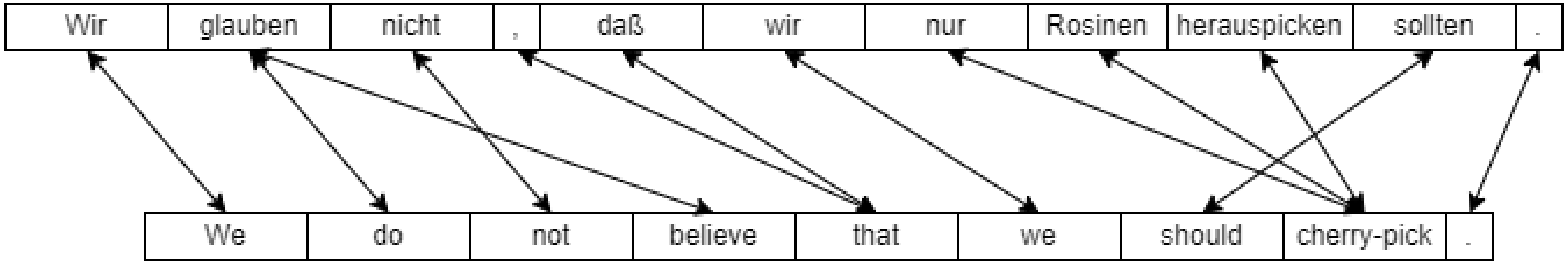
Multilingual Word Alignment with OT

Joshua Hong, University of California, San Diego
Associate Professor Yuki Arase, Onizuka Lab, Graduate School of Information Sciences and Technology



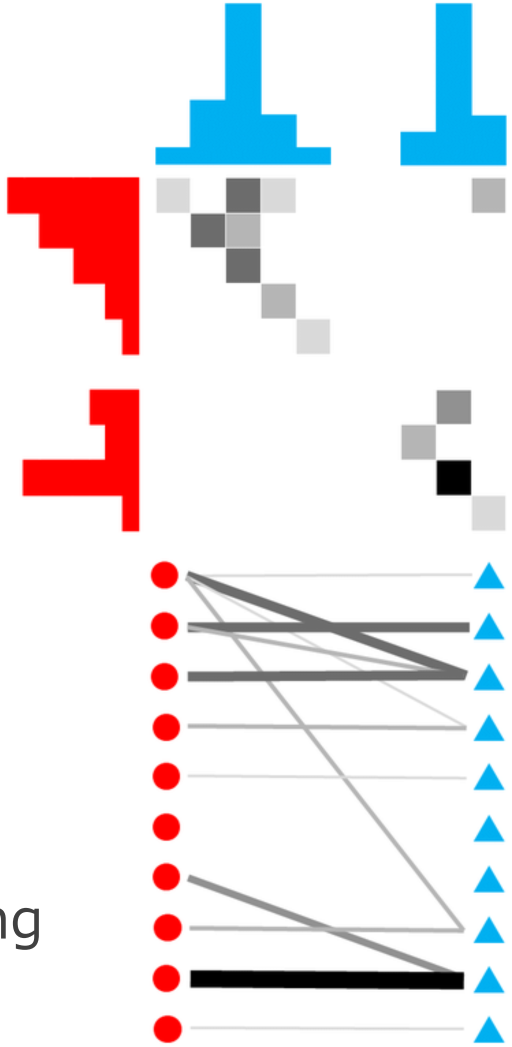
Wir glauben nicht , daß wir nur Rosinen herauspicken sollten .

We do not believe that we should cherry-pick .

- Multilingual word alignment, monolingual word alignment
- Useful for many downstream natural language processing tasks
  - Machine translation, extending datasets for low-resource languages
- Past approaches
  - Statistical aligners
  - Recently, neural word aligners that use pre-trained large language models and probablistic extraction methods
  - How to improve neural word aligners to better align words?

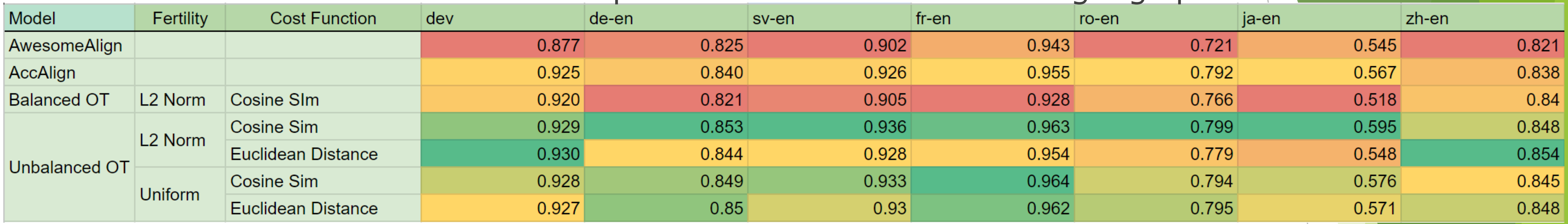
Optimal Transport

- Optimal Transport: Given two distributions and the cost between two points
  - Compute best mapping to transfer "mass" while minimizing cost
- Applying to Word Alignment
  - Treat sentences as a distribution
  - Measure of similarity between words as the cost
- Variations: unbalanced OT and partial OT
  - Relax constraints for the optimal mapping



Results

F1 scores for selected experiments across different language pairs



Model	Fertility	Cost Function	dev	de-en	sv-en	fr-en	ro-en	ja-en	zh-en
AwesomeAlign			0.877	0.825	0.902	0.943	0.721	0.545	0.821
AccAlign			0.925	0.840	0.926	0.955	0.792	0.567	0.838
Balanced OT	L2 Norm	Cosine SIm	0.920	0.821	0.905	0.928	0.766	0.518	0.84
Unbalanced OT	L2 Norm	Cosine Sim	0.929	0.853	0.936	0.963	0.799	0.595	0.848
		Euclidean Distance	0.930	0.844	0.928	0.954	0.779	0.548	0.854
	Uniform	Cosine Sim	0.928	0.849	0.933	0.964	0.794	0.576	0.845
		Euclidean Distance	0.927	0.85	0.93	0.962	0.795	0.571	0.848

- Experiments with variations of optimal transport as well as OT cost formulations and sentence distribution
- Optimal Transport competitive with other methods in an unsupervised setting
  - Outperforms AccAlign (current state of the art) on unseen language pairs
- Future work
  - Further exploration into supervised setting and why unsupervised results don't transfer
  - Additional cost and sentence distribution methods to address common errors