Evaluating Word Prediction: Framing Keystroke Savings

Keith Trnka and Kathleen F. McCoy
University of Delaware

Background

- word prediction
  - application of ***language modeling*** to ***disabilities research***

Word prediction

- reduces the amount of typing effort
- increases communication rate

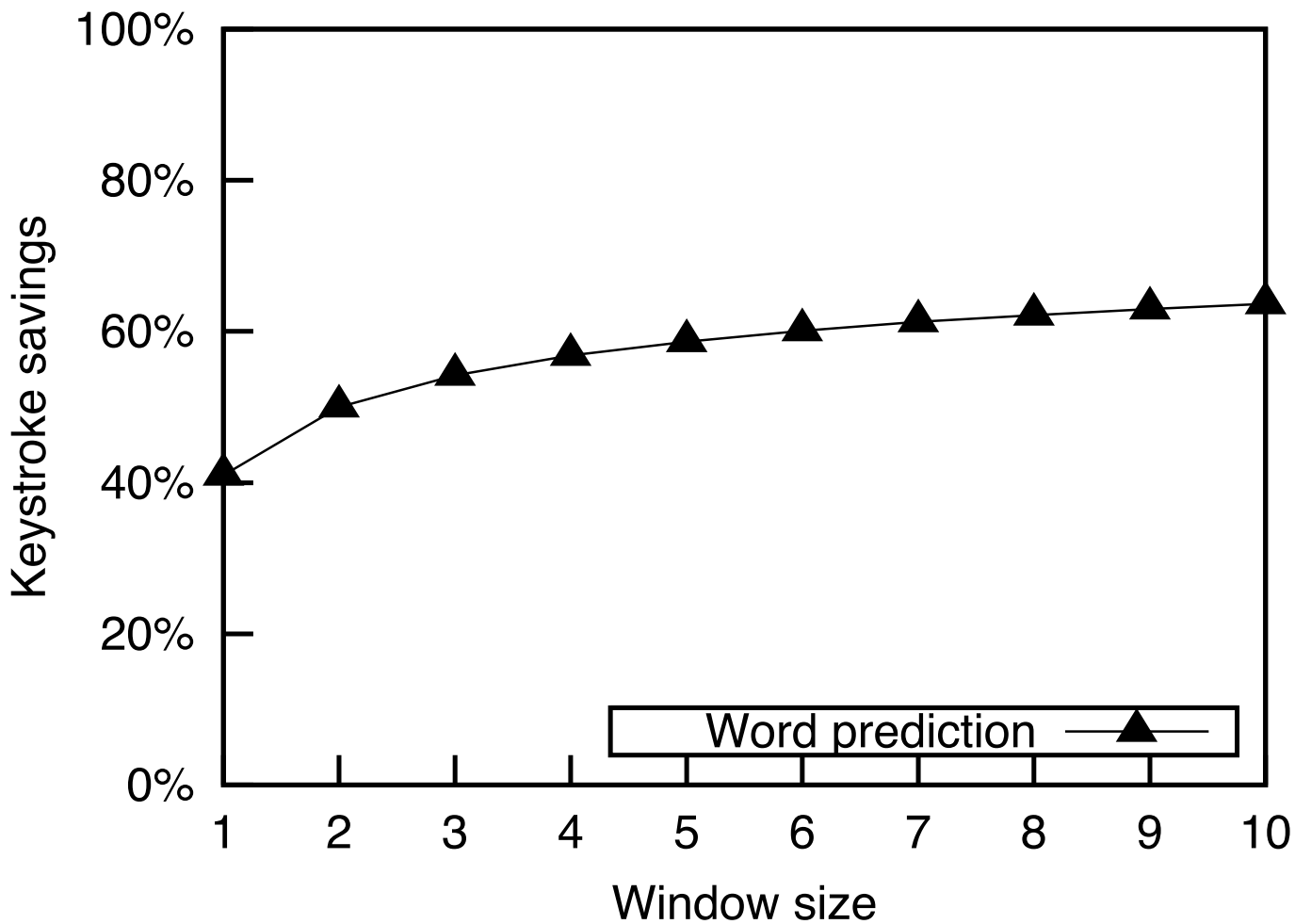
Evaluating word prediction

$$KS = \frac{keystrokes_{letter\text{-}by\text{-}letter} - keystrokes_{with\ prediction}}{keystrokes_{letter\text{-}by\text{-}letter}} \times 100\%$$

- ***keystroke savings***
- the percentage improvement in the number of keys typed

Keystroke savings

- affected by the number of predictions



Research problems

- What do the values mean?
  - Is 60% good or bad?
  - How well can we do?
- How to determine the number of keys under each entry method?

