

Model $p_\theta(t, \ell, s, f)$ jointly

We define a *latent-variable* graphical model over *POS tags* t, lemmata ℓ, slots s, and forms f:



$$p_\theta(s \mid t) \propto \exp(\underbrace{\overset{\text{NEURAL}}{\mathrm{NN}_\theta(\mathbf{v}_{t,s})}}\ )$$

UNIFORM 0 FREE $\boldsymbol{\theta}_{t,s}$ LINEAR $\boldsymbol{\theta}^\top \mathbf{v}_{t,s}$

$$\text{e.g., } \mathbf{v}_{\text{V,PL;GEN}} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\begin{array}{cccccccc} \text{N} & \text{V} & \text{SG} & \text{PL} & \text{NOM} & \text{GEN} & \text{ACC} & \text{DAT} \end{array}$$

For $p_\theta(s \mid t)$, we evaluate three ablations, $p(t)$ and $p(\ell \mid t)$ are unrestricted distributions with support on all seen values, and $p(f \mid t, \ell, s)$ is 1 iff $\langle \ell, f, ts \rangle$ is in the lexicon.

We train all distributions to maximize observed form counts:

$$\sum_f c(f) \log p_\theta(f) = \sum_f c(f) \log \sum_{t,\ell,s} p_\theta(t) p_\theta(s \mid t) p_\theta(f \mid t, \ell, s),$$

↖ finite (as the lexicon is) and tractable!

Assign "counts" to forms using the conditional

Extracting the conditional $p_\theta(t, \ell, s \mid f)$ from the modeled joint is easy:

$$p_\theta(t, \ell, s \mid f) = p_\theta(t, \ell, s, f) \ / \ \underbrace{p_\theta(f)}_{=c(f)/\sum_{f'} c(f')}$$

Then we only need to multiply with *form counts* (denoted $c(\cdot)$; can be taken from unlabeled text):

$$c(t, \ell, s, f) = c(f) \cdot p_\theta(t, \ell, s \mid f)$$

e.g., assuming that $c(\text{Wörter}) = 6$:

$$c(\text{N}, \text{Wort}, \text{PL;GEN}, \text{Wörter})$$
$$= c(\text{Wörter}) \cdot p_\theta(\text{N}, \text{Wort}, \text{PL;GEN} \mid \text{Wörter})$$

1.7	Wort	Wort	N;SG;NOM
2	Wort	Wortes	N;SG;GEN
1.3	Wort	Wort	N;SG;ACC
4	Wort	Worte	N;SG;DAT
3.4	Wort	Wörter	N;PL;NOM
1.3	Wort	Wörter	N;PL;GEN
1.3	Wort	Wörter	N;PL;ACC
3	Wort	Wörtern	N;PL;DAT
8	Herr	Herr	N;SG;NOM
1.3	Herr	Herrn	N;SG;GEN
1.2	Herr	Herrn	N;SG;ACC
1.5	Herr	Herrn	N;SG;DAT
1.2	Herr	Herren	N;PL;NOM
1.3	Herr	Herren	N;PL;GEN
1.4	Herr	Herren	N;PL;ACC
1.1	Herr	Herren	N;PL;DAT
…	…	…	

Count-annotated UniMorph

Note: Imputing the counts in this way can also be seen as the E-step of EM training, an alternative way to maximize our log-likelihood objective.

Languages

Czech German Finnish Hebrew Swedish

Chosen to be high-resource, so we can compare to gold data.

Data provenance

- UniMorph (Kirov et al., 2018)
  → Only type-level, no counts
- Wikipedia
  → read off form counts
  (possible for any language; unsupervised)
  → Lemmatize & tag with UDPipe (Straka et al., 2016)
  (only used for eval, requires high-resource language)
    → Convert to UniMorph format (discard up to 31%)

