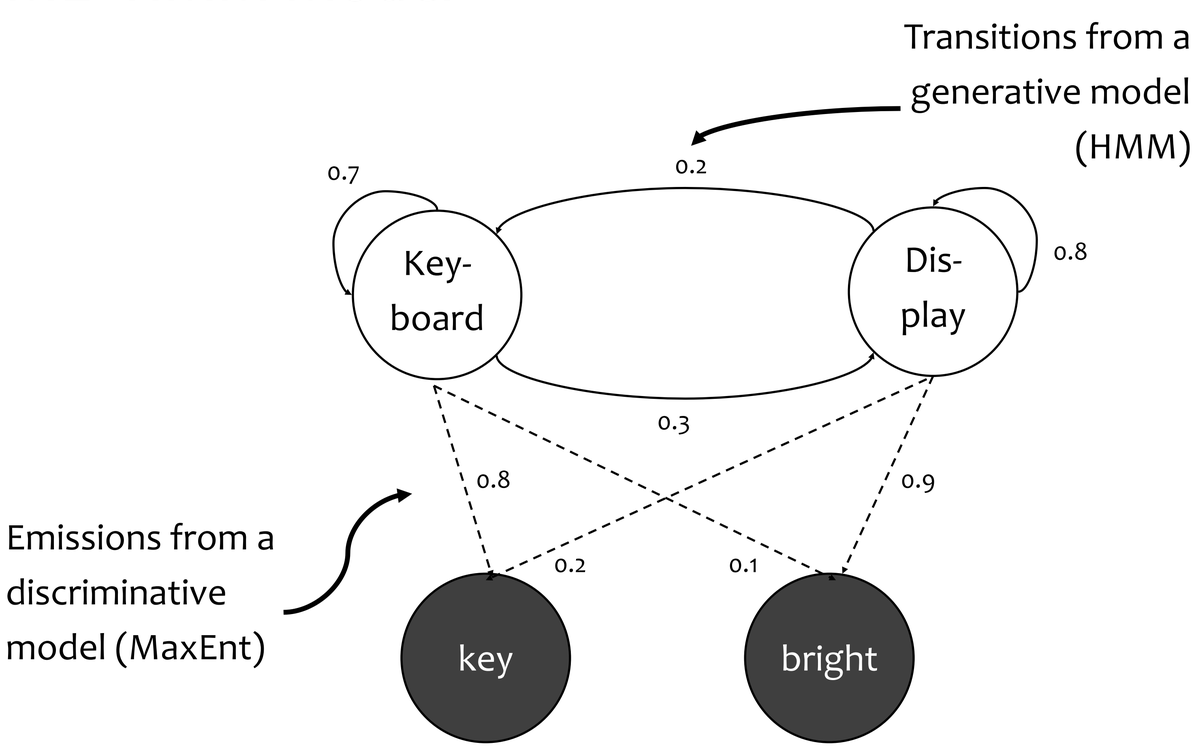
Semi-discriminative Topic Classification on a Small Dataset

Cornelia Ferner and Stefan Wegenkittl
Salzburg University of Applied Sciences

This work aims at identifying sequences of words related to specific product components (topics) in online product reviews. By assuming independence over topics, a Max Entropy classifier reliably solves the classification problem. However, the reviews exhibit an inherent structure on document level allowing to frame the task as sequence classification problem. Due to the limited amount of available training data, powerful models such as Conditional Random Fields are not competitive. We propose using a Hidden Markov Model and decouple the training of transition and emission probabilities. The discriminative power of the Max Entropy approach is used for the latter. Besides outperforming both standalone methods as well as more generic models, the combined classifier assigns topics on subsentence level although labeling in the training data is only available on sentence level.

ME+HMM Model



Token-level Evaluation

Gold labels

Otherwise, the approx. 3.3 kilogram heavy case didn't actually knock our socks off: design, workmanship and materials are only second rate. The input devices could also be a lot better (small touchpad, clattery keyboard, single-rowed enter, etc.). The main point of complaint is the enormous noise development, typical for a gamer: the fan is clearly audible during load.

Build/Case Keyboard Touchpad Noise

ME+HMM

Otherwise, the approx. 3.3 kilogram heavy case didn't actually knock our socks off: design, workmanship and materials are only second rate. The input devices could also be a lot better (small touchpad, clattery keyboard, single-rowed enter, etc.). The main point of complaint is the enormous noise development, typical for a gamer: the fan is clearly audible during load.

Sentence-level Evaluation

MaxEnt [2]

Max Entropy

Discriminative sentence-level classifier that has no "memory" to take the typical review structure into account.

70 %

Sentence level Accuracy

HMM [3]

Hidden Markov Model

Purely generative token-level sequence model that estimates the joint probabilities of hidden states and observed words.

60 %

Sentence level Accuracy

CRF [4]

Conditional Random Field

Discriminative token-level sequence model that models the joint probability of the entire state sequence conditioned on the observed sequence.

40 %

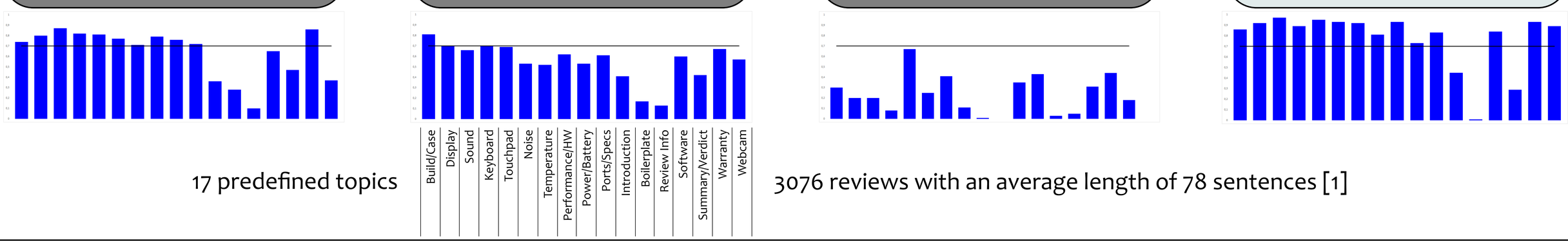
Sentence level Accuracy

ME+HMM

Generative token-level sequence model with a discriminative estimate of emission probabilities.

77 %

Sentence level Accuracy



17 predefined topics

3076 reviews with an average length of 78 sentences [1]

ECMLPKDD Würzburg 2019

[1] fact.ai: Aggregated Text Corpus of Laptop Expert Reviews with Annotated Topics. Online (2018), https://github.com/factai/corpus-laptop-topic.
[2] Klein, D., Manning, C.: Conditional Structure versus Conditional Estimation in NLP Models. In: Proceedings of the ACL Conference on EMNLP, ACL (2002).
[3] Sutton, C., McCallum, A., et al.: An Introduction to Conditional Random Fields. Foundations and Trends ® in Machine Learning 4(4), 267-373 (2012).
[4] Lafferty, J.D., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th ICML, (2001).

Affiliation:
Salzburg University of Applied Sciences
Information Technology & Systems Management
Urstein Sued 1
5412 Urstein/Salzburg, Austria
its.fh-salzburg.ac.at

cornelia.ferner@fh-salzburg.ac.at
stefan.wegenkittl@fh-salzburg.ac.at

FH Salzburg
Informationstechnik & System-Management