How many keystrokes?

$$KS = \frac{keystrokes_{letter\text{-}by\text{-}letter} - keystrokes_{with\ prediction}}{keystrokes_{letter\text{-}by\text{-}letter}} \times 100\%$$

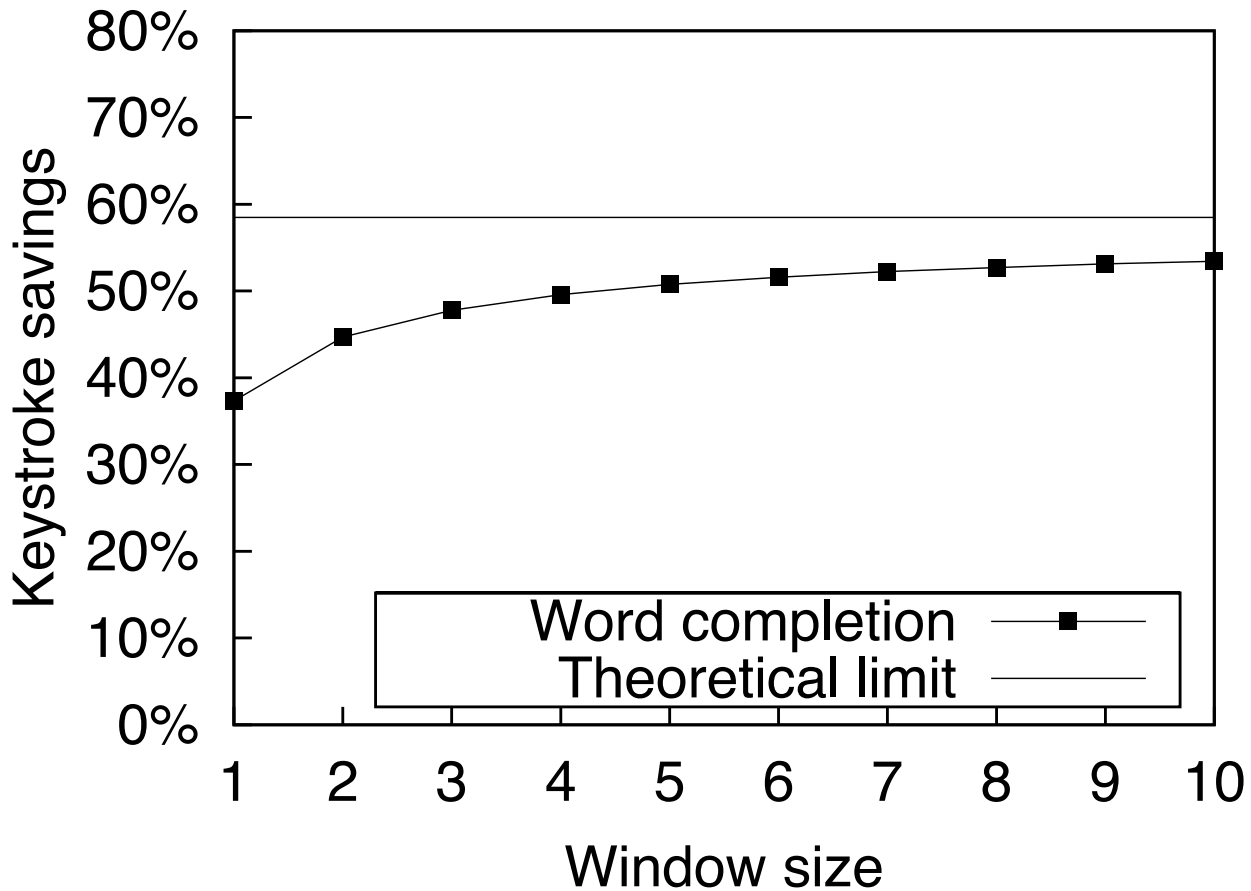
- Which keystrokes do we count?
  - assume all characters take one keystroke, include spaces, newlines
- Best number of keystrokes with prediction
  - assume the user doesn't miss any predictions
  - assume that only single words are predicted

What is "good"?

