Leveraging transliterations from multiple languages

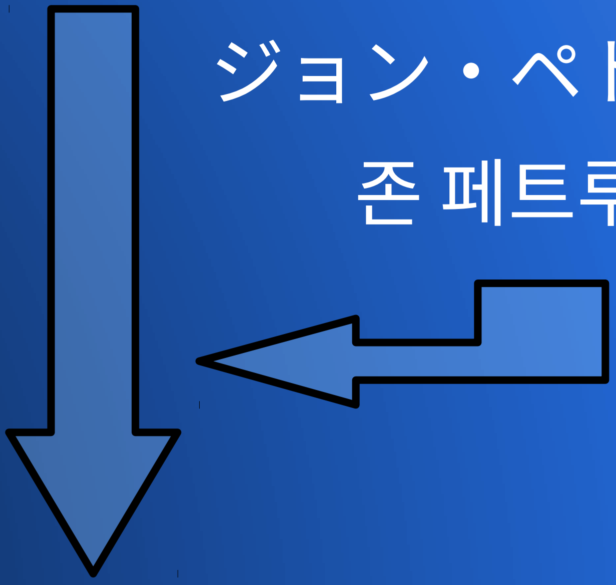
Aditya Bhargava, Bradley Hauer, and Grzegorz Kondrak
University of Alberta

NEWS 2011

Introduction

http://www.flickr.com/photos/rodrigofavera/2327817974
/

John Petrucci



EnHi

ジョン・ペトルーシ

존 페트루치

जॉन?प?ट?रूची

Introduction

- Most previous work on transliteration has focused on a single language
    - English to Hindi, English to Japanese, Arabic to English, etc.
- But data from other languages can be helpful
- Improve existing model's results using supplemental data

Introduction

- Also experiment with:
  - Incorporating other models
  - Hindi romanization
  - English-to-Chinese alignment lengths

