Translation of Unknown Words in Low Resource Languages

Biman Gujral, **Huda Khayrallah**, and Philipp Koehn

Johns Hopkins University

31 October 2016

This talk was presented at AMTA 2016

It is based on this paper:
http://www.cs.jhu.edu/~huda/papers/gujral2016AMTA.pdf

bib:

http://www.cs.jhu.edu/~huda/papers/gujral2016AMTA.bib

Translation of Unknown Words in Low Resource Languages

Biman Gujral, **Huda Khayrallah**, and Philipp Koehn

Johns Hopkins University

31 October 2016

Out of Vocabulary Words (OOVs)

- Hindi → English:
  - It वूवन पैंट्स, graphic टीज, Polo T शर्टे, शर्टे, शॉट्स, स्कर्टे and bright embroidered jackets etc are included.
- Uzbek→ English:
  - Quvayt o‘yinga how ko‘ryapmiz with the preparation.

Goals

- Generate candidates for each OOV
- Select the best one

How big is this problem?

Data

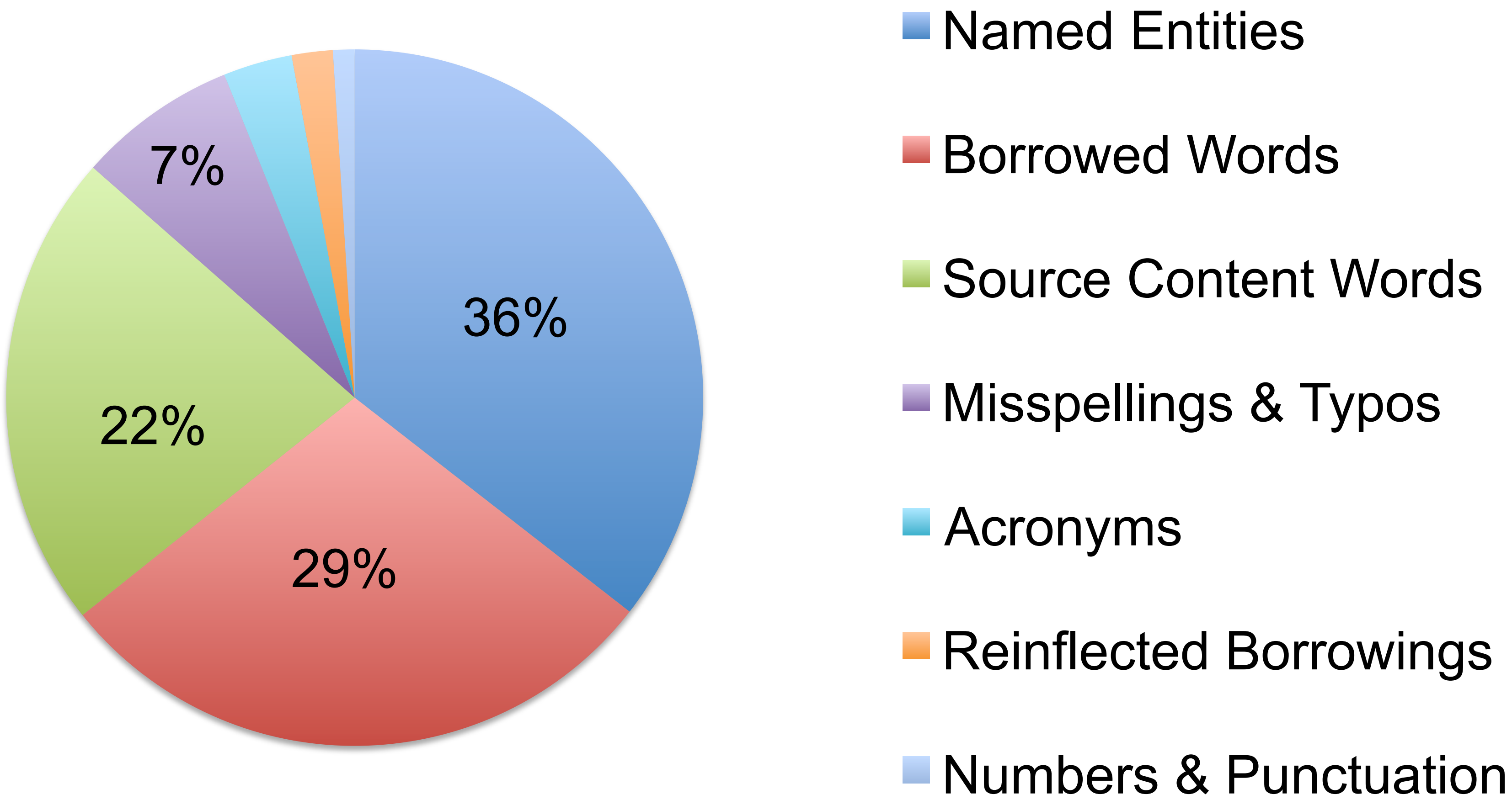
- Hindi → English
  - WMT14
  - News
  - Training
    - 274k sentences
  - Test
    - 2.5k sentences
    - ~1 OOV/sentence
    - ~5% OOVs

- Uzbek → English
  - LORELEI
  - News, Wikipedia, social media
  - Training
    - 55k sentences
  - Test
    - 1k sentences
    - ~4 OOVs/sentence
    - ~20% OOVs

OOV Examples

- Names
  - هدى → Huda
- Misspellings
  - grammer/grammar
- Inflections
  - play/plays/playing
- Borrowed words
  - हैलोवीन → Halloween
- Reinflected Borrowings
  - स्कर्टे → skirts
  - *Googlear* → to Google
- Content words
  - अटकलें → speculation

Distribution of OOVs



MT System

- Moses (Koehn et al. 2007)
  - Phrase Based
- Large English language model
  - WMT English ’07-’12

Methods

Methods

- Transliteration
- Levenshtein distance
- Word Embeddings

Transliteration

- هدى → Huda
- हैलोवीन → Halloween
- Unsupervised Moses mode (Durrani et al. 2014)
  - Character translation model
  - Incorporate larger English language model
- Uzbek is already written in Latin script, keep original spelling
- Generate 1 candidate

Levenshtein distance

- grammer/grammar
- play/plays
- Minimum number of:
  - insertions
  - deletions
  - substitutions

