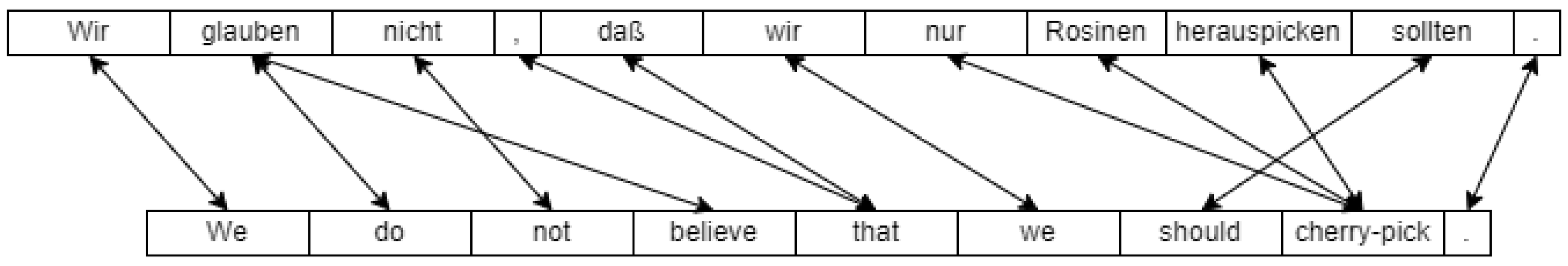
Multilingual Word Alignment with Optimal Transport

Joshua Hong[1], Yuki Arase[2]
1. University of California, San Diego,
2. Graduate School of Information Science and Technology, Osaka University

Introduction



Multilingual word alignment is useful for many downstream tasks including:

- Annotation projection[4]: Technique to transfer annotations/alignments from one language pair to another, extending limited datasets for low resource domains
- Machine Translation

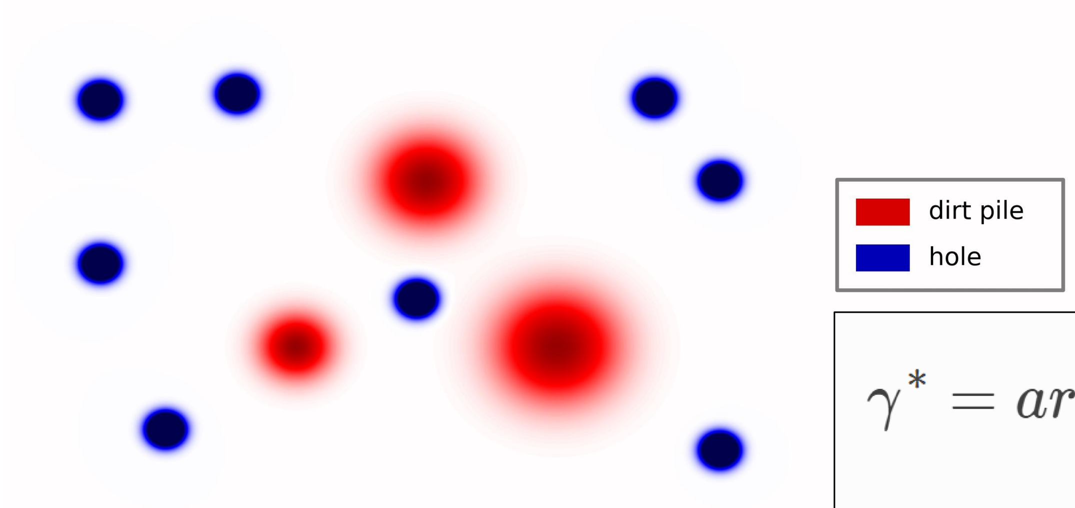
Recently, word alignment has shifted away from statistical methods towards neural alignment

- Statistical methods include Giza++[3], which uses the EM algorithm
- Neural Alignment Methods
  - Uses large language models (LLMs) to learn vector embeddings for sentences and words
  - Recent examples include AwesomeAlign[1] and AccAlign[5], which use embeddings to judge similarity between words

Proposal: Use Optimal Transport to extract word alignments from vector embeddings

Methods

Optimal Transport (OT) is the problem of moving mass from one distribution to another while minimizing the total cost of moving the mass.



2D Toy Example of OT[6]

- $M_{i,j}$ is the cost to move mass from a_i to b_j
- a and b are discrete distributions

$$\gamma^* = arg \min_{\gamma \in \mathbb{R}_+^{m \times n}} \sum_{i,j} \gamma_{i,j} M_{i,j}$$

$$s.t. \gamma 1 = a; \gamma^T 1 = b; \gamma \geq 0$$

Using optimization methods, this problem can be solved to obtain the optimal mass transitions. To adapt OT for word alignment, we

- Treat each parallel sentence as a distribution, where each word has some "mass"
- Define the cost to be some measure of similarity between words
  - The more similar the words, the lower the cost to move mass between them should be

What measure of similarity should we use?

Cosine Similarity

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

Euclidean Distance

$$d(x,\ y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

What initial distribution should we use for each sentence?

- Uniform distribution – Each word is given equal mass
- L^2-Norm – Each word is given mass corresponding to its vector embedding's magnitude

What variation of OT is best for word alignment?

