Generalizing Word Lattice Translation

2008 Paper by Christopher Dyer, Smaranda Muresan, and Philip Resnik
on Arabic/Chinese → English translation

Presentation for TAU 2014 MT seminar by Ariel Ben-Yehuda

Prior Art: Phrase-based Decoding

- We already seen it in class
- Independently match source language phrases to target language phrases
- Phrases are contiguous groups of words

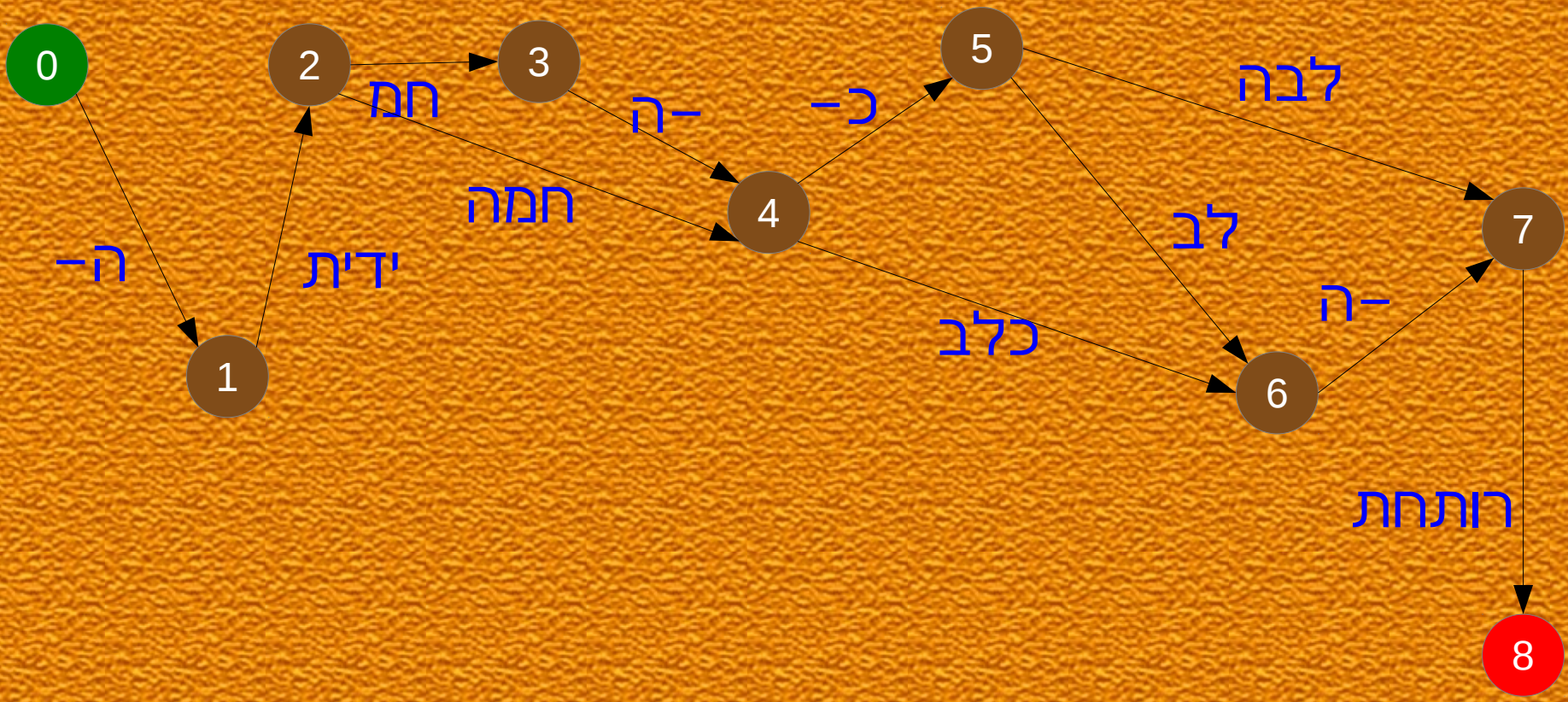
Word-Splitting

- In Arabic (also Hebrew), words contain morphology
  - e.g. in Hebrew, לבנה can be both
    - to her son
    - white
- In Chinese, no spaces between words
  - I don't understand Chinese, so no examples :-(

Word-Splitting cont.