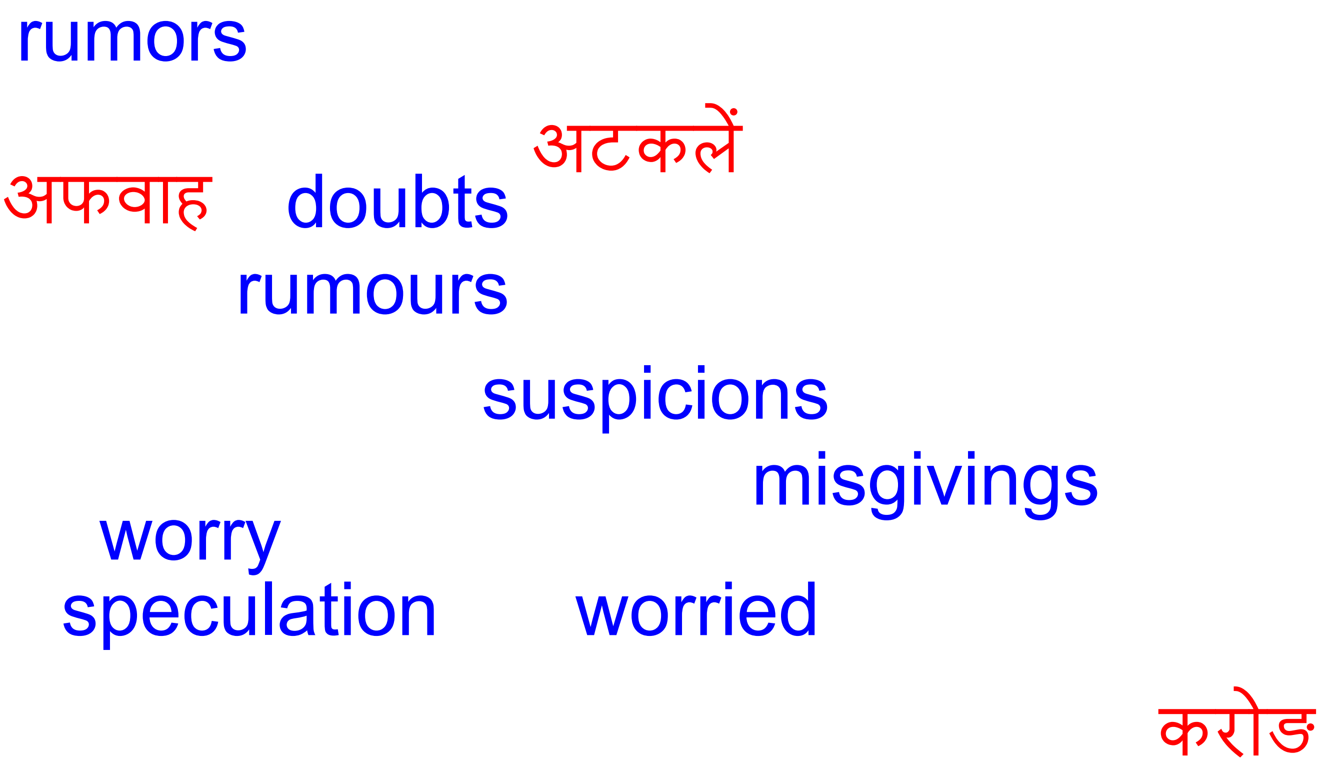
Levenshtein distance

- qilyapmiz → doing
- qilyapsiz → doing

Levenshtein distance = 1

- Find source words with distance ≤ 2 from OOV
  - Use their English translation as translation candidate
- Generate 18 candidates on average