- compare to a gold standard for word prediction
- theoretical best = one keystroke per word (plus one extra per sentence)
  - assumes a perfect language model
  - translates to the ***theoretical (keystroke savings) limit***

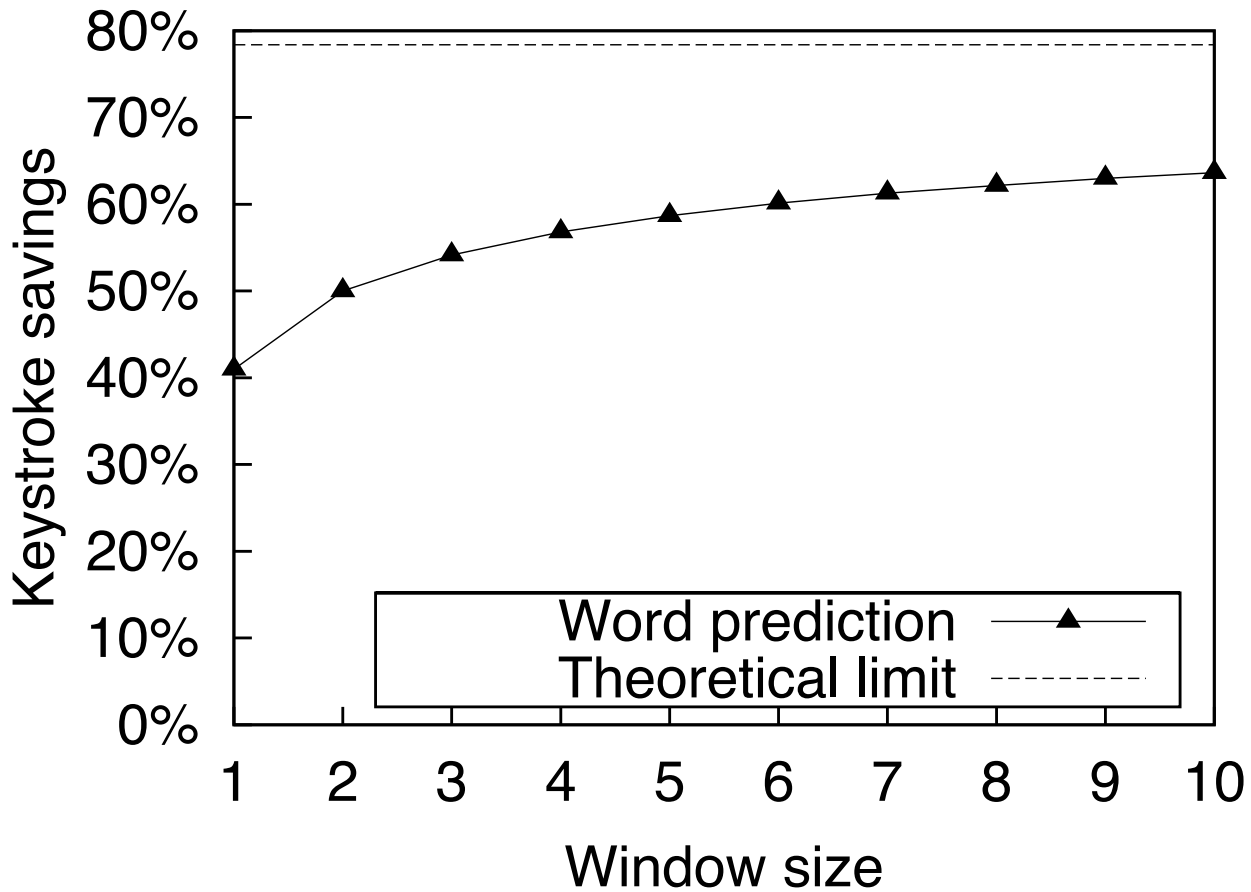
Testing the limit

- theoretical limit vs. trigram model for word *completion* (must type one letter)