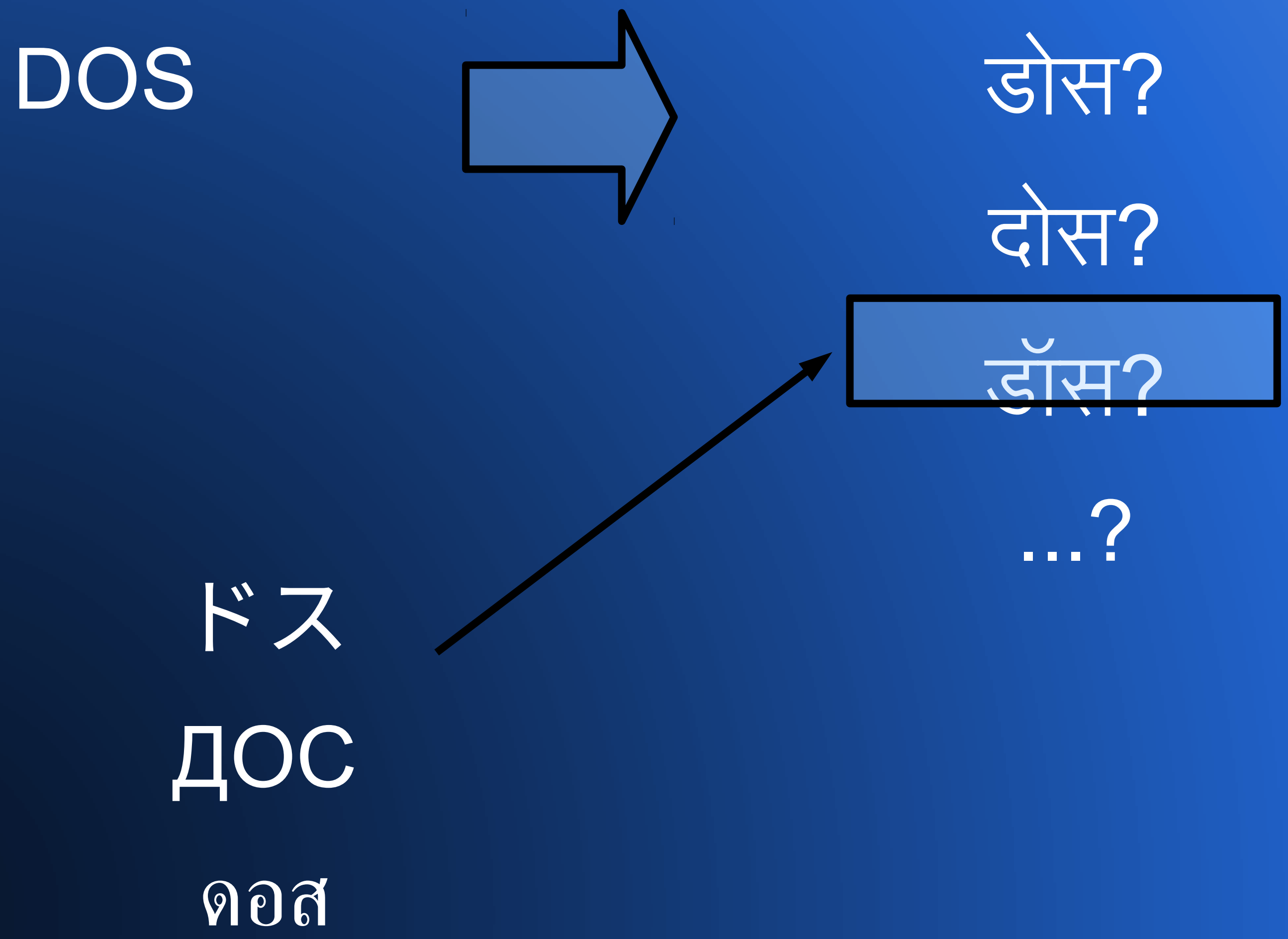
Previous work

- Previous NEWS
  - DirecTL/DirecTL+ (Jiampojamarn et al., 2009/2010)
    - Discriminative, online, max-margin
  - Sequitur + SMT combination (Finch and Sumita, 2010)
    - Sequitur is generative, joint n-gram
- Applying supplemental transliterations to G2P (Bhargava and Kondrak, 2011)
  - We apply this method verbatim
  - Based on SVM re-ranking

Test data overlap

Language	Test set size	Test set overlap
EnBa	1,000	498
EnCh	2,000	2,000
EnHe	1,000	525
EnHi	1,000	889
EnJa	1,815	734
EnKa	1,000	883
EnKo	609	608
EnPe	2,000	1,049
EnTa	1,000	884
EnTh	2,000	1,564

Re-ranking



DOS

डोस?

दोस?

डॉस?

...?