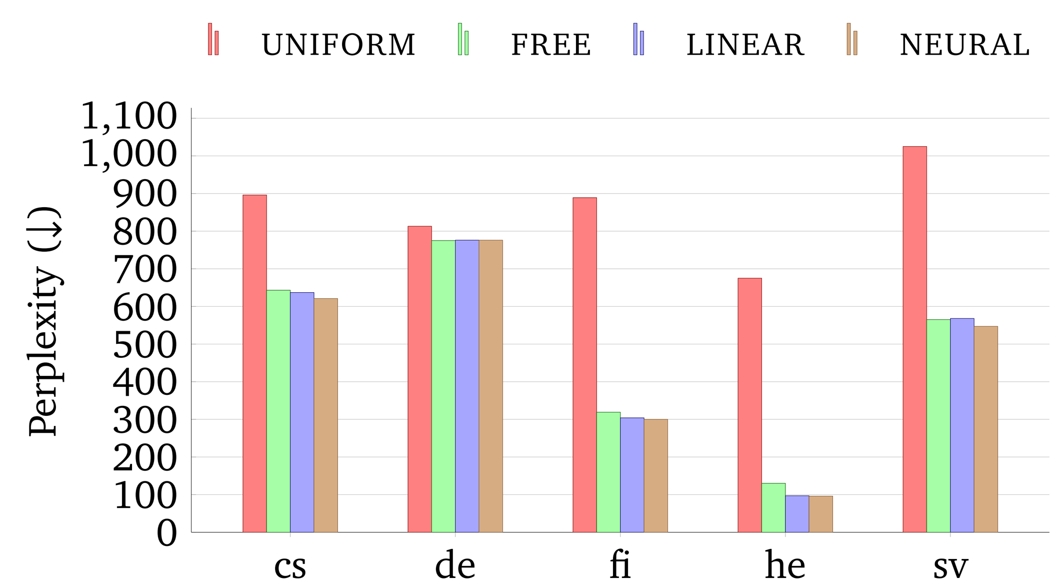
Evaluation

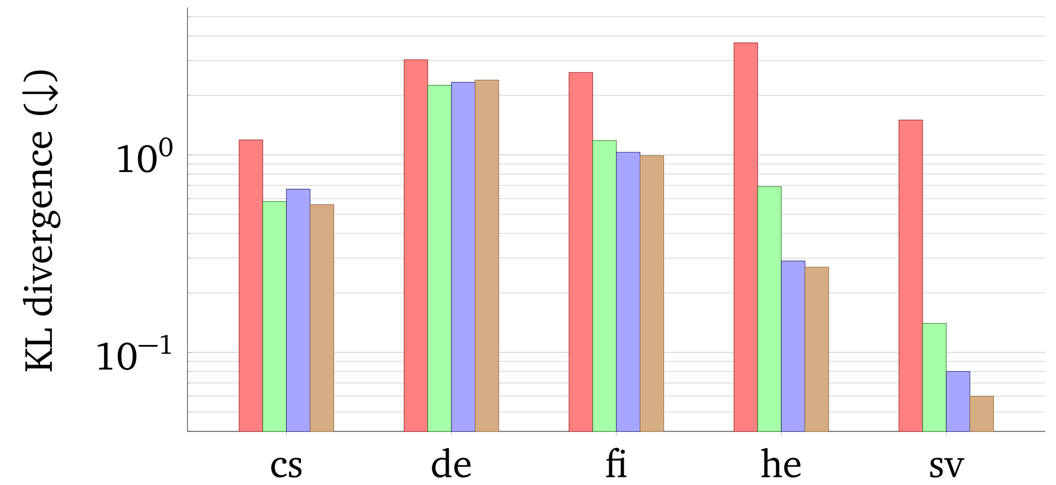
Perplexity (↓) of our generative model on held-out tokens.

KL-divergence (↓) between our *unsupervised* distribution p_θ (trained only on unigram form counts and the unweighted lexicon) and the *supervised* distribution $\hat{p}$ (trained on lemma- and slot-annotated text produced by an existing supervised and contextual tagger):

$$\mathbb{E}_{f \sim \hat{p}} \operatorname{KL}(\hat{p}(\cdot \mid f) \,||\, p_\theta(\cdot \mid f)) = \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{\hat{p}(t_i, \ell_i, s_i \mid f_i)}{p_\theta(t_i, \ell_i, s_i \mid f_i)}$$

UNIFORM would have KL $= 0$ if all forms were either unambiguous or uniformly ambiguous, the fact that it doesn't means the task is nontrivial. NEURAL matches the supervised distributions reasonably closely, achieving an average KL of < 1 bit on all languages but German.





So what?

Convert artificial lexicons into "real" data, bridging the gap from *types* to *tokens*.

Our "counts" can be used for weighting in many morphology-aware NLP tasks (e.g., smoothing of embeddings). Specifically, we use it for the SIGMORPHON 2018 shared task to split dataset by frequency (simulating real application scenarios).

Code and data

`sjmielke.com/papers/syncretism`