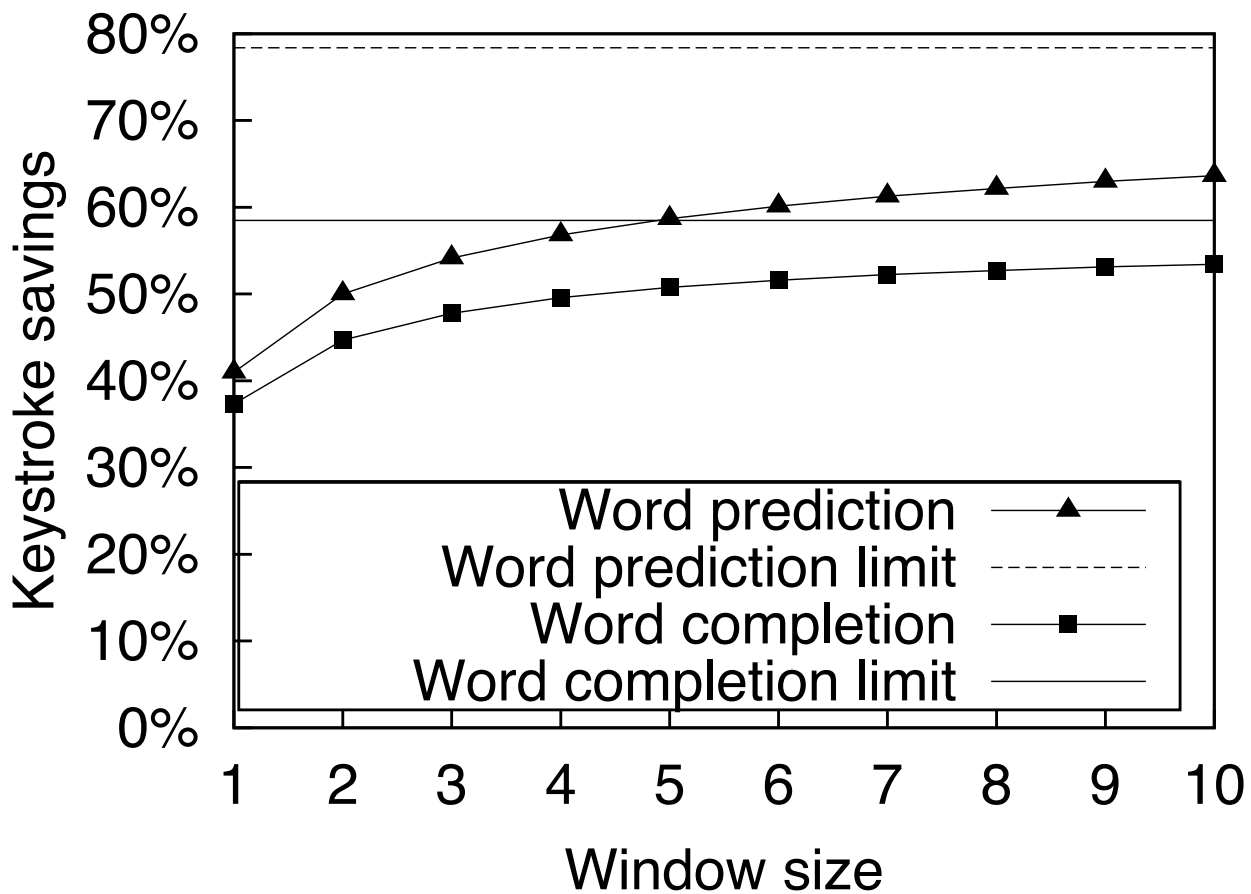
Testing the limit

- theoretical limit vs. trigram model for word *prediction*



Testing the limit

- word completion vs. prediction (and limits)



Problem with the limit

- no practical way to reach the limit
- how can we account for the limitations of language modeling?
  - limits of training data
  - a first step: vocabulary differences
- the ***vocabulary limit***
  - only perfect prediction for words in training text

Testing the new limit

	trigram	**vocab. limit**	**theor. limit**
Switchboard training/testing	60.4%	80.3%	82.6%
Micase training/testing	49.0%	69.2%	84.1%
AAC Email training/testing	48.9%	61.9%	84.8%

Testing the new limit

	trigram	**vocab. limit**	**theor. limit**
Switchboard training/testing	60.4%	80.3%	82.6%
Micase training/testing	49.0%	69.2%	84.1%
AAC Email training/testing	48.9%	61.9%	84.8%

Testing the new limit

	trigram	**vocab. limit**	**theor. limit**
Switchboard training/testing	60.4%	80.3%	82.6%
Micase training/testing	49.0%	69.2%	84.1%
AAC Email training/testing	48.9%	61.9%	84.8%

Testing the new limit

	trigram	**vocab. limit**	theor. limit
Switchboard training/testing	60.4%	80.3%	82.6%
Micase training/testing	49.0%	69.2%	84.1%
AAC Email training/testing	48.9%	61.9%	84.8%

Testing the new limit

	trigram	vocab. limit	theor. limit
Switchboard training/testing	60.4%	80.3%	82.6%
Micase training/testing	49.0%	69.2%	84.1%
AAC Email training/testing	48.9%	61.9%	84.8%

Conclusions

- a step towards understanding the limits of language modeling for word prediction
- gold standards illuminate the limiting factor:
  - low theor. limit ⇒ multi-word prediction
  - low vocab. limit ⇒ expand vocabulary
  - neither low ⇒ improve language modeling