ドス

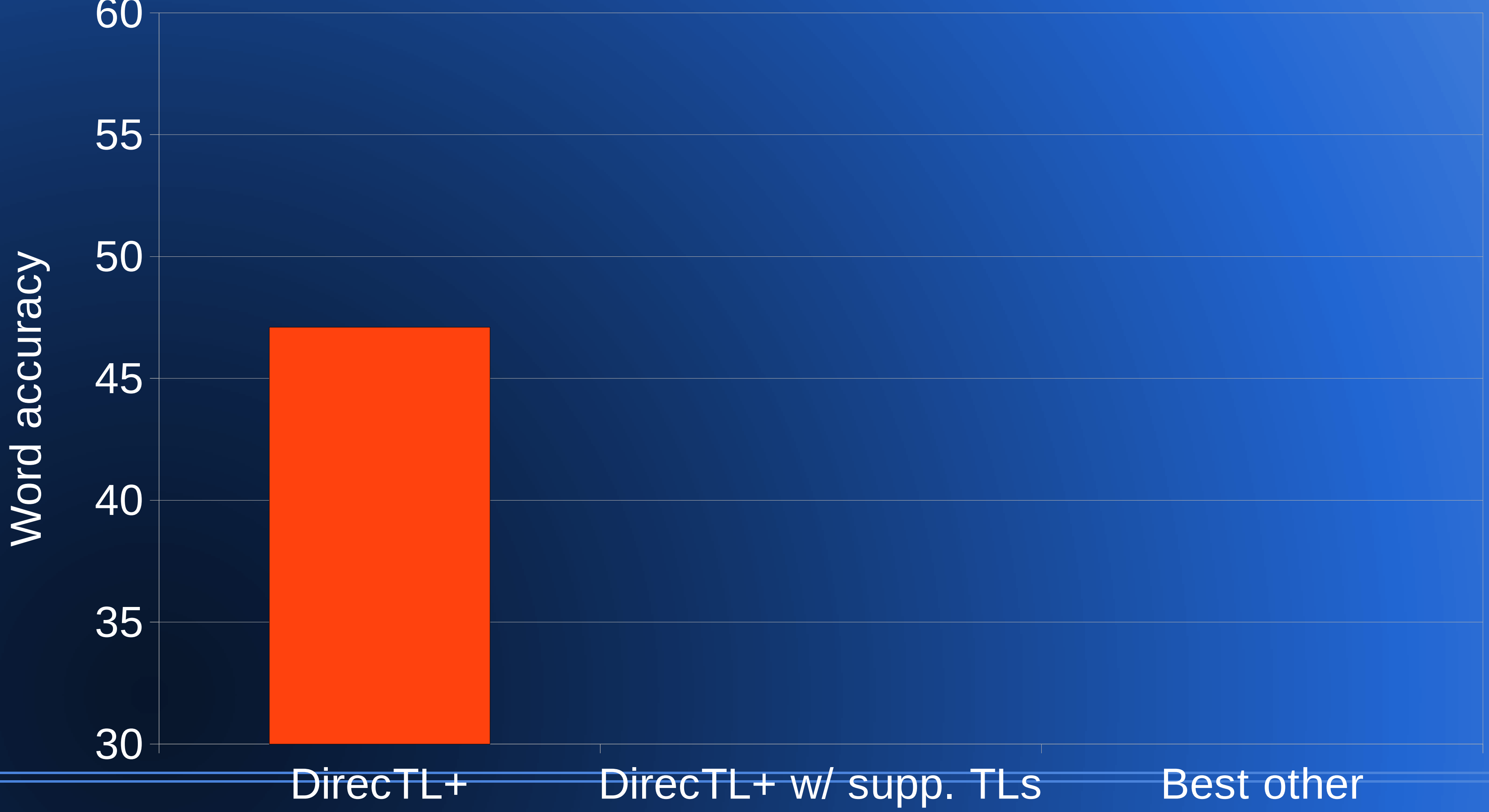
ДОС

ดอส

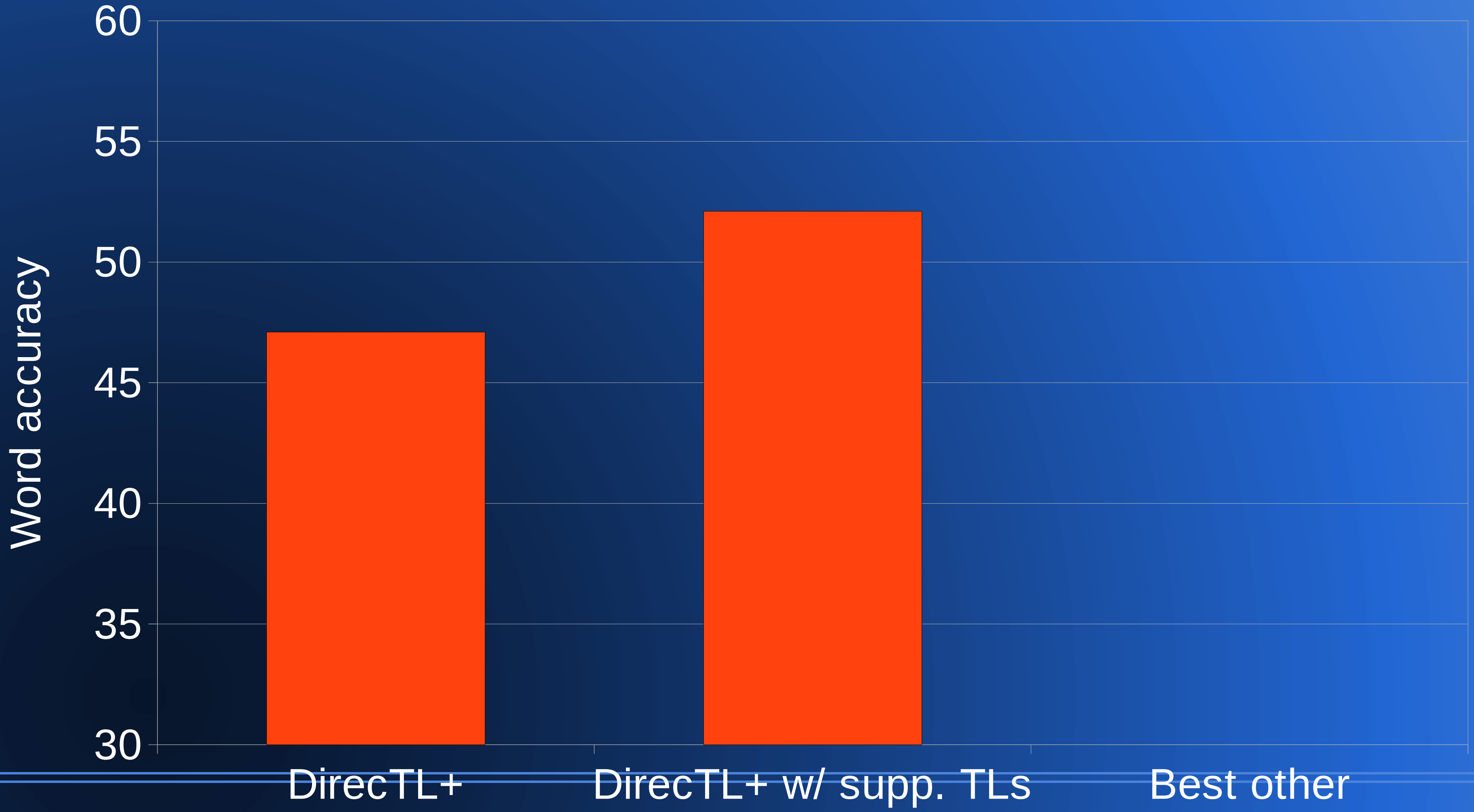
Re-ranking

- SVM re-ranking using all other languages
- Features:
  - N-gram features based on character alignments
  - Similarity features based on alignment scores
- Transliteration data are noisy; handled by:
  - Granular n-gram features
  - Multiple languages
- DirecTL+ baseline