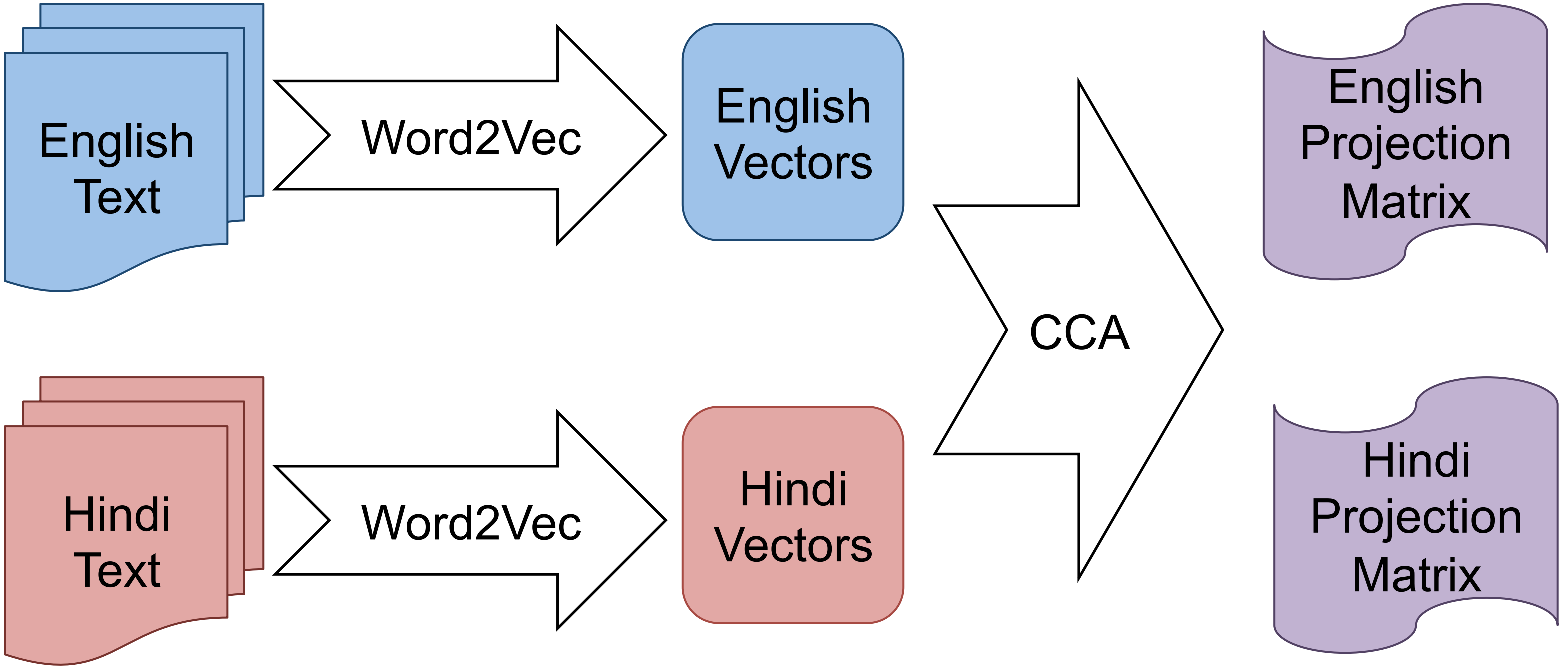
Word Embeddings



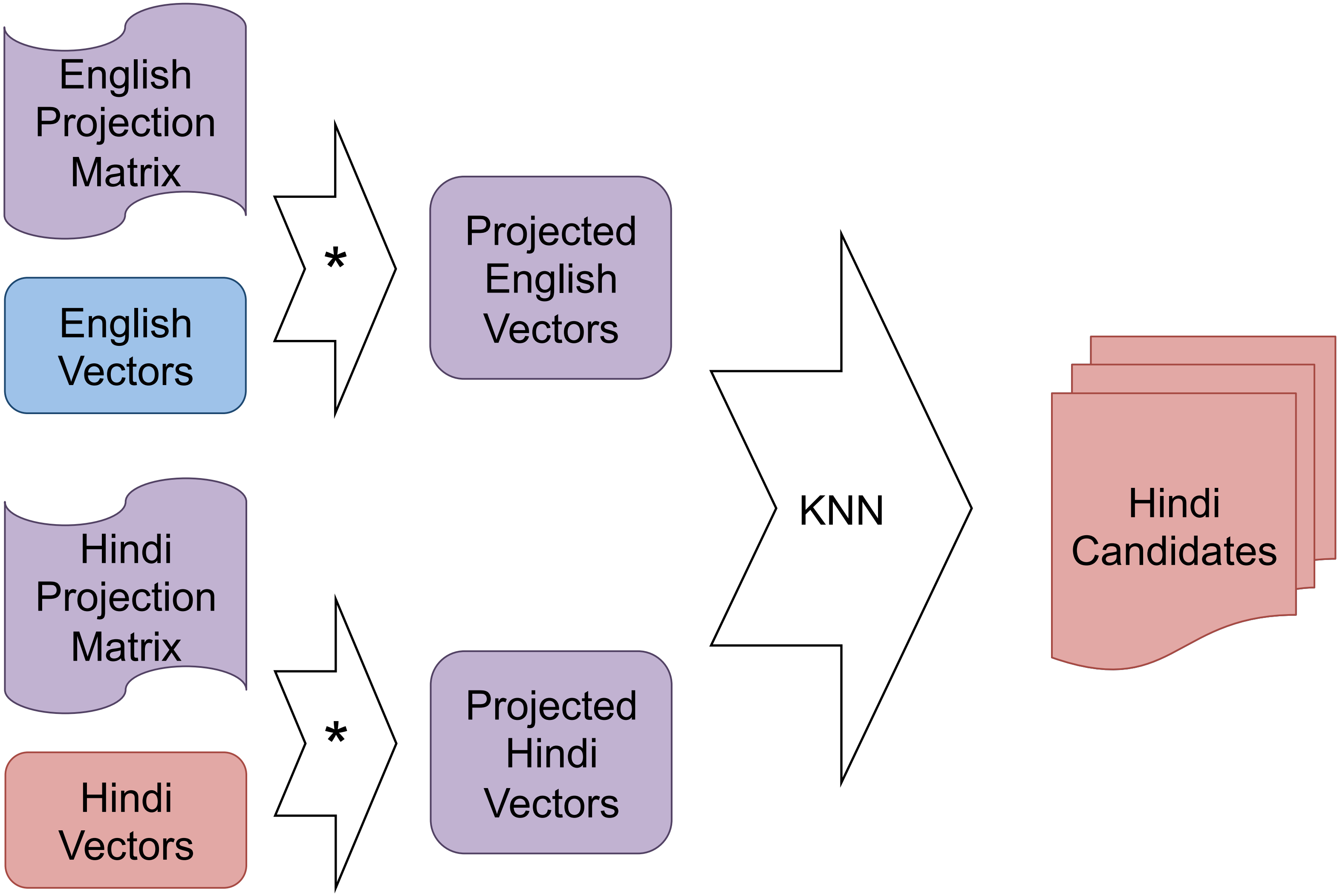
Word Embeddings

- Word2vec (Mikolov et al. 2013)
  - monolingual corpora
- Multilingual word vectors (Faruqui & Dyer 2014)
  - monolingual vectors
  - alignments
  - Canonical Correlation Analysis (CCA)
- Generates 20 candidates