- Balanced Optimal Transport
  - Regularization term allows for new optimization procedures
- Unbalanced Optimal Transport
  - Additional term in optimization function allows for mass deviations from the given distributions
- Partial Optimal Transport
  - Relaxation of OT where only a portion of mass needs to be transported

Additionally, we experiment with different normalization methods before and after applying OT, including matrix min-max value scaling and column/row min-max scaling

Methods

To obtain word embeddings, we use Language-Agnostic BERT (LaBSE[2]), a LLM trained on multilingual sentences.

We test OT in unsupervised and supervised settings.

Unsupervised

- We perform a hyperparameter search on a development dataset, then test on unseen language pairs

Zero-Shot Supervised

- We finetune the LLM on a training dataset, then perform the same steps as the unsupervised setting
- The training dataset has no language pair overlap with the testing datasets

Example word similarities with LaBSE[5]

Results

Model	Fertility	Cost Function	Cost Function Scaling	Alignment Scaling	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	Column-Row Min-max	Matrix Min-max Norm	0.920	0.821	0.905	0.928	0.766	0.518	0.84
Unbalanced OT	L2 Norm	Cosine Sim	Column-Row Min-max	Matrix Min-max Norm	0.929	0.853	0.936	0.963	0.799	0.595	0.848
				Column-Row Min-max	0.9272	0.846	0.93	0.958	0.79	0.6	0.85
			Matrix Min-max Norm	Matrix Min-max Norm	0.918	0.841	0.927	0.958	0.78	0.502	0.83
		Euclidean Distance	Column-Row Min-max	Matrix Min-max Norm	0.930	0.844	0.928	0.954	0.779	0.548	0.854
				Column-Row Min-max	0.9248	0.84	0.927	0.958	0.784	0.584	0.844
			Matrix Min-max Norm	Matrix Min-max Norm	0.9185	0.843	0.924	0.948	0.787	0.55	0.819
	Uniform	Cosine Sim	Column-Row Min-max	Matrix Min-max Norm	0.928	0.849	0.933	0.964	0.794	0.576	0.845
			Matrix Min-max Norm	Matrix Min-max Norm	0.9141	0.837	0.922	0.957	0.774	0.503	0.811
		Euclidean Distance	Column-Row Min-max	Matrix Min-max Norm	0.927	0.85	0.93	0.962	0.795	0.571	0.848

In unsupervised settings, OT is competitive against state of the art (SOTA) techniques, with some variations performing better across different language pairs.

Model	Fertility	Cost Function	Cost Function Scaling	Alignment Scaling	dev	de-en	sv-en	fr-en	ro-en	ja-en	zh-en
AwesomeAlign						0.841	0.932	0.956	0.742	0.581	0.856
AccAlign					0.948	0.862	0.946	0.972	0.791	0.629	0.884
Unbalanced OT, No Adapter	L2 Norm	Cosine Sim	Column-Row Min-max	Matrix Min-max Norm	0.955	0.861	0.938	0.962	0.842	0.62	0.84
	Uniform	Cosine Sim	Column-Row Min-max	Matrix Min-max Norm	0.953	0.86	0.938	0.965	0.845	0.614	0.835
Unbalanced OT, Adapter	L2 Norm	Cosine Sim	Column-Row Min-max	Matrix Min-max Norm	0.938	0.856	0.936	0.962	0.812	0.584	0.848
Unbalanced OT using AccAlign Finetuned Model	L2 Norm	Cosine Sim	Column-Row Min-max	Matrix Min-max Norm	0.958	0.875	0.951	0.972	0.831	0.647	0.87

OT methods perform worse than SOTA methods on most language pairs in the supervised setting. However, OT methods with SOTA method tuning results in better performance.

OT and baseline methods struggle with null alignment (<0.7 F1), when a word has no corresponding word, and many-to-one alignment (<0.4 F1), when a many words are aligned to the same word, when compared to one-to-one alignment (~0.9 F1).

Conclusion

While OT based alignment methods seems to suffer from setting transferability in supervised settings, performance in unsupervised settings is promising.

Many possible directions for further inquiry:

- Alternative formulations for similarity and distribution
- OT alignment methods with different LLMs
- OT alignment methods in a completely supervised setting
- Additional techniques to address difficulties in many-to-one and null alignments.
- Further investigation into OT alignment zero shot supervised performance and issues with finetuning

References

1. Dou, Z. Y., & Neubig, G. (2021). *Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Conference of the European Chapter of the Association for Computational Linguistics (EACL).*
2. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). *Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852.*
3. Franz Josef Och, Hermann Ney. *"A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.*
4. Li, B. (2022). *Word Alignment in the Era of Deep Learning: A Tutorial. arXiv preprint arXiv:2212.00138.*
5. Wang, W., Chen, G., Wang, H., Han, Y., & Chen, Y. (2022). *A Multilingual Word Aligner. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 2952–2963). Association for Computational Linguistics.*
6. Williams, A. (2020, October 9). *Optimal transport 2D toy example. Its Neuronal. http://alexhwilliams.info/itsneuronalblog/code/ot/holes.png*