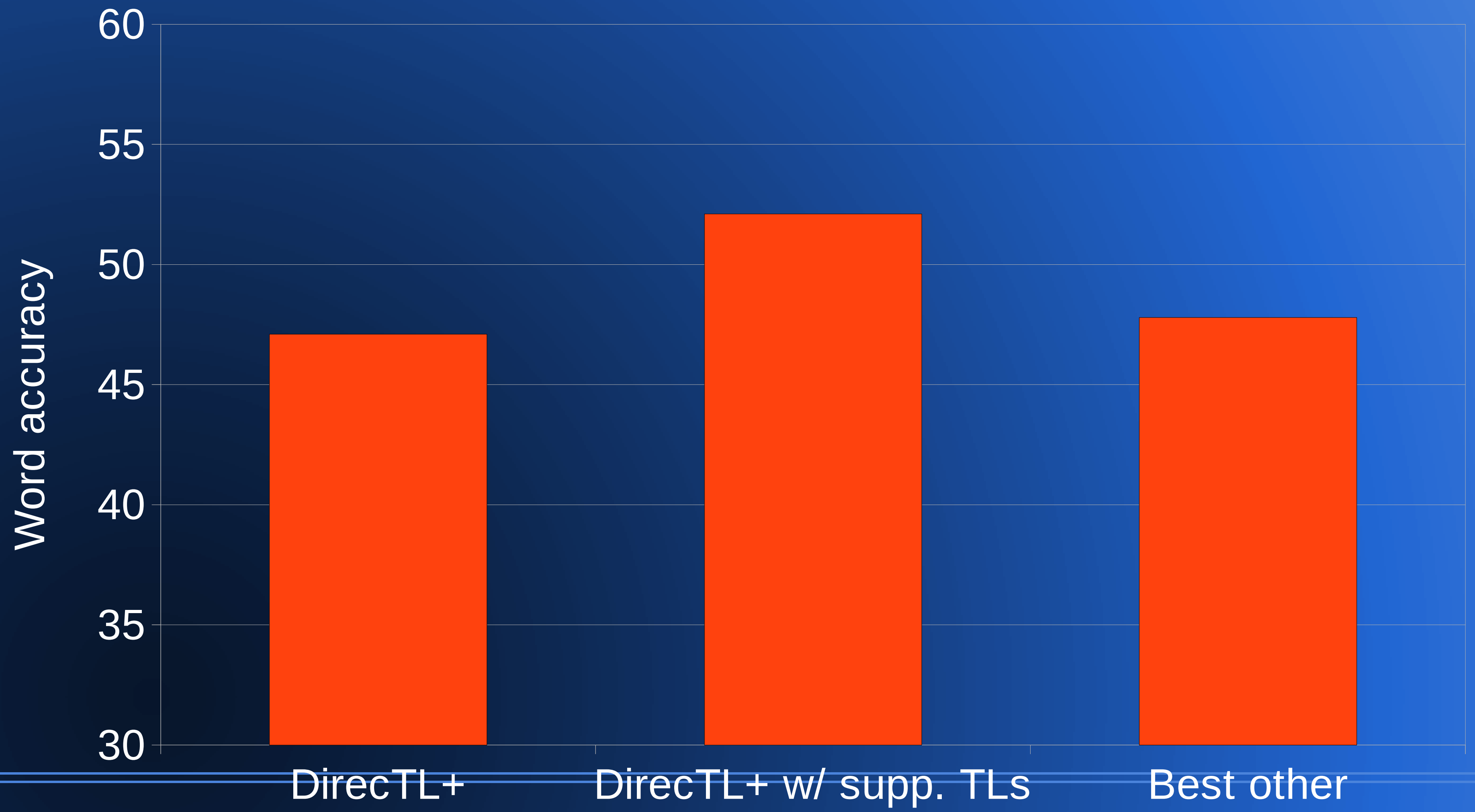
EnHi transliteration re-ranking



EnHi transliteration re-ranking



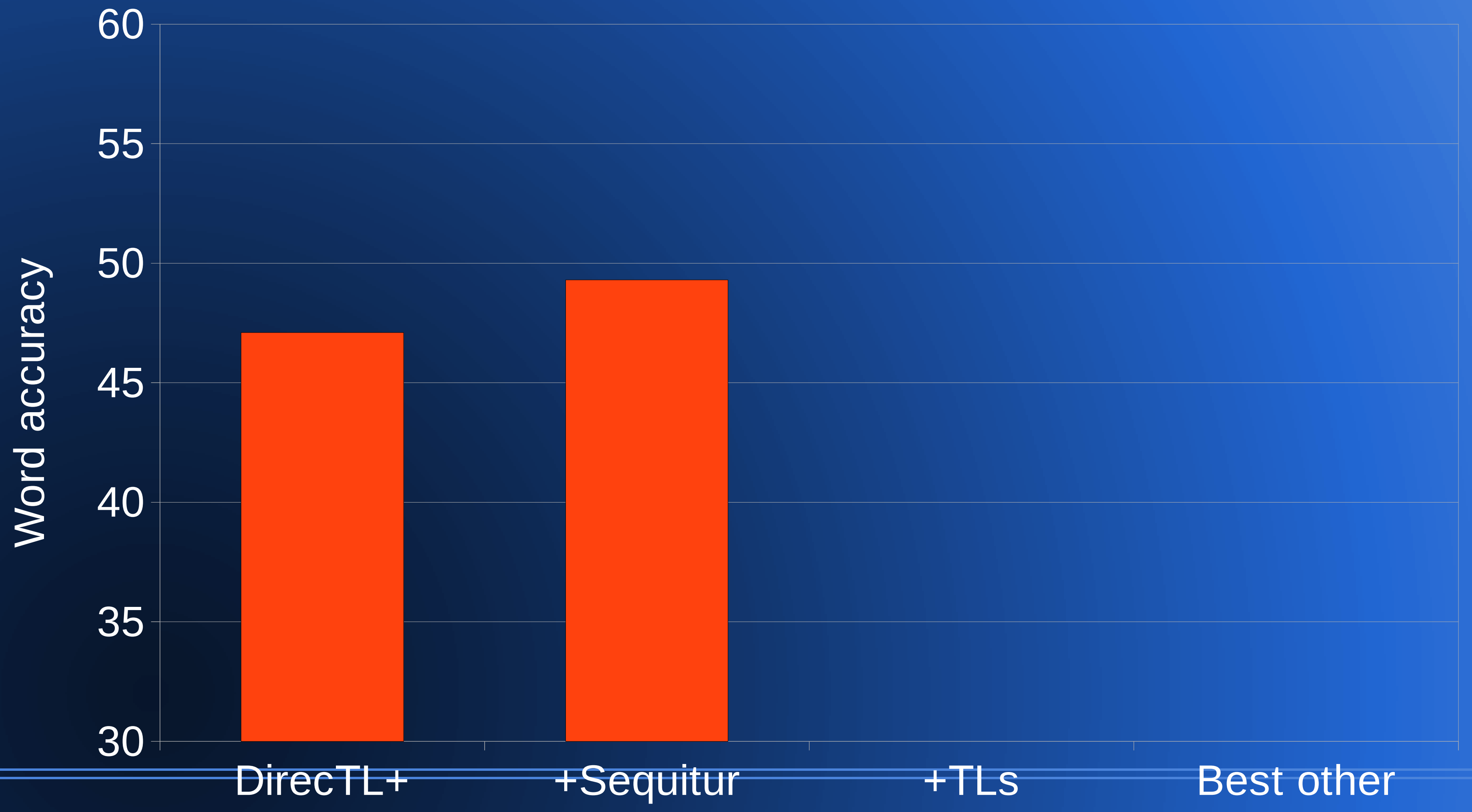
EnHi transliteration re-ranking