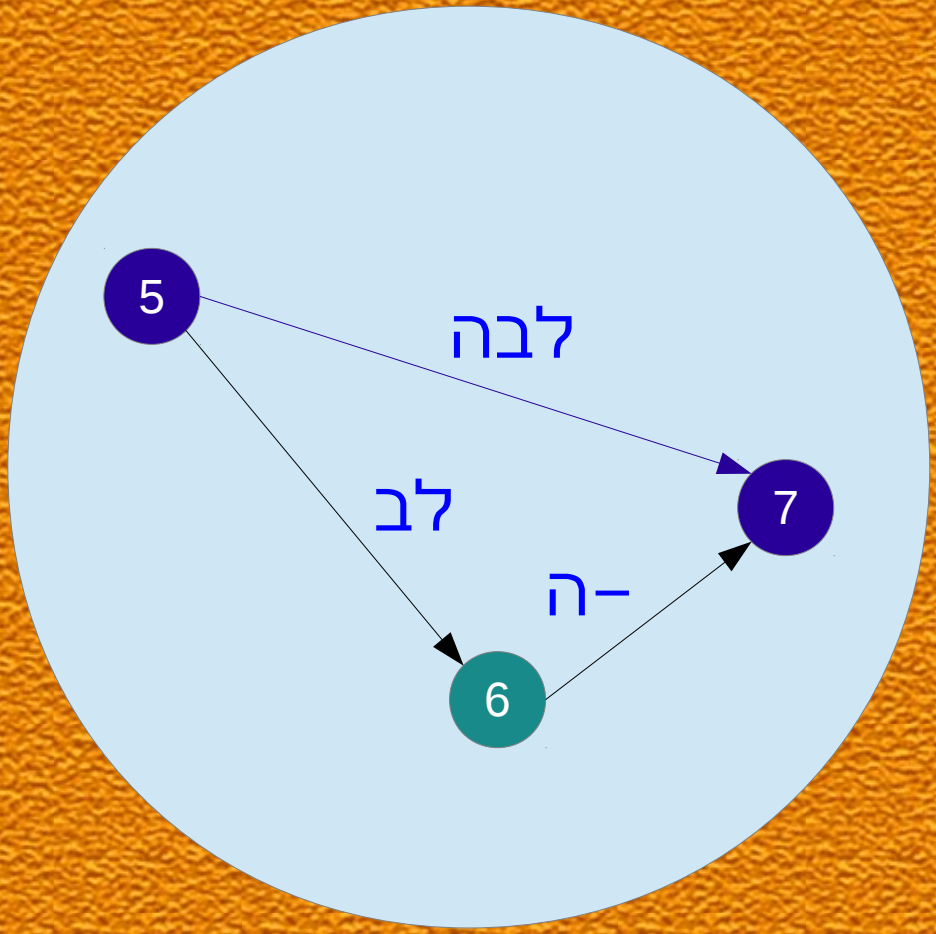
- Split depends on context
  - הידית חמה כלבה רותחת
- Traditional PBD can't handle this
  - Decoding operates **after** splitting
    - Google Translate: “Handle boiling hot dog”