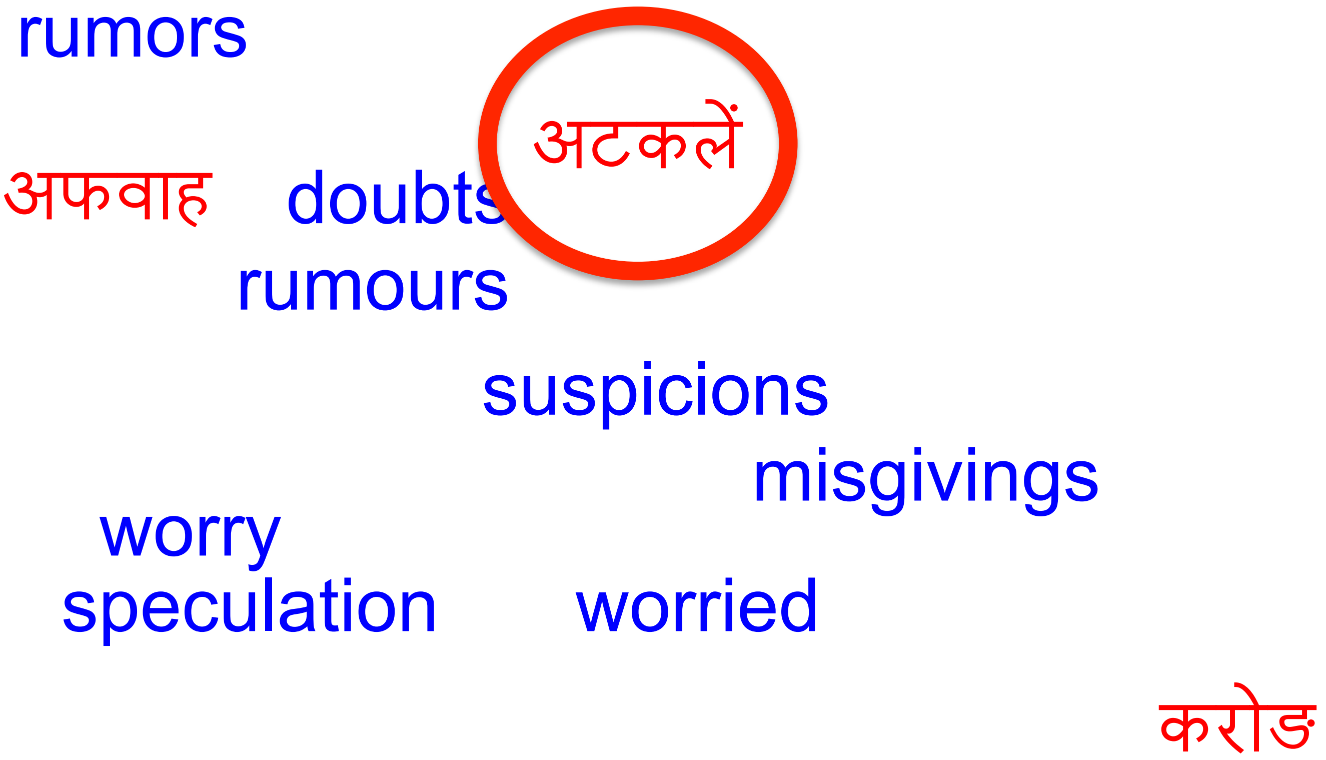
Word Embeddings



Word Embeddings



Word Embeddings



rumors

अटकलें

अफवाह

doubts

rumours

suspicions

misgivings

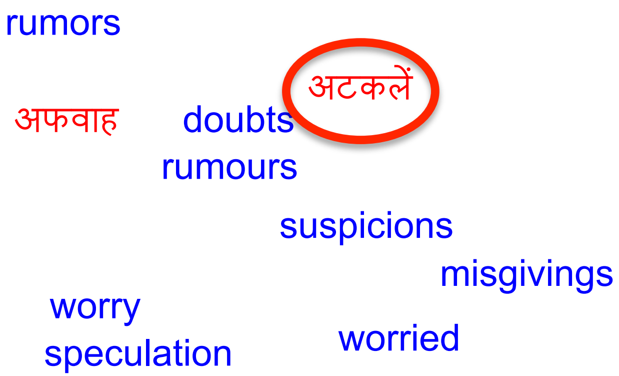
worry

speculation

worried

करोड

Word Embeddings



rumors

अफवाह

doubts

अटकलें

rumours

suspicions

misgivings

worry

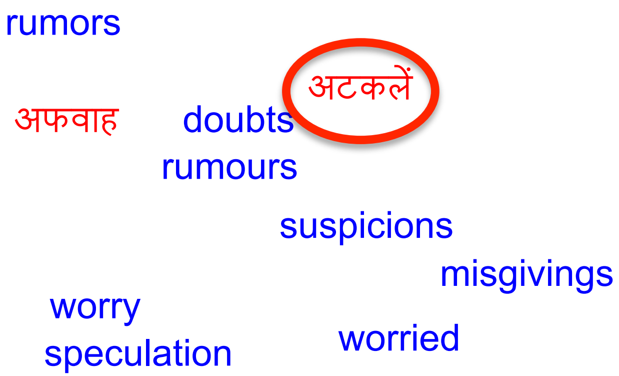
speculation

worried

अटकलें

1) doubts
2) rumours
3) suspicions
4) misgivings
5) worry
6) worried
7) speculation
8) …
9) …
10)…

Word Embeddings


rumors
अफवाह
doubts
अटकलें
rumours
suspicions
misgivings
worry
speculation
worried

अटकलें

1)	अफवाह	→	rumor
2)	करोड़	→	crore
3)	…	→	…
4)	…	→	…
5)	…	→	…
6)	…	→	…
7)	…	→	…
8)	…	→	…
9)	…	→	…
10)	…	→	…

Word Embeddings

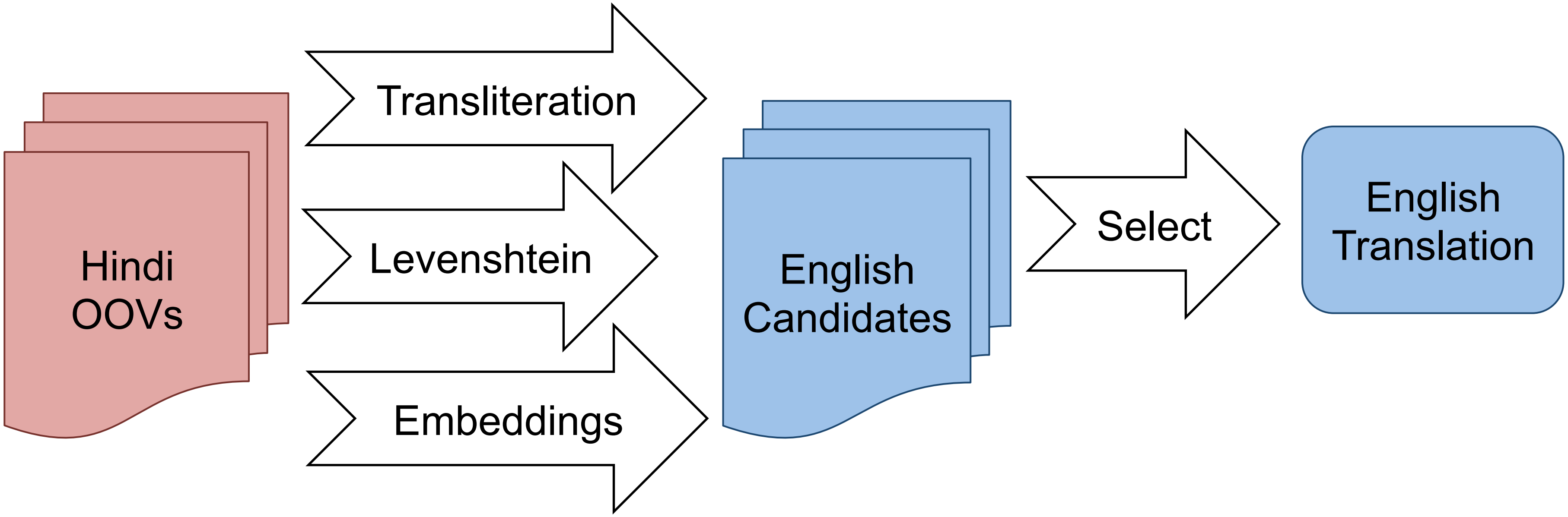
अटकलें

1) doubts
2) rumours
3) suspicions
4) misgivings
5) worry
6) worried
7) speculation
8) …
9) …
10)…