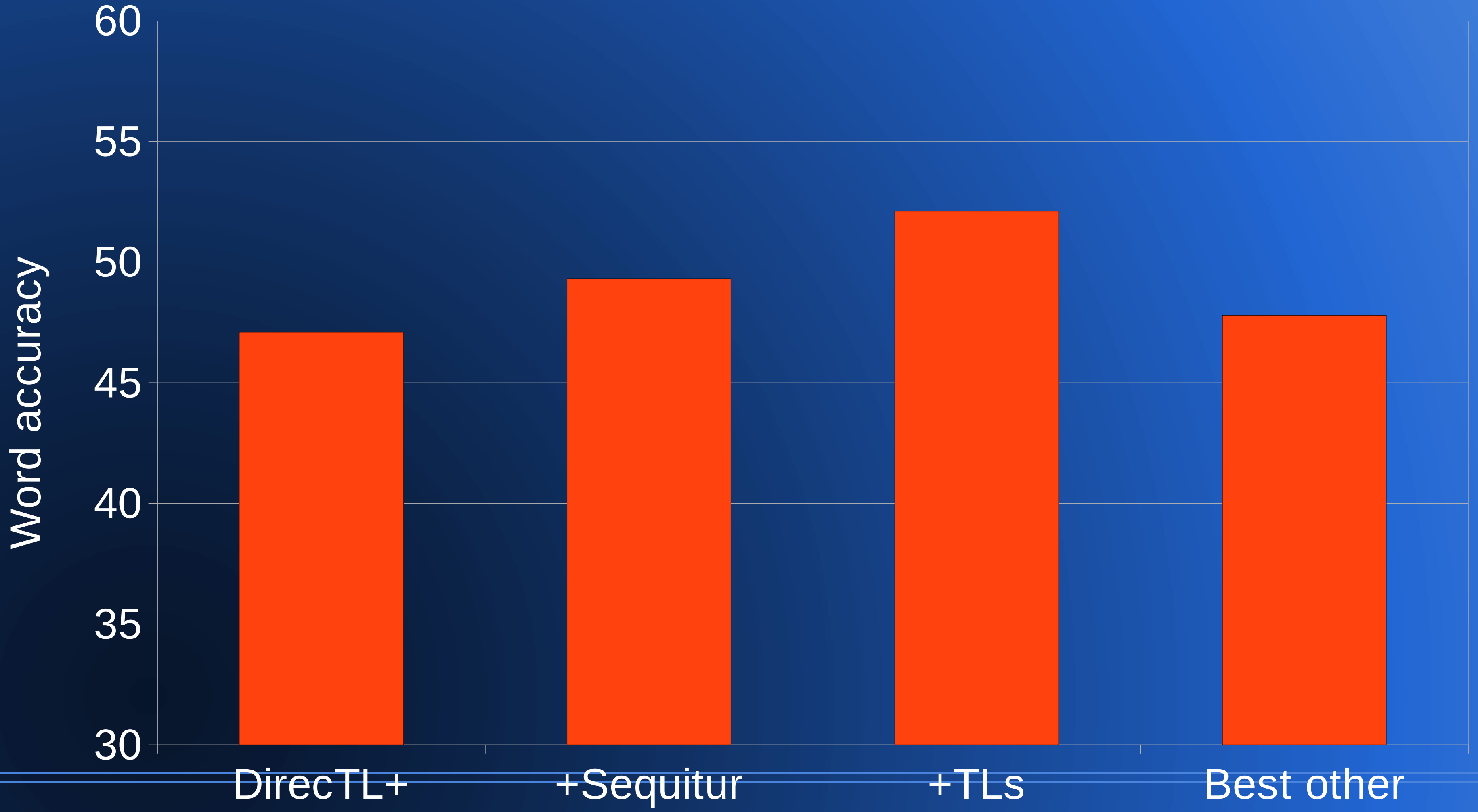
Re-ranking with Sequitur

- Use Sequitur's output for re-ranking
- Exact same features

EnHi Sequitur re-ranking

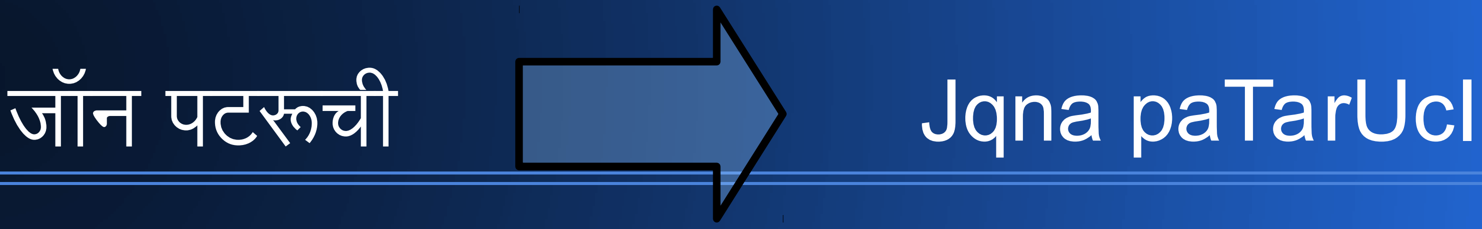


