A Head-to-Head Comparison of Perceptron and Logistic Regression for Morphological Tagging

Saman Rahimi
Stockholm University
sara5578@student.su.se
June 5, 2026

Abstract

This paper compares the Perceptron and Multinomial Logistic Regression (MLR) for morphological feature prediction on the Swedish UniMorph dataset. Both models use identical prefix/suffix and POS-based features and are trained for three epochs. The task is to predict morphological labels such as number, definiteness, and gender. Accuracy is used as the main metric, and a paired bootstrap test assesses statistical significance. Although MLR is theoretically more powerful as a probabilistic model, the Perceptron slightly outperforms it (66.2% vs. 65.8%), though the difference is not statistically significant. These results underline the robustness of mistake-driven learning in sparse-feature settings.

1 Background

Morphological feature classification involves assigning labels that capture grammatical properties such as number (e.g., singular/plural), definiteness, gender, or tense. In richly inflected languages like Swedish, this is a non-trivial task due to the large number of possible tag combinations and the sparsity of some forms in training data.

Two commonly used approaches for classification are the Perceptron and Multinomial Logistic Regression (MLR). The Perceptron is a mistake-driven algorithm that updates weights only when a prediction is incorrect. In contrast, MLR is a probabilistic model that updates weights based on the gradient of a log-likelihood function using all class probabilities (**?**). Both methods can be trained on feature representations such as binary indicators for POS, prefixes, and suffixes.

For evaluation, accuracy is the primary metric, and we apply the paired bootstrap test (**?**) to assess statistical significance in model comparison.

2 Methodology

We use the Swedish UniMorph dataset, where each line contains a wordform, its POS tag, and a set of morphological features. Our task is to predict the morphological features (e.g., SG;DEF) given a word and its POS.

Feature Extraction: For both models, we extract binary features consisting of the part-of-speech (POS), and all prefixes and suffixes of the word up to length five. Each feature is combined with the POS to increase disambiguation capacity. For example, for the word *boken* and POS *NOUN*, the features include `NOUN prefix=b`, `NOUN suffix=en`, etc.

Labels: The label to be predicted is the morphological tag (e.g., SG;DEF), excluding the POS tag.

Models:

Perceptron: A standard online mistake-driven learner. On each error, it updates the weights by ± 1 for the predicted and true classes.

Multinomial Logistic Regression (MLR): A probabilistic classifier that computes softmax scores for all classes and updates weights using the gradient of the log-likelihood function.

Training Setup: Both models are trained for 3 epochs using the same training and test splits, features, and label sets. Learning rate is set to 1.0 for both models, and no shuffling or regularization is applied.

Evaluation: We compute classification accuracy on the test set. To compare the two models, we use the paired bootstrap test (Jurafsky & Martin, 2024, §4.9) with 5000 resampling iterations to determine if differences are statistically significant. tag (e.g., SG;DEF) given the word and its POS.

Features: Binary features are extracted from all prefixes and suffixes (up to length five), combined with POS.

Labels: Only morphological tags, excluding the POS, are used as labels.

Models:

- **Perceptron:** Mistake-driven learner with ± 1 weight updates.
- **MLR:** Probabilistic classifier using softmax and gradient-based updates.

Both models are trained for 3 epochs with learning rate 1.0 on the same data and features. Accuracy is used for evaluation, and the paired bootstrap test (**?**) is applied to assess significance.

3 Experiments and Results

Both models were trained and evaluated on the Swedish UniMorph dataset using the same training and test splits. Each model used POS + prefix/suffix features, with labels consisting only of the morphological features (excluding POS). Models were trained for 3 epochs with a learning rate of 1.0.

Accuracy: Perceptron achieved 66.24% accuracy.
MLR achieved 65.80% accuracy.
Although the Perceptron slightly outperformed MLR, the performance difference was small.

Statistical Significance: To test whether the difference was statistically meaningful, we applied a paired bootstrap test with 5000 iterations. The 95% confidence interval for the difference in accuracy was: [–0.0068, 0.0156].

Since the interval includes zero, we conclude that the difference is not statistically significant.

4 Discussion and Conclusion

Our results indicate that, despite the theoretical strengths of Multinomial Logistic Regression (MLR), the Perceptron achieves slightly higher accuracy in morphological feature classification on the Swedish UniMorph dataset. However, the paired bootstrap test shows that this difference is not statistically significant, suggesting that both models perform comparably in practice.

The competitive performance of the Perceptron highlights the strength of mistake-driven learning in tasks with sparse and binary features, particularly when training data is limited and the label space is large. This aligns with earlier observations in NLP where simpler online algorithms outperform probabilistic models under constrained settings.

For future work, it would be valuable to explore additional features such as character n-grams, use regularization in MLR, and test the models across different languages or morphological paradigms. Incorporating these extensions may reveal more nuanced differences between the algorithms.

References

1. Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing (3rd ed. draft). Sections 4.9, 5.3, 5.8. https://web.stanford.edu/ jurafsky/slp3/ 2. Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The UniMorph 2.0 Project: Universal Morphological Inflection Across Languages. In Proceedings of LREC 2018. 3. Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In EMNLP 2002. 4. Yoav Goldberg. 2017. Neural Network Methods for Natural Language Processing. Morgan Claypool Publishers. (Good overview of both Perceptron and MLR from a modern perspective.) 5. Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In ICML 2008. (Highlights improvements on Perceptron-style algorithms.) 6. Kyunghyun Cho et al. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In EMNLP 2014. (Often cited when comparing classical models like MLR and Perceptron with neural approaches.) 7. Efron, B., Tibshirani, R. J. 1994. An Introduction to the Bootstrap. Chapman Hall. (Canonical reference for the bootstrap method used in significance testing.)