1) rumor
2) crore
3) …
4) …
5) …
6) …
7) …
8) …
9) …
10)…

Integration

Integration



Hindi OOVs

Transliteration

Levenshtein

Embeddings

English Candidates

Select

English Translation

Integration

- Language Model
- Phrase table

Language Model

- Large English language model
- XML markup in Moses (Koehn & Haddow, 2009)
- Selection occurs during decoding

Phrase Table

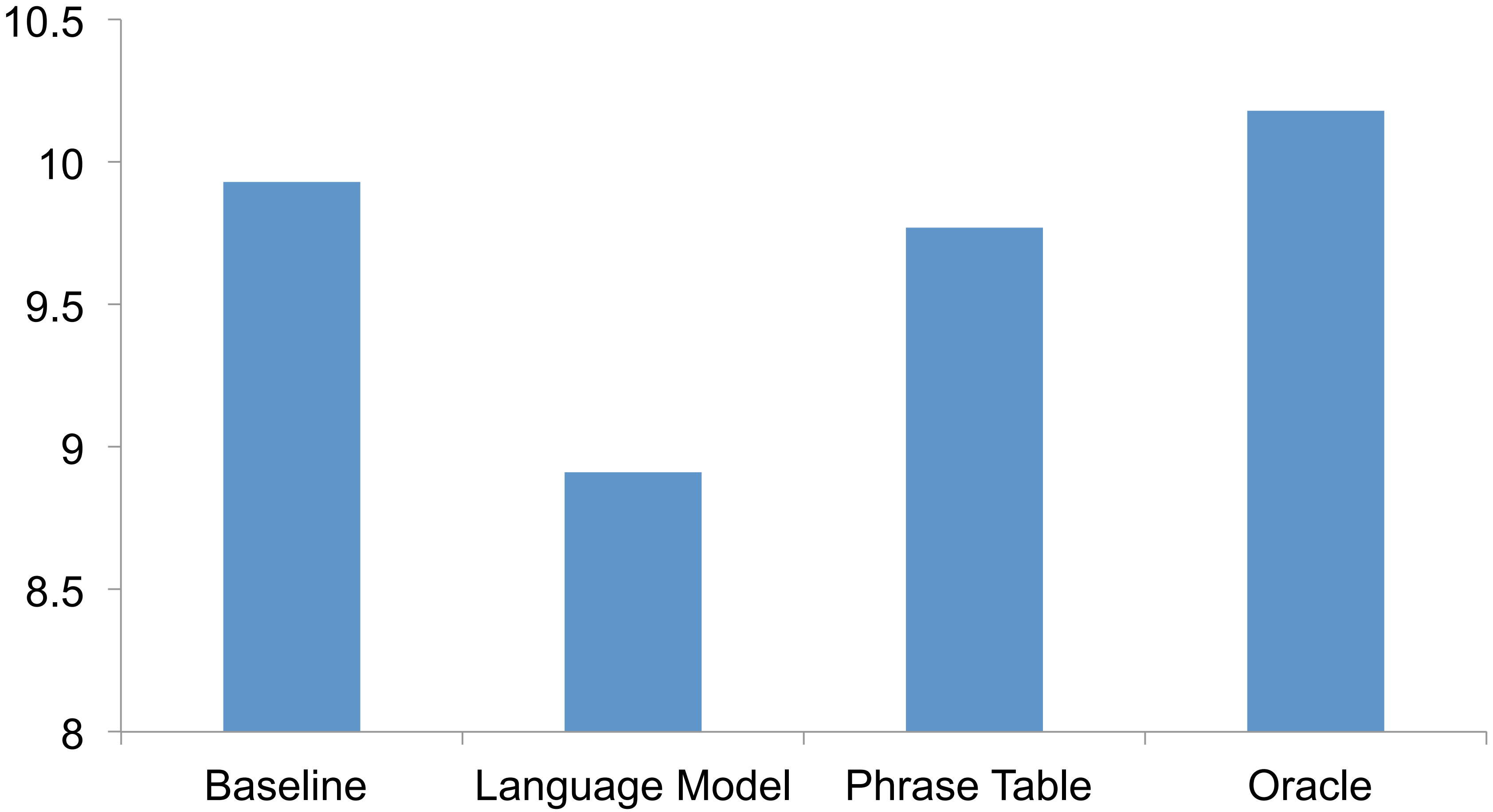
- Secondary Phrase Table only includes OOVs
- Features:
  - Method
  - Word Vector Distance
  - Levenshtein distance
  - Inverse frequency in Monolingual corpus