EnHi Sequitur re-ranking



Hindi romanization

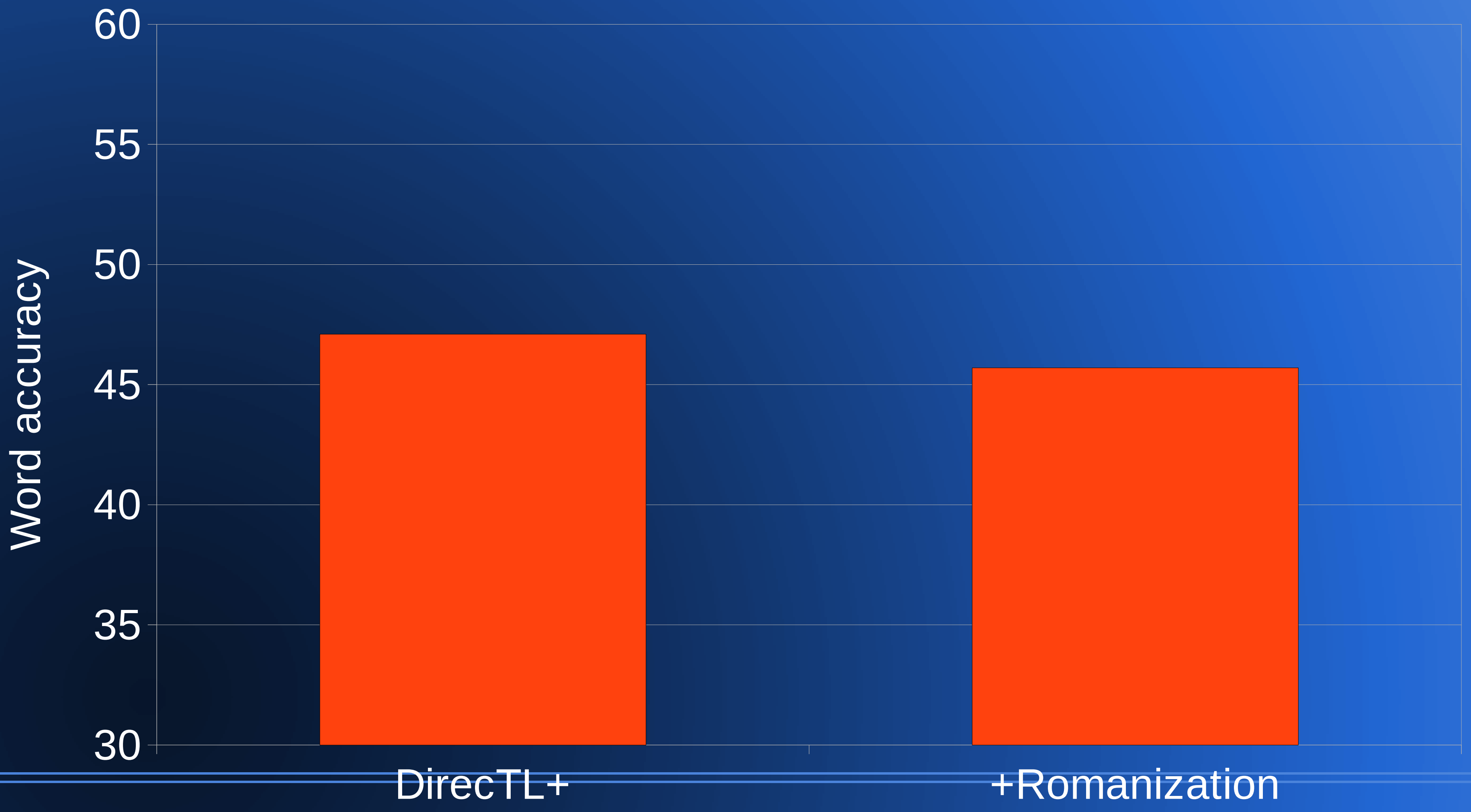
- Devanagari alphabet has combined consonants & vowels
- We experiment with romanizing Hindi
    - Gives DirecTL+ direct individual control
    - Context-sensitive rule-based romanization
    - Use romanized Hindi for training DirecTL+, do testing, then convert outputs to Devanagari