Pass **all** the options to decoding



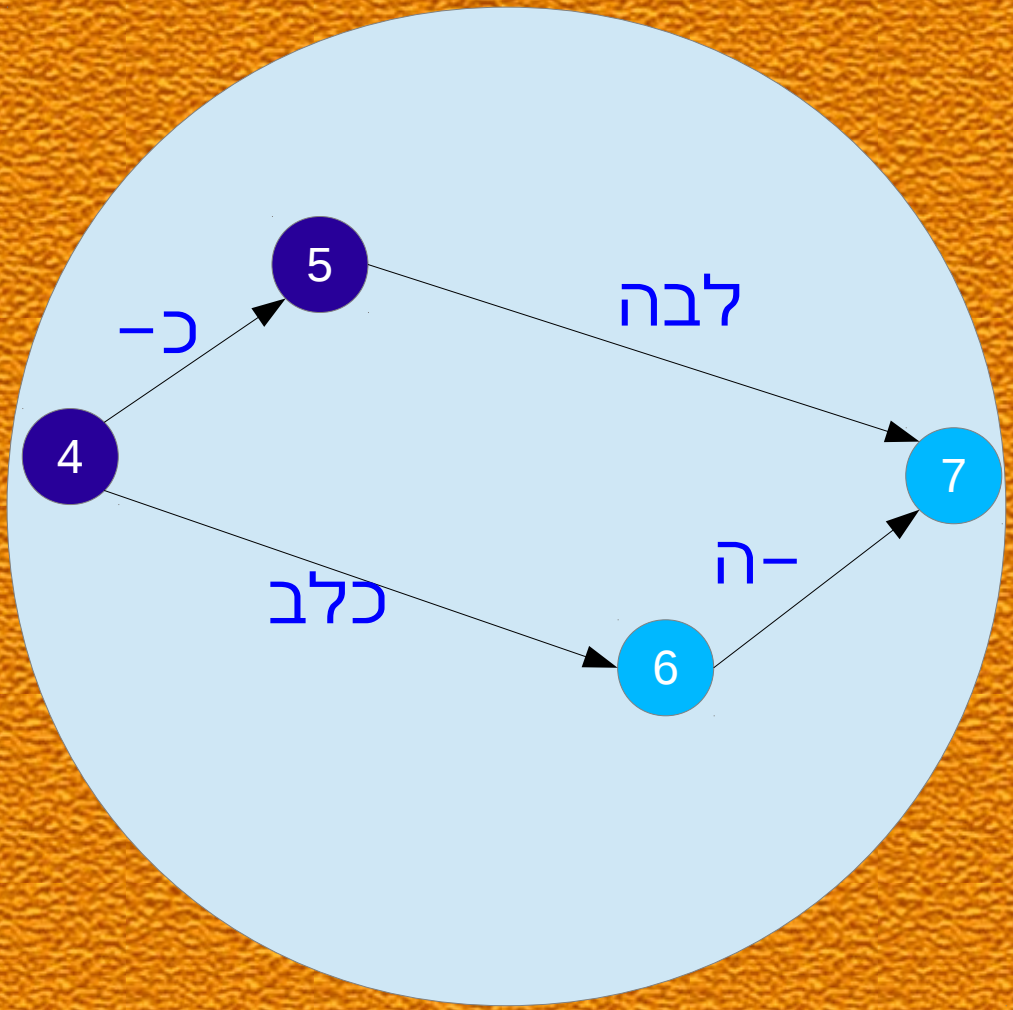
Remove alternatives

- When translating a phrase, mark the nodes in the middle



Non-Monotonicity

- Not a problem here (לב is a word)
  - but if it wasn't...
  - כ——ה is invalid
  - (5) covers (6)



Empirical Results (Chinese)

- Trained and tested on MT05, also tested on MT06
  - ~950k sentences each
- Multiple segmentation algorithms (c,s,h)
- PL uses a better reordering distance
- Both phrase-based and hierarchical decoding

	P-c	P-h	P-s	P-hs	P-hsc	PL-hsc	H-c	H-h	H-s	H-hs	H-hsc
MT05	.2833	.2905	.2894	.2938	.2993	.3072	.2904	.3008	.3071	.3132	.3176
MT06	.2694	.2835	.2801	.2870	.2865	.2992	.2821	.2907	.2964	.3006	.3043

Empirical Results (Arabic)

- Train on MT08 (250M words)
- Evaluate on MT05, MT06
- 2 morphological analysers: surface, morph

	P-surface	P-morph	P-both	H-surface	H-morph	H-both
MT05	.4682	.5087	.5225	.5253	.5377	.5453
MT06	.3512	.3841	.4008	.3991	.4180	.4287

Conclusion

- This seems to significantly improve translation performance
- Chinese: word reordering model is significant