Results

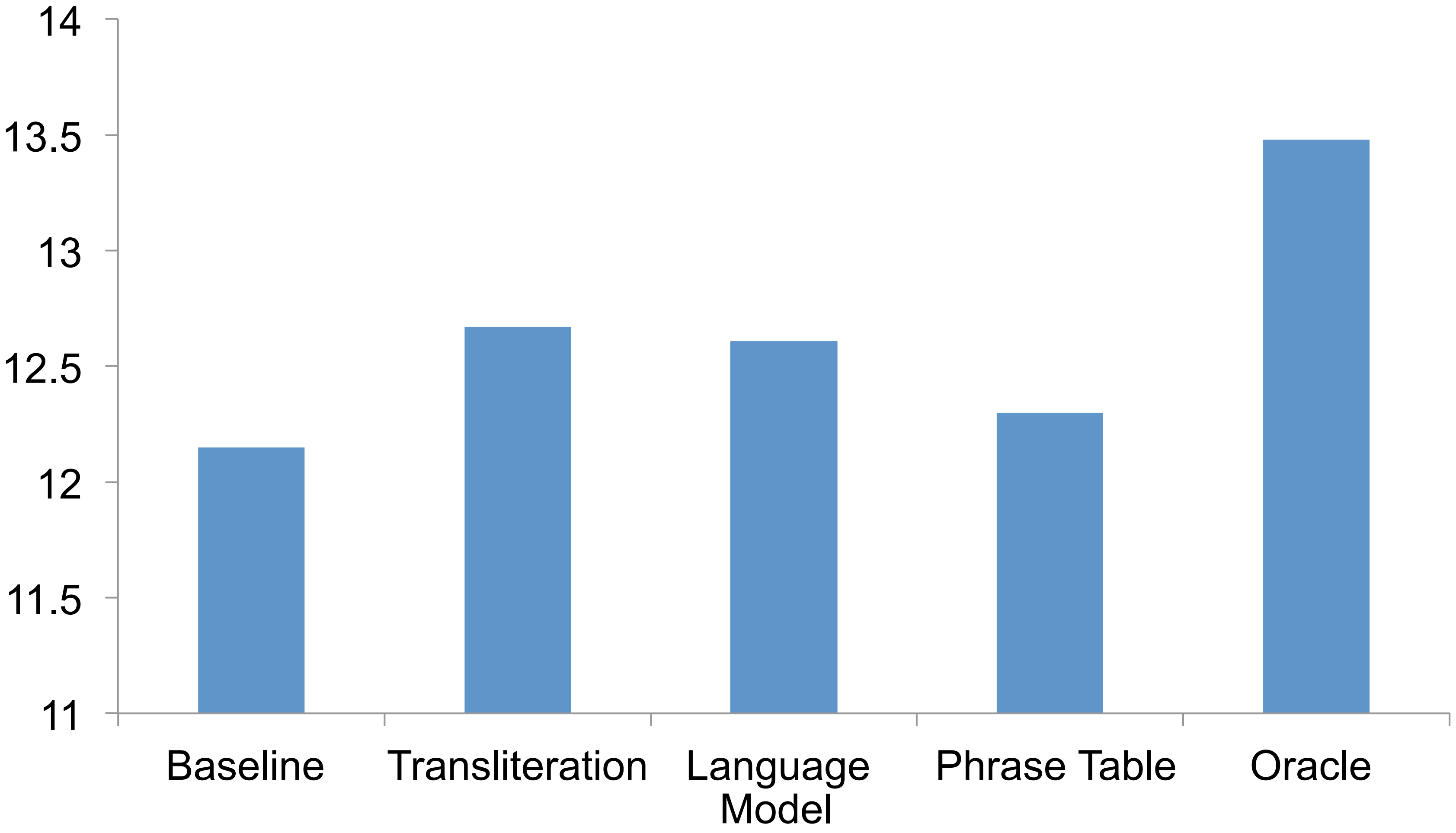
Oracle

- Upper bound on how well a selection method can do given current generation methods
  – Select word from list of candidates that is in the reference

BLEU - Uzbek



BLEU - Hindi


14
13.5
13
12.5
12
11.5
11
Baseline
Transliteration
Language Model
Phrase Table
Oracle

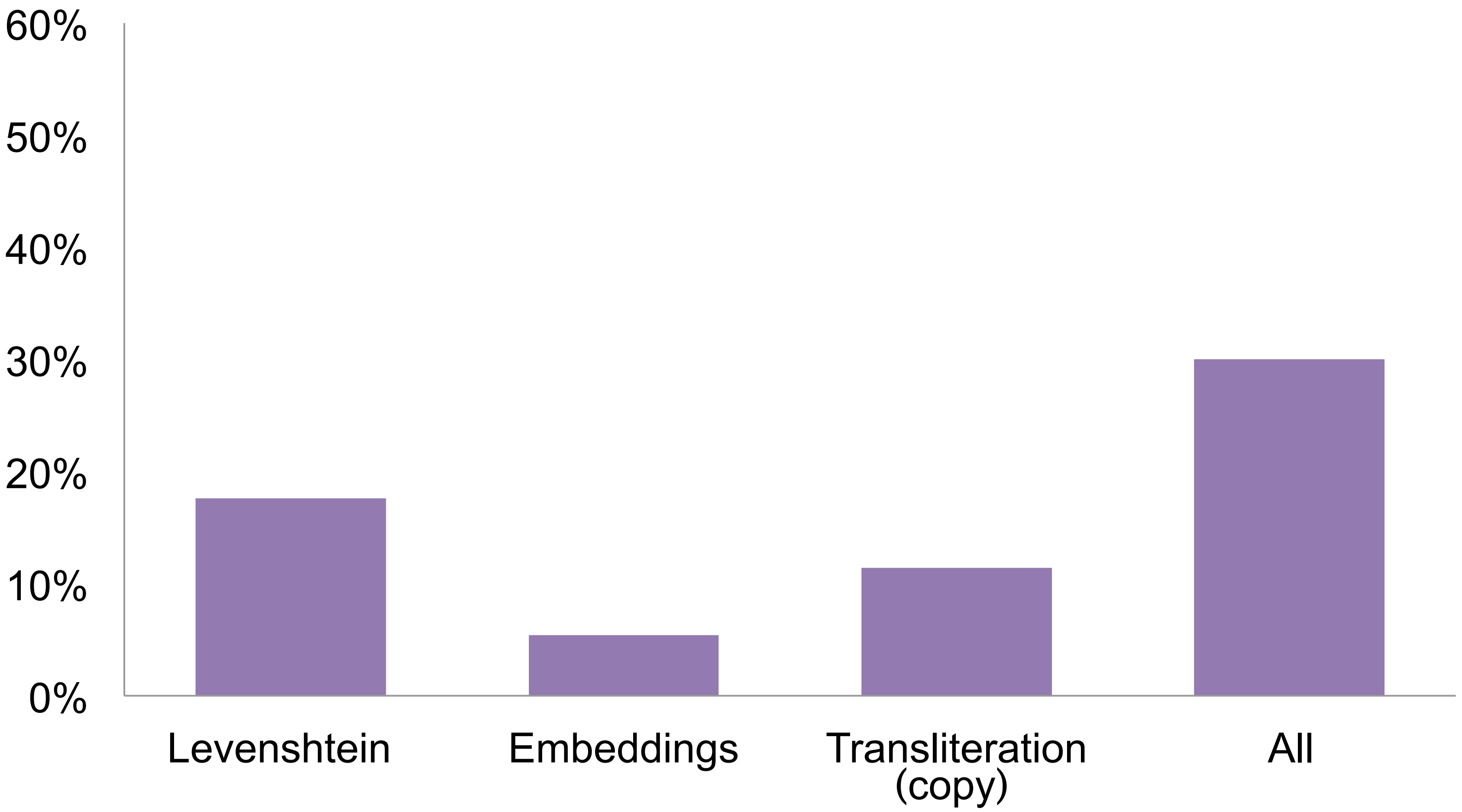
Beyond BLEU

- Goals:
  - generate candidates for each OOV
    - How well can we generate translation candidates?
  - select the best one
    - How well can we select from the translation candidates?