जॉन पटरूची

Jqna paTarUcl

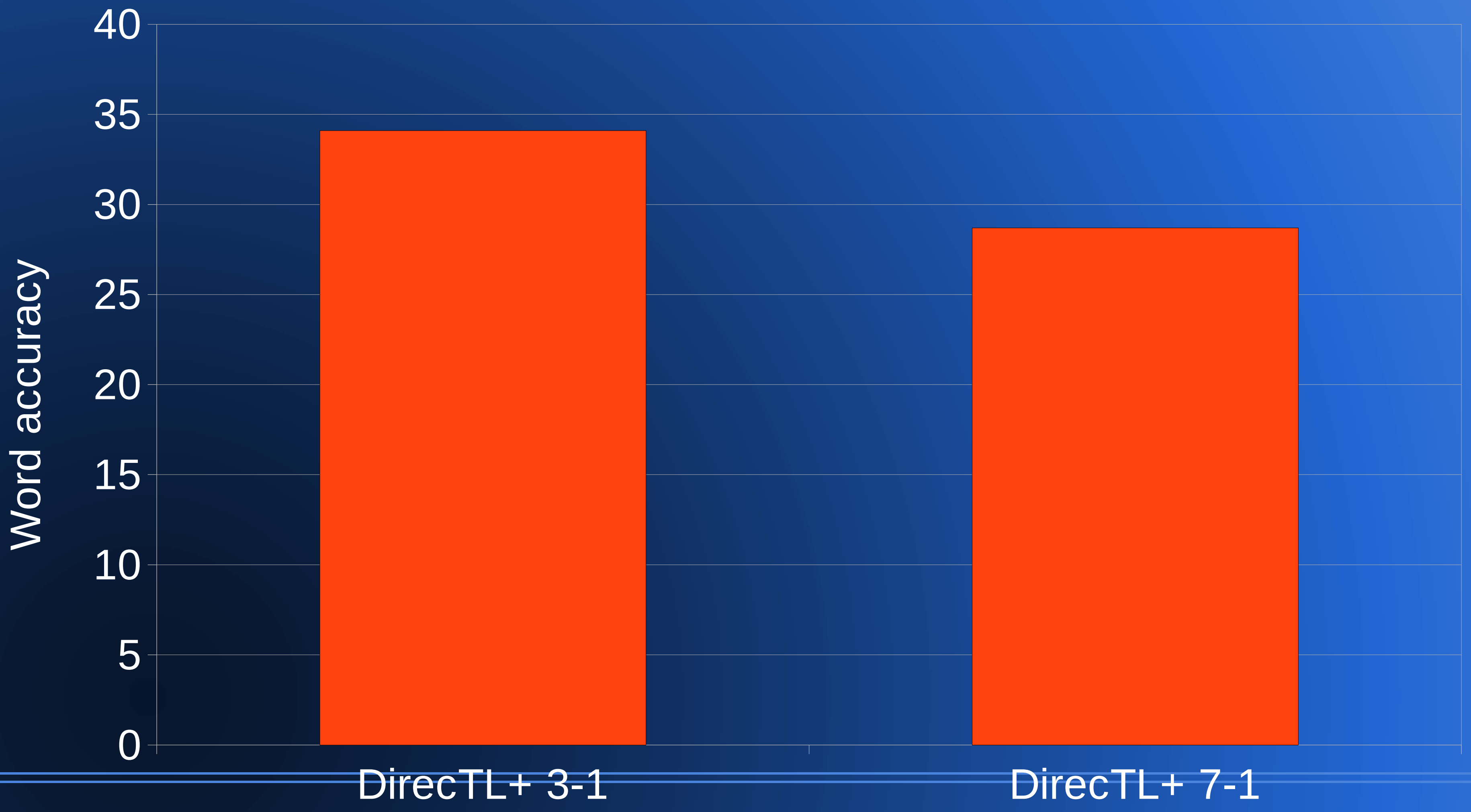
EnHi romanization



Chinese alignment length

- DirecTL+ relies on many-to-many alignments (M2M-Aligner)
- We experiment with maximum alignment length
    - 3-1 vs. 7-1

EnCh alignment length



Conclusion

- SVM re-ranking for transliteration
  - Great improvements with supplemental transliterations
  - Also see improvements for system combination
- Romanization
  - Didn't work for EnHi (unlike EnJa in 2010)
- EnCh alignment lengths
  - Must be careful to choose a good value!