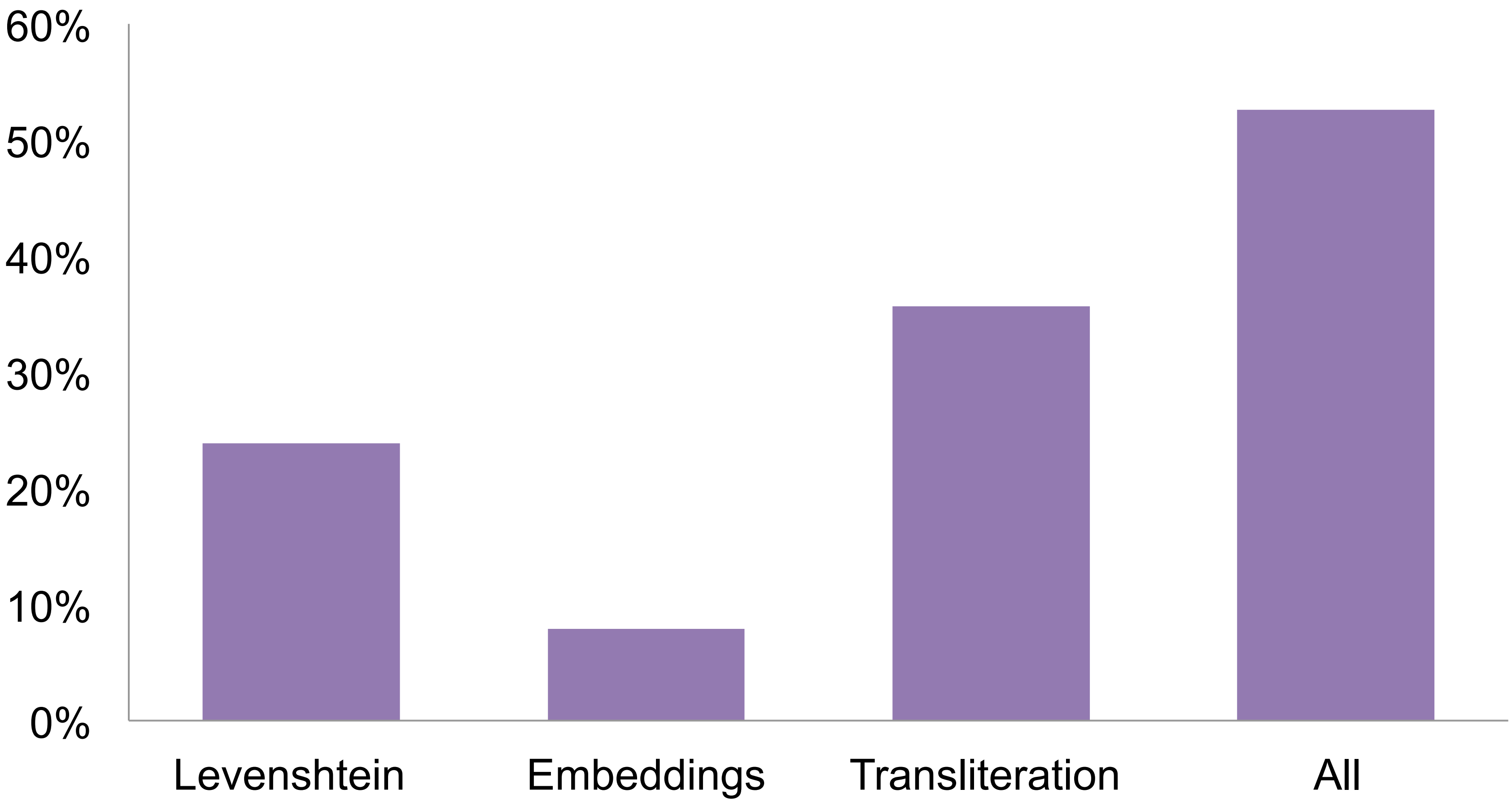
Coverage

- How well can we generate translation candidates?
  - Was one of the candidates **generated** by this method in the reference?

Coverage - Uzbek



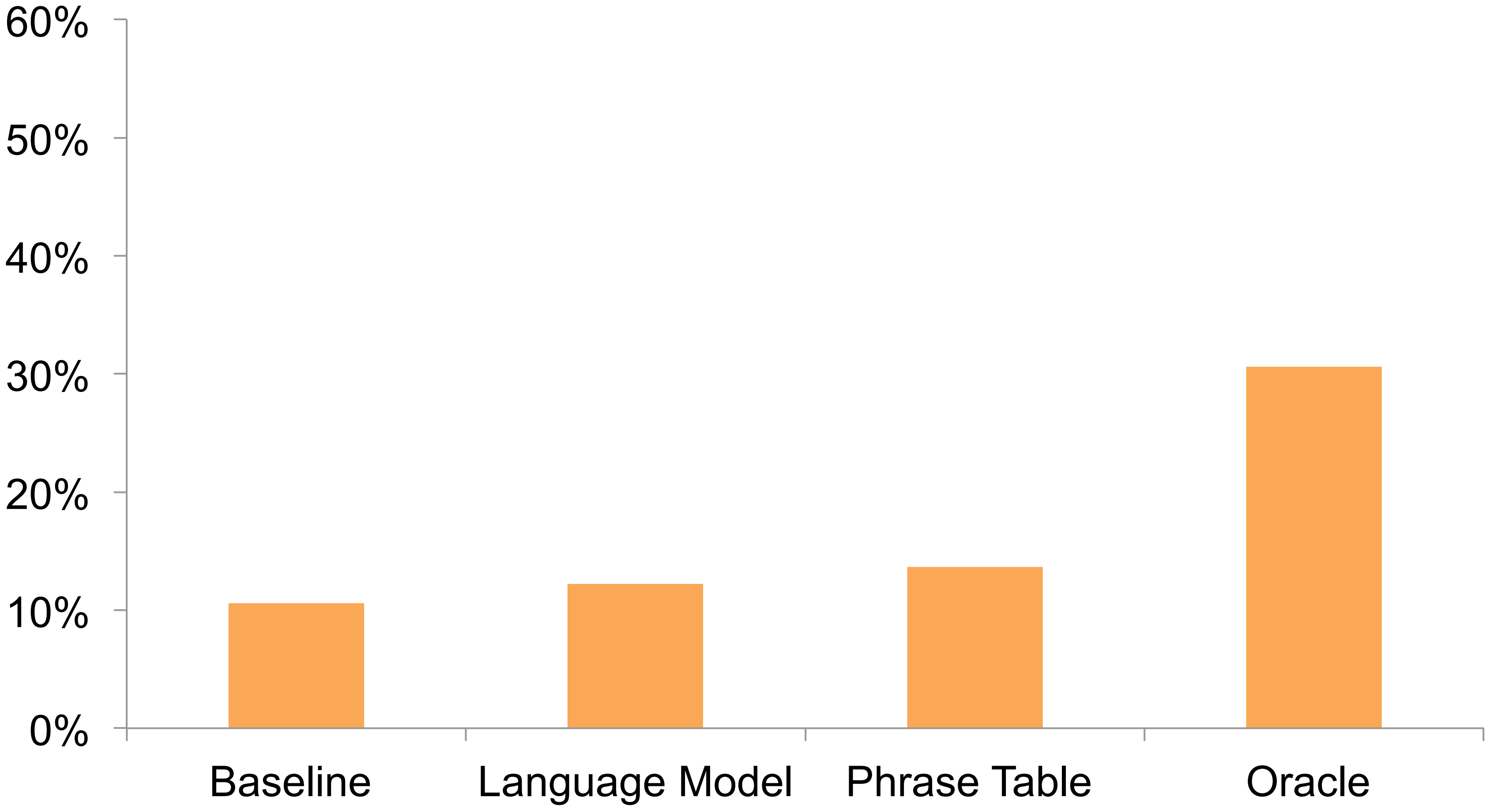
Coverage - Hindi



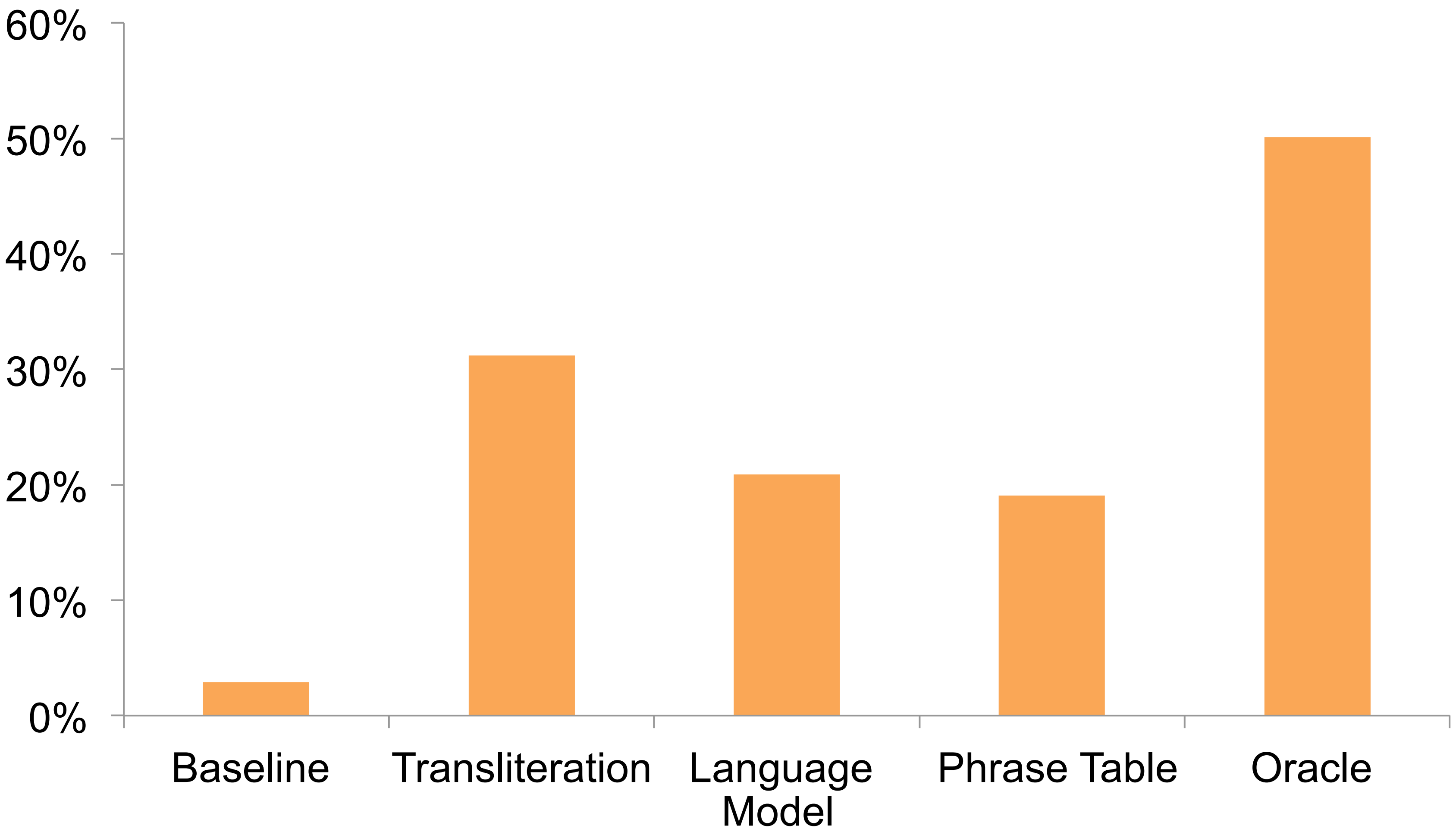
Accuracy

- How well can we select from the translation candidates?
  - Is the word we **selected** in the reference?

Accuracy - Uzbek



Accuracy - Hindi



Conclusion & Future Work

- Generate Quality translations
  - Selection does not perform as well

- Improved selection methods
- More sophisticated embedding projection
- Analysis of what methods work on which types of OOVs

Acknowledgement

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0113.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

Translation of Unknown Words in Low Resource Languages

Biman Gujral, **Huda Khayrallah**, and Philipp Koehn

{bgujral1, **huda**, phi}@jhu.edu

Johns Hopkins University

References

- Durrani, Haddow, Koehn, and Heafield. (2014). Edinburgh's Phrase-Based Machine Translation Systems for WMT-14. *Workshop on Statistical Machine Translation*
- Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst. (2007). Moses: Open source toolkit for Statistical Machine Translation. *ACL Interactive Poster and Demonstration Sessions*
- Koehn and Haddow. Edinburgh's submission to all tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. *Workshop on Statistical Machine Translation*
- Faruqui and Dyer. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. *In Proceedings of EACL.*